Abstract—Machine teaching is an inverse problem of machine learning that aims at steering the student toward its target hypothesis, in which the teacher has already known the student’s learning parameters. Previous studies on machine teaching focused on balancing the teaching risk and cost to find the best teaching examples deriving from the student model. This optimization solver is in general ineffective when the student does not disclose any cue of the learning parameters. To supervise such a teaching scenario, this article presents a distribution matching-based machine teaching strategy via iteratively shrinking the teaching cost in a smooth surrogate, which eliminates boundary perturbations from the version space. Technically, our strategy could be redefined as a cost-controlled optimization process that finds the optimal teaching examples without further exploring the parameter distribution of the student. Then, given any limited teaching cost, the training examples would have a closed-form expression. Theoretical analysis and experiment results demonstrate the effectiveness of this strategy.

Index Terms—Learning parameters, machine teaching, surrogate, teaching cost, teaching risk.

I. INTRODUCTION

A rtificial intelligence (AI) [1] is a broad concept of machines being able to perform science and engineering tasks in ways that we would consider to be intelligent. To automatically improve AI through experience, machine learning [2] studies the advanced computation algorithms, which work on annotated training data, intending to explore the common patterns inside the data. Therefore, machine learning is one of the most important ways we expect to achieve AI. The core issue of machine learning is to construct effective models by inducing feasible parameters [3], where those highly informative or representative training examples accelerate the convergence of the learning model. In the past decades, the progress of machine learning went through about two stages: 1) the diversity of model construction, e.g., unsupervised paradigm [4], supervised paradigm [5], and semisupervised paradigm [6], and 2) the robustness of parameter optimization, e.g., grid search [7], Bayesian optimization [8], gradient descent search [9], and evolutionary optimization [10]. However, the optimal training subset still has no closed-form expression.

The machine learning community, thus, begins to solve this challenging problem. A novel concept termed machine teaching [11] was subsequently proposed. In such a scenario, the model construction process is redefined as a teacher–student interactive process, where the teacher provides necessary parameter information for the model construction of the student. A set of theoretical concepts, such as teaching complexity and teaching dimension [12], was proposed and given meaningful descriptions based on typical machine learning contents. Meanwhile, the core issue of machine teaching also was presented: how to control the machine learning paradigm, if there is a teacher who has already known the learning parameters of the student and wants better training examples to improve its generalization?

This question was studied well by [13], which explores the optimal training data via driving the student to its target hypothesis. It has been shown that many promising paradigms adopt such a teacher–student manner [14], e.g., a curriculum to the human education system in curriculum learning [15], an iterative query algorithm to an annotator in active learning [16]. The key assumption is that the teacher already knows the target parameters of the student model, such as a specified hyperplane in the support vector machine (SVM) classifier, and the geometric properties of clustering centers. Typically, a machine teacher interacts with their student by exploring those teaching examples that minimize the parameter disagreement (difference) of the current training model and its desired one [17], [18].

In simple terms, machine teaching can help to find the best training data for students automatically. With machine teaching, those teaching examples can be closed form to supervise the training model of a student, e.g., teacher–student learning [19] and knowledge distillation [20]. Furthermore,
one unsupervised teacher can find the best teaching examples as the clustering prototypes to depict the inherent skeleton of those clusters [21]; an annotation expert can minimize the labeling cost by picking up those highly informative teaching examples, which can maximally update the classification hypothesis [22]; a transfer learning task, which establishes a source paradigm using the teaching examples, can be adopted in homeomorphic target domains [23]. Moreover, machine teaching can help automate learning to encapsulate an advanced pipeline used in a class of consistent distributions and tasks [24], where automate learning is considered as one key step to achieve AI. As far as we know, the most relevant machine learning study to machine teaching is active learning [25]. Its key assumption is that a student who frequently interacts with an annotator would do better or no worse than other passive learners who randomly solicit the training examples. Generally, active learning forwardly updates the current training model into its target, while the target is always agnostic. We next employ a threshold classifier to further explain active learning and machine teaching.

Considering an interval \([-1, 1]\) with positive and negative labels over a uniform distribution \(P\) (see Fig. 1), the parameterized threshold classifier at \(\theta^* = 0\) stipulates the classification function \(f\) for any \((x, y) \sim P; y = f(x) = +1, \text{s.t. } x \geq \theta^*, y = f(x) = -1, \text{s.t. } x \leq \theta^*\), where \(y\) denotes the label of \(x\), and \(P\) denotes its distribution. Drawing i.i.d. \(n\) samples from \(P\), a passive (random) learner will make \(n\) times of querying from \(f(x)\) that yields a generalization error of \(\|\theta - \theta^*\| = O(1/n)\) due to an average uniform spacing \(1/n\), where \(\theta\) yields a passive learner. Given an error \(\epsilon\), a passive learner needs \(O(1/\epsilon)\) data sent to the classifier. Specifically, the passive learner at least receives 10,000 samples from \(P\) to obtain the desired error of 0.0001. For an active learner who usually employs binary search, the learner halves the remaining interval and removes the data over it, and thereby, receives around \(O(\log 1/\epsilon)\) samples to obtain an error \(\epsilon\) because \(\|\theta - \theta^*\| = O(1/2^n)\). Specifically, the active learner at least receives 13 samples from \(P\) to obtain the desired error of 0.0001. However, a machine teacher who knows \(\theta^*\) only needs two teaching examples to obtain \(\epsilon: (\theta^* - \epsilon/2, -1)\) and \((\theta^* + \epsilon/2, +1)\). Then, for any student, they will easily achieve better performance using those teaching examples.

The minimization on \(\|\theta - \theta^*\|\) is adopted for a linear learner. For a model-agnostic student in teaching, a general strategy is to minimize the teaching risk \(\mathcal{L}(\theta, \theta^*)\) over the student’s parameter disagreement. Its assumption is that the teacher knows \(\theta^*\) and the optimization on \(\mathcal{L}(\theta, \theta^*)\) is natural. However, if the student does not disclose any cue (also termed a black box) on \(\theta^*\), the estimation on \(\mathcal{L}(\theta, \theta^*)\) would be difficult even infeasible, and the teaching optimization would also explore in infinite parameter space. Theoretically, shrinking the training set into an approximately minimal subset for any unknown family of hypothesis is inevitable against a black box, yielding consistent properties as the full training [16]. In this article, we consider this challenging problem from a distribution perspective, which transforms the estimation of \(\mathcal{L}(\theta, \theta^*)\) to distribution matching, i.e., finding an alternative assumption for the transformation of the teaching risk. Motivated by this, we present a comprehensive theoretical analysis and optimization scheme. Contributions of this work are summarized as follows.

1) We introduce the idea of distribution matching for machine teaching a student who does not disclose any cue of the learning parameters, i.e., a black box.

2) We propose to teach a black box in a surrogate, which eliminates boundary perturbations from the version space, yielding consistent properties as with its original hypothesis class. To guarantee this scheme, we prove that teaching in surrogate has safety (w.r.t. Section V-A) and convergence (w.r.t. Section V-B) protocols.

3) We present a cost-controlled optimization (w.r.t. Section VI-B) process that finds the optimal teaching examples without further exploring the parameter distribution of the student.

4) We present a distribution matching-based machine teaching strategy (w.r.t. Section VI-C) by iteratively shrinking the teaching cost.

II. RELATED WORK

Machine teaching studies a complementary problem [26] of active learning: how to teach a student toward a target model with minimal effort. In active learning, the annotation effort is controlled by human experts or an oracle with label information. This usually leads to its target model is suboptimal. From this perspective, machine teaching studies the optimal control of active learning. Related works on the two avenues are presented in this section.

A. Machine Teaching

When the training set of a model consists of plenty of data from an underlying distribution, a teacher always desires to pick up some teaching examples from the original dataset to supervise its students [13]. In this teaching process, the teacher can be generalized as a human expert, a learning algorithm, or a system. Two factors are studied in the teaching process: how to fix the size of the teaching set, and whether the teacher knows the full knowledge of the parameter distribution of the student model?

Early works answered the first problem by studying the teaching complexity. Introducing a novel concept named teaching dimension [27], the complexity of teaching is usually characterized by the minimum number of teaching examples used to identify the desired hypothesis. If the teaching examples are selected independently of the original data pool, the student can cooperate with a teacher who supervises a teaching set. Recently, machine teaching [28] studied the teaching model when the student was a convex minimizer, such as the least square regression, and simple SVM. In [14], they presented a theoretical interpretation against multiple-teacher teaching, rather than a single teacher. Under the joint teaching for conjugate Bayesian learners, Zhu [29] proposed a new concept called class teaching dimension. In this definition, the teacher independently picks up some students as a representation of the whole class of learners. Indeed, it shrinks the original teaching dimension by the representation features of the learner class.
In real-world applications, a variety of learning tasks involving machine teaching were studied. For instance, in curriculum learning [15], [30], a human can intelligently annotate or recognize examples when the teaching examples are not randomly presented but organized in a meaningful order. Those orders help the teachers to optimize a group of teaching sets for the subsequent learning tasks, such as teacher–student training [19] and knowledge distillation [20], [31]. To improve the generalization of an active learner, Dasgupta et al. [16] proposed a black-box teaching scheme to shrink training sets for any family of the classifier by merely serving up the relevant examples beforehand and does not need to observe the feedback from the learner, where “shrink” can be deemed as a typical distribution matching-based machine teaching strategy.

B. Active Learning

Active learning [32], [33] adopts the same sampling goal as machine teaching to find the optimal training data, but forwardly updates the models. In this task, the learners are given access to interactively query the labels of a group of unlabeled data. The learning goal is that the queries can substantially improve the performance of a learning model within a given annotation budget. Theoretically, the learners always try to maintain a version space [34], which covers a series of candidate hypotheses and shrinks its size via querying as little as possible data. However, the version space-based learning theory, unfortunately, has drawbacks of computational intractability [25], i.e., guaranteeing that only hypotheses from this space are returned is intractable for nonlinear classifiers.

To develop a new strategy that addresses the above-mentioned limitations, Balcan [35] and Hanneke [36] constructed learning algorithms to predict which data may significantly affect the subsequent hypothesis, thereby giving different weight coefficients. The convergence guarantees, adopted from a PAC-style, are rigorous and tighter than the generalized bounds of any supervised learning algorithms. The other technique, termed importance-weighted active learning [37], provides an unbiased sampling approach with the loss weighting and has more practical use in observing the changes of generalization error and label complexity (the number of required samples to achieve the desired error).

