Towards Black-box Iterative Machine Teaching

Weiyang Liu [*1] **Bo Dai** [*1] **Xingguo Li** [2] **Zhen Liu** [1] **James M. Rehg** [1] **Le Song** [1 3]

Abstract

In this paper, we make an important step towards the black-box machine teaching by considering the cross-space machine teaching, where the teacher and the learner use different feature representations and the teacher can not fully observe the learner's model. In such scenario, we study how the teacher is still able to teach the learner to achieve faster convergence rate than the traditional passive learning. We propose an active teacher model that can actively query the learner (*i.e.*, make the learner take exams) for estimating the learner's status and provably guide the learner to achieve faster convergence. The sample complexities for both teaching and query are provided. In the experiments, we compare the proposed active teacher with the omniscient teacher and verify the effectiveness of the active teacher model.

1. Introduction

Machine teaching (Zhu, 2015; 2013; Zhu et al., 2018) is the problem of constructing a minimal dataset for a target concept such that a student model (*i.e.*, leaner) can learn the target concept based on this minimal dataset. Recently, machine teaching has been shown very useful in applications ranging from human computer interaction (Suh et al., 2016), crowd sourcing (Singla et al., 2014; 2013) to cyber security (Alfeld et al., 2016; 2017). Besides various applications, machine teaching also has nice connections with curriculum learning (Bengio et al., 2009; Hinton et al., 2015). In traditional machine learning, a teacher usually constructs a batch set of training samples, and provides them to a student in one shot without further interactions. Then the student keeps learning from this batch dataset and tries to learn the target concept. Previous machine teaching paradigm (Zhu, 2013; 2015; Liu et al., 2016) usually focuses on constructing the smallest such dataset, and characterizing the size of such dataset, called the *teaching dimension* of the student model.

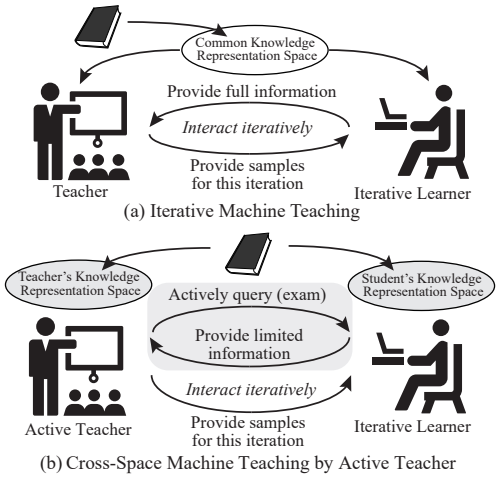

Figure 1: Comparison between iterative machine teaching and cross-space machine teaching by active teacher.

For machine teaching to work effectively in practical scenarios, (Liu et al., 2017a) propose an iterative teaching framework which takes into consideration that the learner usually uses iterative algorithms (e.g. gradient descent) to update the models. Different from the traditional machine teaching framework where the teacher only interacts with the student in one-shot, the iterative machine teaching allows the teacher to interact with the student in every single iteration. It hence shifts the teaching focus from models to algorithms: the objective of teaching is no longer constructing a minimal dataset in one shot but searching for samples so that the student learns the target concept in a minimal number of iterations (*i.e.*, fastest convergence for the student algorithm). Such a minimal number of iterations is called the *iterative teaching dimension* for the student algorithm. (Liu et al., 2017a) mostly consider the simplest iterative case where the teacher can fully observe the student. This case is interesting in theory but too restrictive in practice.

Human teaching is arguably the most realistic teaching scenario in which the learner is completely a black-box to the teacher. Analogously, the ultimate problem for machine teaching is how to teach a black-box learner. We call such problem *black-box machine teaching*. Inspired by the fact that the teacher and the student typically represent the same concept but in different ways, we present a step towards the black-box machine teaching – *cross-space machine teaching*, where the teacher **i)** does not share the same feature representation with the student, and **ii)** can not observe the

[*]Equal contribution [1]Georgia Tech [2]University of Minnesota [3]Ant Financial. Correspondence to: W. L. <wyliu@gatech.edu>.

Proceedings of the 35^{th} International Conference on Machine Learning, Stockholm, Sweden, PMLR 80, 2018.

student model. This setting is interesting in the sense that it can both relax the assumptions for iterative machine teaching and improve our understanding on human learning.

Inspired by a real-life fact, that a teacher will regularly examine the student to learn how well the student has mastered the concept, we propose an active teacher model to address the cross-space teaching problem. The active teacher is allowed to actively query the student with a few (limited) samples every certain number of iterations, and the student can only return the corresponding prediction results to the teacher. For example, if the student uses a linear regression model, it will return to the teacher its prediction $\langle w^t, \tilde{x}\rangle$ where w^t is the student parameter at the t-th iteration and $\tilde{x}$ is the representation of the query example in student's feature space. Under suitable conditions, we show that the active teacher can always achieve faster rate of improvement than a random teacher that feeds samples randomly. In other words, the student model guided by the active teacher can provably achieve faster convergence than the stochastic gradient descent (SGD). Additionally, we discuss the extension of the active teacher to deal with the learner with forgetting behavior, and the learner guided by multiple teachers.

To validate our theoretical findings, we conduct extensive experiments on both synthetic data and real image data. The results show the effectiveness of the active teacher.

2. Related Work

Machine teaching defines a task where we need to find an optimal training set given a learner and a target concept. (Zhu, 2015) describes a general teaching framework which has nice connections to curriculum learning (Bengio et al., 2009) and knowledge distillation (Hinton et al., 2015). (Zhu, 2013) considers Bayesian learners in exponential family and formulates the machine teaching as an optimization problem over teaching examples that balance the future loss of the learner and the effort of the teacher. (Liu et al., 2016) give the teaching dimension of linear learners. Machine teaching has been found useful in cyber security (Mei & Zhu, 2015), human computer interaction (Meek et al., 2016), and human education (Khan et al., 2011). (Johns et al., 2015) extend machine teaching to human-in-the-loop settings. (Doliwa et al., 2014; Gao et al., 2015; Zilles et al., 2008; Samei et al., 2014; Chen et al., 2018) study the machine teaching problem from a theoretical perspective.

Previous machine teaching works usually ignore the fact that a student model is typically optimized by an iterative algorithm (*e.g.*, SGD), and in practice we focus more on how fast a student can learn from the teacher. (Liu et al., 2017a) propose the iterative teaching paradigm and an omniscient teaching model where the teacher knows almost everything about the learner and provides training examples based on the learner's status. Our cross-space teaching serves as a stepping stone towards the black-box iterative teaching.

