Journal of Machine Learning Research 25 (2024) 1-45 Submitted 9/23; Revised 7/24; Published 10/24

Mentored Learning: Improving Generalization and Convergence of Student Learner via Teaching Feedback

Xiaofeng Cao(✉)* XIAOFENG.CAO.UTS@GMAIL.COM
School of Artificial Intelligence
Jilin University, Changchun 130012, China
and
Australian Artifcial Intelligence Institute (AAII)
University of Technology Sydney (UTS), NSW 2007, Australia
Yaming Guo GUOYM21@MAILS.JLU.EDU.CN
School of Artificial Intelligence
Jilin University, Changchun 130012, China
Heng Tao Shen SHENHENGTAO@HOTMAIL.COM
School of Computer Science and Technology
Tongji University, Shanghai 201804, China
and
School of Computer Science and Engineering
University of Electronic Science and Technology of China, Chengdu 611731, China
Ivor W. Tsang IVOR_TSANG@IHPC.A-STAR.EDU.SG
Centre for Frontier AI Research and Institute of High Performance Computing
*Agency for Science, Technology, and Research(A*STAR), Singapore 138632, Singapore*
and
College of Computing and Data Science
Nanyang Technological University, Singapore 639798, Singapore
James T. Kwok JAMESK@CSE.UST.HK
Department of Computer Science and Engineering
The Hong Kong University of Science and Technology, Hong Kong SAR 999077, China

Editor: Laurent Orseau

Abstract

Student learners typically engage in an iterative process of actively updating its hypotheses, like active learning. While this behavior can be advantageous, there is an inherent risk of introducing mistakes through incremental updates including weak initialization, inaccurate or insignificant history states, resulting in expensive convergence cost. In this work, rather than solely monitoring the update of the learner's status, we propose monitoring the disagreement w.r.t. $\mathcal{F}^{\mathcal{T}}(\cdot)$ between the learner and teacher, and call this new paradigm "Mentored Learning", which consists of 'how to teach' and 'how to learn'. By actively incorporating feedback that deviates from the learner's current hypotheses, convergence will be much easier to analyze without strict assumptions on learner's historical status, then deriving tighter generalization bounds on error and label complexity. Formally, we introduce an approximately optimal teaching hypothesis, $h^{\mathcal{T}}$, incorporating a tighter slack

*. Preliminary work was done when Xiaofeng Cao was a Research Assistant at AAII, UTS.

term $\left(1+\mathcal{F}^{\mathcal{T}}(\widehat{h}_t)\right)\Delta_t$ to replace the typical $2\Delta_t$ used in hypothesis pruning. Theoretically, we demonstrate that, guided by this teaching hypothesis, the learner can converge to tighter generalization bounds on error and label complexity compared to non-educated learners who lack guidance from a teacher: 1) the generalization error upper bound can be reduced from $R(h^*)+4\Delta_{T-1}$ to approximately $R(h^{\mathcal{T}})+2\Delta_{T-1}$, and 2) the label complexity upper bound can be decreased from $4\theta\left(TR(h^*)+2O(\sqrt{T})\right)$ to approximately $2\theta\left(2TR(h^{\mathcal{T}})+3O(\sqrt{T})\right)$. To adhere strictly to our assumption, self-improvement of teaching is proposed when $h^{\mathcal{T}}$ loosely approximates h^*. In the context of learning, we further consider two teaching scenarios: instructing a white-box and black-box learner. Experiments validate this teaching concept and demonstrate superior generalization performance compared to fundamental active learning strategies, such as IWAL (Beygelzimer et al., 2009), IWAL-D (Cortes et al., 2019b), etc.

Keywords: Machine Teaching, Hypothesis Pruning, Active Learning, Error Disagreement, Convergence, Generalization Error, Label Complexity.

1. Introduction

The teaching model's exceptional generalization ability (Goldman and Kearns, 1995) is widely acknowledged within the realm of large language models, exemplified by systems like Chat-GPT (Chen et al., 2024), and artificial intelligence agents such as Tesla's Full Self-Driving Computer (Talpes et al., 2020). In these advanced contexts, the pre-trained model assumes the role of a teacher, actively engaging with a student learner (also referred to as the learner) model to facilitate its improvement. The underlying assumption posits that learners who employ self-paced learning no longer surpass those who benefit from teaching demonstrations in terms of learning ability. This transformative shift has reverberated throughout traditional learning communities, yielding significant ramifications in diverse domains like natural language processing (Norouzi et al., 2020), embodied intelligence (Gupta et al., 2021), and autonomous driving (Bhattacharyya et al., 2023).

The machine learning community has gradually recognized the significant benefits and performance enhancements offered by large interactive models. Consequently, various iterations of a novel paradigm have emerged, wherein these large models assume the role of teachers, guiding and supervising the progress of learners. This development has compelled theoretical researchers in the learning community to tackle the generalization challenges associated with teachers. In fact, the theoretical learning community has long acknowledged the importance of teaching and its role in the learning process, and the concept of teaching was introduced early on and subsequently formalized as machine teaching (Simard et al., 2017; Zhu, 2015). Meanwhile, this concept has also garnered attention and exploration within the theoretical learning community, especially about hypothesis pruning.

We begin by first reviewing a theoretical description of learning. In learning theory, hypothesis-pruning (Kääriäinen et al., 2004) interactively trims a pre-specified hypothesis class (space) $\mathcal{H}$ to find a desired output, aiming to enhance the convergence of any learning algorithm using as few labels as possible, such as active learning (Settles, 2009). In this typical scenario, the learner has access to a pool of unlabeled data and can request labels from human annotators for these unlabeled instances. Typically, the hypotheses are

generated based on a functional assumption, such as MLP, CNN, etc. If the hypotheses don't rely on any specific functional assumption, it becomes an agnostic scenario (Balcan et al., 2009), exploring the theoretical performance of achieving a parameterized error by controlling the label complexity bound (Hanneke, 2007a). On the theoretical front, a range of hypothesis update methods and analyses of label complexity bounds have been presented, for example, (Hanneke, 2012) and (Beygelzimer et al., 2009). In practical applications, active learning has already demonstrated benefits in image annotation (Beluch et al., 2018), semantic segmentation (Siddiqui et al., 2020), and more. A common assumption in active learning, whether in theoretical explorations or practical applications, is that an infinite hypothesis class exists, containing the optimal hypothesis that can be iteratively updated from a random initialization. With this assumption, Hanneke et al. introduced an error disagreement coefficient (Hanneke et al., 2014) to regulate the hypothesis updates. The policy dictates that any disagreement arising from the candidate hypothesis exceeding the predefined coefficient is considered feasible and results in positive updates (Cao and Tsang, 2021a). Otherwise, it is deemed an insignificant update. To minimize the label complexity of updating costs, Zhang et al. introduced a tighter bound using a new term called confidence rate (Zhang and Chaudhuri, 2014); Golovin et al. (Golovin et al., 2010) proposed a near-optimal Bayesian policy; Yan et al. (Yan and Zhang, 2017) presented near-optimal label complexity bounds for both bounded and adversarial noise conditions, etc.

Question: Learners typically engage in an iterative process of actively updating its hypotheses. While this behavior can be advantageous, there is an inherent risk of introducing mistakes through incremental updates including weak initialization, inaccurate or insignificant history states, resulting in expensive convergence cost. In short, the existing theoretical results may not robustly guarantee the convergence of these incremental updates in the hypothesis class. In essence, obtaining the optimal hypothesis h^* from these updates may not be easily achievable without explicit guidance and information from h^*. Therefore, can teaching relieve the issues of learning?

For teaching, it refers to a framework and methodology that focuses on the design and optimization of instructional strategies to facilitate effective learning in machine learning systems. Remarkably, the emphasis is placed on the perspective of the teacher, who aims to impart knowledge and guide the learning process of the learner model. Technically, it involves various techniques such as curriculum design, active learning, and example selection, among others. A general goal is to provide the learner model with informative and well-structured training data that leads to improved performance and generalization. By leveraging teaching, researchers and practitioners aim to enhance the efficiency, robustness, and interpretability of machine learning. This approach acknowledges the crucial role of the teacher in shaping the learning process and optimizing the learning outcomes. Theoretical analysis about teaching complexity, teaching dimension, and teaching convergence have been well investigated, such as (Doliwa et al., 2014; Liu et al., 2017). Recently, we (Zhang et al., 2023a,b) delved into iterative teaching in nonparametric settings, where the target models are defined as functions without dependencies on learners' parameters. In contrast to typical parametric optimization, which mainly operates for pruning parameter space, we explore functional optimization in the function space, invoking iterative target approximation. To further extend this idea to multiple learners, each learner is required to focus on learning a scalar-valued target model

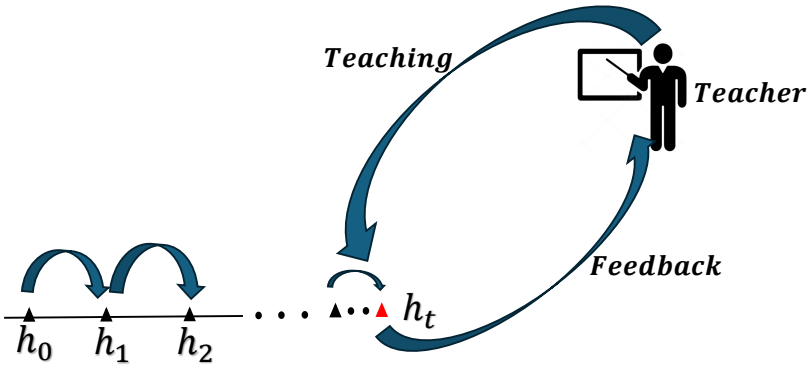


Figure 1: Hypothesis pruning of Learning vs. Teaching. Learning focuses on 'how to learn', and teaching focuses on 'how to teach'. With teaching, the general learning shifts into "Mentored Learning", which consists of the dual questions.

that can be decomposed in a reproducing kernel Hilbert space. In this setting, teaching with implicit information shows potential. (These two works are accepted in ICML and NeurIPS 2023)

Teaching vs. Learning. Machine teaching is a framework that underscores the pivotal role of the instructor in steering the learning process. This approach involves the meticulous design of instructional strategies and the careful selection of informative training data, aimed at enhancing the learning outcomes of machine learning models. Unlike machine learning, which predominantly focuses on the development of algorithms and models that can autonomously learn from data, machine teaching is concerned with the systematic methodologies for effectively training these models. By optimizing the learning process, improving generalization, and enhancing overall performance, machine teaching seeks to significantly elevate the efficacy of machine learning systems. Theoretically, teaching can be seen as a more advanced form of "Mentored Learning." While both teaching and learning involve guidance from a knowledgeable source, teaching goes beyond simply providing guidance and encompasses the intentional design of instructional strategies to optimize the learning process. Refer to Figure 6. In pursuit of the shared objective of facilitating learner improvement, teaching refines the hypothesis through the feedback received from the learner, distinguishing itself from the historical practice of hypothesis pruning in learning. The disparity lies in the contrast between 'how to learn' and 'how to teach'. With teaching, the general learning shifts into "Mentored Learning" which consists of the dual questions.

Motivation Motivated by the teaching advantages, this paper relaxes the traditional assumptions regarding the learning target and introduces an approximately optimal teacher as a target (Dasgupta et al., 2019; Liu et al., 2018). In such a setting, this teacher provides guidance to the learner without disclosing any internal cues, such as parameter distributions or convergence conditions. Theoretically, this model envisions a teacher that offers direction while maintaining opacity concerning its internal mechanisms. Unlike conventional teaching models where the learner might access the teacher's internal workings, the approximately optimal teacher remains a black box, withholding specific details about its parameters and expected convergence conditions. Consequently, the learner must rely exclusively on the guidance provided, devoid of insights into the underlying decision-making processes

of the teacher. This methodology aims to create a more realistic and challenging learning environment, compelling the learner to adapt and improve based solely on external feedback. Such an approach is anticipated to foster a more robust and adaptable learning framework.

In this paper, we call this new learning paradigm by teaching "Mentored Learning". In this way, the teacher provides an approximately optimal hypothesis $h^{\mathcal{T}}$, maintaining a fair teaching scenario compared to non-educated learners who do not receive any guidance from a teacher. With $h^{\mathcal{T}}$, an active learner can easily replace the infeasible h^* and select unlabeled data that maximize the disagreement of the feedback between the teacher and the learner, rather than maximizing the disagreement between the current and subsequent hypotheses as in typical pruning of hypothesis. Our contributions are summarized as follows.

- We propose a new perspective of introducing machine teaching to guide an active learner, which guarantees a desired convergence to an approximated teaching hypothesis, not the typical infeasible optimal hypothesis. We call this new paradigm "Mentored Learning", which consists of 'how to teach' and 'how to learn'.
- We theoretically prove that, under the guidance of the teaching hypothesis, the learner can converge into tighter generalization error and label complexity bounds than those non-educated learners without teacher guidance. This involves "how to teach". To further improve its generalization, we then consider two scenarios: teaching a white-box and black-box learner, where the self-improvement of teaching is firstly proposed to improve the initial teaching hypothesis. This involves "how to learn".
- We present an Approximately Optimal Teaching-based Mentored Learning (ATML) algorithm, which spends fewer annotations to converge, yielding more effective performance than those typical active learning strategies.

Organization. Section 3 presents the related work. Section 4 elaborates the error disagreement-based active learning. Section 5 explains our approximately optimal teaching idea. Section 6 improves teaching when the instructor has only a rough approximation of the optimal hypothesis. Section 7 employs this idea to guide an active learner. Experiments are presented in Section 8. We conclude this work in Section 9. See the main structure below.

- Section 4 presents "what are fundamental concepts" of learning.
- Section 5 introduces "how to teach" in Mentored Learning.
- Section 6 studies "how to teach better" in Mentored Learning.
- Section 7 introduces "how to learn" in Mentored Learning.

Notation We introduce the set of notations used throughout the paper. We denote by $\mathcal{X}$ the input space and by $\mathcal{Y}$ the output space. Let $\mathcal{D}$ be an data distribution over $\mathcal{X} \times \mathcal{Y}$, and $\mathcal{D}_{\mathcal{X}}$ be the marginal distribution of $\mathcal{D}$ over $\mathcal{X}$. We consider the on-line active learning scenario: for each time $t \in [T] = \{1, ..., T\}$, the learner receives an input sample x_t drawn i.i.d. according to $\mathcal{D}_{\mathcal{X}}$ and has to decide whether to query its label.

We denote by $\mathcal{H} = \{h : \mathcal{X} \to \mathcal{Z}\}$ the hypothesis space, where $\mathcal{Z}$ is a prediction space. Let $\ell(h(x), y)$ denotes the loss function which operates $\mathcal{Z} \times \mathcal{Y} \to [0, 1]$. For any hypothesis

h, we denote $R(h)$ to be the generalization error: $R(h) = \mathop{\mathbb{E}}_{(x,y)\sim\mathcal{D}}[\ell(h(x), y)]$, and denote $h^* = \mathop{\text{argmin}}_{h\in\mathcal{H}} R(h)$ to be the optimal hypothesis in $\mathcal{H}$. We also denote by $L_t(h)$ the importance-weighted empirical error of h, defined by the weighted loss of query samples w.r.t. Eq. (3). Let H_t denote the candidate hypothesis set of the learner at t-time, where $H_1 = \mathcal{H}$.

We use $h^{\mathcal{T}}$ to denote the teaching hypothesis w.r.t. Assumption 1, which replaces the infeasible h^* to guide the active learner. Then, we use $\mathcal{H}^{\mathcal{T}}$ to denote the teaching-hypothesis-class w.r.t. Definition 6, which is an efficient approximation to $\mathcal{H}$. To avoid any confusion, we denote by $H_t^{\mathcal{T}}$ the candidate hypothesis set at t-time with respect to the teaching hypothesis $h^{\mathcal{T}}$. At t-time, we define the current empirical optimal hypothesis $\widehat{h}_t = \mathop{\text{argmin}}_{h\in H_t^{\mathcal{T}}} L_t(h)$, which has the minimum importance-weighted empirical error in $H_t^{\mathcal{T}}$.

Notation	**Description**
$\mathcal{X}$	Input space
$\mathcal{Y}$	Output spcae
$\mathcal{D}$	Data distribution over $\mathcal{X}\times\mathcal{Y}$
$\mathcal{D}_{\mathcal{X}}$	Marginal distribution of $\mathcal{D}$ over $\mathcal{X}$
T	Total number of time steps
x_t	Input sample at time t
$\mathcal{H}$	Hypothesis space $\{h : \mathcal{X}\to\mathcal{Z}\}$
$\mathcal{Z}$	Prediction space
$\ell(\cdot,\cdot)$	Loss function operating on $\mathcal{Z}\times\mathcal{Y}\to[0,1]$
$R(h)$	Generalization error: $R(h) = \mathbb{E}_{(x,y)\sim\mathcal{D}}[\ell(h(x), y)]$
h^*	Optimal hypothesis in $\mathcal{H}$: $h^* = \mathop{\text{argmin}}_{h\in\mathcal{H}} R(h)$
$L_t(h)$	Importance-weighted empirical error of h at time t
H_t	Candidate hypothesis set at time t, where $H_1 = \mathcal{H}$
$\widehat{h}_t$	Current empirical optimal hypothesis at time t: $\widehat{h}_t = \mathop{\text{argmin}}_{h\in H_t} L_t(h)$
$h^{\mathcal{T}}$	Teaching hypothesis
$\mathcal{H}^{\mathcal{T}}$	Teaching-hypothesis-class
$H_t^{\mathcal{T}}$	Candidate hypothesis set at time t with respect to teaching hypothesis $h^{\mathcal{T}}$
$\widehat{h}_t$	Current empirical optimal hypothesis at time t: $\widehat{h}_t = \mathop{\text{argmin}}_{h\in H_t^{\mathcal{T}}} L_t(h)$
$\mathcal{F}^{\mathcal{T}}(\widehat{h}_t)$	Teaching feedback function of $\widehat{h}_t$ at time t with respect to $h^{\mathcal{T}}$
H_t'	Candidate generation base at t-time for drawing new hypotheses
$\tilde{h}$	New hypothesis generated from the convex hull of H_t': $\tilde{h} = \sum_j^m \lambda_j h_j, h_j \in H_t'$
$\widetilde{H}_t'$	Hypothesis generation set for self-improvement $\widetilde{H}_t' = \left\{\tilde{h}_i; i\in[n]\right\}$

Table 1: Main notations used in the paper

2. Main Theoretical Results

Main Progress Theoretically, we present the teaching-based hypothesis pruning and its improvements to defend our teaching idea. In detail, 1) we observe whether the teaching-based hypothesis pruning strategy can prune the candidate hypothesis set faster than the error disagreement-based active learning; 2) we observe whether the optimal hypothesis can be usually maintained in the candidate hypothesis set; 3) we also present the generalization error and label complexity bounds of teaching an active learner.