In practical tasks, traditional active learning methods, such as pool-based active learning [38], [39], sample the data that reduces the error rate in the desired change by repeatedly visiting the unlabeled data pool. Usually, the learner is given access to the hypothesis class easily, i.e., can favorably observe the hypothesis updates by estimating the error disagreements (differences). However, when supervising a black-box learner, obtaining precise details of the prior labels and a list of classifier parameters are not available, and only the queries on labels are accessible. This makes many of the traditional strategies, which estimate the error disagreements, not applicable, or at the least unable to work well [40]. To reduce the dependence on a single classifier, query by committee algorithm [41] uses a set of classifiers to evaluate the error rate changes and selects the data which maximizes the disagreement among the committee members. However, estimating the error rate changes in settings of single or multiple classifiers will cost expensively on time and space complexities. Besides the fundamental active learning, there are also some related supervised and unsupervised learning paradigms that also can be adopted into finding the teaching set, e.g., linear discriminant analysis [42] and deep multimodal hashing [43], [44].

III. Motivations

Machine teaching mainly studies the optimal teaching set under the parameter distribution of the student. Section III-A presents the general optimization paradigm of teaching, and Section III-B proposes our proposal for teaching a black box.

A. Optimization of Teaching

In realizable scenarios, machine teaching usually discloses sufficient information for model selection and hyperparameter optimization, such as the linear model [45], sequential model [46], and Bayesian models of the exponential family [29]. From a machine learning perspective, (1) first presents a general strategy of active learning: optimizing a model parameter \( \theta \) with current training data \( D_0 \subset D \), subsequently updated by queries \( D_q \)

\[
\min_{\theta \in \Theta} R(\theta, D_0 \cup D_q) + \Omega(\theta)
\]

s.t. \( D_q \in \arg \max_{D_q \subset D \setminus D_0} \left| R(\theta, D_0) - R(\theta, D_0 \cup D_q) \right| \quad (1)
\]

where \( R(\cdot, \cdot) \) denotes the empirical risk function, \( \Omega(\cdot) \) denotes the regularization constraint, \( D \) denotes the training data, \( D_q \) denotes the queries from \( D \setminus D_0 \), and \( \Theta \) denotes the parameter space. By maximizing the risk disagreement in the second line of (1), \( D_q \) is selected as the update of \( D_0 \), and the training data is updated as \( D_0 \cup D_q \). Subsequent iterations also follow this rule until the learning model achieves the desired convergence or performance.

Given a teacher who has already known the student’s learning parameters, machine teaching inversely optimizes the training data by estimating parameter disagreements over the generalization models. Zhu [11] then proposed a more general machine teaching formula based on (1)

\[
\min_{\mathcal{E} \subseteq D} \mathcal{L}(\hat{\theta}, \theta^*) + \eta \mathcal{M}(\mathcal{E})
\]

s.t. \( \hat{\theta} \in \arg \min_{\theta \in \Theta} R(\theta, \mathcal{E}) + \Omega(\theta) \quad (2)
\]

where \( D \) denotes the complete or whole search space of \( D \) that covers all candidate training subsets (\( D \) may not be the optimal teaching set), \( \mathcal{E} \) denotes an element of \( D \), \( \mathcal{L}(\cdot, \cdot) \) denotes the teaching risk over the student’s parameter disagreement, \( \mathcal{M}(\cdot) \) denotes the teaching cost of the input set, and \( \eta \) denotes a balance coefficient. Typically, \( \mathcal{L}(\theta, \theta^*) \) can be simply defined as the indication disagreement \( E_{x \sim P} 1_{(\theta(x) \neq \theta^*(x))} \), where \( E \) denotes the indication function, and \( \theta \) denotes the desired parameter. Teaching cost can be simply defined as the cardinality of \( \mathcal{E} \): \( \mathcal{M}(\mathcal{E}) = |\mathcal{E}| \).
**B. Teaching a Black Box**

When the parameter information is limited, the student degenerates into a black box. How to find the optimal teaching set for curriculum learning without class information [47], active learning with weakly initial hypothesis [48], Bayesian inference with poor prior [49], [50], and so on, needs a new teaching strategy. Technically, our main proposal for solving (2) in a black-box scenario is to optimize an approximate distribution for teaching.

1) *Why Teach in a Surrogate:* From the perspective of hypothesis class, the teaching set is required to independently draw those significant hypotheses, which present effective model generalizations. Perturbations around the decision boundaries, including the noises [51], out-of-distribution data [52], and adversarial examples [53], usually produce insignificant hypotheses. It is, thus, smoothing those boundary perturbations that can substantially eliminate those unexpected insignificant hypotheses. Then, the teaching hypotheses generated from a surrogate with smooth boundary are required to tightly cover those significant hypotheses (see Definitions 1 and 2), which further results in safety protocols on generalization error and label complexity.

From the perspective of teaching optimization, the learning parameter is decided by its associated hypothesis, and any model-agnostic hypothesis is generated from an associated distribution or training set. We, thus, assume that \( \theta^* \) is generalized from the optimal hypothesis \( h^* \), i.e., \( R(\theta^*, D) = R(h^*, D) \). \( \hat{\theta} \) is generalized from the desired hypothesis \( h \), i.e., \( R(\hat{\theta}, D) = R(h, D) \). Let \( D' \subset D \) be the optimal surrogate with respect to \( h^* \). \( \hat{h} \subset D' \) be the desired training set with respect to \( h \), a distribution matching-based machine teaching algorithm which transfers the disagreement estimation of parameters into hypotheses, thereby approximating the hypothesis to distribution in metric space \( \mathbb{H} \), is then established

\[
\min_{\hat{\theta} \in \Theta} \| \hat{\theta} - \theta^* \| = \min_{h \in \mathbb{H}} \| \hat{h} - h^* \| \approx \min_{\hat{D}, D'} \| \hat{D} - D' \|_{\mathbb{H}} \tag{3}
\]

where \( \| \hat{D} - D' \|_{\mathbb{H}} := f_{\mathbb{H}}(\hat{D}, D') \), and \( f_{\mathbb{H}}(\cdot, \cdot) \) denotes a distribution metric. In other words, our purpose is to find a reasonable representation \( \hat{D} \) for \( D' \in D \). Therefore, we use the distribution metric \( f_{\mathbb{H}}(\cdot, \cdot) \) to minimize their distance.

2) *How to Teach in a Surrogate:* With the proposal of teaching in a surrogate, recalling (2), (3) is simplified into

\[
\mathcal{L}(\hat{\theta}, \theta^*) \approx f_{\mathbb{H}}(\hat{D}, D'). \tag{4}
\]

We next present the distribution matching-based teaching strategy to implement the transformation of (4). From an optimization perspective, there are two steps: 1) transform \( R(\theta, D) \) into a surrogate \( R(\theta, D') \) by shrinking \( D \) into its surrogate \( D' \) with smooth boundary and 2) then transform \( \mathcal{L}(\hat{\theta}, \theta^*) \) as a distribution metric optimized in \( D' \) that assumes \( \hat{D} \) is w.r.t. \( \hat{\theta} \) and \( D' \) is w.r.t. \( \theta^* \). From a hypothesis perspective, the alternatives steps are: 1) transform \( R(h, D) \) into a surrogate \( R(h, D') \) by shrinking \( D \) into its surrogate \( D' \) with smooth boundary and 2) then transform \( \mathcal{L}(h, h^*) \) as a distribution metric optimized in \( D' \) that assumes \( \hat{D} \) is w.r.t. \( h \) and \( D' \) is w.r.t. \( h^* \).

**IV. THEORETICAL FRAMEWORK OF TEACHING**

To guarantee the feasibility of the proposed teaching scheme in Section III-B, Section IV-A introduces importance sampling to shrink \( D \) into \( D' \) with \( T \) times of sampling, and Section IV-B presents preliminaries for safety analysis in hypothesis class.

A. *Importance Sampling*

Importance sampling [37] uses importance weighting to correct sampling bias and rigorously observe the online error change for an active learning model. Considering that active learning has a natural connection to machine teaching, we still follow this theoretical framework to present analysis for the shrinking of \( D \).

In importance sampling, the machine learning algorithm assigns an unlabeled data \( x_t \in D \) with a probability \( p_t \) to query its label \( y_t \), where \( p_t \) is a probability associated with the incremental updates of the hypothesis before \( t \)-time sampling and the underlying rule is: if \( x_t \) is selected for querying at \( t \)-time of sampling, its weight is set to \((1/p_t)\). Let \( \mathcal{Y} \) denote the label space of \( D \), \( h \) be a classification hypothesis \( h : D \times \mathcal{Y} \rightarrow r \), given \( f(\cdot) \) denotes a loss to measure the risk of classifying with \( h \), suppose \( \text{err}_D(h_T) \) denotes the expected error over \( D \) of a hypothesis \( h_T \) at query time \( T \), the learning risk with \( T \) times of sampling from \( D \) is defined as

\[
R(h_T, D) := \text{err}_D(h_T) = \frac{1}{T} \sum_{t=1}^{T} q_t f(h(x_t), y_t) \tag{5}
\]

where \( q_t \) denotes a Bernoulli distribution with \( q_t \in [0, 1], y_t \in \mathcal{Y} \), it denotes the true label of \( x_t \), and \( t \) iterates from 1 to \( T \). Importance sampling uses the probability weights to eliminate the sampling bias with rigorous label complexity bounds to accelerate the convergence of the minimization on \( R(h, D) \).

Remark 1: In optimization of machine teaching, \( R(\theta, D) \) denotes the risk of training on \( D \) with a given parameter \( \theta \in \Theta \). From a perspective of hypothesis, \( R(h, D) \) denotes the risk of training on \( D \) with a given hypothesis \( h \in \mathbb{H} \), which is characterized by \( \text{err}_D(h) \). With \( T \) times of importance sampling, the associated risk of hypothesis \( h_T \) is defined as \( R(h_T, D) \), which is then characterized by \( \text{err}_D(h_T) \).