3. Cross-Space Iterative Machine Teaching

The cross-space iterative teaching paradigm is different from the standard iterative machine teaching in terms of two major aspects: **i)** the teacher does not share the feature representation with the student; **ii)** the teacher cannot observe the student's current model parameter in each iteration. Specifically, we consider the following teaching settings:

Teacher. The teacher model observes a sample $\mathcal{A}$ (e.g. image, text, etc.) and represents it as a feature vector $x_{\mathcal{A}} \in \mathbb{R}^d$ and a label $y \in \mathbb{R}$. The teacher knows the model (*e.g.*, loss function) and the optimization algorithm (including the learning rate[1]) of the learner, and the teacher preserves an optimal parameter v^* of this model in its own feature space. We denote the prediction of the teacher as $\hat{y}_{v^*} = \langle v^*, x\rangle$[2].

Learner. The learner observes the same sample $\mathcal{A}$ and represents it as a vectorized feature $\tilde{x}_{\mathcal{A}} \in \mathbb{R}^s$ and a label $\tilde{y} \in \mathbb{R}$. The learner uses a linear model $\langle w, \tilde{x}\rangle$ where w is its model parameter and updates it with SGD (if guided by a passive teacher). We denote the prediction of the student model as $\hat{y}_w^t = \langle w^t, \tilde{x}\rangle$ in t-th iteration.

Representation. Although the teacher and learner do not share the feature representation, we still assume their representations have an intrinsic relationship. For simplicity, we assume there exists a unknown one-to-one mapping $\mathcal{G}$ from the teacher's feature space to the student's feature space such that $\tilde{x} = \mathcal{G}(x)$. However, the conclusions in this paper are also applicable to injective mappings. Unless specified, we assume that $y = \tilde{y}$ by default.

Interaction. In each iteration, the teacher will provide a training example to the learner and the learner will update its model using this example. The teacher cannot directly observe the model parameter w of the student. In this paper, the active teacher is allowed to query the learner with a few examples every certain number of iterations. The learner can only return to the teacher its prediction $\langle w^t, \tilde{x}\rangle$ in the regression scenario, its predicted label $\text{sign}(\langle w^t, \tilde{x}\rangle)$ or confidence score $S(\langle w^t, \tilde{x}\rangle)$ in the classification scenario, where w^t is the student's model parameter at t-th iteration and $S(\cdot)$ is some nonlinear function. Note that the teacher and student preserve the same loss function $\ell(\cdot, \cdot)$.

Similar to (Liu et al., 2017a), we consider three ways for the teacher to provide examples to the learner:

Synthesis-based teaching. In this scenario, the space of provided examples is

$$\mathcal{X} = \{x \in \mathbb{R}^d, \|x\| \leq R\}$$
$$\mathcal{Y} = \mathbb{R} \text{ (Regression) or } \{-1, 1\} \text{ (Classification).}$$

Combination-based teaching. In this scenario, the space

[1]For simplicity, the teacher is assumed to know the learning rate of the learner, but this prior is not necessary, as discussed later.

[2]For simplicity, we omit the bias term throughout the paper. It is straightforward to add them back.

of provided examples is ($\alpha_i \in \mathbb{R}$)

$\mathcal{X} = \{x | \, \|x\| \leq R, x = \Sigma_{i=1}^{k} \alpha_i x_i, x_i \in \mathcal{D}\}, \mathcal{D} = \{x_1, \ldots, x_k\}$

$\mathcal{Y} = \mathbb{R}$ (Regression) or $\{-1, 1\}$ (Classification)

Rescalable pool-based teaching. This scenario further restrict the knowledge pool for samples. The teacher can pick examples from $\mathcal{X} \times \mathcal{Y}$:

$\mathcal{X} = \{x | \, \|x\| \leq R, x = \gamma x_i, x_i \in \mathcal{D}, \gamma \in \mathbb{R}\}, \mathcal{D} = \{x_1, \ldots\}$

$\mathcal{Y} = \mathbb{R}$ (Regression) or $\{-1, 1\}$ (Classification)

We also note that the pool-based teaching (without rescalability) is the most restricted teaching scenario and it is very close to the practical settings.

4. The Active Teaching Algorithm

To address the cross-space iterative machine teaching, we propose the active teaching algorithm, which actively queries its student for its prediction output. We first describe the general version of the active teaching algorithm. Then without loss of generality, we will discuss three specific examples: least square regression (LSR) learner for regression, logistic regression (LR) and support vector machine (SVM) learner for classification (Friedman et al., 2001).

4.1. General Algorithm

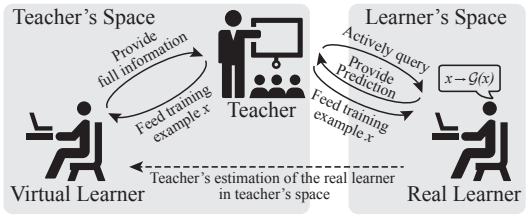

Figure 2: The cross-space teaching by active teacher. The real learner receives training example x but will perceive it as $\mathcal{G}(x)$.

Inspired by human teaching, we expand the teacher's capabilities by enabling the teacher to actively query the student. The student will return its predictions to the teacher. Based on the student's feedback, The teacher will estimate the student's status and determine which example to provide next time. The student's feedback enables the active teacher to teach without directly observing the student's model.

The active teacher can choose to query the learner with a few samples in each iteration, and the learner will usually report the prediction $F(\langle w, \tilde{x} \rangle)$ where $F(\cdot)$ denotes some function of the inner product prediction. For example, we usually have $F(z) = z$ for regression and $F(z) = \text{sign}(z)$ or $F(z) = \frac{1}{1+\exp(-z)}$ for classification. Based on our assumption that there is an unknown mapping from teacher's feature to student's feature, there also exists a mapping from the model parameters of the teacher to those of the student. These active queries enables the teacher to estimate the student's corresponding model parameter "in the teacher's space" and maintain a *virtual learner*, the teacher's estimation of the real learner, in its own space. The teacher will decide which example to provide based on its current virtual learner model. The ideal virtual learner v will have the same prediction output as the real learner, *i.e.* $\langle v, x \rangle = \langle w, \tilde{x} \rangle$ where $\tilde{x} = \mathcal{G}(x)$. Equivalently, $v = \mathcal{G}^{\top}(w)$ always holds for the ideal virtual learner, where $\mathcal{G}^{\top}$ is the conjugate mapping of $\mathcal{G}$. Note that for the purpose of analysis, we assume that $\mathcal{G}$ is a generic linear operator, though our analysis can easily extends to general cases. In fact, one of the most important challenges in active teaching is to recover a virtual student that approximates the real leaner as accurately as possible. The estimation error of the teacher may affect the quality of training examples that the teacher provides for the real learner. Intuitively, if we can recover the virtual learner with an appropriate accuracy, then we can still achieve faster teaching speed than that of passive learning. Fig. 2 shows the pipeline of the cross-space teaching.