Main Assumption For any hypothesis class $\mathcal{H}$, we assert the existence of a specific hypothesis, $h^{\mathcal{T}}$, termed the teaching hypothesis. This hypothesis $h^{\mathcal{T}}$ possesses the capacity to tolerate a defined error bias (gap) to h^* denoted by ϵ. The accommodation of ϵ signifies that $h^{\mathcal{T}}$ remains a valid and functional hypothesis within $\mathcal{H}$, even in the presence of an error rate up to ϵ. Elaborating further, the teaching hypothesis $h^{\mathcal{T}}$ assumes the role of a benchmark or guiding principle within the context of teaching-based hypothesis pruning. The parameter ϵ delineates an acceptable deviation from the optimal hypothesis h^*, indicating that while $h^{\mathcal{T}}$ needs not be flawless, it should uphold an error rate within the prescribed bounds of ϵ. Formally, it is denoted by

$$\mathcal{L}\left(h^*, h^{\mathcal{T}}\right) = \mathop{\mathbb{E}}_{x\sim\mathcal{D}_{\mathcal{X}}}\left[\max_{y}\left|\ell(h^*(x), y) - \ell(h^{\mathcal{T}}(x), y)\right|\right] < \epsilon,$$

where h^* denotes the optimal hypothesis in $\mathcal{H}$, and the disagreement of hypothesis invokes Eq. (1).

This concept is fundamental in the analysis and refinement of teaching strategies within machine learning frameworks. It forms the bedrock for evaluating the effectiveness and efficiency of teaching-based approaches in refining the hypothesis space, safeguarding the optimal hypothesis, and ultimately refining the learning process.

Main Technique We still follow the pruning manner of IWAL w.r.t. Eq. (4) to supervise the updates of the candidate hypothesis set, where the main difference is that we introduce a teaching hypothesis $h^{\mathcal{T}}$ to control the slack constraint of hypothesis pruning. Specifically, the slack constraint $2\Delta_t$ is tightened as $\left(1 + \mathcal{F}^{\mathcal{T}}(\widehat{h}_t)\right)\Delta_t$ by invoking the guidance of a teacher, where $\mathcal{F}^{\mathcal{T}}(\widehat{h}_t)$ denotes disagreement feedback with the teacher w.r.t. current empirical optimal hypothesis $\widehat{h}_t$. With such operation, the candidate hypothesis set $H_{t+1}^{\mathcal{T}}$ at $t+1$-time is updated by

$$H_{t+1}^{\mathcal{T}} = \left\{h \in H_t^{\mathcal{T}} : L_t(h) \leq L_t(\widehat{h}_t) + \left(1 + \mathcal{F}^{\mathcal{T}}(\widehat{h}_t)\right)\Delta_t\right\},$$

where $H_1^{\mathcal{T}} = \mathcal{H}^{\mathcal{T}}$, and $\Delta_t = \sqrt{(2/t)\log(2t(t+1)|\mathcal{H}^{\mathcal{T}}|^2/\delta)}$ for some fixed confidence parameter $\delta > 0$. Therefore teaching-based hypothesis pruning is more aggressive in shrinking the candidate hypothesis set, resulting in better learning guarantees.

Main Theorem 0.1 *For any teaching-hypothesis-class $\mathcal{H}^{\mathcal{T}}$, the instruction of an active learner is conducted within it. Given any $\delta > 0$, with a probability at least $1 - \delta$, for any $T \in \mathbb{N}^+$, the following holds:*

1) *the generalization error holds*

$$R(\widehat{h}_T) \leq R(h^*) + \left(2 + \mathcal{F}^{\mathcal{T}}(\widehat{h}_{T-1}) + \mathcal{F}^{\mathcal{T}}(\widehat{h}_T)\right)\Delta_{T-1} + \epsilon;$$

2) *if the learning problem has disagreement coefficient θ, the label complexity is at most*

$$\tau_T \leq 2\theta\left(2TR(h^*) + \big(3 + \mathcal{F}^{\mathcal{T}}(\widehat{h}_{T-1})\big)\big)O(\sqrt{T}) + 2T\epsilon\right).$$

The theorem outlined above establishes the boundaries for generalization error and label complexity concerning approximately optimal teaching within an active learning framework. The efficacy of such teaching hinges upon two pivotal determinants:

- Efficiency of Learning: This factor is contingent upon the extent of the teacher's feedback disagreement, denoted as $\mathcal{F}(\widehat{h})$, imparted to the learner. Effective active learning necessitates substantial feedback from the teacher to guide the learning process effectively.
- Quality of the Teacher: The teaching ability is epitomized by the maximal discrepancy to the optimal hypothesis, denoted as ϵ, which affects the learner's performance.

Specifically, when $\widehat{h}_{T-1}$ and $\widehat{h}_T$ exhibit close proximity to $h^{\mathcal{T}}$, and ϵ remains within an acceptable error threshold, significant enhancements presents:

- The upper bound on generalization error is tighten from $R(h^*) + 4\Delta_{T-1}$ to approximately $R(h^{\mathcal{T}}) + 2\Delta_{T-1}$.
- Likewise, the upper bound on label complexity is tighten from $4\theta\left(TR(h^*) + 2O(\sqrt{T})\right)$ to approximately $2\theta\left(2TR(h^{\mathcal{T}}) + 3O(\sqrt{T})\right)$.

Stricter assumption : *If the teaching hypothesis is loosely approximated to the optimal hypothesis, i.e. ϵ is large, how do we guarantee the convergence of approximately optimal teaching? We thus design self-improvement of teaching.*

The statement addresses the challenge of achieving convergence towards approximately optimal teaching when there is a significant disparity between the teaching hypothesis and the optimal hypothesis, indicated by a large ϵ value.

To tackle this challenge, "self-improvement of teaching" is introduced. This way entails devising mechanisms within the teaching process to adapt and refine itself iteratively. By incorporating feedback and evaluation, the teaching process undergoes iterative adjustments to reduce the gap between the teaching hypothesis and the optimal hypothesis, which improves the safety and stability of the subsequent learning performance.

In essence, self-improvement of teaching involves an iterative refinement process aimed at enhancing teaching effectiveness and facilitating convergence towards approximately optimal teaching outcomes.

Main Theorem 0.2 *For any teaching-hypothesis-class $\mathcal{H}^{\mathcal{T}}$, the instruction of an active learner is conducted within $\mathcal{H}^{\mathcal{T}}$. If the self-improvement of teaching is applied, given any*

$\delta > 0$, with a probability at least $1-\delta$, for any $T \in \mathbb{N}^+$, the following holds: 1) for any $t \in [T]$, holds $h_t^{\mathcal{T}} \in H_t^{\mathcal{T}}$;

2) *the generalization error holds*

$$R(\widehat{h}_T) \leq R(h^*) + \left(2 + \mathcal{F}_{T-1}^{\mathcal{T}}(\widehat{h}_{T-1}) + \mathcal{F}_{T-1}^{\mathcal{T}}(\widehat{h}_T)\right) \Delta_{T-1} + \epsilon_{T-1};$$

3) *if the learning problem has disagreement coefficient θ, the label complexity is at most*

$$\tau_T \leq 2\theta \left(2TR(h^*) + \left(3 + \mathcal{F}_{T-1}^{\mathcal{T}}(\widehat{h}_{T-1})\right) O(\sqrt{T}) + 2T\epsilon_{T-1}\right).$$

The Theorem shows that the optimal hypothesis of $\bigcup_{k=1}^{t} H_k^{\mathcal{T}}$ is maintained in the candidate hypothesis set with a high probability at any t-time. With Corollary 15, we have $\epsilon_{T-1} \leq \epsilon = \epsilon_1$, which shows that self-improvement of teaching strategy can further reduce the generalization error and label complexity bounds of the learner w.r.t. Theorem 11. Moreover, the improvement of the active learner is decided by the improvement of the approximately optimal teacher.

Given any t-time, the results highlight the theorem's findings regarding the persistence of the optimal hypothesis, generalization error inequality, and label complexity inequality for learning. The below explains these results:

Theorem Findings:

- The theorem establishes that the optimal hypothesis persists within a candidate hypothesis set throughout different learning stages. This forms safety guarantees for the learning process.
- Regardless of the stage denoted by t, the optimal hypothesis remains prominently featured within $\bigcup_{k=1}^{t} H_k^{\mathcal{T}}$ with a high probability. This ensures that the process of self-improvement can uphold the optimal hypothesis within the given class.

Corollary Implications:

- Referring to Corollary 15, it reveals the inequality of $\epsilon_{T-1} \leq \epsilon = \epsilon_1$, that is, with self-improvement, the teacher's gap to the optimal hypothesis could be decreased. It shows the success of the self-improvement.
- This discovery suggests the potential of self-improvement in teaching strategies to further tighten the generalization error and label complexity bounds for the learner.

Alignment with Theorem:

- The optimization through self-improvement aligns with the assurances provided by Theorem 11.

Efficacy of Teaching:

- The statement emphasizes the direct correlation between the refinement of teaching strategies and the efficacy of the active learner's progression.

- Achieving a state of approximately optimal teaching significantly influences learning outcomes and overall performance.

In essence, the statement emphasizes how the iterative refinement of the teaching strategy through the generation of new hypotheses directly contributes to improving the learning guarantees for the active learner. This highlights the significance of self-improvement mechanisms in optimizing the teaching process and facilitating more effective learning experiences.

3. Related Work

First, we introduce active learning, encompassing both its theoretical explorations and practical applications. Subsequently, we delve into machine teaching, which involves supervising both white-box and black-box learners.

3.1 Active Learning

Active learning encompasses two branches: theoretical explorations (Hanneke, 2009) and practical applications (Settles, 2009). The theoretical branch focuses on the generalization analysis of error and label complexity bounds within the hypothesis class. In practical applications, these theoretical results are extended to weakly-supervised sampling (Rasmus et al., 2015), Bayesian approximation (Pinsler et al., 2019), adversarial training (Sinha et al., 2019), and other related areas.

Theoretical explorations Theoretical active learning is approached from two perspectives: agnostic bound convergence and version space shrinking. Agnostic active learning is derived from the standard PAC framework (Denis, 1998), while version space shrinking (Dasgupta, 2004; Tong and Koller, 2001) can be generalized from a hypothesis pruning view (Cortes et al., 2019b; Cao and Tsang, 2020). In the analysis of the linear perceptron (Gonen et al., 2013), Dasgupta et al. (Dasgupta, 2011) presented a series of upper and lower bounds on label complexity, maintaining consistent convergence with the query-by-committee algorithm (Gilad-Bachrach et al., 2006), which involves multiple learners. Hanneke later extended these bounds for more general settings (Hanneke, 2007a, 2012), enhancing the efficiency of the error disagreement coefficient. In a uniform framework, Balcan et al. summarized these theoretical results as the agnostic scenario (Balcan et al., 2009). However, these results often assume a uniform distribution and a noise-free setting. For bounded and adversarial noise, Yan et al. (Yan and Zhang, 2017) presented label complexity bounds. With a consistent assumption about support vectors, Tong et al. (Tong and Koller, 2001) utilized the notion of the version space to shrink its volume by maximizing the minimum distance to any of the delineating hyperplanes. Other similar works can also be found in (Warmuth et al., 2001; Golovin and Krause, 2010; Ailon et al., 2012; Krishnamurthy et al., 2017). To shrink the version space into the minimal covering on the optimal hypothesis, Cortes et al. presented a region-splitting algorithm to ensure that pruning in the hypothesis class can converge to the optimal hypothesis, as demonstrated in works such as (Cortes et al., 2019a, 2020).

Practical applications To prune the hypothesis class, following the error disagreement coefficient, incremental optimization is a common approach. This involves iteratively updating

the current learning model by maximizing its uncertainty. Within this paradigm, various baselines have been proposed, including maximizing error reduction (Roy and McCallum, 2001) and maximizing mean standard deviation (Kampffmeyer et al., 2016), among others. In statistical optimization, active learning can also be redefined as experimental design (Wong, 1994), which includes A, D, E, and T optimal design. In this context, the A-optimal design minimizes the average variance of the parameter estimates, the D-optimal design maximizes the differential Shannon information content of the parameter estimates, the E-optimal design maximizes the minimum eigenvalue of the information matrix, and T-optimal design methods maximize the trace of the information matrix. In the Bayesian setting, active learning is defined as Bayesian approximation on the likelihood (Orekondy et al., 2019) or maximizing the information gain (Kirsch et al., 2019), among other approaches. In recent years, propelled by the powerful modeling capabilities of deep neural networks, deep active learning has emerged, sparking new interest. Examples include Monte-Carlo dropout with active learning (Gal et al., 2017), deep active annotation (Huijser and van Gemert, 2017), adversarial training with an active querying set (Sinha et al., 2019), dual adversarial networks for deep active learning (Wang et al., 2020a), and consistency-based semi-supervised active learning (Gao et al., 2020), among others.

Remark 1 *Active learning prunes the predefined hypothesis class into a desired one. Through an iterative labeling process, the initially broad hypothesis class, which encompasses potential hypotheses or models, gradually narrows down or refines to a more specific one aligned with the desired model characteristics or performance criteria. Essentially, active learning assists in selecting and refining the most relevant hypotheses from the predefined class, thereby guiding the learning process towards achieving the desired outcome more efficiently. This forms the structured perspective of 'how to learn', i.e., the framework of fundamental learning manner of "Mentored Learning".*

3.2 Machine Teaching

Machine teaching (Zhu et al., 2018) focuses on an inverse problem of machine learning, specifically finding the optimal teaching examples when the teacher already knows the learning parameters. Machine teaching is categorized into two scenarios: teaching a white-box and teaching a black-box (Dasgupta et al., 2019; Liu et al., 2018).

Teaching white-box Machine teaching assumes that the teacher is aware of the optimal learning parameters of the learner. It provides theoretical analyses for various types of learners, such as those using linear regression, logistic regression, and support vector machines (SVMs), to identify the best teaching examples. These examples can then adjust a random initial training parameter to its optimal state. Essentially, machine teaching offers optimal control over parameter exploration for a learner. To enhance theoretical guarantees, Goldman et al. (Goldman and Kearns, 1995) introduced a comprehensive set of theoretical concepts, including teaching dimension (Liu et al., 2016; Doliwa et al., 2014), and teaching complexity (Hanneke, 2007b). Zhu et al. (Zhu et al., 2017) subsequently extended teaching theories to multiple learners. In practical scenarios, this teaching approach has found widespread application in teacher-student learning models, as evidenced by works such as those by Wang et al. (Wang et al., 2020b), Matiisen et al. (Matiisen et al., 2019), Meng et al. (Meng

et al., 2018, 2019), and Wang et al. (Wang et al., 2021). Recently, Zhang and Cao et al. (Zhang et al., 2023a,b) studied iterative teaching in nonparametric settings, where the teacher models are defined as functions independent of specific parameters.

Teaching black-box A more challenging problem arises when the teacher is unable to disclose any cues regarding the distribution of the learning parameters, rendering the learner essentially a black box. In this context, Liu et al. (Liu et al., 2018) explored cross-space machine teaching, which involves distinct feature representations for the teacher and the student. Dasgupta et al. (Dasgupta et al., 2019) proposed reducing the training sets for any classifier family by identifying an approximately minimal subset of training instances that maintains the consistent properties of the original hypothesis class. Cicalese et al. (Cicalese et al., 2020) investigated scenarios where the teacher aims for the learner to converge to a well-available approximation of the optimal hypothesis. Cao et al. (Cao and Tsang, 2021b) suggested employing iterative distribution matching to instruct a black-box learner. Orekondy et al. (Orekondy et al., 2019) utilized model functionality feedback to guide a black-box learner, thereby approximating their parameter distributions. Wang et al. (Wang, 2021) introduced a knowledge distillation method capable of extracting knowledge from a decision-based black-box model without accessing the model's internal structure or parameters. This approach facilitates the transfer of knowledge from complex, opaque models into a more interpretable and compact student model. Nguyen et al. (Nguyen et al., 2022) considered the situation where the teacher can only aim for the black-box learner to converge to the best available approximation of the optimal hypothesis, rather than the exact optimal hypothesis itself.