B. *Safety Guarantee*

In importance sampling, teaching in surrogate \( D' \) can keep consistent properties for the machine learning model but eliminates the noisy perturbations from the boundary of the distribution, i.e., its conceptual version space. Theoretically, the desired safety guarantee [37] expects that the performance of a machine learning algorithm keeps a provable consistency on its inherent optimal hypothesis.

Given an agnostic distribution \( P \) maintaining a training set \( D \) for sampling, we define a subset \( D' \) with a smooth boundary as its surrogate. Following [37], Theorem 1 presents the risk disagreement between \( R(h, D) \) and \( R(h_T, D) \).
Theorem 1: Given a finite hypothesis class \( \mathcal{H} \) with a VC dimension bound \( d \) that is uniquely associated with \( D \). Assume that \( p_t > \phi \) for any \( 1 \leq t \leq T \), with a probability \( \delta > 0 \), then

\[
\Pr \left( \max_{h \in \mathcal{H}} | R(h_T, D) - R(h, D) | \leq \frac{\sqrt{2}}{\phi} \left( \frac{\ln d + \ln^2\delta}{T} \right) \right) \leq \delta
\]

(6)

where \( R(h, D) \) denotes the minimizer on \( R(h_T, D) \) given any \( T \).

The proof of [37, Th. 1] can be adopted here. With (6), a hypothesis \( h \) generated from \( \mathcal{H} \) needs to adopt a maximum risk disagreement of \( (\sqrt{2}/\phi)((\ln d + \ln(2/\delta))/T)^{1/2} \). This can be defined as the diameter [55] of hypothesis class \( \mathcal{H} \). Meanwhile, any of its covered surrogates from a distribution perspective also satisfies this property. To legalize a surrogate, the diameter disagreement of the original hypothesis class \( \mathcal{H} \) and its surrogate class \( \mathcal{H}' \) should be bounded by \( (\sqrt{2}/\phi)((\ln d + \ln(2/\delta))/T)^{1/2} \). Specifically, to free the varying parameter \( \phi \) and \( T \), we let \( \phi = (1/\sqrt{T}) \), the diameter disagreement \( (\sqrt{2}/\phi)((\ln d + \ln(2/\delta))/T)^{1/2} \) then relaxes into \( (2\ln d + 2\ln(2/\delta))/T)^{1/2} \).

To clarify the concept of diameter, the following presents a definition of the diameter of the hypothesis class.

**Definition 1 (Diameter of Hypothesis Class):** Given a hypothesis class \( \mathcal{H} \) over \( D \), its diameter is the maximum disagreement of any hypothesis \( h \) and \( h' \)

\[
d_{\mathcal{H}}(D) := \max_{h, h' \in \mathcal{H}} \left| \| \Theta h_D(h) - \Theta h_D(h') \| \right|_{\mathcal{H}}.
\]

(7)

**Diameter at t-Time:** Definition 1 presents a general definition on hypothesis diameter. At t-time, given \( h_t^+ \) and \( h_t^- \) be a subsequent hypothesis after adding and annotating \( x_t \) with ground truth and pseudolabel, respectively, the updated hypothesis diameter is written as

\[
d_{\mathcal{H}}(D) := \max_{h_t^+, h_t^- \in \mathcal{H}} \left| \| \Theta h_D(h_t^+) - \Theta h_D(h_t^-) \| \right|_{\mathcal{H}}.
\]

(8)

With the same annotation assumption, for a surrogate hypothesis class \( \mathcal{H}' \) over the surrogate \( D' \), the hypothesis diameter of \( \mathcal{H}' \) is written as

\[
d_{\mathcal{H}}(D') := \max_{h'_t^+, h'_t^- \in \mathcal{H}'} \left| \| \Theta h'_D(h'_t^+) - \Theta h'_D(h'_t^-) \| \right|_{\mathcal{H}'}.
\]

(9)

We, thus, are ready to present the definition for surrogate \( D' \).

**Definition 2 (Surrogate \( D' \) of \( D \)):** Given a finite hypothesis class \( \mathcal{H} \) with a VC dimension bound \( d \) that is uniquely associated with \( D \). Let \( D' \) be a surrogate of \( D \) and \( \mathcal{H}' \) be the shrunk hypothesis class over \( D' \), assume that \( d_{\mathcal{H}}(D) \) [w.r.t. (8)] and \( d_{\mathcal{H}'}(D') \) [w.r.t. (9)] are the hypothesis diameters (maximum hypothesis disagreement) of \( \mathcal{H} \) and \( \mathcal{H}' \), respectively, for any probability \( \delta \), surrogate \( D' \) is one subset from \( D \) with smooth boundary that tightly enclose \( D \), satisfying

\[
d_{\mathcal{H}}(D) - d_{\mathcal{H}'}(D') \leq \sqrt{2 \left( \ln d + \ln^2\delta \right)}
\]

s.t. \( D' \subseteq D \), \( |D'| \leq n' < |D| = n \).

(10)

In Definition 2, the surrogate \( D' \) has the consistent property of hypothesis diameter as with its original hypothesis class. However, will this surrogate always exist, and does teaching in surrogate has any sufficient analysis? We follow the consistency analysis of IWAL to explore a safety discussion. A preliminary definition of this safety concept is presented.

**Definition 3 (Safety Guarantee):** Given \( \mathbb{E}_{h \in \mathcal{H}} \left( R(h, D') \right) \) be the expected empirical risk over surrogate \( D' \), let \( d \) be the finite VC dimension bound that is uniquely associated with \( D \), for any probability \( \delta \), if

\[
\Pr \left( |R(h, D) - \mathbb{E}_{h \in \mathcal{H}} R(h, D')| \leq \sqrt{2 \left( \ln d + \ln^2\delta \right)} \right) \approx 1
\]

(11)

any machine learning model that minimizes \( R(h, D) \) is guaranteed safely on minimizing \( R(h, D') \).

V. MAIN THEORETICAL RESULTS

We present our main theoretical results in this section. Specifically, Section V-A provides the safety protocol for teaching in surrogate \( D' \), and Section V-B presents the convergence protocol by analyzing the label complexity bounds of minimizing \( R(h, D') \) over a surrogate. To improve the generalization of teaching in surrogate, Section V-C shrinks \( D \) into \( D' \) in hyperbolic geometry, and then presents a case study for the shrinking. Proof sketches are presented in the Appendix in the Supplementary Material.

A. Safety Protocol

Following Section IV-B, we present a safety protocol for teaching in surrogate. The goal is to prove that teaching in surrogate can keep consistent properties as with teaching in the original hypothesis class. To promote our theoretical analysis, a sufficient assumption on the learning risk \( R(h, D) \) is emphasized.

**Assumption 1:** With importance sampling, assume that \( \theta \in \Theta \) is respected to \( h \in \mathcal{H} \), recalling (4), transforming \( R(\theta, D) \) into \( R(\theta, D') \) is equivalent to transforming \( R(h, D) \) into \( R(h_T, D) \), where \( R(h_T D) \) denotes error risk of \( T \) times of importance sampling w.r.t. (5), \( R(h, D) \) denotes the risk of the hypothesis \( h \) over \( D \), which solicits the expectation over the loss measure \( f \), i.e., \( R(h, D) = \mathbb{E}_{i \in D} f(h(x_i), y_i) \) without importance sampling.

We next explain how teaching in a surrogate converges into accessible bound. Theorem 1 observes the ground-truth risk disagreement and its expectation, where the risk disagreement is over the full training data and its surrogate.

**Theorem 2:** With Assumption 1, given the training set \( D \), for all finite hypothesis class \( \mathcal{H} \) with a VC dimension bound \( d \), for any probability \( \delta > 0 \) and \( p_t > \phi \), if a learning algorithm samples \( T \) times to obtain a surrogate of \( D \), let

\[\text{2Vapnik-Chervonenkis dimension. It is a measure of the capacity, such as the complexity of a space of functions, that can be learned by a classification learning algorithm. In VC theory [54], the VC bound is defined as the cardinality of the largest set of input training data that a learning algorithm can shatter.}\]
R be the ground-truth risk disagreement of the surrogate and its full training data that stipulates $R = |R(h_T, D) − R(h, D)|$, $R$ be the expected risk disagreement that stipulates $:\hat{R} = E_{h \in \mathcal{H}}(R(h_T, D) − R(h, D))$, with Definition 3, and the generalization probability bound of achieving a safe surrogate is

$$\Pr \left\{ |R − \hat{R}| ≤ \sqrt{2 \left( \ln n + \frac{\ln 2}{\delta} \right)} \right\} \leq \exp \left( \frac{-4 \left( \ln n + \frac{\ln 2}{\delta} \right)}{T \phi^{-2}} \right). \quad (12)$$

In brief, Theorem 2 shows that there exist nearly consistent hypothesis diameters between the full training data and its surrogate, where the diameter of the surrogate is over its expectation. This demonstrates the consistency of Definitions 1 and 2. Corollary 1 presents a more accessible explanation for Theorem 2.

**Corollary 1:** In Theorem 2, $R$ denote the maximum risk disagreement between the $T$ times of importance sampling and the full training, and $\hat{R}$ denotes its expectation. For a given hypothesis class $\mathcal{H}$ which covers all feasible hypotheses, the maximum error disagreement is close to the hypothesis diameter $d_{\mathcal{H}}$ of $\mathcal{H}$. If the expected hypothesis distance of a subhypothesis class over $D'$ is close to it, we say sampling in $D'$ yields consistency as sampling in $D$. Therefore, with Definition 3, Theorem 1 has another equivalent form

$$\Pr \left\{ |d_{\mathcal{H}}(D) − E_{D' \subseteq D}d_{\mathcal{H}}(D')| ≤ \sqrt{2 \left( \ln n + \frac{\ln 2}{\delta} \right)} \right\} \approx 1. \quad (13)$$

Specifically, the probability bound of (13) tends to 1. With Assumption 1, transforming $R(h, D)$ into $R(h, D')$ achieves safety guarantee for any $h \in \mathcal{H'}$ over $D'$.

**B. Convergence Protocol**

Considering that teaching in surrogate can maintain consistent properties as with the original hypothesis class, Theorem 3 presents the convergence protocol on label complexity [35] of teaching in surrogate by minimizing the risk $R(h, D')$. Specifically, Assumption 2 presents the condition of the analysis and Theorem 3 presents the label complexity bound.