With full access to the obtained virtual learner in the teacher's space, the teacher can perform omniscient teaching as in (Liu et al., 2017a). Specifically, the active teacher will optimize the following objective:

$$\underset{x \in \mathcal{X}, y \in \mathcal{Y}}{\operatorname{argmin}} \; \eta_t^2 \left\| \frac{\partial \ell(\langle v^t, x \rangle, y)}{\partial v^t} \right\|_2^2 - 2\eta_t \left\langle v^t - v^*, \frac{\partial \ell(\langle v^t, x \rangle, y)}{\partial v^t} \right\rangle \quad (1)$$

where ℓ is a loss function and v^t is the teacher's estimation of $\mathcal{G}^{\top}(w^t)$ after the teacher performs an active query in t-th iteration (*i.e.*, the current model parameter of the virtual learner). η_t is the learning rate of the virtual learner. The learning rate of the student model is not necessarily needed. The general teaching algorithm is given in Algorithm 1.

Particularly, different types of feedback (*i.e.*, the form of $F(\cdot)$) from learners contain different amount of information, resulting in different levels of difficulties in recovering the parameters of the learner's model. We will discuss two general ways to recover the virtual learner for two types of frequently used feedbacks in practice.

Exact recovery of the virtual learner. We know that the learner returns a prediction in the form of $F(\langle w, \tilde{x} \rangle)$. In general, if $F(\cdot)$ is an one-to-one mapping, we can exactly recover the ideal virtual learner (*i.e.* $\mathcal{G}^{\top}(w)$) in the teacher's space using the system of linear equations. In other words, the recovery of virtual learner could be exact as long as there is no information loss from $\langle w, \tilde{x} \rangle$ to $F(\langle w, \tilde{x} \rangle)$. Specifically, we have $\langle v, q_j \rangle = \langle w, \tilde{q_j} \rangle$ where q_j is the j-th query for the learner. Because $\langle w, \tilde{q_j} \rangle$ is given by the real learner, we only need to construct d queries (d is the dimension of the teacher space) and require $\{q_1, q_2, \cdots, q_d\}$ to be linearly independent to estimate v. Without no numerical error, we can exactly recover v. Since the recovery is exact, we have $\mathcal{G}^{\top}(w) = v$. Note that there are cases that we can achieve exact recovery without $F(\cdot)$ being an one-to-one mapping. For example, $F(z) = \max(0, z)$ (hinge function) is not an one-to-one mapping but we can still achieve exact recovery.

Approximate recovery of the virtual learner. If $F(\cdot)$ is not an one-to-one mapping (*e.g.*, sign$(\cdot)$ which provides

Algorithm 1 The active teacher

Randomly initialize the student parameter w^0;
Set $t = 1$, exam = True (*i.e.*, whether we make the student takes exams) and maximal iteration number T;
while v^t has not converged or $t < T$ **do**
  if $\mathcal{G}^\top \mathcal{G} \neq I$ and exam = True **then**
    Obtain an estimation $\hat{\mathcal{G}}^\top(w^t)$ of the student model in the teacher's space using the virtual learner construction Algorithm 2;
    $v^t = \hat{\mathcal{G}}^\top(w^t)$;
  else if $\mathcal{G}^\top \mathcal{G} = I$ and exam = True **then**
    Perform the one-time "background" exam using Algorithm 2 and set exam to False;
  end if
  Solve the optimization for the virtual learner (*e.g.* pool-based teaching):

$$(x^t, y^t) = \operatorname*{argmin}_{x\in\mathcal{X}, y\in\mathcal{Y}} \eta_t^2 \left\| \frac{\partial \ell\left(\left\langle v^{t-1}, x\right\rangle, y\right)}{\partial v^{t-1}} \right\|^2 - 2\eta_t \left\langle v^{t-1} - v^*, \frac{\partial \ell\left(\left\langle v^{t-1}, x\right\rangle, y\right)}{\partial v^{t-1}} \right\rangle$$

  if exam = False **then**
    Use the selected example (x^t, y^t) to perform the update of the virtual learner in the teacher's space:

$$v^t = v^{t-1} - \eta_t \frac{\partial \ell\left(\left\langle v^{t-1}, x^t\right\rangle, y^t\right)}{\partial v^{t-1}}.$$

  end if
  Use the selected example $(\tilde{x}^t, \tilde{y}^t)$ where $\tilde{x} = \mathcal{G}(x), \tilde{y} = y$ to perform the update of the real learner in the student's space:

$$w^t = w^{t-1} - \eta_t \frac{\partial \ell\left(\left\langle w^{t-1}, \tilde{x}^t\right\rangle, \tilde{y}^t\right)}{\partial w^{t-1}}.$$

  $t \leftarrow t + 1$;
end while

1-bit feedback), then generally we may not be able to exactly recover the student's parameters. Therefore, we have to develop a more intelligent technique (*i.e.* less sample complexity) to estimate $\mathcal{G}^\top(w)$. In this paper, we use active learning (Settles, 2010) to help the teacher better estimate $\mathcal{G}^\top(w)$ for the virtual learner. One of the difficulties is that the active learning algorithm obtains the parameters of a model based on the predicted labels on which the norm of the weights has no effect. It becomes ambiguous which set of weights the teacher should choose. Therefore, the active teacher also needs to have access to the norm of the student's weights for recovering the virtual learner. In the following sections, we will develop and analyze our estimation algorithm for the virtual learner based on the existing active learning algorithms with guarantees on sample complexity (Balcan et al., 2009; Ailon, 2012; Hanneke, 2007; Schein & Ungar, 2007; Settles, 2010).

4.2. Least Square Regression Learner

For the LSR learner, we use the following model:

$$\min_{w\in\mathbb{R}^s, b\in\mathbb{R}} \frac{1}{n}\sum_{i=1}^{n} \frac{1}{2}(\langle w, \tilde{x}_i\rangle - \tilde{y}_i)^2. \tag{2}$$

Algorithm 2 The virtual learner construction

if The feedback function $F(z)$ is an one-to-one mapping or a hinge function **then**
  Perform one-time exam by actively query multiple examples;
  Solve a system of linear equations to obtain the exact recovery of the ideal virtual learner;
else
  Apply acitve learning algorithms to perform an approximate recovery of the ideal virtual learner (in this case, the teacher will need to know the norm of the student model);
end if

Because $F(\langle w, \tilde{x}\rangle) = \langle w, \tilde{x}\rangle$, the LSR learner belongs to the case where the active teacher can exactly recover the ideal virtual learner. When $\mathcal{G}^\top\mathcal{G} = I$, the teacher only need to perform active exam once. It can be viewed as a "background exam" for the teacher to figure out how well the student has mastered the knowledge at the beginning, and the teacher can track the dynamics of students exactly later. Otherwise, for a general one-to-one mapping $\mathcal{G}$, the teacher needs to query the student in each iteration. Still, the teacher can reuse the same set of queries in all iterations.

4.3. Logistic Regression Learner

For the LR learner, we use the following model (without loss of generality, we consider the binary classification):

$$\min_{w\in\mathbb{R}^s, b\in\mathbb{R}} \frac{1}{n}\sum_{i=1}^{n} \log\left(1 + \exp\{-\tilde{y}_i(\langle w, \tilde{x}_i\rangle)\}\right) \tag{3}$$

We discuss two cases separately: (1) the learner returns the probability of each class (*i.e.* $F(z) = S(z)$ where $S(\cdot)$ denotes a sigmoid function); (2) the learner only returns the predicted label (*i.e.* $F(z) = \text{sign}(z)$).