Remark 2 *Machine teaching involves supervising and guiding the learner to refine its hypothesis class to achieve a desired outcome. Through this supervision, the initial broad hypothesis class is narrowed down or refined to better fit the specific requirements or goals. Unlike learning, teaching involves the presence of a "teacher" entity, which actively guides and instructs the learner throughout the learning process. The teacher's role is crucial in facilitating the learner's improvement by providing guidance, feedback, and instruction. In contrast to general learning approaches that may rely on incremental updates and be potentially prone to mistakes or inaccuracies, machine teaching aims to mitigate these issues. By providing direct guidance and supervision, machine teaching helps to avoid incremental errors and ensures a more accurate and efficient learning process. This forms the structured perspective of 'how to teach', i.e., the framework of the fundamental teaching manner of "Mentored Learning".*

4. Error Disagreement-based Active Learning: Fundamental Concepts

In this section, we introduce the error disagreement and its generalized learning algorithm with guarantees. This constitutes the fundamental concepts for learners in mentored learning. Within this prototype, the learner can be generalized across various classifiers and examined under a range of noise models. For a thorough exploration of introduced and established results in error disagreement-based active learning, refer to Hanneke et al. (2014).

4.1 Error Disagreement

Given a hypothesis class $\mathcal{H}$, active learning tries to reduce the maximum disagreement of hypothesis in $\mathcal{H}$ by invoking a disagreement function $\mathcal{L}(\cdot,\cdot)$ (Cortes et al., 2019b).

For any hypothesis pair $\{h, h'\} \subseteq \mathcal{H}$, $\mathcal{L}(h, h')$ measures their disagreement by the error disagreements, i.e.,

$$\mathcal{L}(h, h') = \mathop{\mathbb{E}}_{x\sim\mathcal{D}_{\mathcal{X}}} \left[\max_{y} \left| \ell(h(x), y) - \ell(h'(x), y) \right| \right], \tag{1}$$

where $\ell(h(x), y)$ denotes the loss function which operates $\mathcal{X} \times \mathcal{Y} \to [0, 1]$. The calculation of error disagreement w.r.t. Eq. (1) does not require labels, i.e., it can be calculated over the unlabeled dataset. Given an i.i.d. sample $x_1, x_2, \ldots, x_n$ from $\mathcal{D}_{\mathcal{X}}$, the error disagreement is the empirical average $\mathcal{L}(h, h') = \frac{1}{n}\sum_{i=1}^{n} \left[\max_y \left| \ell(h(x_i), y) - \ell(h'(x_i), y) \right| \right]$. As an example of binary classification, we solve for $\max\left\{ \left| \ell(h(x_i), +1) - \ell(h'(x_i), +1) \right|, \left| \ell(h(x_i), -1) - \ell(h'(x_i), -1) \right| \right\}$ to obtain the error disagreement between h and h' over sample x_i.

4.2 Learning Algorithm

Importance weighted active learning (IWAL) (Beygelzimer et al., 2009) invokes the error disagreement to prune the hypothesis class $\mathcal{H}$, which is a typical error disagreement-based active learning algorithm.

Given an initial candidate hypothesis set $H_1 = \mathcal{H}$, IWAL receives $x_t \in \mathcal{X}$ drawn i.i.d. according to $\mathcal{D}_{\mathcal{X}}$. At t-time, the algorithm decides whether to query the label of x_t and prunes the candidate hypothesis set H_t to H_{t+1}.

Query At t-time, IWAL does a Bernoulli trial Q_t with success probability p_t, where p_t is the maximum error disagreement of H_t over x_t:

$$p_t = \max_{h,h'\in H_t} \max_{y} \left| \ell(h(x_t), y) - \ell(h'(x_t), y) \right|. \tag{2}$$

If $Q_t = 1$, the algorithm queries the label y_t of x_t.

Hypothesis pruning Let $L_t(h)$ be the importance-weighted empirical error of hypothesis $h \in \mathcal{H}$, there is:

$$L_t(h) = \sum_{k=1}^{t} \frac{Q_k}{p_k} \ell\left(h\left(x_k\right), y_k\right), \tag{3}$$

where its minimizer is $\widehat{h}_t = \mathop{\operatorname{argmin}}_{h\in H_t} L_t(h)$. With the expectation taken over all the random variables, we know $\mathbb{E}\left[L_t(h)\right] = R(h)$. At t-time, IWAL prunes H_t to H_{t+1} through $L_t(\widehat{h}_t)$ and an allowed slack $2\Delta_t$:

$$H_{t+1} = \left\{ h \in H_t : L_t(h) \le L_t(\widehat{h}_t) + 2\Delta_t \right\}, \tag{4}$$

where $\Delta_t = \sqrt{(2/t)\log(2t(t+1)|\mathcal{H}|^2/\delta)}$ for a fixed confidence parameter $\delta > 0$. At T-time, IWAL returns the current empirical optimal hypothesis $\widehat{h}_T$ as the final hypothesis output.

We add some remarks on evaluating the quality of active learning algorithms. The following Remark 3 presents the necessary conditions for a feasible active learning algorithm.

Remark 3 *Whether the optimal hypothesis h^* can usually be maintained in the candidate hypothesis set H_t is a necessary condition for the success of an active learning algorithm.*

The following Remark 4 presents two factors for evaluating the quality of an active learning algorithm.

Remark 4 *Two factors measure the quality of an active learning algorithm: 1) tighter bound on generalization error $R(\widehat{h}_T)$, where $\widehat{h}_T$ is the hypothesis returned by the algorithm after T rounds, and 2) tighter bound on label complexity τ_T, where τ_T is the expected value of label numbers queried by the active learning algorithm within T rounds.*

With Remarks 3 and 4, to guarantee a high-quality learning performance, any active learning algorithm needs to satisfy the three factors, including 1) maintaining the optimal hypothesis, 2) tighter bound on generalization error, and 3) tighter bound on label complexity.

4.3 Learning Guarantees

We present the learning guarantees analysis for IWAL. Firstly, we introduce another definition of the disagreement with respect to hypothesis. For any two hypotheses h, h', let $\rho(h, h')$ denote their disagreement:

$$\rho(h, h') = \mathop{\mathbb{E}}_{(x,y)\sim\mathcal{D}} \left[\left| \ell(h(x), y) - \ell(h'(x), y) \right| \right]. \tag{5}$$

The new disagreement $\rho(\cdot, \cdot)$ can derive a more favorable learning guarantees for the error disagreement-based active learning. Cortes et al. (2019b) shows that the new disagreement $\rho(\cdot, \cdot)$ removes a constant K_ℓ from the label complexity bound of IWAL compared to the error disagreement $\mathcal{L}(\cdot, \cdot)$ w.r.t. Eq. (1). Based on the new disagreement $\rho(\cdot, \cdot)$, we can define a ball with respect to the hypothesis. Given $r > 0$, let $B(h^*, r)$ denote a ball centered in $h^* \in \mathcal{H}$ with the radius r: $B(h^*, r) = \{h \in \mathcal{H} : \rho(h^*, h) \leq r\}$, where h^* is the optimal hypothesis of $\mathcal{H}$. The error disagreement coefficient is then defined as the minimum value of θ for all $r > 0$:

$$\theta \geq \mathop{\mathbb{E}}_{x\sim\mathcal{D}_\mathcal{X}} \left[\max_{h\in B(h^*,r)} \max_y \frac{\left| \ell(h(x), y) - \ell(h^*(x), y) \right|}{r} \right]. \tag{6}$$

The error disagreement coefficient θ is a complexity measure widely used for label complexity analysis in disagreement-based active learning. See Hanneke et al. (2014) for more analysis of disagreement coefficient in active learning. Based on the error disagreement coefficient, guarantees of the learning algorithm is proved by Beygelzimer et al. (2009) and improved by Cortes et al. (2019b).

Theorem 5 *For any hypothesis class $\mathcal{H}$, we perform* IWAL *within it. Given any $\delta > 0$, with probability at least $1 - \delta$, for any $T \in \mathbb{N}^+$, the following holds: 1) for any $t \in [T]$, holds $h^* \in H_t$; 2) the generalization error holds $R(\widehat{h}_T) \leq R(h^*) + 4\Delta_{T-1}$; 3) if the learning problem has a disagreement coefficient θ, the label complexity is at most $\tau_T \leq 4\theta \left(TR(h^*) + 2O(\sqrt{T}) \right)$.*

Theorem 5 guarantees the following facts. 1) The optimal hypothesis h^* is maintained in the candidate hypothesis set H_t with high probability, which is the key to the success of IWAL. 2) As time T increases, Δ_{T-1} gradually tends to zero, leading to a tighter approximation of $\widehat{h}_T$ to h^* in terms of $R(\widehat{h}_T) - R(h^*)$. 3) The upper bound on the number of query labels of IWAL depends on the disagreement coefficient θ.

5. Approximately Optimal Teaching: How to Teach

Error disagreement-based active learning may not easily prune the candidate hypotheses into their optimum. We thus introduce a teaching hypothesis that guides an active learner to converge with tighter bounds on generalization error and label complexity.

Theoretical analysis prove that 1) we introduce a teaching hypothesis to guide the hypothesis pruning, which results in faster pruning speed but always retains the optimal hypothesis in the candidate hypothesis set; 2) to improve the initial teaching hypothesis, self-improvement is applied and shows better learning guarantee than any initialization on the teacher. Related proofs are presented in Appendix A.

5.1 Teaching Assumption

The primary assumption of our approximately optimal teaching idea is formed as follows.

Assumption 1 *For any hypothesis class $\mathcal{H}$, assume that there exists a teaching hypothesis $h^{\mathcal{T}}$ such that tolerates an error bias ϵ:*

$$\mathcal{L}\left(h^*, h^{\mathcal{T}}\right) = \mathop{\mathbb{E}}_{x\sim\mathcal{D}_{\mathcal{X}}}\left[\max_{y}\left|\ell(h^*(x), y) - \ell(h^{\mathcal{T}}(x), y)\right|\right] < \epsilon,$$

where h^* is the optimal hypothesis in $\mathcal{H}$, and the disagreement of hypothesis invokes Eq. (1).

Note that Assumption 1 presents a formal description for our teaching idea, and we also consider a loose approximation of $h^{\mathcal{T}}$ in Section 6, i.e., ϵ is large. In real-world scenarios, it is a more practical problem and can help to improve the credibility of our assumption.

With Assumption 1, we then construct an approximation to the hypothesis class $\mathcal{H}$.

Definition 6 ***Teaching-hypothesis-class.*** *For any hypothesis class $\mathcal{H}$, $h^{\mathcal{T}}$ is a teaching hypothesis that satisfies Assumption 1. If there exists a hypothesis class $\mathcal{H}^{\mathcal{T}}$ such that $h^{\mathcal{T}} = \mathop{\operatorname{argmin}}_{h\in\mathcal{H}^{\mathcal{T}}} R(h)$, then $\mathcal{H}^{\mathcal{T}}$ is called the teaching-hypothesis-class of $\mathcal{H}$.*

By introducing a approximately optimal teaching hypothesis $h^{\mathcal{T}}$, we define a new hypothesis class $\mathcal{H}^{\mathcal{T}}$ related to $\mathcal{H}$, which uses $h^{\mathcal{T}}$ to replace the infeasible h^*. The following two feasible corollaries show the validity of Definition 6.

Corollary 7 *For any hypothesis class $\mathcal{H}$ and given $h^{\mathcal{T}}$, $\mathcal{H}^{\mathcal{T}}$ is a teaching-hypothesis-class of $\mathcal{H}$. Then we have the inequality $\mathop{\mathbb{E}}_{(x,y)\sim\mathcal{D}}\left[\ell(h^*(x), y) - \ell(h^{\mathcal{T}}(x), y)\right] \leq 0$, which requires that there are at least $\tau \geq 0$ hypotheses with tighter generalization errors than $h^{\mathcal{T}}$. Therefore, the teaching-hypothesis-class $\mathcal{H}^{\mathcal{T}}$ has fewer candidate hypotheses than $\mathcal{H}$, that is, $|\mathcal{H}^{\mathcal{T}}| \leq |\mathcal{H}|$.*

For any learning algorithm, Corollary 7 shows that hypothesis pruning in $\mathcal{H}^{\mathcal{T}}$ may have lower complexity than that of $\mathcal{H}$. Corollary 7 gives the validity of $\mathcal{H}^{\mathcal{T}}$ in terms of complexity, and the following corollary gives the validity of $\mathcal{H}^{\mathcal{T}}$ in terms of error.

Corollary 8 *For any hypothesis class $\mathcal{H}$ and given $h^{\mathcal{T}}$, $\mathcal{H}^{\mathcal{T}}$ is a teaching-hypothesis-class of $\mathcal{H}$. Based on properties of expectation, we have $\left|R(h^{\mathcal{T}}) - R(h^*)\right| \leq \mathcal{L}(h^*, h^{\mathcal{T}}) < \epsilon$.*

For any learning algorithm, Corollary 8 shows that the error of hypothesis pruning in $\mathcal{H}^{\mathcal{T}}$ is almost equal to the error of hypothesis pruning in $\mathcal{H}$. In conclusion, Corollary 7-8 initially demonstrates the validity of our teaching idea. The subsequent theorems in this paper strictly give the improved bounds on generalization error and label complexity.

5.2 Teaching Model

Before precisely presenting our theoretical results, we set some notes and explain the approximately optimal teaching model in more detail. We use $\mathcal{F}^{\mathcal{T}}(\cdot) = \mathcal{L}(h^{\mathcal{T}}, \cdot)$ to denote a disagreement feedback function with operation $\mathcal{H} \to [0, 1]$.

Teacher: the teacher has a teaching hypothesis $h^{\mathcal{T}}$, which only can provide the disagreement feedback $\mathcal{F}^{\mathcal{T}}(\cdot)$ to the **Learner**.

Learner: the learner has a teaching-hypothesis-class $\mathcal{H}^{\mathcal{T}}$, which prunes $\mathcal{H}^{\mathcal{T}}$ by identifying the disagreement feedback $\mathcal{F}^{\mathcal{T}}(\cdot)$ with the **Teacher**.

At t-time, the learner receives a sample x_t and decides whether to query the label y_t of x_t. Then the learner prunes the candidate hypothesis set based on the disagreement feedback $\mathcal{F}^{\mathcal{T}}(\cdot)$ with the teacher. The goal of the learner is to return a desired hypothesis $\widehat{h}$ from $\mathcal{H}^{\mathcal{T}}$ by using fewer labeled samples, where $\widehat{h}$ has the minimum generalization error on the input dataset $\mathcal{X}$.

The approximately optimal teaching scenario we consider is simple and practical, which merely necessitates the teacher's ability to provide the learner with disagreement feedback. In this setting, the teacher is required to be an end-to-end model which only provides output as the feedback of the input and does not know the model configuration. Therefore, the learner only requires very limited information from the teacher, which maintains a fair teaching scenario compared to those non-educated learners who do not receive any guidance from a teacher.

We follow the rules of notations used in a standard hypothesis pruning like IWAL of Section 4. Let $H_t^{\mathcal{T}}$ denote the candidate hypothesis set of the learner at t-time, where $H_1^{\mathcal{T}} = \mathcal{H}^{\mathcal{T}}$. We denote by $\widehat{h}_t = \underset{h \in H_t^{\mathcal{T}}}{\operatorname{argmin}} L_t(h)$ the current empirical optimal hypothesis, which has the minimum importance-weighted empirical error in $H_t^{\mathcal{T}}$. At T-time, the algorithm returns the current empirical optimal hypothesis $\widehat{h}_T$ as the final hypothesis output.

5.3 Teaching Improves Hypothesis Pruning

We below present the teaching-based hypothesis pruning and its theoretical improvements to defend our teaching idea. In detail, 1) we observe whether the teaching-based hypothesis

pruning strategy can prune the candidate hypothesis set faster than the error disagreement-based active learning; 2) we observe whether the optimal hypothesis can be usually maintained in the candidate hypothesis set; 3) we also present the generalization error and label complexity bounds of teaching an active learner.

Teaching-based hypothesis pruning We still follow the pruning manner of IWAL w.r.t. Eq. (4) to supervise the updates of the candidate hypothesis set, where the main difference is that we introduce a teaching hypothesis $h^{\mathcal{T}}$ to control the slack constraint of hypothesis pruning. Specifically, the slack constraint $2\Delta_t$ is tightened as $\left(1+\mathcal{F}^{\mathcal{T}}(\widehat{h}_t)\right)\Delta_t$ by invoking the guidance of a teacher, where $\mathcal{F}^{\mathcal{T}}(\widehat{h}_t)$ denotes disagreement feedback with the teacher w.r.t. current empirical optimal hypothesis $\widehat{h}_t$. With such operation, the candidate hypothesis set $H_{t+1}^{\mathcal{T}}$ at $(t+1)$-time is updated by

$$H_{t+1}^{\mathcal{T}}=\left\{h\in H_t^{\mathcal{T}} : L_t(h)\leq L_t(\widehat{h}_t)+\left(1+\mathcal{F}^{\mathcal{T}}(\widehat{h}_t)\right)\Delta_t\right\}, \tag{7}$$

where $H_1^{\mathcal{T}}=\mathcal{H}^{\mathcal{T}}$, and $\Delta_t=\sqrt{(2/t)\log(2t(t+1)|\mathcal{H}^{\mathcal{T}}|^2/\delta)}$ for some fixed confidence parameter $\delta>0$. Therefore teaching-based hypothesis pruning is more aggressive in shrinking the candidate hypothesis set, resulting in better learning guarantees.