**Assumption 2:** Let $n$ denote the sample amount in $D$, its VC dimension bound $d$ is at most $2^n$. By using importance sampling, $D'$ is with a VC bound $2^T$ or $2^n$.

With Assumption 2, an upper bound of the label complexity of minimizing $R(h, D')$ is presented.

**Theorem 3:** Given the slope asymmetry $K_f$ that bounds the loss function $f(h(x), y)$ w.r.t. (5) for any hypothesis $h$ over $D'$: $K_f = sup_{x_i, x_i \in D}[((\max \ell(h(x_i), y) − \ell(h(x_i'), y))/(\min \ell(h(x_i), y) − \ell(h(x_i'), y))]$, considering a disagreement coefficient $\delta = sup_{x_i \in D} sup_{h \in \mathcal{H}(h, D)} \{(\ell(h(x_i), y) − \ell(h(x_i'), y))/r\}$, if the learning algorithm uses $\delta$ to smooth those data of $D$ with smaller hypothesis disagreements than $r$, with a probability $1 − \delta$, at $t$-time, minimizing $R(h, D')$ into $R(h^*, D')$, i.e., updating the current hypothesis $h$ into the optimal hypothesis $h^*$ in surrogate $D'$, costs at most $4\theta \times K_f \times (R(h^*, D') + 2((8/0.5 − t)\ln((2t^2 + t)(d/2^n − T))/\delta))^{1/2}$.

Note that $\delta$ is an error disagreement parameter used to perform the importance sampling. Any hypothesis $h$ holding a hypothesis disagreement smaller than $r$ in terms of $\delta$ will be considered as an insignificant hypothesis that presents insignificant influence for updating the current model, thereby being smoothed from the candidate hypothesis class. More related analysis based on this class of error disagreement parameters can refer to Hanneke et al.'s [36], [56] work.

Note that $K_f$ is a constant that satisfies $K_f ≥ 1$. Based on the importance sampling of [37], $K_f$ affects the label complexity bound due to its “sensitivity.” For example, given a 0–1 loss for $f(h(x), y)$, $K_f$ will be 1. However, for a hinge loss, $K_f$ will be $∞$. Therefore, for a sensitive loss function, the learning algorithm will require a large number of importance sampling times to obtain the desired hypothesis, then may lead to many ineffective queries. In other words, the sensitive loss function usually presents a coarse estimation of hypothesis disagreements. We here present a lemma to improve the generalization of $K_f$.

**Lemma 1:** Let $h$ be generalized as a logistic hypothesis that stipulates $\ell(h(x), y) := \ln(1 + e^{−2y})$, assume that the label space $\mathcal{Y} \in \{-1, +1\}$, if $x \in [-M, M]$, $K_f$ can be as large as $(1 + e^M)$.

**C. Case Study of Teaching in Surrogate**

To explain our theoretical results, we compare the clustering performance of teaching in the full training set and its surrogate.

1) **Density Estimation:** Based on the work of [57], perturbations around the boundary of cluster usually are characterized by low-density observations. Thus, we perform the density estimation of each data constrained within a fixed hypersphere

$$S := \{v, \forall v \in D, \text{s.t. } \psi(u, v) = 1\}$$

s.t. $\psi(u, v) = \begin{cases} 0, & f_{\mathcal{H}}(u, v) > r \\ 1, & 0 \leq f_{\mathcal{H}}(u, v) \leq r \end{cases}$

where $r$ denotes the radius of the $d$-dimensional hypersphere $S$ centered with $u$, $f_{\mathcal{H}}(u, v)$ denotes the distance metric. Then, a more general equation that applies hypersphere $S$ to observe the density on $u$ is presented

$$f_s(u) = \frac{1}{|S|} \sum_{v \in S} \frac{1}{\sqrt{2\pi d}} \exp \left[ -\frac{1}{2} \left( \frac{f_{\mathcal{H}}(u, v)}{r} \right)^2 \right]. \quad (15)$$

2) **Specification on $f_{\mathcal{H}}(u, v)$:** The Poincaré–distance [58] of hyperbolic (non-Euclidean) geometry has presented effective improvements in latent hierarchical tasks compared with the Euclidean distance, such as feature ranking [59], graph embedding [60], and non-Euclidean gradient descending [61]. Due to its nonzero curvature measure, it can individually separate and distinguish the features in a more discrete manner than the Euclidean distance with zero curvature. For example, given points of $p_1$, $p_2$, $p_3$, and $\lambda_E = (\|p_2 − p_1\|/\|p_3 − p_1\|) < 1$ in Euclidean geometry, there

Authorized licensed use limited to the terms of the applicable license agreement with IEEE. Restrictions apply.
exists $\lambda_E = ((f_{\mathbb{H}}(p_2, p_1))/f_{\mathbb{H}}(p_3, p_1)) < \lambda_E$ in hyperbolic geometry; if $\lambda_E = \|p_2 - p_1\|/\|p_3 - p_1\| > 1$ in Euclidean geometry, there exists $\lambda_E = (f_{\mathbb{H}}(p_2, p_1))/f_{\mathbb{H}}(p_3, p_1)) > \lambda_E$. With this special distance property, Poincaré distance would be a better generalization for $f_{\mathbb{H}}(\hat{D}, D')$ than the general Euclidean distance. To implement the iterative halving (w.r.t. Remark 2 of Section VI-B) of distribution matching-based machine teaching, we, thus, employ the Poincaré distance [58] to nonlinearly rank the transformed teaching risk. Specifically, let $B^d = \{x \in R^d, \|x\| < 1\}$ be an open $d$-dimensional unit Poincaré sphere, $u$ and $v$ be any two data in the sphere, i.e., $u, v \in B^d$, the Poincaré distance between them is defined as

$$f_{\mathbb{H}}(u, v) := \arccosh \left( 1 + 2 \frac{\|u - v\|^2}{(1 - \|u\|^2)(1 - \|v\|^2)} \right).$$

(16)

3) Clustering in Surrogate: The case study collects three real-world datasets and then compares the clustering performance of three typical clustering baselines in $D'$, where $f_S(u)$ is used to shrink $D$ into $D'$. To show the advantages of Poincaré distance in ranking, $f_S(u)$ is also generalized into $F_S(u)$ characterized with the Euclidean distance, which is further used to compare (14)

$$F_S(u) = \frac{1}{|S|} \sum_{i \in S} \frac{1}{\sqrt{2\pi} d} \exp \left[ -\frac{1}{2} \left( \frac{\|u - v\|_2}{r} \right)^2 \right].$$

(17)

4) Data Selection of the Case Study: Datasets of the case study are digit, USPS, and FashionMnist, where all the features of the data are scaled within a numerical unit of $10^{-5}$ to satisfy $\|x_i\|_2 < 1$, $\forall x_i \in D$. The sizes of these datasets are 3823 × 65, 9298 × 257, and 70 000 × 784, respectively. $R(\theta, D)$ is generalized as [1-adjusted rand index (ARI)] and [1-mutual information (MI)] coefficients.

5) Protocol of the Case Study: Fig. 2 presents the $R(\theta, D)$ values that yield: 1) minimizing $R(h, D)$ by clustering baselines and 2) minimizing $R(h, D')$ by clustering baselines with $F_S(u)$, and minimizing $R(h, D')$ by clustering baselines with $f_S(u)$, where the selected clustering baselines are the k-means, hierarchical, and spectral clustering algorithms, the parameters $r$ of (15) and (17) are defined as 0.4, and $|D'| - |D'| = 0.05n$, i.e., transform $D$ into $D'$ by eliminating perturbations from 0.05n boundary examples. Specifically, the kernel function of spectral clustering is set as RBF, driving a kernel parameter as 0.1 to construct an affinity matrix, where a k-means clustering is used to assign labels in the embedding space of the kernel.

6) Results of the Case Study: As shown in Fig. 2, we present the (1-ARI) and (1-MI) coefficients of clustering on the three datasets. On the x-axis of each subfigure, we perform the operation of clustering in $D$ and $D'$, respectively. To shrink $D$ into $D'$, density estimation is employed to remove boundary perturbations of $D$. Specifically, $D'$ with (15) denotes that the shrinking process invokes $f_S(\cdot)$ of (15), and $D'$ with (17) denotes that the shrinking process invokes $F_S(\cdot)$ of (17). There are two findings from the figures presented: 1) clustering in $D'$ has lower risk than that of the original $D$ in terms of their (1-ARI) and (1-MI) coefficients and 2) shrinking $D$ into $D'$ using $f_S(u)$ achieves lower risks than that of $F_S(u)$ due to its nonzero curvature measure on nonlinear ranking. This further demonstrates that the density observation employing the non-Euclidean distance may obtain a more accurate surrogate than the Euclidean distance in terms of eliminating those boundary perturbations.

VI. DISTRIBUTION MATCHING-BASED MACHINE TEACHING

Section VI-A generalizes the surrogate in algorithmic paradigm. Section VI-B presents the cost-controlled optimization scheme. Section VI-C describes the distribution matching-based machine teaching algorithm.

A. Generalization of Surrogate

When teaching a black box, any estimation on parameter disagreements will derive an insignificant hypothesis, and it is an invalid strategy. The theoretical results of Sections V-A and V-B have proved that teaching in surrogate can maintain safety protocol as with original hypothesis class and converge into finite label complexity bound, respectively.

With these theoretical guarantees, we begin to generalize surrogates for teaching a black box. Recalling (4), we here present a more formal assumption: transfer the disagreement estimation of parameters into hypotheses, thereby approximating the hypothesis to distribution.

Assumption 3: Assume that $\theta^*$ is generalized from the optimal hypothesis $h^*$, i.e., $R(\theta^*, D) = R(h^*, D)$, $\hat{D}$ is generalized from the hypothesis $\hat{h}$, i.e., $R(\hat{h}, D) = R(\hat{h}, D)$, let $D' \subset D$ be the optimal surrogate with respect to $h^*$, let $\hat{D} \subset D$ be the desired training set with respect to $\hat{h}$, with the proposal of (4), we further have

$$\min_{\theta \in \Theta} \|\hat{\theta} - \theta^*\| = \min_{h \in \mathcal{H}} \|\hat{h} - h^*\| \approx \min_{D' \subset D} \|\hat{D} - D'\|_{\mathbb{H}}$$

(18)

where $\|\hat{D} - D'\|_{\mathbb{H}} := f_{\mathbb{H}}(\hat{D}, D')$.