In the first case where $F(\cdot)$ is a sigmoid function, we can exactly recover the ideal virtual learner. This case is essentially similar to the LSR learner where we need only one "background exam" if $\mathcal{G}^\top\mathcal{G} = I$ and we can reuse the queries in each iteration for a general one-to-one mapping $\mathcal{G}$ ($\mathcal{G}^\top\mathcal{G} \neq I$). In the second case where $F(\cdot)$ is a sign function, we can only approximate the ideal virtual learner with some error. In this case, we use active learning to do the recovery.

4.4. Support Vector Machine Learner

For the SVM learner, we use the following model for the binary classification:

$$\min_{w\in\mathbb{R}^s, b\in\mathbb{R}} \frac{1}{n}\sum_{i=1}^{n} \max(1 - y_i(w^T\tilde{x}_i + b), 0) \tag{4}$$

Similarly, we have two cases: (1) the learner returns the hinge value of each class (*i.e.* $F(z) = \max(0, z)$; (2) the learner only returns the label (*i.e.* $F(z) = \text{sign}(z)$).

In the first case where $F(\cdot)$ is a hinge function, we can still recover the ideal virtual learner. Although the hinge function is not a bijective mapping (only half of it is one-to-one),

we prove that it can still achieve exact recovery with slightly more query samples. For $\mathcal{G}^\top\mathcal{G} = I$, we need only one "background exam" as in the case of the LR learner. Otherwise, we still need to query the student in each iteration. In the second case where $F(\cdot)$ is a sign function, we can only approximate the ideal virtual learner with some error.

5. Theoretical Results

We define an important notion of being "exponentially teachable" to characterize the teacher's performance.

Definition 1 *Given $\epsilon > 0$, the loss function ℓ and feature mapping $\mathcal{G}$, $(\ell, \mathcal{G})$ is **exponentially teachable** (ET) if the number of total samples (teaching samples and query samples) is $t = \mathcal{O}(\text{poly}(\log \frac{1}{\epsilon}))$ for a learner to achieve ϵ-approximation, i.e., $\left\|\mathcal{G}^\top(w^t) - v^*\right\| \leq \epsilon$.*

Note that the potential dependence of t on the problem dimension is omitted here, which will be discussed in detail in the following. We summarize our theoretical results in Table 1. Given a learner that is exponentially teachable by the omniscient teacher, we find that the learner is not exponentially teachable by the active teacher only when $F(\cdot)$ is not an one-to-one mapping and the teacher uses rescalable pool-based teaching.

$F(\cdot)$	Synthesis teaching	Combination teaching	Rescalable pool teaching
One-to-one or hinge function	√	√	√
The other function	√	√	×

Table 1: The exponential teachability by active teacher. Assume that the learner is exponentially teachable by omniscient teacher.

5.1. Synthesis-Based Active Teaching

We denote $\sigma_{\max} = \max_{x^\top x=1} \mathcal{G}^\top(x)\mathcal{G}(x)$ and $\sigma_{\min} = \min_{x^\top x=1} \mathcal{G}^\top(x)\mathcal{G}(x) > 0$ ($\mathcal{G}$ is invertible). We first discuss the teaching algorithm when the teacher is able to exactly recover the student's parameters. A generic theory for synthesis-based ET is provided as follows.

Theorem 2 *Suppose that the teacher can recover $\mathcal{G}^\top(w^t)$ exactly using m samples at each iteration. If for any $v \in \mathbb{R}^d$, there exists $\gamma \neq 0$ and $\hat{y}$ such that $\hat{x} = \gamma(v - v^*)$ and*

$$0 < \gamma \nabla_{\langle v^t, \hat{x}\rangle} \ell\left(\langle v^t, \hat{x}\rangle, \hat{y}\right) < \frac{2\sigma_{\min}}{\eta\sigma_{\max}^2},$$

then $(\ell, \mathcal{G})$ is ET with $\mathcal{O}\left((m+1)\log\frac{1}{\epsilon}\right)$ samples.

Existence of the exponentially teachable $(\ell, \mathcal{G})$ via exact recovery. Different from (Liu et al., 2017a) where the condition for synthesis-based exponentially teaching is only related to the loss function ℓ, the condition for the cross-space teaching setting is related to both loss function ℓ and feature mapping $\mathcal{G}$. The spectral property of $\mathcal{G}$ is involved due to the differences of feature spaces, leading to the mismatch of parameters of the teacher and student. It is easy to see that $\exists\, \mathcal{G}$ such that the commonly used loss functions, *e.g.*, absolute loss, square loss, hinge loss, and logistic loss, are ET with exact recovery, i.e., $\mathcal{G}^\top(w^t) = v^t$. This can be shown by construction. For example, if the $\frac{\sigma_{\min}}{\sigma_{\max}^2} = \frac{1}{2}$, the ET condition will be the same for both omniscient teacher (Liu et al., 2017a) and active teacher.

Next we present generic results of the sample complexity m required to recover $\mathcal{G}^\top$, which is a constant to ϵ (i.e., $(\ell, \mathcal{G})$ is ET), shown as follows.

Lemma 3 *If $F(\cdot)$ is bijective, then we can exactly recover $\mathcal{G}^\top(w) \in \mathbb{R}^d$ with d samples.*

Lemma 4 *If $F(\cdot) = \max(0, \cdot)$, then we can exactly recover $\mathcal{G}^\top(w) \in \mathbb{R}^d$ with $2d$ samples.*

Lemma 3 and 4 cover $F(\cdot) = I(\cdot)$, $F(\cdot) = S(\cdot)$, or $F(\cdot) = \max(0, \cdot)$, where I denotes the identity mapping and S denotes some sigmoid function, *e.g.*, logistic function, hyperbolic tangent, error function, etc. If the student's answers to the queries via these student feedbacks $F(\cdot)$ in the exam phase, then we can exactly recover $v = \mathcal{G}^\top(w) \in \mathbb{R}^d$ with arbitrary d independent data, omitting the numerical error. Also note that the query samples in Lemma 3 and 4 can be reused in each iteration, thus the query sample complexity is $m = \mathcal{O}(d)$, which is formalized as follows.

Corollary 5 *Suppose that the student answers questions in query phase via $F(\cdot) = I(\cdot)$, $F(\cdot) = S(\cdot)$, or $F(\cdot) = \max(0, \cdot)$, then $(\ell, \mathcal{G})$ is ET with $\mathcal{O}\left(\log\frac{1}{\epsilon}\right)$ teaching samples and $\mathcal{O}(d)$ query samples via exact recovery.*

Here we emphasize that the number of query samples (*i.e.* active queries) does not depend on specific tasks. For both regression and classification, as long as the student feedbacks $F(\cdot)$ are bijective functions, then Corollary 5 holds. The loss function only affects the synthesis or selection of the teaching samples.