Pruning speed With a fast hypothesis pruning speed, the candidate hypothesis set $H_t^{\mathcal{T}}$ is shrunk rapidly, which reduces the learning difficulty, easily converting into $h^{\mathcal{T}}$. The primary determinant of pruning speed is the pruning slack term, i.e., $\left(1+\mathcal{F}^{\mathcal{T}}(\widehat{h}_t)\right)\Delta_t$ of Eq. (7). With Eqs. (4) and (7), there is $\left(1+\mathcal{F}^{\mathcal{T}}(\widehat{h}_t)\right)\Delta_t\leq 2\Delta_t$, which means that the teaching-based hypothesis pruning employs a tighter slack term to shrink $H_t^{\mathcal{T}}$ than IWAL. It then leads to a faster pruning speed for our teaching strategy. Therefore, our teaching-based hypothesis pruning may be easier to prune the candidate hypotheses into their optimum than the error disagreement-based hypothesis pruning.

Retain the teaching hypothesis To evaluate Remark 3 of teaching-based hypothesis pruning, we present our analysis. The following lemma relates importance-weighted empirical error to the generalization error.

Lemma 9 *For any teaching-hypothesis-class $\mathcal{H}^{\mathcal{T}}$, the instruction of an active learner is conducted within $\mathcal{H}^{\mathcal{T}}$, where the sequence of candidate hypothesis sets satisfies $H_{t+1}^{\mathcal{T}}\subseteq H_t^{\mathcal{T}}$ with $H_1^{\mathcal{T}}=\mathcal{H}^{\mathcal{T}}$. Given any $\delta>0$, with a probability at least $1-\delta$, for any $T\in\mathbb{N}^+$ and for all $h,h'\in H_T^{\mathcal{T}}$, the following inequality holds:*

$$\left|L_T(h)-L_T(h')-\big(R(h)-R(h')\big)\right|\leq\big(1+\mathcal{L}(h,h')\big)\Delta_T,$$

where $\Delta_T=\sqrt{(2/T)\log(2T(T+1)|\mathcal{H}^{\mathcal{T}}|^2/\delta)}$.

Lemma 9 indicates that the generalization error is concentrated near its importance-weighted empirical error for every pair $\{h,h'\}\subseteq\mathcal{H}$. Based on Lemma 9, we can derive the Theorem 10, which connects the importance-weighted empirical error of the teacher and the learner.

Theorem 10 *For any teaching-hypothesis-class $\mathcal{H}^{\mathcal{T}}$, the instruction of an active learner is conducted within it. Given any $\delta>0$, with a probability at least $1-\delta$, for any $t\in\mathbb{N}^+$, the*

following inequality holds:

$$L_t(h^{\mathcal{T}}) - L_t(\widehat{h}_t) \leq \left(1 + \mathcal{F}^{\mathcal{T}}(\widehat{h}_t)\right) \Delta_t.$$

Theorem 10 shows that the teaching hypothesis $h^{\mathcal{T}}$ satisfies the pruning rule with a high probability at any t-time. And $h^{\mathcal{T}}$ is the optimal hypothesis in the teaching-hypothesis-class $\mathcal{H}^{\mathcal{T}}$. Thus teaching-based hypothesis pruning maintains the optimal hypothesis in the candidate hypothesis set with a high probability.

Learning guarantees To demonstrate the improvement of teaching an active learner (w.r.t. Remark 4), we present the learning guarantee for teaching-based hypothesis pruning in Theorem 11.

Theorem 11 *For any teaching-hypothesis-class $\mathcal{H}^{\mathcal{T}}$, the instruction of an active learner is conducted within it. Given any $\delta > 0$, with a probability at least $1 - \delta$, for any $T \in \mathbb{N}^+$, the following holds:*

1) the generalization error holds

$$R(\widehat{h}_T) \leq R(h^*) + \left(2 + \mathcal{F}^{\mathcal{T}}(\widehat{h}_{T-1}) + \mathcal{F}^{\mathcal{T}}(\widehat{h}_T)\right) \Delta_{T-1} + \epsilon;$$

2) if the learning problem has disagreement coefficient θ, the label complexity is at most

$$\tau_T \leq 2\theta \left(2TR(h^*) + \left(3 + \mathcal{F}^{\mathcal{T}}(\widehat{h}_{T-1})\right)\right) O(\sqrt{T}) + 2T\epsilon\Big).$$

Theorem 11 shows the generalization error and label complexity bounds of approximately optimal teaching an active learner. The performance of approximately optimal teaching relies on two key factors: firstly, the effectiveness of active learning, i.e., the magnitude of the teacher's disagreement feedback $\mathcal{F}(\widehat{h})$ of the learner; and secondly, the quality of the teacher, which is determined by the maximum level of disagreement ϵ between the teaching hypothesis and the optimal hypothesis. In particular, when $\widehat{h}_{T-1}, \widehat{h}_T$ are sufficiently close to $h^{\mathcal{T}}$ and ϵ is a tolerable error, the generalization error upper bound can be reduced from $R(h^*) + 4\Delta_{T-1}$ to approximately $R(h^{\mathcal{T}}) + 2\Delta_{T-1}$, and the label complexity upper bound can be decreased from $4\theta \left(TR(h^*) + 2O(\sqrt{T})\right)$ to approximately $2\theta \left(2TR(h^{\mathcal{T}}) + 3O(\sqrt{T})\right)$.

In conclusion, by improving hypothesis pruning, approximately optimal teaching guides an active learner to converge into tighter bounds on generalization error and label complexity.

6. Self-improvement of Teaching: How to Teacher Better

Section 5.3 demonstrates that approximately optimal teaching an active learner is effective. However, for Assumption 1, if the teaching hypothesis is loosely approximated to the optimal hypothesis, i.e. ϵ is large, how do we guarantee the convergence of approximately optimal teaching? We thus design a self-improvement of teaching strategy, which generates new hypotheses after each hypothesis pruning and determines whether to update the teacher.

We then observe the improvement of teaching performance and further analyze gains for the active learner of the bounds on generalization error and label complexity.

New hypotheses Since hypothesis pruning is a process of shrinking the candidate hypothesis set, generating new hypotheses should not interrupt this process. More specifically, we require that the sequence of candidate hypothesis sets satisfy $\mathrm{Conv}(H_{t+1}^{\mathcal{T}}) \subseteq \mathrm{Conv}(H_t^{\mathcal{T}})$, where $\mathrm{Conv}(\cdot)$ is the convex hull of a set. To avoid any confusion, we denote by H_t' the candidate hypothesis set after pruning at t-time, i.e., $H_t' = \left\{ h \in H_t^{\mathcal{T}} : L_t(h) \leq L_t(\widehat{h}_t) + \left(1 + \mathcal{F}^{\mathcal{T}}(\widehat{h}_t)\right) \Delta_t \right\}$ w.r.t. Eq. (7). This indicates the candidate generation base for the new hypotheses, which is aligned with the teacher's hypothesis and ensures the reliability of the subsequent generation. Therefore, after pruning the hypothesis set from $H_t^{\mathcal{T}}$ to H_t' at t-time, we generate new hypotheses $\tilde{h}$ from the convex hull of H_t', that is, drawing $\tilde{h}$ from $\mathrm{Conv}(H_t')$:

$$\tilde{h} = \sum_{j}^{m} \lambda_j h_j, h_j \in H_t', \tag{8}$$

subjected to $\sum_j^m \lambda_j = 1, \lambda_j \in [0,1]$, where m denotes the size of H_t', and this ensures $\tilde{h} \in \mathrm{Conv}(H_t')$. We use Eq. (8) to draw n hypotheses for obtaining the hypothesis generation set $\widetilde{H}_t' = \left\{ \tilde{h}_i; i \in [n] \right\}$ and combine it with H_t' as the candidate hypothesis set next time: $H_{t+1}^{\mathcal{T}} = H_t' \cup \widetilde{H}_t'$.

Remark 12 *In other words, our teaching scheme incorporates a dual-loop mechanism, consisting of: 1) outer candidate hypothesis pruning, and 2) inner hypothesis generation. In the inner loop, the candidate hypothesis set $H_t^{\mathcal{T}}$ is redefined as the candidate generation base H_t' after t iterations of pruning for subsequent hypothesis generation. Within this framework, the hypothesis generation set is denoted as $\widetilde{H}_t'$ w.r.t. Eq. (8). In the outer loop, the candidate hypothesis set is iteratively pruned, and the hypothesis generation set also follows. Consequently, the updated of $H_{t+1}^{\mathcal{T}} \leftarrow H_t^{\mathcal{T}}$ is expressed as $H_{t+1}^{\mathcal{T}} = H_t' \cup \widetilde{H}_t'$.*

Self-improvement The new hypotheses may perform better than the teaching hypothesis, that is, $R(\tilde{h}) \leq R(h^{\mathcal{T}})$. By adding a restriction to the loss function, we give a condition for determining whether the teacher improves or not. We assume $\ell(h(x), y) = \phi\left(yh(x)\right)$, where ϕ is functional non-increasing and convex. In short, $\ell(h(x), y)$ can be specified as 0-1, hinge, logistic loss functions, etc. Under the additional assumptions of the loss function, the following lemma reveals the variation of the maximum error disagreement in the candidate hypothesis set.

Note that the loss function $\ell(h(x), y)$ is modeled using a non-increasing and convex function ϕ, which operates on the product $yh(x)$ of the true label and the predicted value. This structure ensures that better and more confident predictions lead to lower loss values, making the optimization process more effective. Common examples of such loss functions include the 0-1 loss, hinge loss, and logistic loss, each suited to different types of machine learning models and optimization strategies.

By understanding the role and properties of ϕ, we can design loss functions that are both theoretically sound and practically useful for various learning algorithms.

Lemma 13 *For any teaching-hypothesis-class $\mathcal{H}^{\mathcal{T}}$, the instruction of an active learner is conducted within it. If the loss function can be rewritten to form $\ell(h(x), y) = \phi(yh(x))$ and the function ϕ is non-increasing and convex, for any candidate hypothesis set $H_t^{\mathcal{T}}$ and for all $x \in \mathcal{X}$, the following equation holds:*

$$\max_{h,h' \in \mathrm{Conv}(H_t^{\mathcal{T}})} \max_y \left|\ell(h(x), y) - \ell(h'(x), y)\right| = \max_{h,h' \in H_t^{\mathcal{T}}} \max_y \left|\ell(h(x), y) - \ell(h'(x), y)\right|,$$

where $\mathrm{Conv}(H_t^{\mathcal{T}})$ *is the convex hull of the hypothesis set* $H_t^{\mathcal{T}}$.

Lemma 13 shows that the maximum error disagreement at a certain fixed sample x will not increase despite the learning algorithm generating new hypotheses in the convex hull of $H_t^{\mathcal{T}}$. Based on this property, Theorem 14 presents the lower bound analysis for the generalization error difference between the teaching hypothesis $h^{\mathcal{T}}$ and new hypotheses $\tilde{h}$.

Theorem 14 *For any teaching-hypothesis-class $\mathcal{H}^{\mathcal{T}}$, the instruction of an active learner is conducted within it, where the sequence of candidate hypothesis sets satisfies* $\mathrm{Conv}(H_{t+1}^{\mathcal{T}}) \subseteq \mathrm{Conv}(H_t^{\mathcal{T}})$ *with $H_1^{\mathcal{T}} = \mathcal{H}^{\mathcal{T}}$. For any $t \in \mathbb{N}^+$, given any $\delta > 0$, with a probability at least $1 - \delta$, for any $\tilde{h} \in \widetilde{H}_t'$, the following inequality holds:*

$$R(h^{\mathcal{T}}) - R(\tilde{h}) \geq L_t(h^{\mathcal{T}}) - L_t(\tilde{h}) - \left(1 + \mathcal{F}^{\mathcal{T}}(\tilde{h})\right) \Delta_t. \tag{9}$$

Theorem 14 elaborates on the determining conditions necessary for self-improvement:: if $\beta_i^{(t)} = L_t(h^{\mathcal{T}}) - L_t(\tilde{h}_i) - \left(1 + \mathcal{F}^{\mathcal{T}}(\tilde{h}_i)\right) \Delta_t > 0$, i.e., $R(\tilde{h}_i) < R(h^{\mathcal{T}})$, then the teaching hypothesis of $\mathcal{H}^{\mathcal{T}}$ is updated to $h^{\mathcal{T}} = \widetilde{h}_i$. Thus self-improvement of teaching strategy reduces generalization error of the teaching hypothesis without excessive additional calculations.

Improvement of teaching performance Self-improvement of teaching strategy obtain a teaching hypothesis sequence $\left\{h_1^{\mathcal{T}}, ..., h_T^{\mathcal{T}}\right\}$, where $h_t^{\mathcal{T}}$ denote the optimal hypothesis in $\bigcup_{k=1}^{t} H_k^{\mathcal{T}}$. Based on Theorem 14, Corollary 15 gives the improvement of teaching performance.

Corollary 15 *For any teaching-hypothesis-class $\mathcal{H}^{\mathcal{T}}$, tthe instruction of an active learner is conducted within it. Let $\alpha_t = \max\{\max_i \beta_i^{(t)}, 0\}$ and $h_1^{\mathcal{T}}$ be the initial teaching hypothesis. If the self-improvement of teaching is applied, given any $\delta > 0$, with a probability at least $1 - \delta$, for any $T \in \mathbb{N}^+$, we have an inequality $R(h_T^{\mathcal{T}}) \leq R(h_1^{\mathcal{T}}) - \sum_{t=1}^{T-1} \alpha_t$.*

Corollary 15 guarantees that, with high probability, self-improvement of teaching can reduce the generalization error of the initial teaching hypothesis by at least $\sum_{t=1}^{T-1} \alpha_t$. Moreover, assuming ϵ is the initial approximation error of the teaching hypothesis to the optimal hypothesis, we have an inequality $R(h_T^{\mathcal{T}}) \leq R(h^*) + \epsilon_T$, where $\epsilon_T := \epsilon - \sum_{t=1}^{T-1} \alpha_t (T > 1)$ and $\epsilon_1 = \epsilon$. Thus the self-improvement of teaching alleviates the loose approximation of the teaching hypothesis to the optimal hypothesis w.r.t. ϵ of Assumption 1.

Learning guarantees With the improvement of the teacher, the improvement of the learner is natural. Recalling Theorem 11, we then present the learning guarantees for the self-improvement of teaching. The primary motivation is to replace the pre-defined teaching hypothesis $h^{\mathcal{T}}$ by a teaching hypothesis sequence $\left\{h_1^{\mathcal{T}}, ..., h_T^{\mathcal{T}}\right\}$. At t-time, we denote by

$\mathcal{F}_t^{\mathcal{T}}(\cdot) := \mathcal{L}(h_t^{\mathcal{T}}, \cdot)$ the disagreement feedback with latest teaching hypothesis $h_t^{\mathcal{T}}$. Because the disagreement coefficient θ w.r.t. Eq. (6) is defined based on the varying candidate hypothesis set $H_t^{\mathcal{T}}$, it varies with time t. To make the theoretical results more concise, we assume that θ is stable for smooth distribution and does not change dramatically as $H_t^{\mathcal{T}}$ changes. The learning guarantees of Theorem 11 are then re-derived.

Theorem 16 *For any teaching-hypothesis-class $\mathcal{H}^{\mathcal{T}}$, the instruction of an active learner is conducted within it. If the self-improvement of teaching is applied, given any $\delta > 0$, with a probability at least $1-\delta$, for any $T \in \mathbb{N}^+$, the following holds: 1) for any $t \in [T]$, holds $h_t^{\mathcal{T}} \in H_t^{\mathcal{T}}$;*

2) the generalization error holds

$$R(\widehat{h}_T) \leq R(h^*) + \left(2 + \mathcal{F}_{T-1}^{\mathcal{T}}(\widehat{h}_{T-1}) + \mathcal{F}_{T-1}^{\mathcal{T}}(\widehat{h}_T)\right)\Delta_{T-1} + \epsilon_{T-1};$$

3) if the learning problem has disagreement coefficient θ, the label complexity is at most

$$\tau_T \leq 2\theta\left(2TR(h^*) + \left(3 + \mathcal{F}_{T-1}^{\mathcal{T}}(\widehat{h}_{T-1})\right)O(\sqrt{T}) + 2T\epsilon_{T-1}\right).$$

Theorem 16 shows that the optimal hypothesis of $\bigcup_{k=1}^{t} H_k^{\mathcal{T}}$ is maintained in the candidate hypothesis set with a high probability at any t-time. Recalling Corollary 15, we have $\epsilon = \epsilon_1 \leq \epsilon_{T-1}$, which shows that self-improvement of teaching strategy can further reduce the generalization error and label complexity bounds of the learner w.r.t. Theorem 11. Moreover, the improvement of the active learner is decided by the improvement of the approximately optimal teacher.

In conclusion, by generating new hypotheses, self-improvement of teaching strategy tightens the approximation of the teaching hypothesis to the optimal hypothesis, which provides more favorable learning guarantees for an active learner.

7. Teaching-based Mentored Learning: How to Learn

Based on the theoretical results of Section 5, we present the Approximately Optimal Teaching-based Mentored Learning algorithm (ATML), which guides a white-box learner. To guide a black-box learner, we then extend ATML into ATML$^+$.

7.1 Teaching a White-box Learner

We here consider the teaching for a white-box learner who discloses its hypothesis class information to the teacher. In this setting, the learner prunes the teaching-hypothesis-class $\mathcal{H}^{\mathcal{T}}$ by querying the sample labels and finally outputs a desired hypothesis. In each round, ATML includes three stages: 1) query, 2) hypothesis pruning, and 3) self-improvement. Its pseudo-code is presented in Algorithm 1.