Another expression of (18) is $\mathcal{L}(\hat{\theta}, \theta^*) = \min_{\delta \in \Theta} \|\hat{\theta} - \theta^*\| \approx f_{\mathbb{H}}(\hat{D}, D')$. We, thus, have the following remark.

Remark 2: With Assumption 3, the optimization of distribution matching-based machine teaching is generalized into $\min_{\hat{D} \subset D} \|\hat{D} - D'\|_{\mathbb{H}}$. We, thus, iteratively halve $M(D')$, i.e., $|D'|$, which linearly reduces the teaching cost. With iterative halving, $M(D')$ varies from $n'/2$ into $n'/4$, $n'/8$, $n'/16$, . . . , $n'/2^L$, where $l$ denotes the halving frequency, $n'/2^l = [n'/2^l]$, and the remaining examples after the
Ith halving are the final teaching test if the student does not control the output \( M(D') \). Specifically, the halving process is implemented with the Poincaré distance of hyperbolic geometry. With iterative halving on \( D' \), the final update on \( D' \) is defined as the teaching set \( \hat{D} \).

B. Cost-Controlled Optimization

With Remark 2, the optimization of distribution matching-based machine teaching over-generalized \( \min_{\hat{D} \in D'} \| \hat{D} \subset D' \|_H \) is solved by controlling the teaching cost of \( D' \), that is, including a continuous algorithmic halving on \( D' \), where the final update on \( D' \) is the desired target \( \hat{D} \).

1) Machine Teaching a Black Box: Following Assumption 3, the parameter \( \theta \) can be associated with the training set \( D \). The typical machine teaching algorithm, thus, begins by generalizing \( R(\theta, D) \). In machine teaching, the parameter disagreement \( \| \hat{\theta} - \theta^* \| \) is with a linear relationship. To generalize our assumption, the risk function also follows this construction manner. Therefore, let \( D' = \{ x_1, x_2, \ldots, x_n \} \) with \( n' \) samples, where \( x_i \) denotes the ith sample, considering a linear relationship of \( R(\theta, x_i) := \theta x_i + e_i \) given a bias \( e_i \), machine teaching with a black box is to optimize

\[
\min_{D' \subset D} \left\{ \| \hat{\theta} - \theta^* \|^2 := \sum_{i=1}^{n'} \| \hat{\theta} x_i - R(\theta, x_i) \|^2 \right\}
\]

s.t. \( \hat{\theta} \in \arg \min_{\theta \in \Theta} R(\theta, D) \) (19)

where minimizing the parameter disagreement of \( \| \hat{\theta} - \theta^* \| \) presents a formal description for machine teaching. If the student is a black box who does not disclose any cue of the distribution of the parameters, we follow Assumption 3 to transform the parameter estimation into distribution matching, that is, \( \min_{\hat{D} \in D'} \sum_{i=1}^{n'} \| \hat{\theta} x_i - R(\theta, x_i) \|^2 \). Recalling (1), we add a regularization constraint \( \Omega(\theta) := \| \theta \|^2 \) to improve the generalization ability of (19), that is, reducing its risk of converging with overfitting

\[
\min_{D' \subset D} \left\{ \| \theta - \theta^* \|^2 + \Omega(\theta) := \sum_{i=1}^{n'} \| \hat{\theta} x_i - R(\theta, x_i) \|^2 + \| \theta \|^2 \right\}
\]

s.t. \( \hat{\theta} \in \arg \min_{\theta \in \Theta} R(\theta, D) + \eta \| \theta \|^2 \). (20)

Remark 3: Assumption 3 has presented our main proposal of teaching a black box, i.e., estimating the parameter disagreement can be transferred into distribution matching, which requires redefining \( R(\theta, x_i) \) into distribution metric of \( x_i \) over \( D' \). Let \( \theta = \{ \alpha_1, \alpha_2, \ldots, \alpha_n \} \), the relationship between \( \theta \) and \( x_i \) is transformed as the calculation between \( \alpha_j, j = 1, 2, 3, \ldots, n' \) and \( x_i \), where \( \alpha_j \) is associated with the distribution metric of \( x_j \) to \( x_i \). Then, we have \( R(\theta, x_i) = \sum_{j=1}^{n'} \alpha_j f_H(x_j, x_i) \). To find the optimal \( \hat{D} \) in \( D' \), we invoke an iterative halving strategy which stipulates \( R(\theta, x_i) = \sum_{j=1}^{n'/2} \alpha_j f_H(x_j, x_i) \) in the iterations, where halving is a theoretically feasible strategy in distribution shattering [48]. If we shatter more samples to match the distribution in each iteration, the final update on \( D' \) does not easily converge into the desired teaching set. If we shatter fewer samples to match the distribution in each iteration, the convergence of the optimization costs excessively.

2) Iterative Halving-Based Cost-Control: Equation (18) has defined the optimization goal of machine teaching a black box. With Remark 3, we use iterative halving to control the teaching cost to find a final update of the teaching set. Specifically, we generalize \( f_H(\cdot, \cdot) \) as Poincaré distance based on the results of the case study of Section V-C. Proposition 1 then presents the specification.

Proposition 1: Based on Assumption 3, estimating the parameter disagreement can be transferred into distribution matching. We next introduce the Poincaré distance \( f_H(\cdot, \cdot) \) as the risk metric, which estimates error risk by distribution distance between \( x_i \) to halved \( D' \), that is \( R(\theta, x_i) := \sum_{j=1}^{n'/2} f_H(x_j, D') = \sum_{j=1}^{n'/2} \alpha_j f_H(x_j, x_i) \), where \( \theta \) is associated with \( \alpha_i \). With halving operation, \( \hat{\theta} x_i \) then can be defined as \( \hat{\theta} x_i := f_H(\theta, D') = \sum_{j=1}^{n'/2} \alpha_i f_H(x_j, x_i) \). Then, taking a regularized least squares, \( \| \theta \|^2 \) is transformed as \( \| \theta \|^2 = \sum_{j=1}^{n'/2} \alpha_j^2 \). Due to the above-mentioned transformations, (20) is redefined as

\[
\min_{\hat{\theta}} \sum_{i=1}^{n'/2} \sum_{j=1}^{n'/2} \| \hat{\theta} f_H(x_j, x_i) - \alpha_i f_H(x_j, x_i) \|^2 + \eta \sum_{i=1}^{n'/2} \alpha_i^2 \| \theta \|^2
\]

where \( \theta = \{ \alpha_1, \alpha_2, \ldots, \alpha_{n'/2} \} \).

In the rest of this article, we redefine \( D \) as both the set of \( \{ x_1, x_2, \ldots, x_n \} \) and the matrix \( [x_1, x_2, \ldots, x_n]^T \in \mathbb{R}^{n \times dim} \), where \( dim \) denotes the dimensions. Then, \( D' \) and \( \hat{D} \) also follow such dual expressions. Let \( \hat{D} = \{ \hat{x}_1, \ldots, \hat{x}_{n'/2} \} \), also with \( \hat{D} = \{ \hat{x}_1, \ldots, \hat{x}_{n'/2} \}^T \), there exists \( \hat{H}_{ij} := f_H(x_i, x_j) \), and then \( \hat{H}_{D'D'} \| \hat{D} = \| \hat{D} = n'/2 \). With [62, Definition 3.1], (21) is transferred as \( \min_{D' \subset D'} \| H_{D'D'}(\cdot, \cdot) - \eta C \|^{-1} H_{D'D} \) that can be solved by transductive optimization

\[
\min_{D' \subset D'} \text{Tr} \left[ H_{D'D'}(H_{D'D'} + \eta C)^{-1} H_{D'D} \right]
\]

where \( C \propto D'D'^T \). In transductive optimization, \( C \) usually is approximated by \( D'D'^T \) regularized by \( \eta \), which bridges the incremental matrix of \( H_{D'D'} \) and \( H_{D'D} \), where \( \propto \) denotes “proportional.” To optimize \( \hat{D} \), let \( D \) be the last selected teaching example, \( \hat{D} \) subsequently is obtained by

\[
\hat{D} := \arg \min_{D' \subset D'} \text{Tr} \left[ H_{D'D'}(H_{D'D} + \eta C)^{-1} H_{D'D} \right]
\]

where \( H = H - H_{D'D'}(H_{D'D} + \eta C)^{-1} H_{D'D} \).

C. Distribution Matching-Based Machine Teaching Algorithm

Our distribution matching-based machine teaching algorithm is presented in Algorithm 1. Based on our experience
in cluster boundary detection (see [57, Fig. 1]), the number of those cluster boundary points (i.e., the data outside the surrogate), usually occupy about 5%–30% of the entire dataset. Besides this, our previous work of [63] and [64] also has consistent insights. Note that cluster boundary detection is an explainable scenario for those data outside the surrogate of the full training data. We, hence, set $\mathcal{M}(D') = n' = 0.95|D|$ as a default constraint, where $l$ denotes the frequency of performing the halving operation on $D'$. Lines 2–4 transform $R(\theta, D)$ into $R(\theta, D')$ by shrinking $D$ into its surrogate $D'$. Lines 5–12 perform the iterative halving process on $D'$. The final update on $D'$ after $l$ times halving is the machine teaching set $\hat{D}$. If the student controls the output $\mathcal{M}(\hat{D})$, k-medoids are performed on the final update of $\hat{D}$ to satisfy the student’s request.

1) Computational Complexity: In the proposed algorithm, shrinking $D$ into its surrogate $D'$ w.r.t. lines 2–4 costs $\mathcal{O}(n'n)$, and the iterative halving w.r.t. lines 5–12 costs $\mathcal{O}(n'^2)$. It is, thus, the computational complexity of Algorithm 1 is about $\mathcal{O}(n'n)$.

2) Convergence: Lines 5–12 of Algorithm 1 perform the iterative halving by invoking an argmin optimization. After $l$ times of halving, line 13 outputs the final teaching set with a size of $\mathcal{M}(\hat{D})$. It is, thus, Algorithm 1 can properly converge within a computational complexity of $\mathcal{O}(n'n)$.