In both regression and classification, if $F(\cdot) = \text{sign}(\cdot)$ which only provides 1-bit feedback, F^{-1} no longer exists and the exact recovery of $\mathcal{G}^\top(w)$ may not be obtained. In such case, the teacher may only approximate the student's parameter using active learning. We first present the generic result for ET via approximate recovery as follows.

Theorem 6 *Suppose that the loss function ℓ is L-Lipschitz smooth in a compact domain $\Omega_v \subset \mathbb{R}^d$ containing v^* and sample candidates (x, y) are from bounded set $\mathcal{X} \times \mathcal{Y}$, where $\mathcal{X} = \left\{x \in \mathbb{R}^d, \|x\| \leq R\right\}$. Further suppose at t-th iteration, the teacher estimates the student $\epsilon_{est} := \left\|\mathcal{G}^\top(w^t) - v^t\right\| = \mathcal{O}(\epsilon)$ with probability at least $1-\delta$ using $m(\epsilon_{est}, \delta)$ samples. If for any $v \in \Omega_v$, there exists $\gamma \neq 0$ and $\hat{y}$ such that for $\hat{x} = \gamma(v - v^*)$, we have*

$$0 < \gamma \nabla_{\langle v^t, \hat{x}\rangle} \ell\left(\langle v^t, \hat{x}\rangle, \hat{y}\right) < \frac{2(1-\lambda)\sigma_{\min}}{\eta\sigma_{\max}^2},$$
$$\text{with } 0 < \lambda < \min\left(\frac{\kappa(\mathcal{G}^\top\mathcal{G})}{\sqrt{2}}, 1\right),$$

then the student can achieve ϵ-approximation of v^* with $\mathcal{O}\left(\log\frac{1}{\epsilon}\left(1 + m\left(\lambda\epsilon, \frac{\delta}{\log\frac{1}{\epsilon}}\right)\right)\right)$ samples with probability at least $1-\delta$. If $m(\epsilon_{est}, \delta) = \mathcal{O}(\log\frac{1}{\epsilon})$, then $(\ell, \mathcal{G})$ is ET.

Existence of exponentially teachable $(\ell, \mathcal{G})$ via approximate recovery. $m(\epsilon_{\text{est}}, \delta)$ is the number of samples needed for approximately recovering $\mathcal{G}^\top(w^t)$ in each iteration. Different from the exact recovery setting where m only depends on the feature dimension, $m(\epsilon_{\text{est}}, \delta)$ here also depends on how accurately the teacher wants to recover $\mathcal{G}^\top(w^t)$ in each iteration (ϵ_{est} denotes the estimation error of $\mathcal{G}^\top(w^t)$). The condition for exponentially teachable with approximate recovery is related to both $(\ell, \mathcal{G})$ and the approximation level of the student parameters, *i.e.*, the effect of λ. For example, if the $\frac{\sigma_{\min}}{\sigma^2_{\max}} = 1$ and $\lambda = \frac{1}{2}$, the exponentially teachable condition will be the same for both the omniscient teaching (Liu et al., 2017a) and active teaching with exact recovery.

For $F(\cdot) = \text{sign}(\cdot)$, if the student provides $\text{sign}(\langle w, \mathcal{G}(x)\rangle)$ for the query x, it is unlikely to recover $\mathcal{G}^\top(w)$ unless we know $\left\|\mathcal{G}^\top(w)\right\|$. This leads to the following assumption.

Assumption 1 *The feedback is 1-bit,* i.e. $F(\cdot) = \text{sign}(\cdot)$, *and the norm of $\mathcal{G}^\top(w)$ is known to teacher.*

Assumption 1 is necessary because $\text{sign}(\cdot)$ is scale invariant. We cannot distinguish between $\mathcal{G}^\top(w)$ and $k \cdot \mathcal{G}^\top(w)$ for any $k \in \mathbb{R}^+$ only with their signs. The following theorem provides the query sample complexity in this scenario.

Theorem 7 *Suppose that Assumption 1 holds. Then with probability at least $1-\delta$, then we can recover $\mathcal{G}^\top(w) \in \mathbb{R}^d$ with $\tilde{\mathcal{O}}\left(\left(d^2 + d\log\frac{1}{\delta}\right)\log\frac{1}{\epsilon}\right)$ query samples.*

Combining Theorem 6 with Theorem 7, we have the results for the 1-bit feedback case.

Corollary 8 *Suppose Assumption 1 holds. Then then $(\ell, \mathcal{G})$ is ET with $\mathcal{O}\left(\log\frac{1}{\epsilon}\right)$ teaching samples and $\tilde{\mathcal{O}}\left(\log\frac{1}{\epsilon}\log\frac{1}{\lambda\epsilon}\left(d^2 + d\log\frac{\log\frac{1}{\epsilon}}{\delta}\right)\right)$ query samples.*

Trade-off between teaching samples and query samples. There is a delicate trade-off between query sample complexity (in the exam phase) and teaching sample complexity. Specifically, with $\epsilon_{\text{est}} = \mathcal{O}\left(\frac{1}{t^2}\right)$ and $m\left(\mathcal{O}\left(\frac{1}{t^2}\right)\right)$ query samples, we can already achieve the conclusion that $\left\|\mathcal{G}^\top(w^{t+1}) - v^*\right\|^2$ converges in rate $\mathcal{O}\left(\frac{1}{t}\right)$, which makes the number of teaching samples to be $\mathcal{O}\left(\frac{1}{\epsilon^2}\right)$. We emphasize that this rate is the same with the convergence of SGD minimizing strongly convex functions. Note that the teaching algorithm can achieve at least this rate for general convex loss. Compared to the number of teaching samples in Corollary 8, although the query samples is less, this setting requires much more effort in teaching. Such phenomenon is reasonable in practice in the sense that if the examination is not accurate, the teacher provides the student less effective samples and hence has to teach for more iterations when the teacher cannot accurately evaluate student's performance.

We remark that if $\mathcal{G}$ is a unitary operator, *i.e.*, $\mathcal{G}^\top\mathcal{G} = I$, we can show that the teacher need only *one* exam. The key insight is that after the first "background exam", the teacher can replace the following exams by updating the virtual learner via the same dynamic of the real learner. This is formalized as follows.

Lemma 9 *Suppose that $\mathcal{G}$ is a unitary operator. If $\left\|\mathcal{G}^\top(w^0) - v^0\right\| \le \epsilon$, then $\left\|\mathcal{G}^\top(w^{t+1}) - v^{t+1}\right\| \le \epsilon$.*

Therefore, with a unitary feature mapping, we only need one exam in the whole teaching procedure. It follows that the query sample complexity in theorem 6 will be reduced to $\tilde{\mathcal{O}}\left(\log\frac{1}{\lambda\epsilon}\left(d^2 + d\log\frac{\log\frac{1}{\epsilon}}{\delta}\right)\right)$ via approximate recovery.