Query (Steps 3-6) On the setting of white-box learner, ATML adopts a similar label query strategy as IWAL w.r.t. Eq. (2), with a slightly different hypothesis class. At t-time, ATML

Algorithm 1 ATML$(\mathcal{H}^{\mathcal{T}}, h^{\mathcal{T}}, T, n)$

Initialize: $H_1^{\mathcal{T}} = \mathcal{H}^{\mathcal{T}}, h_1^{\mathcal{T}} = h^{\mathcal{T}}, \widetilde{H}_t' = \emptyset$
for $t \in [T]$ **do**
  $p_t = \max_{h,h' \in H_t^{\mathcal{T}}} \max_y |\ell(h(x_t), y) - \ell(h'(x_t), y)|$ w.r.t. Eq. (10)
  $Q_t \in \{0, 1\}$ with $Q_t \sim \mathcal{B}(1, p_t)$
  if $Q_t = 1$ **then**
    $y_t \leftarrow \text{LABEL}(x_t)$
    $\widehat{h}_t = \operatorname{argmin}_{h \in H_t^{\mathcal{T}}} L_t(h)$ w.r.t. Eq. (3)
    $H_t' = \left\{ h \in H_t^{\mathcal{T}} : L_t(h) \leq L_t(\widehat{h}_t) + \left(1 + \mathcal{F}^{\mathcal{T}}(\widehat{h}_t)\right) \Delta_t \right\}$ w.r.t. Eq. (7)
    for $i \in [n]$ **do**
      $\tilde{h} = \sum_j^m \lambda_j h_j$ where $h_j \in H_t'$ w.r.t. Eq. (8)
      $\widetilde{H}_t' = \widetilde{H}_t' \cup \tilde{h}$
      $h_{t+1}^{\mathcal{T}} = \tilde{h}$ **if** $R(\tilde{h}) < R(h^{\mathcal{T}})$ **else** $h_t^{\mathcal{T}}$ w.r.t. Eq. (9)
    end for
    $H_{t+1}^{\mathcal{T}} = H_t' \cup \widetilde{H}_t'$
  end if
end for
Return $\widehat{h}_T$

does a Bernoulli trial Q_t with a success probability p_t:

$$p_t = \max_{h,h' \in H_t^{\mathcal{T}}} \max_y \left| \ell(h(x_t), y) - \ell(h'(x_t), y) \right|. \tag{10}$$

If $Q_t = 1$, the algorithm queries the label y_t of x_t.

Hypothesis pruning (Step 7-8) At any t-time, ATML maintains a candidate hypothesis set $H_t^{\mathcal{T}}$ with $H_1^{\mathcal{T}} = \mathcal{H}^{\mathcal{T}}$. After querying the label, ATML updates the current empirical optimal hypothesis $\widehat{h}_t = \operatorname{argmin}_{h \in H_t^{\mathcal{T}}} L_t(h)$ w.r.t. Eq. (3). Then the algorithm prunes the candidate hypothesis set from $H_t^{\mathcal{T}}$ to hypothesis generation base H_t' according to Eq. (7). At T-time, ATML returns the hypothesis $\widehat{h}_T$ as the final hypothesis output. Note that $L_t(h)$ of Eq. (3) denotes the importance-weighted empirical error of hypothesis $h \in \mathcal{H}$, and it is incrementally updated with additional sampling of (x_t, y_t).

Self-improvement (Steps 9-14) After the hypothesis pruning, ATML will generate new hypotheses to improve the performance of the teaching hypothesis. At t-time, ATML generates new hypotheses $\tilde{h}_i$ from the convex hull of H_t' according to Eq. (8) and obtains the generation hypothesis set $\widetilde{H}_t' = \left\{ \tilde{h}_i; i \in [n] \right\}$. Next, the algorithm updates the teaching hypothesis according to Eq. (9) and uses $H_{t+1}^{\mathcal{T}} = H_t' \cup \widetilde{H}_t'$ as the candidate hypothesis set at $(t+1)$-time. More detailed contents can be found in the part of "New hypothesis" of Section 6 and Remark 12. Note that this is an augmentation of the candidate hypothesis class due to the generation of new hypotheses.

7.2 Teaching a Black-box Learner

Here, we consider a more challenging problem: the learner is also a black-box who can not disclose its hypothesis class information to the teacher. In this setting, the teaching-hypothesis-class $\mathcal{H}^{\mathcal{T}}$ of learner is non-transparent. Therefore, the learner tries to converge to the optimal hypothesis from an initial hypothesis $\widehat{h}_0$ by incremental updates. We extend ATML into ATML$^+$ for teaching a black-box learner. In each round, ATML$^+$ includes three stages: 1) query, 2) hypothesis pruning, and 3) self-improvement. Its pseudo-code is presented in Algorithm 2.

Algorithm 2 ATML$^+(h^{\mathcal{T}}, T, n)$

1: **Initialize:** Teacher $h_1^{\mathcal{T}} = h^{\mathcal{T}}$, Learner $\widehat{h}_0$
2: **for** $t \in [T]$ **do**
3: $p_t = \max_y |\ell(h_t^{\mathcal{T}}(x), y) - \ell(\widehat{h}_{t-1}(x), y)|$ w.r.t. Eq. (11)
4: $Q_t \in \{0, 1\}$ with $Q_t \sim \mathcal{B}(1, p_t)$
5: **if** $Q_t = 1$ **then**
6: $y_t \leftarrow \text{LABEL}(x_t)$
7: $\widehat{h}_t = \text{argmin } L_t(h)$ with SGD w.r.t. Eq. (3)
8: **if** $L_{t-1}(\widehat{h}_t) > L_{t-1}(\widehat{h}_{t-1}) + \Phi_{t-1}$ **then**
9: $\widehat{h}_t = \widehat{h}_{t-1}$
10: **end if**
11: **for** $i \in [n]$ **do**
12: $\tilde{h} \leftarrow \lambda h_i^{\mathcal{T}} + (1 - \lambda)\widehat{h}_t$
13: $h_{t+1}^{\mathcal{T}} = \tilde{h}$ **if** $R(\tilde{h}) < R(h^{\mathcal{T}})$ **else** $h_t^{\mathcal{T}}$ w.r.t. Eq. (9)
14: **end for**
15: **end if**
16: **end for**
17: **Return** $\widehat{h}_T$

Query (Steps 3-6) In the setting of black-box learner, the maximum error disagreement of the candidate hypothesis set $H_t^{\mathcal{T}}$ cannot be obtained. We thus re-characterize query probability p_t by the maximum error disagreement between teacher and learner:

$$p_t = \max_y \left| \ell(h_t^{\mathcal{T}}(x), y) - \ell(\widehat{h}_{t-1}(x), y) \right|. \tag{11}$$

Formally, p_t should be defined as $p_t = \max_y \left| \ell(h_t^{\mathcal{T}}(x), y) - \ell(\widehat{h}_t(x), y) \right|$. Before the update on $\widehat{h}_t$ at t-time, $\widehat{h}_{t-1}$ is used to approximate $\widehat{h}_t$. At t-time, ATML$^+$ does a Bernoulli trial Q_t with success probability p_t to decide whether to query the label of x_t.

Hypothesis pruning(Steps 7-10) Since the teaching-hypothesis-class $\mathcal{H}^{\mathcal{T}}$ is non-transparent, hypothesis pruning is generalized as a constraint in incremental updates. Specifically, we present a backtracking approach to ensure that the current empirical optimal hypothesis $\widehat{h}_t$ is maintained in the candidate hypothesis set $H_t^{\mathcal{T}}$. After the learner updates to $\widehat{h}_t$ using a standard optimization algorithm such as stochastic gradient descent (SGD) (Ruder, 2016) at t-time, we judge whether the hypothesis pruning rule of $\widehat{h}_t$ is satisfied at $(t-1)$-time by

the following inequality:

$$L_{t-1}(\widehat{h}_t) \leq L_{t-1}(\widehat{h}_{t-1}) + \Phi_{t-1}, \tag{12}$$

where $\Phi_{t-1} = \left(1 + \mathcal{F}^{\mathcal{T}}_{t-1}(\widehat{h}_{t-1})\right) \Delta_{t-1}$ represents the slack term. If Eq. (12) does not satisfy, it means that $\widehat{h}_t$ has already been pruned, and we backtrack the hypothesis $\widehat{h}_t = \widehat{h}_{t-1}$. The backtracking approach forces the learner not to be updated too far for each update, which prevents the learner to be disordered when updating towards a subsequent hypothesis.

Note that we can not prune the candidate hypothesis $H^{\mathcal{T}}_t$ to H'_t as Line 8 of Algorithm 1 due to the black-box learner setting, i.e., $H^{\mathcal{T}}_1 = \mathcal{H}^{\mathcal{T}}$ cannot be initialized. Alternatively, we employ the loss inequality of Eq. (12) to determine whether $\widehat{h}_t$ has been pruned.

Self-improvement(Steps 11-14) Because the teaching-hypothesis-class $\mathcal{H}^{\mathcal{T}}$ of the learner is non-transparent, we cannot generate new hypotheses from the convex hull of the candidate hypothesis set. We suggest generating new hypotheses by a linear combination of the teaching hypothesis $h^{\mathcal{T}}_t$ and the current empirical optimal hypothesis $\widehat{h}_t$:

$$\tilde{h} = \lambda h^{\mathcal{T}}_t + (1 - \lambda)\widehat{h}_t. \tag{13}$$

At t-time, ATML$^+$ generates n new hypotheses by Eq. (13) and determines whether updating the teacher according to Eq. (9).

Note that we can not determine whether the candidate hypothesis sets satisfy $\mathrm{Conv}(H^{\mathcal{T}}_{t+1}) \subseteq \mathrm{Conv}(H^{\mathcal{T}}_t)$ due to the black-box setting. Thus, we cannot perform the new hypothesis generation of Line 10 of Algorithm 1. Alternatively, we introduce the teacher hypothesis $h^{\mathcal{T}}_t$ for new hypothesis generation according to Eq. (13).

8. Experiments

To demonstrate our teaching idea of Section 5, we present the empirical studies for teaching-based hypothesis pruning of Section 5.3, and the self-improvement of teaching of Section 6. With their guarantees, we then present real-world studies for ATML of Section 7.1 and ATML$^+$ of Section 7.2.

Dataset We experimented with algorithms on 7 binary classification datasets: *skin, shuttle, magic04, covtype, nomao, jm1 and mnist*. Table 2 shows the summary statistics for all datasets used in our experiment. We denote by N the number of samples, by Dim the number of features, and R is the relative size of the minority class. For the high-dimensional datasets (*covtype, nomao, jm1*), we only keep the first 10 principal components of its original features. For the multi-class datasets (*shuttle, covtype*), we set the majority class as positive classes and all the remaining classes as negative classes. For *mnist* dataset, we set the digit 3 as the positive class and the digit 5 as the negative class. For all datasets, we normalize each feature to $[0, 1]$.

8.1 Empirical Studies

We present the following empirical studies on six UCI binary classification datasets: 1) whether the teaching-based hypothesis pruning of ATML can prune the candidate hypothesis

Table 2: Dataset summary in experiments.

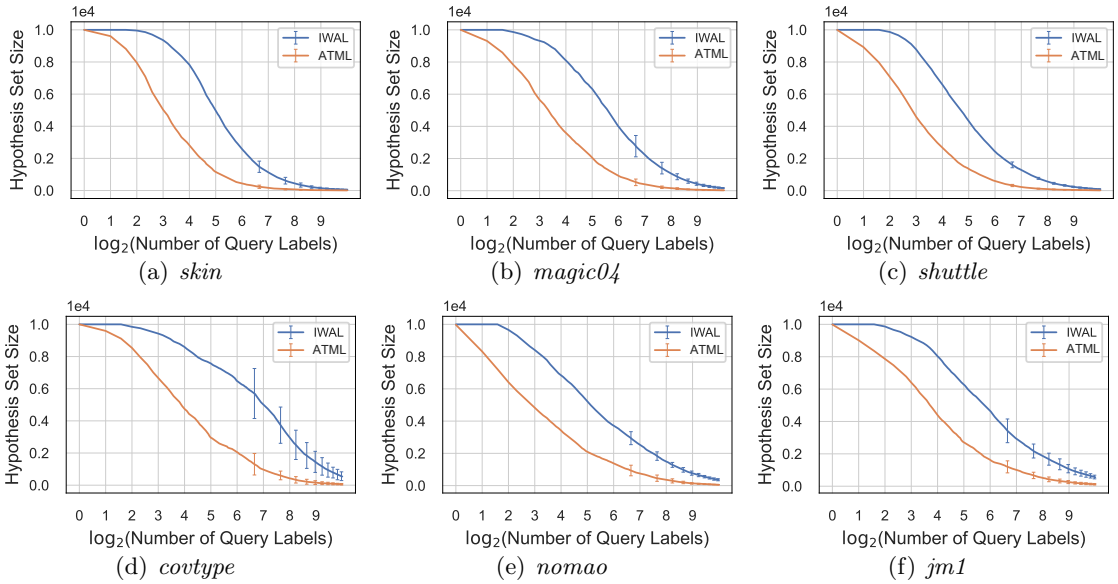
Dataset	N	Dim	R
skin	245,057	3	0.208
magic04	19,020	10	0.352
shuttle	43,500	9	0.216
covtype	581,012	54	0.488
nomao	34,465	118	0.286
jm1	10,880	21	0.193
mnist	11,552	784	0.469

set faster than hypothesis pruning of IWAL; 2) whether self-improvement of teaching strategy of ATML can reduce the generalization error of teaching hypothesis.

Setting In our empirical studies, we randomly generate $10,000$ hyperplanes with bounded norms as the initial hypothesis class $\mathcal{H}^{\mathcal{T}}$ and set the teaching hypothesis as that hypothesis with the minimum empirical error from $\mathcal{H}^{\mathcal{T}}$. For a Dim-dimensional dataset, the sample can be described as $\vec{x} = (x_1, ..., x_{Dim})$. Correspondingly, the generated hyperplanes are $Dim + 1$-dimensional and can be parameterized as $\vec{w} = (w_1, ..., w_{Dim}, b)$, where b is the bias term. Thus the prediction of the hypothesis is $h(x) = \sum_{n=1}^{Dim} w_n x_n + b$. For all $(x, y) \in \mathcal{X} \times \mathcal{Y}$, the loss function is written as $\ell(h(x), y) = \log\big(1 + \exp\big(-yh(x)\big)\big)$, and we use function $g\left(\ell(h(x), y)\right) = 2/\left(1 + \exp\left(-\ell(h(x), y)\right)\right) - 1$ to normalize the output of $\ell(h(x), y)$ to $[0, 1]$. The hypothesis pruning strategy of IWAL follows Section 4.2, and ATML follows Section 5.3. To reduce computation, we use 10% of unlabeled samples of $\mathcal{X}$ to calculate approximately $\mathcal{L}(\cdot, \cdot)$ w.r.t. Eq. (1). For example, at t-time, for all $x \in S$, we solve for the disagreement feedback of teacher and learner by traversing the label y, where S is the unlabeled data subset of $\mathcal{X}$ s.t. $|S| = 10\% \times |\mathcal{X}|$. If the dataset is split into training set and test set, we use 10% of training set for calculating $\mathcal{L}(\cdot, \cdot)$ approximately to prevent leakage of test set information. We repeat the empirical studies 20 times on each dataset and collect the average results with standard error.

Teaching-based hypothesis pruning To analyze the hypothesis pruning performance of our teaching idea, we employ IWAL to compare our proposed ATML in the specified $\mathcal{H}^{\mathcal{T}}$. The size of the candidate hypothesis set written as $|H_t^{\mathcal{T}}|$ is generalized as a feasible measure to show the pruning speed. We thus present the dynamic change of $|H_t^{\mathcal{T}}|$ with the number of query labels (on $\log_2$ scale) in Figure 2. Since ATML applies a tighter hypothesis pruning slack term (w.r.t. Eq. (7)) under the guidance of the approximately optimal teacher, its pruning speed is naturally faster than that of IWAL in terms of the $|H_t^{\mathcal{T}}|$.

Self-improvement of teaching To verify the effectiveness of the self-improvement of teaching, we observe changes in the generalization error of the teaching hypothesis for ATML. For each dataset, we randomly select 50% of the data as the training set and approximate the generalization error of the teaching hypothesis by the empirical error on the remaining data. The results are presented in Figure 3. With self-improvement of teaching, ATML gradually tightens the approximation of the teaching hypothesis to the optimal hypothesis.

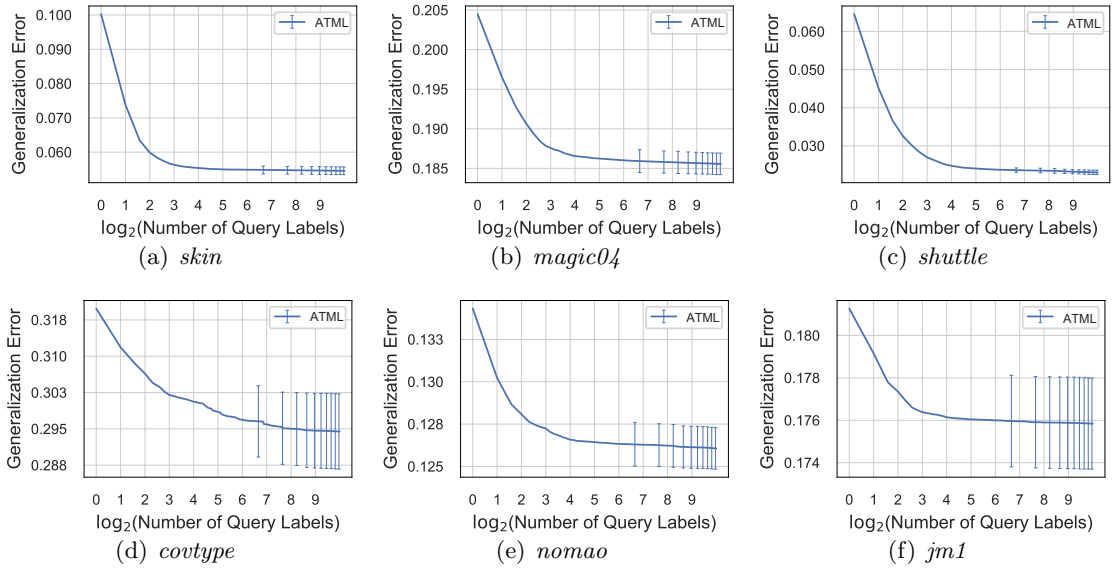


(a) *skin* (b) *magic04* (c) *shuttle*

(d) *covtype* (e) *nomao* (f) *jm1*

Figure 2: The size of the candidate hypothesis set of IWAL and ATML vs. the number of query labels (log 2 scale).