VII. EXPERIMENTS

Typical machine teaching algorithms estimate the parameter disagreement of models to generalize the teaching risk, where the teacher knows the desired parameter. When teaching a black box, parameter estimations may be inefficient due to improper parameter disagreement or inestimable parameter space. We, thus, select a series of supervised and unsupervised machine learning baselines, which can be generalized as typical machine teaching, to compare our distribution-based machine teaching algorithm.

To solve (2) of general machine teaching, there exist three conditions which can simplify its optimization process: 1) control $\mathcal{M}(D)$ with (1)’s solver of active learning; 2) reduce the search space for limited risk minimization; and 3) fix $\mathcal{M}(D)$ with unsupervised machine learning. To realize these conditions, three groups of experiments are presented.

1) Datasets: The datasets used in the first two groups of experiments are the full training data of Adult, Phishing, Satimage, and MNIST datasets, where $||R(\theta, D) - R(\theta^*, D)||$ is over those training data. The sizes of these datasets are $11055 \times 68$, $1605 \times 14$, $4435 \times 36$, and $60000 \times 780$, respectively. The datasets used in the third group of the experiment are CIFAR10 and CIFAR100, where $||R(\theta, D) - R(\theta^*, D)||$ is over their test data. The sizes of the two datasets are all $60000$ with $32 \times 32$ pixels.

2) Baselines: Four supervised learning algorithms that regulate $\mathcal{M}(D)$ to minimize $||R(\theta, D) - R(\theta^*, D)||$ are selected, including expected error reduction (ERR) [65], prelearning [66], transductive experimental design (TED) [62], and self-paced active learning (SPAL) [67]. Specifically, they are active learning algorithms. Three typical unsupervised machine learning algorithms that minimize $||R(\theta, D) - R(\theta^*, D)||$ with quantitative $\mathcal{M}(D)$ are selected: k-medoids, hierarchical, and spectral clustering. Those baselines are finally used in the experiment of teaching a deep neural network. A case study of teaching Gaussian data is first presented before the experiments. Note that distribution-based machine teaching is denoted as DM-based machine teaching in all experimental figures.

3) Specification of $||R(\theta, D) - R(\theta^*, D)||$: Our main assumption is to transfer $||R(\theta, D) - R(\theta^*, D)||$ into distribution matching. To verify the effectiveness of our machine teaching strategy, we specify $||R(\theta, D) - R(\theta^*, D)||$ over the ground-truth error risk. Specifically, given a classification model, we set $R(\theta, D) = 1 - err_D(\theta)$ and $R(\theta^*, D) = 1 - err_D(\theta^*)$, where $\theta$ and $\theta^*$ are generated from the teaching set and full training set, respectively. In short, $R(\theta^*, D)$ and $R(\theta, D)$ are specified as the optimal and current classification errors, respectively.

A. Case Study: Teaching on a 2-D Gaussian Dataset

Fig. 3 presents a case study of distribution matching-based machine teaching on the 2-D Gaussian dataset. The 2-D visualizations dynamically show the iterative halving process on surrogate $D'$: 1) Fig. 3(a) draws the full training Gaussian data $D$, where $\mathcal{M}(D) = 1400$; 2) Fig. 3(b) draws the surrogate of $D$, where the circled 170 data are boundary.
examples, the remaining blue points are the data of $D'$, and the parameter settings are $r = 0.4$, $\eta = 1.0e^{-4}$; and 3) Fig. 3(c)-(h) shows the iterative halving process, where $M(D')$ is continuously halved. The presented teaching sets with different $M(D')$ properly match the distribution of $D$ without noisy perturbations around the boundary.

Specifically, all the teaching examples are distributed inside the clusters with high densities due to the smooth boundary of the surrogate $D'$ (w.r.t. lines 1–4 of Algorithm 1). The iterative halving (w.r.t. lines 5–12) is performed on the last update of $D'$, which keeps consistent distribution properties as its previous. Therefore, all the teaching sets with different $M(D')$ yield consistent distributions as the original distribution of $D$.

B. Regulate $M(D)$ to Minimize $\| R(\theta, D) - R(\theta^*, D) \|$  

Regulating $M(D)$ is important for both machine teacher and student due to over-fitting or computational overhead. With Assumption 3, the goal of machine teaching is to minimize $\| R(\theta, D) - R(\theta^*, D) \|$, where $\theta^*$ is with respect to $D$. The experimental datasets are Adult, Phishing, Satimage, and MNIST. We assume that $h$ is generated from an SVM classifier with an RBF kernel. That means, $R(\theta^*, D) = 0.1200$ on Adult, $R(\theta^*, D) = 0.1141$ on Phishing, $R(\theta^*, D) = 0.0875$ on Satimage, and $R(\theta^*, D) = 0.0009$ on MNIST, where each $h^*$ is over the full training data.

The compared machine learning baselines are typical active learning algorithms, including ERR, preclustering, TED, and SPAL, where ERR maximizes the EER over an SVM classifier, preclustering employs the hierarchical clustering and the pruning budget is set from 100 to 1000 with a step of 100, TED uses a hyperparameter $\sigma = 1.8$ (kernel bandwidth parameter) to generate the kernel matrix and varies $\lambda$ (kernel ridge regression) from 0.01 to 1 with a step of 0.01, and SPAL sets the paced learning parameter from 0.01 to 1 with a step of 0.01, and so on. Before running those machine learning baselines, we randomly select ten data from each dataset to train an initial hypothesis for them. For our distribution matching-based machine teaching algorithm, $M(D') = n'$, $n'/n \in [0.85, 0.95]$, $r = [0.1, 0.5]$ $\eta = 10e^{-4}$, and $l$ is constrained by $M(D')$ that satisfies $(n'/2^l) \geq M(D')$.

Fig. 4 draws the learning curves of regulating $M(D)$ to minimize $\| R(\theta, D) - R(\theta^*, D) \|$ into an expected risk across the best parameter candidates of each baseline. From the test results in Fig. 4, we find that the machine teaching algorithm can regulate $M(D)$ better than the machine learning baselines, i.e., spend fewer training data to obtain an expected learning risk. Especially at the beginning of those curves, $M(D)$ of machine teaching is much smaller than that of the machine learning baselines.

C. Shrink the Search Space of $D$

Shrinking the search space of $D$ can relieve the minimization costs of (1) and (2) because estimating whether one data can be picked up as a teaching example needs to access the whole unlabeled data pool. If the teacher needs to give feedback to the student in a limited budget cost, e.g., time...
and space, the teaching algorithm must help the teacher make a decision on which example should be selected.

Given an access budget of \( \lceil n/2 \rceil \) to the unlabeled data pool one time, the machine learning and teaching algorithms have to return one best candidate teaching example. Fig. 5 draws the \( \mathcal{M}(D) \) curves by progressively minimizing \( \| R(\theta, D) - R(\theta^*, D) \| \). Compared with the results in Fig. 4 with a full access budget, the \( \mathcal{M}(D) \) of machine learning baselines arising rapidly due to the greedy updates on \( \theta \) to \( \hat{\theta} \) cannot always be optimal. This forces the learning algorithm to request more data to reach an expected learning risk. However, our machine teaching algorithm backward and iteratively halves \( \theta^* \) to \( \hat{\theta} \) without greedy search in \( D \), and thereby, lower perturbations to the limited access budget are presented. Therefore, the generalized distribution matching-based machine teaching algorithm could trust a black-box student with inestimable teaching loss in real-teaching tasks.

D. Minimize \( \| R(\theta, D) - R(\theta^*, D) \| \) With Quantitative \( \mathcal{M}(D) \)

Optimizing (1) and (2) with a quantitative \( \mathcal{M}(D) \) is also a possible condition to simplify the minimization process. Then, (2) can be solved in an unsupervised way. Therefore, unsupervised machine learning algorithms, such as clustering, can be deemed as a special class of candidate teaching methods with quantitative \( \mathcal{M}(D) \).

In this group of experiments, we collect the learning risk change of \( \| R(\theta, D) - R(\theta^*, D) \| \) with quantitative \( \mathcal{M}(D) \) settings. The compared three unsupervised algorithms are k-medoids, hierarchical, and spectral clustering, where \( \mathcal{M}(D) \) is set as the clustering numbers. Specifically, the kernel function of spectral clustering is set as RBF, driving a kernel parameter as 0.1 to construct an affinity matrix, where a k-means clustering is used to assign labels in the embedding space of the kernel. The whole collected teaching results of

Authorized licensed use limited to the terms of the applicable license agreement with IEEE. Restrictions apply.
the three baselines are drawn in Fig. 6. We intuitively find

they show sensitive change on the performance of all the clustering algorithms is very unstable. In real-world scenarios, realizable h may have perturbations to this assumption. In other terms, the test set of Adult and Phishing since they are binary classification datasets without strong clustering structures. We also find the hierarchical clustering algorithm cannot decrease \(\|R(\theta, D) - R(\theta^*, D)\|\) when setting \(\mathcal{M}(D)\) lower than 700. This is because the two datasets have no intuitive tree structures. For the Satimage and MNIST datasets with clear clustering structure, unsupervised machine learning algorithms achieve better performance on minimizing \(\|R(\theta, D) - R(\theta^*, D)\|\), even better than machine teaching on Satimage.

Overall, the unsupervised machine learning approaches can be applied in teaching a black box, but show very unstable performance in minimizing learning risk due to their local convergence conditions. A global strategy should be considered to minimize \(\|R(\theta, D) - R(\theta^*, D)\|\) with quantitative \(\mathcal{M}(D)\). This also is the inherent reason why our proposed distribution-based algorithm can be adopted in machine teaching with inestimable teaching risk.

To visualize the distribution of the output teaching set of distribution matching-based machine teaching algorithm, Fig. 7 presents the 2-D embeddings of distribution matching-based machine teaching on MNIST with different \(\mathcal{M}(D')\). The results show those teaching sets can properly draw the ten separable classes.