5.2. Combination-Based Active Teaching

We discuss how the results for synthesis-based active teaching can be extended to the combination-based active teaching. In this scenario, we assume both training and query samples are constructed by linear combination of k samples in $\mathcal{D} = \{x_i\}_{i=1}^k$. We have the following corollaries for both exact recovery and approximate recovery in the sense of

$$\langle v_1, v_2\rangle_{\mathcal{D}} := \sqrt{v_1^\top \mathcal{D}\left(\mathcal{D}^\top\mathcal{D}\right)^+ \mathcal{D}^\top v_2}, \text{ and}$$
$$\|v\|_{\mathcal{D}} := \langle v, v\rangle_{\mathcal{D}}.$$

Note that with the introduced metric, for $v \in \mathbb{R}^d$, we only consider its component in $\text{span}(\mathcal{D})$ and the components in the null space will be ignored. Therefore, $\forall\ v_1, v_2 \in \text{span}(\mathcal{D})$ such that $\|v_1\|_{\mathcal{D}} = \|v_2\|_{\mathcal{D}}$, we have $v_1^\top x = v_2^\top x = \langle v_1, x\rangle_{\mathcal{D}}$ for all $x \in \mathbb{R}^d$. Then we have the result via exact recovery as follows.

Corollary 10 *Suppose the learner gives feedbacks in query phase by $F(\cdot) = I(\cdot)$ or $F(\cdot) = S(\cdot)$, and $\mathcal{G}^\top(w^0), v^* \in \text{span}(\mathcal{D})$. Then $(\ell, \mathcal{G})$ is ET with $\mathcal{O}\left(\log\frac{1}{\epsilon}\right)$ teaching samples and $\text{rank}(\mathcal{D})$ query samples for exact recovery.*

The result via approximate recovery holds analogously to synthesis-based active teaching, given as follows.

Corollary 11 *Suppose Assumption 1 holds, the student answers questions in query phase via $F(\cdot) = I(\cdot)$ or $F(\cdot) = S(\cdot)$ and $\mathcal{G}^\top(w^0), v^* \in \text{span}(\mathcal{D})$. Then $(\ell, \mathcal{G})$ is ET with $\mathcal{O}\left(\log\frac{1}{\epsilon}\right)$ teaching samples and $\tilde{\mathcal{O}}\left(\log\frac{1}{\epsilon}\log\frac{1}{\lambda\epsilon}\left(d^2 + d\log\frac{\log\frac{1}{\epsilon}}{\delta}\right)\right)$ query samples via approximate recovery.*

5.3. Rescaled Pool-Based Active Teaching

In this scenario, the teacher can only pick examples from a fixed sample candidate pool, $\mathcal{D} = \{x_i\}_{i=1}^k$, for teaching and active query. We still evaluate with the metric $\|\cdot\|_{\mathcal{D}}$ defined in (5.2). We first define *pool volume* to characterize the richness of the pool (Liu et al., 2017a).

Definition 12 (Pool Volume) *Given the training example pool $\mathcal{X} \in \mathbb{R}^d$, the volume of $\mathcal{X}$ is defined as*

$$\mathcal{V}(\mathcal{X}) := \min_{w \in \text{span}(\mathcal{D})} \max_{x \in \mathcal{X}} \frac{\langle w, x\rangle_{\mathcal{D}}}{\|w\|_{\mathcal{D}}^2}.$$

Then the result via exact recovery is given as follows.

Theorem 13 *Suppose that the student answers questions in the exam phase via $F(\cdot) = I(\cdot)$ or $F(\cdot) = S(\cdot)$*

and $\mathcal{G}^\top(w^0), v^* \in \mathrm{span}(\mathcal{D})$. *If* $\forall\ \mathcal{G}^\top(w) \in \mathrm{span}(D)$, *there exist* $(x,y) \in \mathcal{D} \times \mathcal{Y}$ *and* γ *such that for* $\hat{x} = \frac{\gamma \|\mathcal{G}^\top(w) - v^*\|_{\mathcal{D}}}{\|x\|_{\mathcal{D}}} x$, $\hat{y} = y$, *we have*

$$0 \le \gamma \nabla_{\langle v^t, \hat{x}\rangle} \ell\left(\langle v^t, \hat{x}\rangle, \hat{y}\right) \le \frac{2\mathcal{V}(\mathcal{X})\,\sigma_{\min}}{\eta \sigma_{\max}^2},$$

then $(\ell, \mathcal{G})$ *is ET with* $\mathcal{O}\left(\log \frac{1}{\epsilon}\right)$ *teaching samples and* $\mathrm{rank}(\mathcal{D})$ *query samples.*

For the approximate recovery case, the active learning is no longer able to achieve the desired accuracy for estimating the student's parameter in the restricted pool scenario. Thus the active teacher may not achieve exponential teaching.

6. Discussions and Extensions

The active teacher need not know the learning rate. To estimate the learning rate, the active teacher should first estimate the student's initial parameters $w_1 \in R^d$, and then feed the student with one random sample (x_r, y_r). Once the updated student's parameter w_2 is estimated by the teacher, the learning rate η can be computed by $\eta = \frac{1}{d}\sum\left((w_1 - w_2)./\nabla_w \ell(w_1^T x, y)\right)$ where $./$ denotes the element-wise division and the sum is over all the dimensions in w_1. The number of samples for estimating η will be $2m+1$, where m denotes the samples used in estimating student's parameter. Even if the learning rate is unknown, the teacher only needs $2m+1$ more samples to estimate it. Most importantly, it will not affect the exponential teachability.

Teaching with forgetting. We consider the scenario where the learner may forget some knowledge that the teacher has taught, which is very common in human teaching. We model the forgetting behavior of the learner by adding a deviation to the learned parameter. Specifically in one iteration, the learner updates its model with $w^{t+1} = w^t + \nabla_w \ell(\langle w^t, x\rangle, y)$, but due to the forgetting, its truly learned parameter $\hat{w}^{t+1}$ is $w^{t+1} + \epsilon_t$ where ϵ_t is a random deviation vector. Based on Theorem 6, we can show that such forgetting learner is not ET with a teacher that only knows the learner's initial parameter and can not observe the learner along iteration. However, the active teacher can make the forgetting learner ET via the active query strategy. More details and experiments are provided in Appendix D.

Teaching by multiple teachers. Suppose multiple teachers sequentially teach a learner, a teacher can not guide the learner without knowing its current parameter. It is natural for the teacher to actively estimate the learner. Our active teaching can be easily extended to multiple teacher scenario.

7. Experiments

General settings. Detailed settings are given in Appendix B. We mainly evaluate the practical pool-based teaching (without rescaling) in the experiments. Still, in the exam stage, our active teacher is able to synthesize novel query examples as needed. The active teacher works in a different feature space from the learner's space, while the omniscient teacher (Liu et al., 2017a) can fully observe the learner and works in the same feature space as the learner. The omniscient teacher serves as a baseline (possibly an upper bound) in our experiments. For active learning, we use the algorithm in (Balcan et al., 2009; Schein & Ungar, 2007).

Evaluation. For synthetic data, we use two metrics to evaluate the convergence performance: the objective value and $\left\|\mathcal{G}^\top(w^t) - v^*\right\|_2$ w.r.t. the training set. For real images, we further use accuracy on the testing set for evaluation. We put the experiments of forgetting learner in Appendix D.