It then leads to the continuous and steady decreases in the generalization error curve of the teaching hypothesis for ATML.



(a) *skin* (b) *magic04* (c) *shuttle*

(d) *covtype* (e) *nomao* (f) *jm1*

Figure 3: The generalization error of teaching hypothesis for ATML vs. the number of query labels (log 2 scale).

8.2 Real-world Studies

We present the performance of ATML and ATML$^+$ in real-world studies. We first report the performance of ATML in the setting of white-box learner, where IWAL(Beygelzimer et al., 2009) and IWAL-D(Cortes et al., 2019b) are used as the baseline. We then report the performance of ATML$^+$ in the setting of black-box learner, where MVR(Freeman, 1965), ME(Shannon, 2001), and Random(Gal et al., 2017) are used as the baseline. The reason for the different baselines in the two settings is that these algorithms are not directly applicable to each other.

White-box learner In this setting, we compare the performance of IWAL, IWAL-D, and ATML on six UCI binary classification datasets. For all algorithms, we adopt the same settings as in Section 8.1, including 1) the initialization of the hypothesis set, 2) the loss function, and 3) the calculation method of $\mathcal{L}(\cdot, \cdot)$. For each dataset, we randomly select 70% of the data as the training set and the remaining data as the test set. We run the three algorithms 20 times and collect the average results with standard error.

Firstly, we compare the performance of the hypothesis $\widehat{h}_T$ returned by IWAL, IWAL-D, and ATML. Figure 4 presents the error rate of $\widehat{h}_T$ on the test dataset against the number of query labels (on $\log_2$ scale). The $\widehat{h}_T$ returned by IWAL and IWAL-D are subjected to the initial hypothesis class, so the error rate of $\widehat{h}_T$ is almost the same. However, the self-improvement of teaching strategy for ATML can generate new hypotheses in the candidate hypothesis set, so $\widehat{h}_T$ has a lower error rate. This verifies the Theorem 16 that the learner guided by an approximately optimal teacher can converge into a tighter generalization error than those non-educated learners.

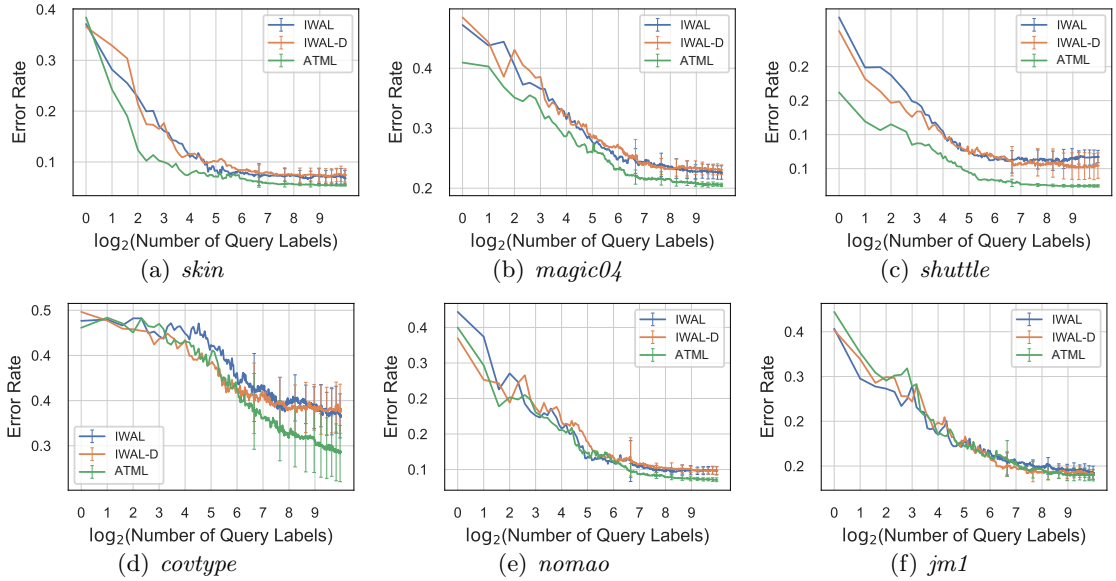


Figure 4: The error rate of IWAL, IWAL-D, and ATML on the test dataset vs. the number of query labels (log 2 scale).

Secondly, we compare the number of query labels of IWAL, IWAL-D, and ATML. Figure 5 presents the relationship between the number of query labels and the number of samples seen.

IWAL-D uses the error disagreement of the learner for hypothesis pruning and thus spends fewer the number of query labels than IWAL. ATML uses the error disagreement between the teacher and the learner for hypothesis pruning, thus spending the fewest number of query labels. This further verifies the Theorem 16 that the learner guided by an approximately optimal teacher can converge into a tighter label complexity than those non-educated learners.

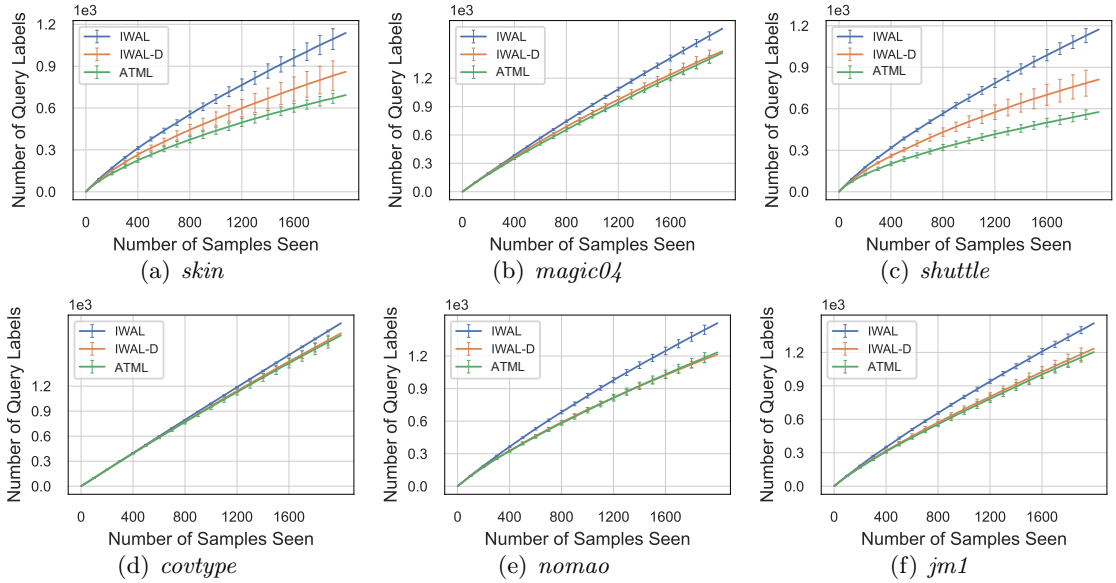


Figure 5: The number of query labels of IWAL, IWAL-D, and ATML vs. the number of samples seen.

Black-box learner In this setting, we test ATML$^+$ on the digits 3 and 5 of *mnist* dataset(Crammer et al., 2009) as a binary classification task, where 70% of dataset is randomly selected as the training set and the remaining data as the test set. As a comparison, we examine the performance of several standard active learning algorithms: 1) maximize variation ratios (MVR), 2) max entropy (ME), and 3) random (Random).

In all algorithms, we use the CNN network as a classifier following the structure of convolution-relu-convolution-relu-max pooling-dropout-dense-relu-dropout-dense, with the loss function: $\log(1+\exp(-yh(x))$ and normalize to $[0,1]$. In ATML$^+$, the teaching hypothesis is specified as a pre-trained CNN model. We repeat the learning algorithm 20 times and collect the average results with standard error.

Figure 6 presents the relationship of the test accuracy and the number of query labels, where ATML$^+$ wins the traditional active learning baselines. The reasons are two-fold: 1) the traditional active learning algorithms have an unclear purpose, which leads to convergence of incremental updates that is usually infeasible; 2) ATML$^+$ focuses on the disagreement between learner and teacher, and thus its convergence benefits from the pre-trained teaching hypothesis. This shows that teaching a black-box learner is also effective.

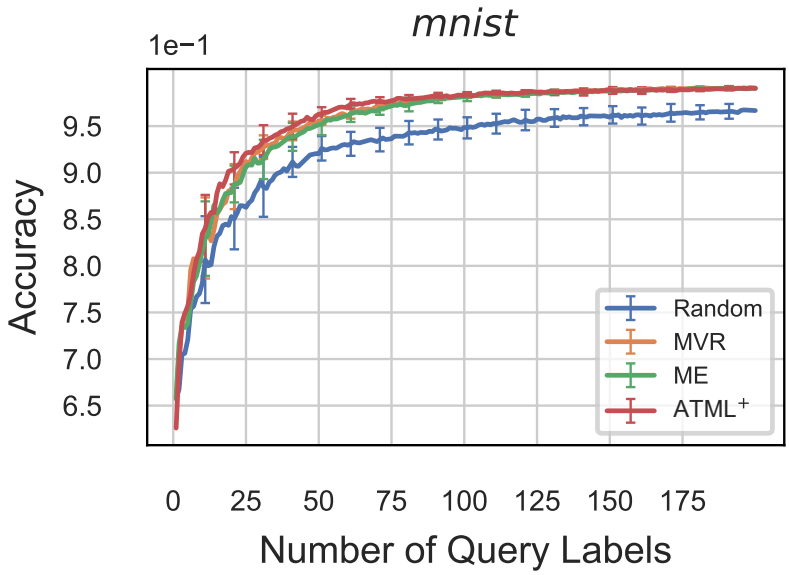


Figure 6: The accuracy of Random, MVR, ME, and ATML$^+$ on the test dataset vs. the number of query labels.

8.3 Open Discussions

The above empirical experiments evaluate the theoretical advantages of teaching. However, one question remains: does a warm start to the learning process yield better performance than teaching? We thus present the below discussions.

Our teaching scenario Student learners typically engage in an iterative process of actively updating its hypotheses, like active learning. While this behavior can be advantageous, there is an inherent risk of introducing mistakes through incremental updates including weak initialization, inaccurate or insignificant history states, resulting in expensive convergence cost.

Theoretical feasibility Incorporating teaching methods into learning endeavors helps reduce the potential risks associated with errors, inefficient learning trajectories, poor generalization, and ineffective hypothesis selection. By providing structured guidance and feedback, teaching enhances the robustness and reliability of the learning process, leading to more successful outcomes. Therefore, teaching reduces the potential risk of learning. This is theoretically feasible and statistically significant, as demonstrated in our continued work presented at ICML and NeurIPS.

Warm start accelerates learning With appropriate configurations, such as warm starting, the learning process can exhibit accelerated convergence rates. This phenomenon stems from the establishment of a robust learning foundation, which inherently facilitates improved convergence outcomes. The efficacy of this approach is contingent upon meticulous setup of the learning configuration. Consequently, in theoretical contexts, the amalgamation of robust learning mechanisms with supplementary information holds promise for enhancing the overall learning process. While acknowledging the validity of this assertion, it is pertinent to note that our primary research focus lies elsewhere. This experimental problem, characterized by non-absolute statistical significance, is intrinsically related to various configurations.

In summary, we acknowledge that a solid learning framework can markedly improve learning performance. Introducing teaching as a mentor for learning proves instrumental in mitigating ineffective or unexpected processes. While our theoretical insights establish a foundational understanding, our empirical experiments serve to validate these concepts. Nevertheless, for experimental configurations such as warm-start vs. teaching, the absence of rigorous theoretical proofs remains a notable gap.

9. Conclusion

Teaching the learner for mentored learning is a novel concept for the traditional hypothesis space theory in the machine learning community. To maintain fair teaching in error disagreement-based learners, we designate the machine teacher as approximately optimal teaching, providing only disagreement feedback to the learner. With this assumption, we introduce a teaching hypothesis to enhance hypothesis pruning, resulting in tighter bounds on generalization error and label complexity. Recognizing that the teaching hypothesis may be a loose approximation to the optimal hypothesis, we also present the self-improvement of teaching. Supported by our theoretical insights, we explore teaching for both white-box and black-box learners. Rigorous analysis and robust experiments demonstrate the effectiveness of our teaching approach.

Appendix A. Proof

Lemma 9 *For any teaching-hypothesis-class $\mathcal{H}^{\mathcal{T}}$, the instruction of an active learner is conducted within it, where the sequence of candidate hypothesis sets satisfies $H_{t+1}^{\mathcal{T}} \subseteq H_t^{\mathcal{T}}$ with $H_1^{\mathcal{T}} = \mathcal{H}^{\mathcal{T}}$. Given any $\delta > 0$, with a probability at least $1-\delta$, for any $T \in \mathbb{N}^+$ and for all $h, h' \in H_T^{\mathcal{T}}$, the following inequality holds:*

$$\left|L_T(h) - L_T(h') - \big(R(h) - R(h')\big)\right| \leq \big(1 + \mathcal{L}(h, h')\big)\,\Delta_T, \tag{14}$$

where $\Delta_T = \sqrt{(2/T)\log(2T(T+1)|\mathcal{H}^{\mathcal{T}}|^2/\delta)}$.

Proof Pick any $T \in \mathbb{N}^+$ and a pair of $h, h' \in H_T^{\mathcal{T}}$. We define a sequence of random variables $\{U_1, \cdots, U_T\}$, where $U_t (t \in [T])$ with respect to h, h':

$$U_t = \frac{Q_t}{p_t}\left[\ell(h(x_t), y_t) - \ell(h'(x_t), y_t)\right] - [R(h) - R(h')],$$

We then solve for the expectation of the random variable U_t with respect to the past:

$$\begin{aligned}
&\mathbb{E}[U_t | U_1, \cdots, U_{t-1}] \\
&= \mathop{\mathbb{E}}_{(x_t, y_t)\sim\mathcal{D}} \frac{Q_t}{p_t}[\ell(h(x_t), y_t) - \ell(h'(x_t), y_t)] - [R(h) - R(h')] \\
&= 0.
\end{aligned}$$

This indicates that U_t has zero expectation of the past, i.e., the sequence of random variables $\{U_1, \cdots, U_T\}$ is a martingale difference sequence.

In order to use the Azuma's inequality, we also need to prove the individual U_t are bounded. We split U_t into two parts and prove that each is bounded separately.

We first prove that $|\ell(h(x_t), y_t) - \ell(h'(x_t), y_t)|$ is bounded. For all hypothesis pruning strategies, the sequence of candidate hypothesis sets satisfies $H_T^{\mathcal{T}} \subseteq H_{T-1}^{\mathcal{T}} \subseteq \cdots \subseteq H_1^{\mathcal{T}} = \mathcal{H}^{\mathcal{T}}$. Thus for all $t \leq T$, combine the definition of p_t, we have:

$$\begin{aligned}
&|\ell(h(x_t), y_t) - \ell(h'(x_t), y_t)| \\
&\leq \max_y \left|\ell(h(x_t), y) - \ell(h'(x_t), y)\right| \\
&\leq \max_{h_1, h_2 \in H_T^{\mathcal{T}}} \max_y |\ell(h_1(x_t), y) - \ell(h_2(x_t), y)| \\
&\leq \max_{h_1, h_2 \in H_t^{\mathcal{T}}} \max_y |\ell(h_1(x_t), y) - \ell(h_2(x_t), y)| \\
&= p_t.
\end{aligned} \tag{15}$$

Considering $H_T^{\mathcal{T}} \subseteq H_t^{\mathcal{T}}$, given the loss disagreement of $\max_{h_1,h_2} \max_y |\ell(h_1(x_t), y) - \ell(h_2(x_t), y)|$, the upper bound subjected to "$h_1, h_2 \in H_T^{\mathcal{T}}$" is tighter than that of "$h_1, h_2 \in H_t^{\mathcal{T}}$". From the perspective of hypothesis diameter, the dual-max optimization elucidates the diameter of the hypothesis set, and we find the hypothesis diameter of $H_T^{\mathcal{T}}$ is smaller than that of $H_t^{\mathcal{T}}$. Thus, the above inequality holds.