E. Teach a Deep Neural Network

We compare the deep learning performance of our distribution matching-based machine teaching algorithm to the supervised and unsupervised machine learning models. Fig. 8 presents the learning curves of regulating \(\mathcal{M}(D)\) to minimize \(\|R(\theta, D) - R(\theta^*, D)\|\) following the experiments of Sections VII-A and VII-B. The deep neural network is ResNet20 and the tested datasets are CIFAR10 and CIFAR100. The hyperparameters of the network architecture are batch size = 32, epochs = 200, depth = 20, learning rate = 0.001, filter number = 16, and so on. The network architecture was implemented by Keras 2.2.3. The results show our machine teaching algorithm can still minimize \(\|R(\theta, D) - R(\theta^*, D)\|\) faster than the compared supervised and unsupervised machine learning baselines, where \(R(\theta^*, D) = 0.9200\) over CIFAR10 and \(R(\theta^*, D) = 0.6729\) over CIFAR100. Fig. 9 presents the 2-D embeddings of distribution matching-based machine teaching sets on CIFAR10 with different \(\mathcal{M}(D')\).

VIII. DISCUSSION ON OUR ASSUMPTION

This study is based on Assumption 3 that stipulates \(R(\theta^*, D) ≈ R(h^*, D)\), in which how to derive a more general hypothesis h may have perturbations to this assumption. In real-world scenarios, realizable h is usually generalized from different classifiers. We, thus, collect diverse classifiers to test the main technical steps of iterative halving in distribution matching-based machine teaching.

With this goal, we apply our machine teaching algorithm to derive teaching sets as labeled data for the subsequent supervised classification. The candidate classifiers are k neighbors regressor (KNR), random forest (RandomForest), multilayer perceptron classifier (MLPClassifier), and SVM. Fig. 10 presents the learning curves of regulating \(\mathcal{M}(D)\) to minimize \(\|R(\theta, D) - R(\theta^*, D)\|\) on digit, USPS, and FashionMnist datasets, where \(\theta^*\) is respect to \(h^*\) generalized from different classifiers, and “∼” denotes “with.”
The training parameters of the four classifiers are described as follows: 1) we set the $k$ nearest number as 10 for KNR; 2) we set the number of trees in the RandomForest as 600; 3) for MLPClassifier, we set the size of the hidden layers as 100, the maximum iteration number as 1000, the $L_2$ penalty (regularization term) parameter as 0.0004, the optimization strategy as stochastic gradient descent, and the learning rate as 0.001; and 4) for the SVM classifier, we set the penalty parameter as 1.0, the kernel type as RBF, the degree of the kernel function as 3, and the tolerance for stopping criterion as 0.003.

Distribution matching-based machine teaching significantly reduces the expected learning risks with the increase of $M(D)$. In the reported results of the four classifiers, perturbations of $\|R(\theta, D) - R(\theta^*, D)\|$, that is,

$$\left( \max R(\theta^*, D) - R(\theta, D) \right) - \left( \min R(\theta^*, D) - R(\theta, D) \right)$$

which yields an interval of $[0.06, 0.14]$, where $M(D)/n$ yields an interval of $[0.1, 0.5]$. Particularly, a part of the learning curves does not keep consistently decreasing with the increase of $M(D)/n$, such as MLPClassifier on digit and RandomForest on FashionMNIST. This explores that the iterative halving may delay the decrease of the expected learning risks, which further reduces the perturbations. Therefore, our machine teaching algorithm keeps a uniform decrease in expected learning risks, which also may delay its decrease to a lower loss. Besides this, SVM achieves the lowest $\|R(\theta, D) - R(\theta^*, D)\|$ and cooperates with a solid classifier may further delay the decrease of the learning risks.

**Remark 4:** Note that with the increase of $M(D)$, $D$ will be tightly approximated by the incremental update on $D$, which leads $\|R(\theta, D) - R(\theta^*, D)\|$ to be nearly zero. Generally, machine teaching aims to find an optimal teaching set to guide the student, where $M(D) < n$.

**IX. CONCLUSION**

In this work, we proposed a distribution matching-based machine teaching algorithm with regard to estimating the teaching risk on distributions against a black box. Theoretical analysis proved that teaching in a surrogate has safety and convergence protocols as the original hypothesis space. The case study further presented support for this theoretical perspective and demonstrated that the Poincaré distance of hyperbolic geometry could derive smoother boundaries for teaching in surrogate than that of the Euclidean distance. We, thus, projected the subsequent iterative halving in this geometry. Experiments demonstrated distribution matching-based machine teaching outperformed the supervised and unsupervised machine learning algorithms in minimizing expected learning risk disagreement. Finally, this work leads to an open question: can we co-teach the disagreement estimations on the distribution and model parameters? Furthermore, more practical teacher-student learning models in real-world tasks, such as computer vision, would be studied in the future.

**REFERENCES**

[1] M. A. Boden, “Creativity and artificial intelligence,” *Artif. Intell.*, vol. 103, nos. 1–2, pp. 347–356, 1998.

[2] T. M. Mitchell, *Machine Learning*. McGraw-Hill, 1997.

[3] M. I. Jordan and T. M. Mitchell, “Machine learning: Trends, perspectives, and prospects,” *Science*, vol. 349, no. 6245, pp. 255–260, 2015.

[4] T. Hofmann, “Unsupervised learning by probabilistic latent semantic analysis,” *Mach. Learn.*, vol. 42, no. 1, pp. 177–196, Jan. 2001.

[5] R. Caruana and A. Niculescu-Mizil, “An empirical comparison of supervised learning algorithms,” in *Proc. 23rd Int. Conf. Mach. Learn. (ICML)*, 2006, pp. 161–168.

[6] X. Zhu and A. B. Goldberg, “Introduction to semi-supervised learning,” *Synth. Lect. Artif. Intell. Mach. Learn.*, vol. 4, no. 1, pp. 1–130, 2009.

[7] J. Bergstra and Y. Bengio, “Random search for hyper-parameter optimization,” *J. Mach. Learn. Res.*, vol. 13, no. 2, pp. 1–25, 2012.

[8] J. Snoek, H. Larochelle, and R. P. Adams, “Practical Bayesian optimization of machine learning algorithms,” in *Proc. NIPS*, 2012, pp. 1–9.

[9] R. Johnson and T. Zhang, “Accelerating stochastic gradient descent using predictive variance reduction,” in *Proc. Adv. Neural Inf. Process. Syst.*, vol. 26, 2013, pp. 315–323.

[10] T. Bäck and H.-P. Schwefel, “An overview of evolutionary algorithms for parameter optimization,” *Evol. Comput.*, vol. 1, pp. 1–23, Dec. 1993.

[11] X. Zhu, “Machine teaching: An inverse problem to machine learning and an approach toward optimal education,” in *Proc. 29th AAAI Conf. Artif. Intell.*, 2015, pp. 1–5.

[12] S. A. Goldman and M. J. Kearns, “On the complexity of teaching,” *J. Comput. Syst. Sci.*, vol. 50, no. 1, pp. 20–31, 1995.

[13] X. Zhu, A. Singla, S. Zilles, and A. N. Rafferty, “An overview of machine teaching,” 2018, arXiv:1801.05927.

[14] W. Liu, B. Dai, X. Li, Z. Liu, J. M. Rehg, and L. Song, “Towards black-box iterative machine teaching,” in *Proc. 35th Int. Conf. Mach. Learn. (ICML)*, Stockholm, Sweden, Jul. 2018, pp. 3147–3155.

[15] T. Matiisen, A. Oliver, T. Cohen, and I. Schulman, “Teacher-student curriculum learning,” *IEEE Trans. Neural Netw. Learn. Syst.*, vol. 31, no. 9, pp. 3732–3740, Sep. 2020.

[16] S. Dasgupta, D. Hsu, S. Pouils, and X. Zhu, “Teaching a black-box learner,” in *Proc. 36th Int. Conf. Mach. Learn. (ICML)*, Long Beach, CA, USA, Jun. 2019, pp. 1547–1555.

[17] Z. Gao, C. Ries, H. U. Simon, and S. Zilles, “Preference-based teaching,” *J. Mach. Learn. Res.*, vol. 18, no. 1, pp. 1012–1043, 2017.

[18] A. Shinohara and S. Miyano, “Teachability in computational learning,” in *New Generat. Comput.*, vol. 8, no. 4, pp. 337–347, Feb. 1991.

[19] T. Gu, Z. Wang, X. Xu, D. Li, H. Yang, and W. Du, “Frame-level teacher-student learning with data privacy for EEG emotion recognition,” *IEEE Trans. Neural Netw. Learn. Syst.*, early access, Apr. 29, 2022, doi: 10.1109/TNNLS.2022.3168935.

[20] S. Li et al., “Distilling a powerful student model via online knowledge distillation,” *IEEE Trans. Neural Netw. Learn. Syst.*, early access, Mar. 7, 2022, doi: 10.1109/TNNLS.2022.3152710.

[21] V. Hauthamaki, P. Nykanen, and P. Franti, “Time-series clustering by approximate prototypes,” in *Proc. 19th Int. Conf. Pattern Recognit.*, Dec. 2008, pp. 1–4.

[22] M. Gao, Z. Zhang, G. Yu, S. O. Ark, L. S. Davis, and T. Pfister, “Consistency-based semi-supervised active learning: Towards minimizing labeling cost,” in *Proc. Eur. Conf. Comput. Vis. Cham, Switzerland: Springer*, 2020, pp. 510–526.

[23] B. Peng, C. Li, J. Li, S. Shyamdeh, L. Liden, and J. Gao, “SOLIOIST: Building task bots at scale with transfer learning and machine teaching,” 2020, arXiv:2005.05298.

[24] D. Weitkamp, E. Harpstead, and K. R. Koedinger, “An interaction design for machine teaching to develop AI tutors,” in *Proc. CHI Conf. Hum. Factors Comput. Syst.*, Apr. 2020, pp. 1–11.

[25] S. Dasgupta, D. J. Hsu, and C. Monteileoni, “A general agnostic active learning algorithm,” in *Proc. Adv. Neural Inf. Process. Syst.*, 2008, pp. 353–361.

[26] T. Pelinola, M. M. Celikok, P. Daes, and S. Kaski, “Machine teaching of active sequential learners,” in *Proc. Adv. Neural Inf. Process. Syst.*, 2019, pp. 1–12.

[27] F. Khan, B. Mutlu, and J. Zhu, “How do humans teach: On curriculum learning and teaching dimension,” in *Proc. Adv. Neural Inf. Process. Syst.*, 2011, pp. 1449–1457.