7.1. Teaching with Synthetic Data

We use Gaussian distributed data to evaluate our active teacher model on linear regression and binary linear classification tasks. We study the LRS learner with $F(\langle w, \tilde{x}\rangle) = \langle w, \tilde{x}\rangle$, LR learner with $F(\langle w, \tilde{x}\rangle)$ being the sigmoid function, LR learner with $F(\langle w, \tilde{x}\rangle) = \mathrm{sign}(\langle w, \tilde{x}\rangle)$. For the first two cases, the active teacher can perform an one-time exam ("background exam") to exactly recover the ideal virtual learner. After recovering the ideal virtual learner, the active teaching could achieve the performance of the omniscient teaching. The experimental results in Fig. 3(a) and Fig. 3(b) meet our expectations. In the initial iterations (on the order of feature dimensions), we can see that the learner does not update itself. In this stage, the active teacher provides query samples to the learner and recover a virtual learner based on the feedbacks of these query samples. After the exact recovery of the virtual learner, one can observe that the active teacher achieves faster convergence compared with the random teacher (SGD). In fact, the active teacher and the omniscient teacher should achieve the same convergence speed if omitting numerical errors.

For the LR learner with $F(\langle w, \tilde{x}\rangle) = \mathrm{sign}(\langle w, \tilde{x}\rangle)$, the teacher could only approximate the learner with the active learning algorithm. Besides, the active teacher needs to know the norm of the student model. We use the algorithm in (Schein & Ungar, 2007) and recover the virtual learner in each iteration such that $\|\hat{\mathcal{G}}^\top(w) - \mathcal{G}^\top(w)\|_2$ becomes small enough. From the results in Fig. 3(c), we can see that due to the approximation error between the recovered virtual learner and the ideal virtual learner, the active teacher can not achieve the same performance as the omniscient teacher. However, the convergence of the active teacher is very close to the omniscient teacher, and is still much faster than SGD. Note that, we remove the iterations used for exams to better compare the convergence of different approaches.

7.2. Teaching with Real Image Data

We apply the active teacher to teach the LR learner on the MNIST dataset (LeCun et al., 1998) to further evaluate the performance. In this experiment, we perform binary classification on the digits 7 and 9. We use two random projections to obtain two sets of 24-dim features for each image: one is for the teacher's feature space and the other is for the

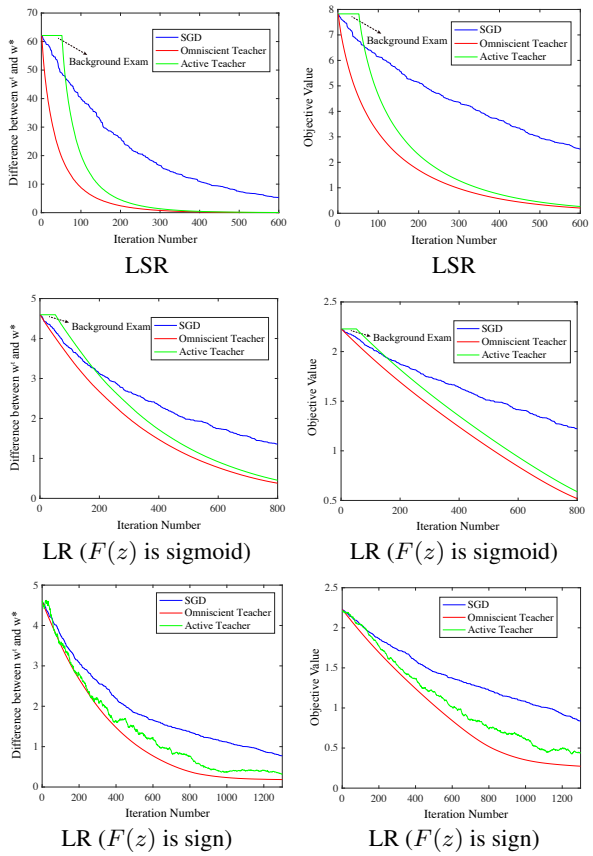

Figure 3: The convergence performance of random teacher (SGD), omniscient teacher and active teacher. As we need to perform the active query in each iteration for logistic regression ($F(z)$ is sign), we remove the iteration for fair comparison. We only show the teaching complexity for fair comparison.

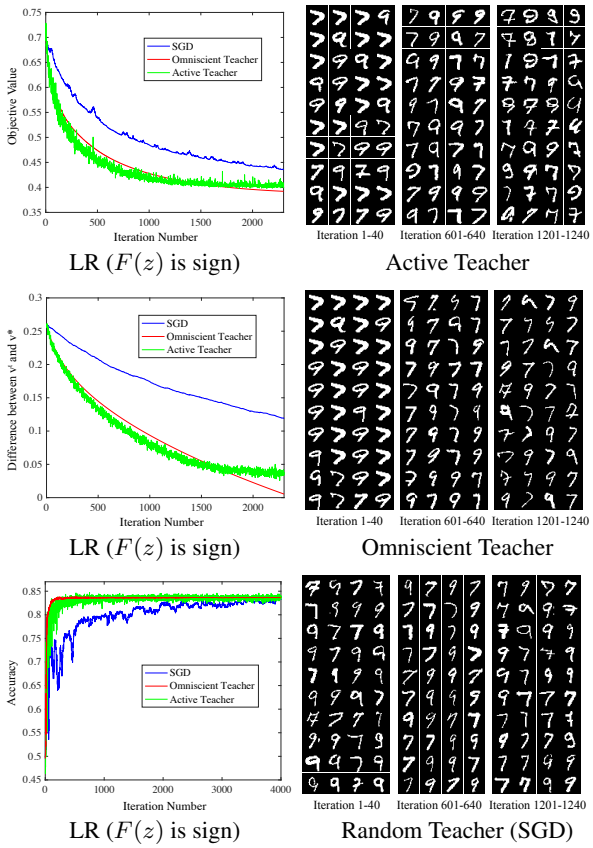

Figure 4: The convergence performance of random teacher (SGD), omniscient teacher and active teacher in MNIST 7/9 classification. Similar to the previous, we only show the teaching complexity for fair comparison. More experiments on the logistic regression with $F(z) = S(z)$ is in Appendix C.

student's feature space. The omniscient teacher uses the student's space as its own space (*i.e.*, shared feature space), while the active teacher uses different feature space with the student. For the LR learner with sign function (*i.e.* 1-bit feedbacks), one can observe that the active teacher has comparable performance to the omniscient teacher, even doing better at the beginning. Because we evaluate the teaching performance on real image data, the omniscient teacher will not necessarily be an upper bound of all the teacher. Still, as the algorithms iterate, the active teacher becomes worse than the omniscient teacher due to its approximation error.

In the right side of Fig.4, we visualize the images selected by the active teacher, omniscient teacher and random teacher. The active teacher preserves the pattern of images selected by the omniscient teacher: starting from easy examples first and gradually shifting to difficult ones, while the images selected by the random teacher have no patterns.