Following the similar principle, we next prove that $|R(h) - R(h')|$ is bounded. Considering the margin distribution $\mathcal{D}_{\mathcal{X}}$ is i.i.d. drawn from $\mathcal{D}$, we have

$$\begin{aligned}
|R(h) - R(h')| &= \left| \mathop{\mathbb{E}}_{(x,y)\sim\mathcal{D}} [\ell(h(x), y) - \ell(h'(x), y)] \right| \\
&\leq \mathop{\mathbb{E}}_{(x,y)\sim\mathcal{D}} \left| \ell(h(x), y) - \ell(h'(x), y) \right| \\
&\leq \mathop{\mathbb{E}}_{x\sim\mathcal{D}_{\mathcal{X}}} \left[\max_{y} |\ell(h(x), y) - \ell(h'(x), y)| \right] \\
&= \mathcal{L}(h, h').
\end{aligned}$$

Using the above inequality, we obtain that $|U_t|$ is bounded for all $t \in [T]$:

$$\begin{aligned}
|U_t| &= \left| \frac{Q_t}{p_t} \left[\ell(h(x_t), y_t) - \ell(h'(x_t), y_t) \right] - [R(h) - R(h')] \right| \\
&\leq \frac{1}{p_t} \left| \ell(h(x_t), y_t) - \ell(h'(x_t), y_t) \right| + |R(h) - R(h')| \\
&\leq 1 + \mathcal{L}(h, h').
\end{aligned}$$

Thus $\{U_1, \cdots, U_T\}$ is a martingale difference sequence with bounded $1 + \mathcal{L}(h, h')$. To make the subsequent proof clearer, let $Z_t = \frac{U_t}{1+\mathcal{L}(h,h')}$. Then $\{Z_1, \cdots, Z_T\}$ is a martingale difference sequence with bounded $|Z_t| \leq 1$. Applying Azuma's inequality to $\sum_{t=1}^{T} Z_t$:

$$\begin{aligned}
&\mathbb{P}\left(|L_T(h) - L_T(h') - R(h) + R(h')| \geq (1 + \mathcal{L}(h, h'))\Delta_T \right) \\
&= \mathbb{P}\left(\frac{1}{T} |\sum_{t=1}^{T} Z_t| \geq \Delta_T \right) \\
&= \mathbb{P}\left(|\sum_{t=1}^{T} Z_t| \geq T\Delta_T \right) \\
&\leq 2\exp(\frac{-T^2\Delta_T^2}{2T}) \\
&= \frac{\delta}{T(T+1)|\mathcal{H}^{\mathcal{T}}|^2}.
\end{aligned}$$

The above probability inequality shows that the probability that Eq. (14) does not hold is less than $\frac{\delta}{T(T+1)|\mathcal{H}^{\mathcal{T}}|^2}$. Note that with the last line of the inequality, we have $\Delta_T = \sqrt{(2/T)\log(2T(T+1)|\mathcal{H}^{\mathcal{T}}|^2/\delta)}$.

Since $H_T^{\mathcal{T}}$ is a random subset of $\mathcal{H}^{\mathcal{T}}$, a union bound over all $T \in \mathbb{N}^+$ and all pairs of $h, h' \in H_T^{\mathcal{T}}$, we can conclude the proof. In short, any measure of hypothesis disagreement yields a given maximum discrepancy. This ensures the safety of hypothesis pruning, preventing the subsequent process from stepping outside the hypothesis space.

■

Theorem 10 *For any teaching-hypothesis-class $\mathcal{H}^{\mathcal{T}}$, the instruction of an active learner is conducted within it. Given any $\delta > 0$, with a probability at least $1-\delta$, for any $t \in \mathbb{N}^+$, the following inequality holds:*

$$L_t(h^{\mathcal{T}}) - L_t(\widehat{h}_t) \leq \left(1 + \mathcal{F}^{\mathcal{T}}(\widehat{h}_t)\right) \Delta_t.$$

Proof Start by assuming that the $1-\delta$ probability event of Lemma 9 holds.

Let $t = T \in \mathbb{N}^+$. By using the absolute value inequality, we have:

$$\begin{aligned}
&L_t(h^{\mathcal{T}}) - L_t(\widehat{h}_t) \\
&\leq R(h^{\mathcal{T}}) - R(\widehat{h}_t) + (1 + \mathcal{L}(h^{\mathcal{T}}, \widehat{h}_t))\Delta_t \\
&\leq (1 + \mathcal{F}^{\mathcal{T}}(\widehat{h}_t))\Delta_t.
\end{aligned}$$

The last inequality follows from the fact that $h^{\mathcal{T}}$ has the minimum generalization error in $\mathcal{H}^{\mathcal{T}}$, i.e., $R(h^{\mathcal{T}}) - R(\widehat{h}_t) \leq 0$.

From the arbitrariness of T, the theorem is proved. Note that this shows the empirical hypothesis $\widehat{h}_t$ is mentored by the teaching hypothesis $h^{\mathcal{T}}$ employing a teaching feedback function $\mathcal{F}^{\mathcal{T}}$. ■

Theorem 11 *For any teaching-hypothesis-class $\mathcal{H}^{\mathcal{T}}$, the instruction of an active learner is conducted within it. Given any $\delta > 0$, with a probability at least $1-\delta$, for any $T \in \mathbb{N}^+$, the following holds:*

1) the generalization error holds

$$R(\widehat{h}_T) \leq R(h^*) + \left(2 + \mathcal{F}^{\mathcal{T}}(\widehat{h}_{T-1}) + \mathcal{F}^{\mathcal{T}}(\widehat{h}_T)\right) \Delta_{T-1} + \epsilon;$$

2) if the learning problem has disagreement coefficient θ, the label complexity is at most

$$\tau_T \leq 2\theta \left(2TR(h^*) + \left(3 + \mathcal{F}^{\mathcal{T}}(\widehat{h}_{T-1})\right)\right) O(\sqrt{T}) + 2T\epsilon\Big).$$

Proof Start by assuming that the $1-\delta$ probability event of Lemma 9 holds.

Firstly, we give the bound of $R(\widehat{h}_T)$. Since $H_T^{\mathcal{T}} \subseteq H_{T-1}^{\mathcal{T}}$, there is $\widehat{h}_T, h^{\mathcal{T}} \in H_{T-1}^{\mathcal{T}}$. To eliminate the importance-weighted empirical error, we consider Eq. (14) with respect to $\widehat{h}_T, h^{\mathcal{T}}$ at $(T-1)$-time:

$$\begin{aligned}
&R(\widehat{h}_T) - R(h^{\mathcal{T}}) \\
&\leq L_{T-1}(\widehat{h}_T) - L_{T-1}(h^{\mathcal{T}}) + (1 + \mathcal{L}(\widehat{h}_T, h^{\mathcal{T}}))\Delta_{T-1} \\
&\leq L_{T-1}(\widehat{h}_{T-1}) + (1 + \mathcal{F}^{\mathcal{T}}(\widehat{h}_{T-1}))\Delta_{T-1} - L_{T-1}(\widehat{h}_{T-1}) + (1 + \mathcal{F}^{\mathcal{T}}(\widehat{h}_T))\Delta_{T-1} \\
&\leq \left(2 + \mathcal{F}^{\mathcal{T}}(\widehat{h}_{T-1}) + \mathcal{F}^{\mathcal{T}}(\widehat{h}_T)\right) \Delta_{T-1},
\end{aligned}$$

where the second to last inequality follows from teaching-based hypothesis pruning rule w.r.t. Eq. (7). Thus, for any $T \in \mathbb{N}^+$, the bound of generalization error for $\widehat{h}_T$ satisfies the following inequality:

$$\begin{aligned} R(\widehat{h}_T) &\leq R(h^{\mathcal{T}}) + \left(2 + \mathcal{F}^{\mathcal{T}}(\widehat{h}_{T-1}) + \mathcal{F}^{\mathcal{T}}(\widehat{h}_T)\right) \Delta_{T-1} \\ &\leq R(h^*) + \left(2 + \mathcal{F}^{\mathcal{T}}(\widehat{h}_{T-1}) + \mathcal{F}^{\mathcal{T}}(\widehat{h}_T)\right) \Delta_{T-1} + \epsilon, \end{aligned}$$

where the last inequality comes from Corollary 8. Note that the upper bound of the empirical loss of $h^{\mathcal{T}}$ could be mentored by the T-time teaching feedback $\mathcal{F}^{\mathcal{T}}(\widehat{h}_T)$. To tighten this bound, the above inequality employs the teaching feedback at $(T-1)$-time since $\mathcal{F}^{\mathcal{T}}(\widehat{h}_T)$ is influenced by $\mathcal{F}^{\mathcal{T}}(\widehat{h}_{T-1})$. This reduces the loose property of the inequality.

Next, we give the upper bound of τ_T. For any $h \in \mathcal{H}^{\mathcal{T}}$ and the teaching hypothesis $h^{\mathcal{T}}$, their disagreement $\rho(h, h^{\mathcal{T}})$ w.r.t Eq. (5) has the upper bound:

$$\begin{aligned} \rho(h, h^{\mathcal{T}}) &= \mathop{\mathbb{E}}_{(x,y)\sim\mathcal{D}} \left|\ell(h(x), y) - \ell(h'(x), y)\right| \\ &\leq \mathop{\mathbb{E}}_{(x,y)\sim\mathcal{D}} \left[\ell(h(x), y) + \ell(h'(x), y)\right] \\ &\leq R(h) + R(h^{\mathcal{T}}). \end{aligned}$$

For any $t \in [T]$, if $h \in H_t^{\mathcal{T}}$, using Lemma 9, we have the upper bound for $R(h)$:

$$R(h) \leq R(h^{\mathcal{T}}) + \left(2 + \mathcal{F}^{\mathcal{T}}(\widehat{h}_{t-1}) + \mathcal{F}^{\mathcal{T}}(h)\right) \Delta_{t-1}. \tag{16}$$

When $h \in H_t^{\mathcal{T}}$, we use Eq. (16) to rewrite the upper bound of $\rho(h, h^{\mathcal{T}})$ as follows:

$$\rho(h, h^{\mathcal{T}}) \leq 2R(h^{\mathcal{T}}) + \left(2 + \mathcal{F}^{\mathcal{T}}(\widehat{h}_{t-1}) + \mathcal{F}^{\mathcal{T}}(h)\right) \Delta_{t-1}.$$

The above inequality shows that there is a common upper bound on the disagreement between any h in $H_t^{\mathcal{T}}$ and the teaching hypothesis $h^{\mathcal{T}}$. Thus, we can construct a ball $B(h^{\mathcal{T}}, r_t)$ such that $H_t^{\mathcal{T}} \subseteq B(h^{\mathcal{T}}, r_t)$ for any t-time, where

$$r_t = 2R(h^{\mathcal{T}}) + \left(2 + \mathcal{F}^{\mathcal{T}}(\widehat{h}_{t-1}) + \max_{h \in H_t^{\mathcal{T}}} \mathcal{F}^{\mathcal{T}}(h)\right) \Delta_{t-1}. \tag{17}$$

Let $\mathcal{O}_t$ denote all the previous observations up to t-time: $\mathcal{O}_t = \{(x_1, y_1, p_1, Q_1), ..., (x_t, y_t, p_t, Q_t)\}$ with $\mathcal{O}_0 = \emptyset$. By using the error disagreement coefficient w.r.t Eq. (6), the expected value of

the query probability p_t is at most:

$$\begin{aligned}
&\mathop{\mathbb{E}}_{x_t\sim\mathcal{D}_\mathcal{X}}[p_t \mid \mathcal{O}_{t-1}] \\
&= \mathop{\mathbb{E}}_{x_t\sim\mathcal{D}_\mathcal{X}} \max_{h,h'\in H_t^\mathcal{T}} \max_y |\ell(h(x_t),y) - \ell(h'(x_t),y)| \\
&\le 2 \mathop{\mathbb{E}}_{x_t\sim\mathcal{D}_\mathcal{X}} \max_{h\in H_t^\mathcal{T}} \max_y |\ell(h(x_t),y) - \ell(h^\mathcal{T}(x_t),y)| \\
&\le 2 \mathop{\mathbb{E}}_{x_t\sim\mathcal{D}_\mathcal{X}} \max_{h\in B(h^\mathcal{T},r_t)} \max_y |\ell(h(x_t),y) - \ell(h^\mathcal{T}(x_t),y)| \\
&\le 2\theta r_t \\
&= 2\theta \left(2R(h^\mathcal{T}) + (2 + \mathcal{F}^\mathcal{T}(\widehat{h}_{t-1}) + \max_{h\in H_t^\mathcal{T}} \mathcal{F}^\mathcal{T}(h))\Delta_{t-1} \right),
\end{aligned}$$

where the first inequality follows from the triangle inequality, the second inequality follows from $H_t^\mathcal{T} \subseteq B(h^\mathcal{T}, r_t)$, and the third inequality follows from the definition of θ.

Summing over $t = 1, ..., T$, we get the upper bound of the label complexity τ_T:

$$\begin{aligned}
\tau_T &= \sum_{t=1}^{T} \mathop{\mathbb{E}}_{x_t\sim\mathcal{D}_\mathcal{X}} [p_t \mid \mathcal{O}_{t-1}] \\
&= \sum_{t=1}^{T} 2\theta \left(2R(h^\mathcal{T}) + (2 + \mathcal{F}^\mathcal{T}(\widehat{h}_{t-1}) + \max_{h\in H_t^\mathcal{T}} \mathcal{F}^\mathcal{T}(h))\Delta_{t-1} \right) \\
&\le \sum_{t=1}^{T} 2\theta \left(2R(h^\mathcal{T}) + (3 + \mathcal{F}^\mathcal{T}(\widehat{h}_{t-1}))\Delta_{t-1} \right) \\
&= 2\theta \left(2TR(h^\mathcal{T}) + (3 + \mathcal{F}^\mathcal{T}(\widehat{h}_{T-1})) \sum_{t=1}^{T} \Delta_{t-1} \right) \\
&\le 2\theta \left(2TR(h^\mathcal{T}) + (3 + \mathcal{F}^\mathcal{T}(\widehat{h}_{T-1}))O(\sqrt{T}) \right),
\end{aligned}$$

where the last inequality uses $\sum_{t=1}^{T} \sqrt{\frac{1}{t}} = O(\sqrt{T})$. Recalling Corollary 8, there exists $R(h^*) \le R(h^\mathcal{T})$ such that

$$\tau_T \le 2\theta \left(2TR(h^*) + \left(3 + \mathcal{F}^\mathcal{T}(\widehat{h}_{T-1})\right))O(\sqrt{T}) + 2T\epsilon \right).$$

Note that recalling Eqs. (4) and (7), there is $\left(1 + \mathcal{F}^\mathcal{T}(\widehat{h}_t)\right) \Delta_t \le 2\Delta_t$, we thus obtain the first inequality. By shifting $\sum_{t=1}^{T} \Delta_{t-1}$ into $O(\sqrt{T})$, the second inequality holds. This serves to integrate inequalities at the function level into the complexity scale. ■

Lemma 13 *For any teaching-hypothesis-class $\mathcal{H}^\mathcal{T}$, the instruction of an active learner is conducted within it. If the loss function can be rewritten to form $\ell(h(x), y) = \phi(yh(x))$ and*

the function ϕ is non-increasing and convex, for any candidate hypothesis set $H_t^{\mathcal{T}}$ and for all $x \in \mathcal{X}$, the following equation holds:

$$\max_{h,h' \in \mathrm{Conv}(H_t^{\mathcal{T}})} \max_y \left| \ell(h(x), y) - \ell(h'(x), y) \right| = \max_{h,h' \in H_t^{\mathcal{T}}} \max_y \left| \ell(h(x), y) - \ell(h'(x), y) \right|, \quad (18)$$

where $\mathrm{Conv}(H_t^{\mathcal{T}})$ is the convex hull of the hypothesis set $H_t^{\mathcal{T}}$.

Proof Let $f(x) = \max_{h,h' \in \mathrm{Conv}(H_t^{\mathcal{T}})} \max_y |\ell(h(x), y) - \ell(h'(x), y)|$ denote the left-hand side of Eq. (18), and $g(x) = \max_{h,h' \in H_t^{\mathcal{T}})} \max_y |\ell(h(x), y) - \ell(h'(x), y)|$ denote the right-hand side of Eq. (18). For all $x \in \mathcal{X}$, we prove $f(x) \geq g(x)$, then prove $f(x) \leq g(x)$, and get $f(x) = g(x)$. Since $\mathrm{Conv}(H_t^{\mathcal{T}}) \supseteq H_t^{\mathcal{T}}$, there is $f(x) \geq g(x)$. We next prove $f(x) \leq g(x)$.

For any $t \in [T]$ and for all hypotheses $h \in \mathrm{Conv}(H_t^{\mathcal{T}})$, h can be linear representation by the hypothesis h_j in $H_t^{\mathcal{T}} = \{h_1, h_2, ..., h_{|H_t^{\mathcal{T}}|}\}$, that is

$$h = \sum_{j=1}^{|H_t^{\mathcal{T}}|} \lambda_j h_j,$$

where $\sum_j^m \lambda_j = 1$ with $\lambda_j \in [0, 1]$.

Based on the additional assumptions of the loss function, $\ell(h(x), y) = \phi(yh(x))$, where ϕ is a non-increasing function. Then, for all hypotheses $h \in \mathrm{Conv}(H_t^{\mathcal{T}})$ and for any $(x, y) \in \mathcal{X} \times \mathcal{Y}$, $\ell(h(x), y)$ has the following upper bound:

$$\begin{aligned} \ell(h(x), y) &= \ell\left(\sum_{j=1}^{|H_t^{\mathcal{T}}|} \lambda_j h_j(x), y \right) \\ &= \phi\left(y \sum_{j=1}^{|H_t^{\mathcal{T}}|} \lambda_j h_j(x) \right) \\ &\leq \max_{h_j \in H_t^{\mathcal{T}}} \phi(y h_j(x)) \\ &= \max_{h_j \in H_t^{\mathcal{T}}} \ell(h_j(x), y). \end{aligned} \quad (19)$$

Note that $\phi\left(y \sum_{j=1}^{|H_t^{\mathcal{T}}|} \lambda_j h_j(x)\right)$ is a linear group of h_j subjected to $\sum_j \lambda_j = 1$. Thus this new generation hypothesis inherits the upper and lower bounds of the loss of h_j. On this reason, $\ell(h(x), y)$ has the following lower bound:

$$\begin{aligned} \ell(h(x), y) &= \ell\left(\sum_{j=1}^{|H_t^{\mathcal{T}}|} \lambda_j h_j(x), y \right) \\ &= \phi\left(y \sum_{j=1}^{|H_t^{\mathcal{T}}|} \lambda_j h_j(x) \right) \\ &\geq \min_{h_j \in H_t^{\mathcal{T}}} \phi(y h_j(x)) \\ &= \min_{h_j \in H_t^{\mathcal{T}}} \ell(h_j(x), y). \end{aligned} \quad (20)$$

With inequalities Eq. (19)-(20), for all hypotheses $h \in \mathrm{Conv}(H_T^{\mathcal{T}})$ over any $(x, y) \in \mathcal{X} \times \mathcal{Y}$, the loss $\ell(h(x), y)$ can be bounded by the loss of hypothesis $h_j \in H_t^{\mathcal{T}}$.