[28] S. Mei and X. Zhu, “Using machine teaching to identify optimal training-set attacks on machine learners,” in *Proc. 29th AAAI Conf. Artif. Intell.*, 2015, pp. 1–7.

[29] J. Zhu, “Machine teaching for Bayesian learners in the exponential family,” in *Proc. Adv. Neural Inf. Process. Syst.*, 2013, pp. 1905–1913.
Y. Bengio, J. Louradour, R. Collobert, and J. Weston, “Curriculum learning,” in Proc. 26th Annul. Int. Conf. Mach. Learn., 2009, pp. 41–48.

Y. Yu, B. Li, Z. Ji, J. Han, and Z. Zhang, “Knowledge distillation classifier generation network for zero-shot learning,” IEEE Trans. Neural Netw. Learn. Syst., early access, Sep. 2, 2021, doi: 10.1109/TNNLS.2021.3112229.

D. Colin, L. Atlas, and R. Ladner, “Improving generalization with active learning,” Mach. Learn., vol. 15, no. 2, pp. 201–221, 1994.

B. Gu, Z. Zhou, C. Deng, and H. Huang, “Efficient active learning by querying discriminative and representative samples and fully exploiting unlabeled data,” IEEE Trans. Neural Netw. Learn. Syst., vol. 32, no. 9, pp. 4111–4122, Sep. 2021.

A. Beygelzimer, D. J. Hsu, J. Langford, and T. Zhang, “Agnostic active learning without constraints,” in Proc. Adv. Neural Inf. Process. Syst., 2010, pp. 99–106.

M.-F. Balcan, A. Beygelzimer, and J. Langford, “Agnostic active learning,” in Proc. 23rd Int. Conf. Mach. Learn. (ICML), 2006, pp. 65–72.

S. Hanneke, “A bound on the label complexity of agnostic active learning,” in Proc. 24th Int. Conf. Mach. Learn., Jun. 2007, pp. 353–360.

A. Beygelzimer, S. Dasgupta, and J. Langford, “Importance weighted active learning,” in Proc. 26th Annu. Int. Conf. Mach. Learn., Jun. 2009, pp. 49–56.

S. Tong and D. Koller, “Support vector machine active learning with applications to text classification,” J. Mach. Learn. Res., vol. 2, pp. 45–66, Nov. 2002.

D. Wu, “Pool-based sequential active learning for regression,” IEEE Trans. Neural Netw. Learn. Syst., vol. 30, no. 5, pp. 1348–1359, May 2019.

N. Rubens, V. Steinman, R. Tomioka, and M. Sugiyama, “Active learning in black-box settings,” Austrian J. Statist., vol. 40, nos. 1–2, pp. 125–135, 2011.

H. S. Seung, M. Opper, and H. Sompolinsky, “Query by committee,” in Proc. 5th Annu. Workshop Comput. Learn. Theory, Jul. 1992, pp. 287–294.

C. Yan et al., “Self-weighted robust LDA for multiclass classification with edge classes,” ACM Trans. Intell. Syst. Technol., vol. 12, no. 1, pp. 1–19, Feb. 2021.

X. Lu, L. Liu, L. Nie, X. Chang, and H. Zhang, “Semantic-driven interpretable deep multi-modal hashing for large-scale multimedia retrieval,” IEEE Trans. Multimedia, vol. 23, pp. 4541–4554, 2021.

T. Hoang, T.-T. Do, T. V. Nguyen, and N.-M. Cheung, “Multimodal mutual information maximization: A novel approach for unsupervised deep cross-modal hashing,” IEEE Trans. Neural Netw. Learn. Syst., early access, Jan. 4, 2022, doi: 10.1109/TNNLS.2021.3135420.

J. Liu and X. Zhu, “The teaching dimension of linear learners,” J. Mach. Learn. Res., vol. 17, no. 1, pp. 5631–5655, 2016.

L. Lessard, X. Zhang, and X. Zhu, “An optimal control approach to sequential machine teaching,” in Proc. 22nd Int. Conf. Artif. Intell. Statist., 2019, pp. 2495–2503.

G. Hacohen and D. Weinshall, “On the power of curriculum learning in training deep networks,” in Proc. Int. Conf. Mach. Learn., 2019, pp. 2535–2544.

X. Cao and I. W. Tsang, “Shattering distribution for active learning,” IEEE Trans. Neural Netw. Learn. Syst., vol. 33, no. 1, pp. 215–228, Jan. 2022.

J. B. Tenenbaum, “Bayesian modeling of human concept learning,” in Proc. Adv. Neural Inf. Process. Syst., 1999, pp. 59–68.

C. Elster and G. Wübbeler, “Bayesian inference using a noninformative prior for linear Gaussian random coefficient regression with inhomogeneous within-class variances,” Comput. Statist., vol. 32, no. 1, pp. 51–69, Mar. 2017.

Y. Sun and J. Cao, “Adaptive lag synchronization of unknown chaotic delayed neural networks with noise perturbation,” Phys. Lett. A, vol. 364, nos. 3–4, pp. 277–285, 2007.

Y.-C. Hsu, Y. Shen, H. Jin, and Z. Kira, “Generalized ODIN: Detecting out-of-distribution image without learning from out-of-distribution data,” in Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR), Jun. 2020, pp. 10951–10960.

I. J. Goodfellow, J. Shlens, and C. Szegedy, “Explaining and harnessing adversarial examples,” in Proc. ICLR, 2015, pp. 1–11.

V. Vapnik, The Nature of Statistical Learning Theory. Springer, 1999.

C. Tosh and S. Dasgupta, “Diameter-based active learning,” in Proc. Int. Conf. Mach. Learn., 2017, pp. 3444–3452.

S. Hanneke, “Theory of disagreement-based active learning,” Found. Trends Mach. Learn., vol. 7, nos. 2–3, pp. 131–309, 2014.

X. Cao, B. Qiu, X. Li, Z. Shi, G. Xu, and J. Xu, “Multidimensional balance-based cluster boundary detection for high-dimensional data,” IEEE Trans. Neural Netw. Learn. Syst., vol. 30, no. 6, pp. 1867–1880, Jun. 2019.

O.-E. Ganea, G. Bécigneul, and T. Hofmann, “Hyperbolic entailment cones for learning hierarchical embeddings,” in Proc. 35th Int. Conf. Mach. Learn. (ICML), 2018, pp. 1646–1655.

L. Vinh Tran, Y. Tay, S. Zhang, G. Cong, and X. Li, “HyperML: A boosting metric learning approach in hyperbolic space for recommender systems,” in Proc. 13th Int. Conf. Web Search Data Mining, Jan. 2020, pp. 609–617.

M. Nickel and D. Kiela, “Learning continuous hierarchies in the Lorentz model of hyperbolic geometry,” in Proc. 35th Int. Conf. Mach. Learn. (ICML), 2018, pp. 3777–3785.

T. Nitta and Y. Kuroe, “Hyperbolic gradient operator and hyperbolic back-propagation learning algorithms,” IEEE Trans. Neural Netw. Learn. Syst., vol. 29, no. 5, pp. 1689–1702, May 2018.

K. Yu, J. Bi, and V. Tresp, “Active learning via transductive experimental design,” in Proc. 23rd Int. Conf. Mach. Learn. (ICML), 2006, pp. 1081–1088.

B. Qiu and X. Cao, “Clustering boundary detection for high dimensional space based on space inversion and Hopkins statistics,” Knowl.-Based Syst., vol. 98, pp. 216–225, Apr. 2016.

X. Cao, B. Qiu, and G. Xu, “BorderShift: Toward optimal MeanShift vector for cluster boundary detection in high-dimensional data,” Pattern Anal. Appl., vol. 22, no. 3, pp. 1015–1027, Aug. 2019.

S. Dasgupta and D. Hsu, “Hierarchical sampling for active learning,” in Proc. 25th Int. Conf. Mach. Learn. (ICML), 2008, pp. 208–215.

Y.-P. Tang and S.-J. Huang, “Self-paced active learning: Query the right thing at the right time,” in Proc. AAAI Conf. Artif. Intell., vol. 33, 2019, pp. 5117–5124.

Xiaofeng Cao received the Ph.D. degree from the Australian Artificial Intelligence Institute, University of Technology Sydney, Ultimo, NSW, Australia, in 2021.

He is currently an Associate Professor with the School of Artificial Intelligence, Jilin University, Changchun, China, where he also leads the Machine Perceptron Research Group, with more than 15 master’s and Ph.D. students. His research interests include probably approximately correct learning theory, agnostic learning algorithm, generalization analysis, and hyperbolic geometry.

Ivor W. Tsang (Fellow, IEEE) was a Professor of artificial intelligence with the University of Technology Sydney, Ultimo, NSW, Australia, and the Research Director of the Australian Artificial Intelligence Institute, Ultimo. He is currently the Director of the Centre for Frontier AI Research, Agency for Science, Technology and Research, Singapore. He is working at the forefront of big data analytics and artificial intelligence. In 2013, he received the ARC Future Fellowship for his outstanding research on big data analytics and large-scale machine learning. His research interests include transfer learning, deep generative models, learning with weak supervision, and big data analytics for data with extremely high dimensions in features, samples, and labels.

Prof. Tsang has conferred the IEEE Fellow for his outstanding contributions to large-scale machine learning and transfer learning in 2021. He was a recipient of the IEEE TNN Outstanding 2004 Paper Award, the Best Student Paper Award at CVPR 2010, and the 2014 IEEE TMM Prize Paper Award. In 2019, he received the International Consortium of Chinese Mathematicians Best Paper Award. In 2020, he was recognized as the AI 2000 AAAI/IJCAI Most Influential Scholar in Australia. He serves as the Senior Area Chair/Area Chair for NeurIPS, ICML, AAAI, and IJCAI, and the steering committee of ACML. He also serves as the Editorial Board Member for the Journal of Machine Learning Research, Machine Learning, the Journal of Artificial Intelligence Research, the IEEE TRANSACTIONS ON PATTERN ANALYSIS AND MACHINE INTELLIGENCE, the IEEE TRANSACTIONS ON ARTIFICIAL INTELLIGENCE, the IEEE TRANSACTIONS ON BIG DATA, and the IEEE TRANSACTIONS ON EMERGING TOPICS IN COMPUTATIONAL INTELLIGENCE.