8. Conclusions and Open Problems

As a step towards the ultimate black-box machine teaching, cross-space teaching greatly relaxes the assumptions of previous teaching scenarios and bridges the gap between the iterative machine teaching and the practical world. The active teaching strategy is inspired by realistic human teaching. For machine teaching to be applicable in practice, we need to gradually remove all the unrealistic assumptions to obtain more realistic teaching scenario. The benefits of more realistic machine teaching are in two folds. First, it enables us make better use of the existing off-the-shelf pretrained models to teach a new model on some new tasks. It is also related to transfer learning (Pan & Yang, 2010). Second, it can improve our understanding on human education and provide more effective teaching strategies for humans.

Rescalable pool-based active teaching with 1-bit feedback. The proposed algorithm may not work the in pool-based teaching setting when the student return 1-bit feedback. We leave the possibility of achieving exponential teachability in this setting as an open problem.

Relaxation for the conditions on $\mathcal{G}$. Current constraints on the operator $\mathcal{G}$ are still too strong to match more practical scenarios. How to relax the conditions on $\mathcal{G}$ is important.

A better alternative to approximate recovery? Is there some other tool other than active learning for our teacher to recover the virtual learner? For example, 1-bit compressive sensing (Boufounos & Baraniuk, 2008) may help.

Acknowledgements

The project was supported in part by NSF IIS-1218749, NSF Award BCS-1524565, NIH BIGDATA 1R01GM108341, NSF CAREER IIS-1350983, NSF IIS-1639792 EAGER, NSF CNS-1704701, ONR N00014-15-1-2340, Intel ISTC, NVIDIA, and Amazon AWS.

References

Ailon, N. An active learning algorithm for ranking from pairwise preferences with an almost optimal query complexity. *Journal of Machine Learning Research*, 13(Jan):137–164, 2012.

Alfeld, S., Zhu, X., and Barford, P. Data poisoning attacks against autoregressive models. In *AAAI*, pp. 1452–1458, 2016.

Alfeld, S., Zhu, X., and Barford, P. Explicit defense actions against test-set attacks. In *AAAI*, 2017.

Balcan, M.-F., Beygelzimer, A., and Langford, J. Agnostic active learning. *Journal of Computer and System Sciences*, 75(1): 78–89, 2009.

Bengio, Y., Louradour, J., Collobert, R., and Weston, J. Curriculum learning. In *ICML*, 2009.

Boufounos, P. T. and Baraniuk, R. G. 1-bit compressive sensing. In *CISS*, 2008.

Chen, Y., Aodha, O. M., Su, S., Perona, P., and Yue, Y. Near-optimal machine teaching via explanatory teaching sets. In *AISTATS*, 2018.

Doliwa, T., Fan, G., Simon, H. U., and Zilles, S. Recursive teaching dimension, vc-dimension and sample compression. *Journal of Machine Learning Research*, 15(1):3107–3131, 2014.

Friedman, J., Hastie, T., and Tibshirani, R. *The elements of statistical learning*, volume 1. Springer series in statistics New York, 2001.

Gao, Z., Simon, H. U., and Zilles, S. On the teaching complexity of linear sets. In *International Conference on Algorithmic Learning Theory*, pp. 102–116. Springer, 2015.

Hanneke, S. A bound on the label complexity of agnostic active learning. In *Proceedings of the 24th international conference on Machine learning*, pp. 353–360. ACM, 2007.

Hinton, G., Vinyals, O., and Dean, J. Distilling the knowledge in a neural network. *arXiv preprint arXiv:1503.02531*, 2015.

Johns, E., Mac Aodha, O., and Brostow, G. J. Becoming the expert-interactive multi-class machine teaching. In *CVPR*, 2015.

Khan, F., Mutlu, B., and Zhu, X. How do humans teach: On curriculum learning and teaching dimension. In *NIPS*, 2011.

LeCun, Y., Bottou, L., Bengio, Y., and Haffner, P. Gradient-based learning applied to document recognition. *Proceedings of the IEEE*, 86(11):2278–2324, 1998.

Liu, J., Zhu, X., and Ohannessian, H. G. The teaching dimension of linear learners. In *ICML*, 2016.

Liu, W., Dai, B., Humayun, A., Tay, C., Yu, C., Smith, L. B., Rehg, J. M., and Song, L. Iterative machine teaching. In *ICML*, 2017a.

Liu, W., Zhang, Y.-M., Li, X., Yu, Z., Dai, B., Zhao, T., and Song, L. Deep hyperspherical learning. In *NIPS*, 2017b.

Liu, W., Liu, Z., Yu, Z., Dai, B., Lin, R., Wang, Y., Rehg, J. M., and Song, L. Decoupled networks. In *CVPR*, 2018.

Meek, C., Simard, P., and Zhu, X. Analysis of a design pattern for teaching with features and labels. *arXiv preprint arXiv:1611.05950*, 2016.

Mei, S. and Zhu, X. Using machine teaching to identify optimal training-set attacks on machine learners. In *AAAI*, 2015.

Nemirovski, A., Juditsky, A., Lan, G., and Shapiro, A. Robust stochastic approximation approach to stochastic programming. *SIAM Journal on optimization*, 19(4):1574–1609, 2009.

Pan, S. J. and Yang, Q. A survey on transfer learning. *IEEE Transactions on knowledge and data engineering*, 22(10):1345–1359, 2010.

Samei, R., Semukhin, P., Yang, B., and Zilles, S. Algebraic methods proving sauer's bound for teaching complexity. *Theoretical Computer Science*, 558:35–50, 2014.

Schein, A. I. and Ungar, L. H. Active learning for logistic regression: an evaluation. *Machine Learning*, 68(3):235–265, 2007.

Settles, B. Active learning literature survey. *University of Wisconsin, Madison*, 52(55-66):11, 2010.

Singla, A., Bogunovic, I., Bartók, G., Karbasi, A., and Krause, A. On actively teaching the crowd to classify. In *NIPS Workshop on Data Driven Education*, number EPFL-POSTER-221572, 2013.

Singla, A., Bogunovic, I., Bartok, G., Karbasi, A., and Krause, A. Near-optimally teaching the crowd to classify. In *ICML*, pp. 154–162, 2014.

Suh, J., Zhu, X., and Amershi, S. The label complexity of mixed-initiative classifier training. In *ICML*, pp. 2800–2809, 2016.

Zhu, X. Machine teaching for bayesian learners in the exponential family. In *NIPS*, 2013.

Zhu, X. Machine teaching: An inverse problem to machine learning and an approach toward optimal education. In *AAAI*, 2015.

Zhu, X., Singla, A., Zilles, S., and Rafferty, A. N. An overview of machine teaching. *arXiv preprint arXiv:1801.05927*, 2018.

Zilles, S., Lange, S., Holte, R., and Zinkevich, M. Teaching dimensions based on cooperative learning. In *COLT*, 2008.