Using the above properties of the loss function w.r.t. Eq. (19)-(20), we can derive the upper bound of error disagreement for any hypothesis pair $\{h, h'\} \subseteq \mathrm{Conv}(H_t^{\mathcal{T}})$ over all $(x, y) \in \mathcal{X} \times \mathcal{Y}$. For any $h, h' \in \mathrm{Conv}(H_t^{\mathcal{T}})$ and for all $(x, y) \in \mathcal{X} \times \mathcal{Y}$, there exists $h, h' \in \mathrm{Conv}(H_t^{\mathcal{T}})$ such that

$$
\begin{aligned}
&\ell(h(x), y) - \ell(h'(x), y) \\
&\leq \max_{h \in H_t^{\mathcal{T}}} \ell(h(x), y) - \min_{h \in H_t^{\mathcal{T}}} \ell(h(x), y) \\
&\leq \max_{h, h' \in H_t^{\mathcal{T}}} |\ell(h(x), y) - \ell(h'(x), y)| \\
&\leq \max_{h, h' \in H_t^{\mathcal{T}}} \max_{y} |\ell(h(x), y) - \ell(h'(x), y)|.
\end{aligned}
$$

For the first and second inequalities, the disagreement of the upper and lower bound of $h(x) \in H_t^{\mathcal{T}}$ must be tighter than that of the maximal discrepancy of $h(x), h'(x) \in H_t^{\mathcal{T}}$, which denotes the hypothesis diameter of $H_t^{\mathcal{T}}$. Observing any class over y, the third inequality holds.

The above inequalities show that the error disagreement in $\mathrm{Conv}(H_t^{\mathcal{T}})$ over a fixed sample x will not exceed the maximum error disagreement in $H_t^{\mathcal{T}}$ over x.

Take the common upper bound on the left side of the inequality to conclude the proof. ■

Lemma 17 *For any teaching-hypothesis-class $\mathcal{H}^{\mathcal{T}}$, the instruction of an active learner is conducted within it, where the sequence of candidate hypothesis sets satisfies $\mathrm{Conv}(H_{t+1}^{\mathcal{T}}) \subseteq \mathrm{Conv}(H_t^{\mathcal{T}})$ with $H_1^{\mathcal{T}} = \mathcal{H}^{\mathcal{T}}$. Given any $\delta > 0$, with a probability at least $1 - \delta$, for any $T \in \mathbb{N}^+$ and for all $h, h' \in \mathrm{Conv}(H_T^{\mathcal{T}})$, the following inequality holds:*

$$
\left| L_T(h) - L_T(h') - \big(R(h) - R(h')\big) \right| \leq \big(1 + \mathcal{L}(h, h')\big)\, \Delta_T, \tag{21}
$$

where $\Delta_T = \sqrt{(2/T)\log(2T(T+1)|\mathcal{H}^{\mathcal{T}}|^2/\delta)}$.

Proof Lemma 17 complements the scenario of Lemma 9, where the satisfying condition is weakened from $H_{t+1}^{\mathcal{T}} \subseteq H_t^{\mathcal{T}}$ to $\mathrm{Conv}(H_{t+1}^{\mathcal{T}}) \subseteq \mathrm{Conv}(H_t^{\mathcal{T}})$. Recalling the process of proving Lemma 9, to prove Lemma 17, we only need to show that Eq. (15) still holds.

For all $t \leq T$ and for any $h, h' \in \mathrm{Conv}(H_T^{\mathcal{T}})$, we have:

$$\begin{aligned}
&|\ell(h(x_t), y_t) - \ell(h'(x_t), y_t)| \\
&\leq \max_y \left|\ell(h(x_t), y) - \ell(h'(x_t), y)\right| \\
&\leq \max_{h_1, h_2 \in \mathrm{Conv}(H_T^{\mathcal{T}})} \max_y |\ell(h_1(x_t), y) - \ell(h_2(x_t), y)| \\
&\leq \max_{h_1, h_2 \in \mathrm{Conv}(H_t^{\mathcal{T}})} \max_y |\ell(h_1(x_t), y) - \ell(h_2(x_t), y)| \\
&= \max_{h_1, h_2 \in H_t^{\mathcal{T}}} \max_y |\ell(h_1(x_t), y) - \ell(h_2(x_t), y)| \\
&= p_t,
\end{aligned}$$

where the second to last inequality follows that maximum error disagreement at a certain fixed sample x will not increase w.r.t. Lemma 13. ■

Theorem 14 *For any teaching-hypothesis-class $\mathcal{H}^{\mathcal{T}}$, the instruction of an active learner is conducted within it, where the sequence of candidate hypothesis sets satisfies $\mathrm{Conv}(H_{t+1}^{\mathcal{T}}) \subseteq \mathrm{Conv}(H_t^{\mathcal{T}})$ with $H_1^{\mathcal{T}} = \mathcal{H}^{\mathcal{T}}$. For any $t \in \mathbb{N}^+$, given any $\delta > 0$, with a probability at least $1 - \delta$, for any $\tilde{h} \in \widetilde{H}_t'$, the following inequality holds:*

$$R(h^{\mathcal{T}}) - R(\tilde{h}) \geq L_t(h^{\mathcal{T}}) - L_t(\tilde{h}) - \left(1 + \mathcal{F}^{\mathcal{T}}(\tilde{h})\right) \Delta_t.$$

Proof Start by assuming that the $1 - \delta$ probability event of Lemma 17 holds.

Let $t = T \in \mathbb{N}^+$. For all new hypotheses $\tilde{h} \in \widetilde{H}_t'$, there is $\tilde{h} \in \mathrm{Conv}(H_t^{\mathcal{T}})$ by the definition of $\tilde{h}$ w.r.t. Eq. (8). At t-time, we use Lemma 17 w.r.t. $h^{\mathcal{T}}, \tilde{h}$ to obtain the following inequality:

$$R(h^{\mathcal{T}}) - R(\tilde{h}) \geq L_t(h^{\mathcal{T}}) - L_t(\tilde{h}) - (1 + \mathcal{F}^{\mathcal{T}}(\tilde{h}))\Delta_t.$$

From the arbitrariness of T, the theorem is proved. Consequently, the new generation hypothesis also follows the teaching constraints as the other hypotheses. ■

Theorem 16 *For any teaching-hypothesis-class $\mathcal{H}^{\mathcal{T}}$, the instruction of an active learner is conducted within it. If the self-improvement of teaching is applied, given any $\delta > 0$, with a probability at least $1 - \delta$, for any $T \in \mathbb{N}^+$, the following holds: 1) for any $t \in [T]$, holds $h_t^{\mathcal{T}} \in H_t^{\mathcal{T}}$;*

2) the generalization error holds

$$R(\widehat{h}_T) \leq R(h^*) + \left(2 + \mathcal{F}_{T-1}^{\mathcal{T}}(\widehat{h}_{T-1}) + \mathcal{F}_{T-1}^{\mathcal{T}}(\widehat{h}_T)\right) \Delta_{T-1} + \epsilon_{T-1};$$

3) if the learning problem has disagreement coefficient θ, the label complexity is at most

$$\tau_T \leq 2\theta \left(2TR(h^*) + \left(3 + \mathcal{F}_{T-1}^{\mathcal{T}}(\widehat{h}_{T-1})\right) O(\sqrt{T}) + 2T\epsilon_{T-1}\right).$$

Proof Start by assuming that the $1-\delta$ probability event of Lemma 17 holds.

We first show that $h_t^{\mathcal{T}} \in H_t^{\mathcal{T}}$ for any $t \in [T]$ by mathematical induction. It obviously applies to $t=1$. Now suppose it holds for $t=k$, that is, $h_k^{\mathcal{T}} \in H_k^{\mathcal{T}}$, let us prove that it is also true for $t=k+1$. By Lemma 17, there is $h_k^{\mathcal{T}} \in H_k^{\mathcal{T}}$ such that

$$\begin{aligned}
&L_k(h_k^{\mathcal{T}}) - L_k(\widehat{h}_k) \\
&\leq R(h_k^{\mathcal{T}}) - R(\widehat{h}_k) + (1 + \mathcal{L}(h_k^{\mathcal{T}}, \widehat{h}_k))\Delta_k \\
&\leq (1 + \mathcal{F}_k^{\mathcal{T}}(\widehat{h}_k))\Delta_k.
\end{aligned}$$

Therefore, we have $L_k(h_k^{\mathcal{T}}) \leq L_k(\widehat{h}_k) + (1 + \mathcal{F}_k^{\mathcal{T}}(\widehat{h}_k))\Delta_k$, which shows that the teaching hypothesis $h_k^{\mathcal{T}}$ satisfies the hypothesis pruning rule, i.e., $h_k^{\mathcal{T}} \in H'_k$. If the self-improvement of teaching strategy does not find a better teaching hypothesis, then $h_{k+1}^{\mathcal{T}} = h_k^{\mathcal{T}} \in H'_k$. This provides a fallback guarantee. If the self-improvement of teaching strategy finds a better teaching hypothesis, then $h_{k+1}^{\mathcal{T}} \in \widetilde{H}'_k$. In both cases, there is always holds $h_{k+1}^{\mathcal{T}} \in H'_k \cup \widetilde{H}'_k = H_{k+1}^{\mathcal{T}}$. Thus $h_t^{\mathcal{T}} \in H_t^{\mathcal{T}}$ holds for any $t \in [T]$ by the mathematical induction. This provides a significant forward guarantee.

Next, we give the bound of $R(\widehat{h}_T)$. Since $\widehat{h}_T \in H_T^{\mathcal{T}} = H'_{T-1} \cup \widetilde{H}'_{T-1}$, we consider $h \in H'_{T-1}$ and $h \in \widetilde{H}'_{T-1}$ separately.

Assuming that $\widehat{h}_T \in H'_{T-1}$, we give an upper bound on the generalization error for any hypothesis h in H'_{T-1}. Since $h \in H'_{T-1} \subseteq H_{T-1}^{\mathcal{T}}$, by Lemma 17, we have:

$$\begin{aligned}
&R(h) - R(h_{T-1}^{\mathcal{T}}) \\
&\leq L_{T-1}(h) - L_{T-1}(h_{T-1}^{\mathcal{T}}) + (1 + \mathcal{L}(h, h_{T-1}^{\mathcal{T}}))\Delta_{T-1} \\
&\leq L_{T-1}(\widehat{h}_{T-1}) + (1 + \mathcal{F}_{T-1}^{\mathcal{T}}(\widehat{h}_{T-1}))\Delta_{T-1} - L_{T-1}(\widehat{h}_{T-1}) + (1 + \mathcal{F}_{T-1}^{\mathcal{T}}(\widehat{h}_T))\Delta_{T-1} \\
&\leq \left(2 + \mathcal{F}_{T-1}^{\mathcal{T}}(\widehat{h}_{T-1}) + \mathcal{F}_{T-1}^{\mathcal{T}}(\widehat{h}_T)\right)\Delta_{T-1},
\end{aligned}$$

where $\mathcal{F}_t^{\mathcal{T}}(\cdot) := \mathcal{L}(h_t^{\mathcal{T}}, \cdot)$ denotes the disagreement feedback with latest teaching hypothesis $h_t^{\mathcal{T}}$ at t-time.

Assuming that $h \in \widetilde{H}'_{T-1}$, the second inequality above is no longer true because h does not necessarily satisfy the hypothesis pruning rule. According to the self-improvement of teaching strategy w.r.t. Section 6, we can express h as $h = \sum_{j=1}^{|H'_{T-1}|} \lambda_j h_j$, where $h_j \in H'_{T-1}$ and $\sum_j^m \lambda_j = 1$ with $\lambda_j \in [0,1]$. Based on the additional assumptions of the loss function,

the following inequality portrays the upper bound of $R(h)$.

$$\begin{aligned} R(h) &= \mathop{\mathbb{E}}_{(x,y)\sim\mathcal{D}}[\ell(\sum_{j=1}^{|H'_{T-1}|}\lambda_j h_j(x), y)] \\ &= \mathop{\mathbb{E}}_{(x,y)\sim\mathcal{D}}[\phi(y\sum_{j=1}^{|H'_{T-1}|}\lambda_j h_j(x))] \\ &\leq \sum_{j=1}^{|H'_{T-1}|}\lambda_j \mathop{\mathbb{E}}_{(x,y)\sim\mathcal{D}}[\phi(y h_j(x))] \\ &\leq \max_{h_j\in H'_{T-1}} \mathop{\mathbb{E}}_{(x,y)\sim\mathcal{D}}[\ell(h_j(x), y)] \\ &= \max_{h_j\in H'_{T-1}} R(h_j). \end{aligned}$$

Note this group of inequalities inherit the similar properties of Eq. (19). It shows that the generalization error of the hypothesis in $\widetilde{H}'_T$ will not be greater than the generalization error of the hypothesis in H'_T. In other words, any new generation hypothesis derived from the convex hull of H'_T will inherit the constraints of H'_T's hypothesis diameter, which requires both an upper and lower bound for the loss.

Thus, for any $T \in \mathbb{N}^+$, the bound of generalization error for $\widehat{h}_T$ satisfies the following inequality:

$$\begin{aligned} R(\widehat{h}_T) &\leq R(h^{\mathcal{T}}_{T-1}) + \left(2 + \mathcal{F}^{\mathcal{T}}_{T-1}(\widehat{h}_{T-1}) + \mathcal{F}^{\mathcal{T}}_{T-1}(\widehat{h}_T)\right)\Delta_{T-1} \\ &\leq R(h^*) + \left(2 + \mathcal{F}^{\mathcal{T}}_{T-1}(\widehat{h}_{T-1}) + \mathcal{F}^{\mathcal{T}}_{T-1}(\widehat{h}_T)\right)\Delta_{T-1} + \epsilon_{T-1}, \end{aligned}$$

where the last inequality comes from the definition of ϵ_{T-1} and Corollary 15.

The proof of label complexity is similar to Theorem 11, which is omitted here. ■

Acknowledgments

This work was supported by National Natural Science Foundation of China, Grant Number: 62476109, 62206108, the A*STAR Centre for Frontier AI Research, and also supported in part by the Research Grants Council of the Hong Kong Special Administrative Region (Grants 16202523 and HKU C7004-22G).

References

Nir Ailon, Ron Begleiter, and Esther Ezra. Active learning using smooth relative regret approximations with applications. In *Conference on Learning Theory*, pages 19–1. JMLR Workshop and Conference Proceedings, 2012.

Maria-Florina Balcan, Alina Beygelzimer, and John Langford. Agnostic active learning. *Journal of Computer and System Sciences*, 75(1):78–89, 2009.

William H Beluch, Tim Genewein, Andreas Nürnberger, and Jan M Köhler. The power of ensembles for active learning in image classification. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 9368–9377, 2018.

Alina Beygelzimer, Sanjoy Dasgupta, and John Langford. Importance weighted active learning. In *Proceedings of the 26th annual international conference on machine learning*, pages 49–56, 2009.

Prarthana Bhattacharyya, Chengjie Huang, and Krzysztof Czarnecki. Ssl-lanes: Self-supervised learning for motion forecasting in autonomous driving. In *Conference on Robot Learning*, pages 1793–1805. PMLR, 2023.

Xiaofeng Cao and Ivor Tsang. Distribution disagreement via lorentzian focal representation. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2021a.

Xiaofeng Cao and Ivor W Tsang. Shattering distribution for active learning. *IEEE transactions on neural networks and learning systems*, 2020.

Xiaofeng Cao and Ivor W Tsang. Distribution matching for machine teaching. *arXiv preprint arXiv:2105.13809*, 2021b.

Jiawei Chen, Hongyu Lin, Xianpei Han, and Le Sun. Benchmarking large language models in retrieval-augmented generation. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 38, pages 17754–17762, 2024.

Ferdinando Cicalese, Eduardo Laber, Marco Molinaro, et al. Teaching with limited information on the learner's behaviour. In *International Conference on Machine Learning*, pages 2016–2026. PMLR, 2020.

Corinna Cortes, Giulia DeSalvo, Claudio Gentile, Mehryar Mohri, and Ningshan Zhang. Region-based active learning. In *The 22nd International Conference on Artificial Intelligence and Statistics*, pages 2801–2809. PMLR, 2019a.

Corinna Cortes, Giulia DeSalvo, Mehryar Mohri, Ningshan Zhang, and Claudio Gentile. Active learning with disagreement graphs. In *International Conference on Machine Learning*, pages 1379–1387. PMLR, 2019b.

Corinna Cortes, Giulia DeSalvo, Claudio Gentile, Mehryar Mohri, and Ningshan Zhang. Adaptive region-based active learning. In *International Conference on Machine Learning*, pages 2144–2153. PMLR, 2020.

Koby Crammer, Alex Kulesza, and Mark Dredze. Adaptive regularization of weight vectors. *Advances in neural information processing systems*, 22, 2009.

Sanjoy Dasgupta. Analysis of a greedy active learning strategy. *Advances in neural information processing systems*, 17, 2004.

Sanjoy Dasgupta. Two faces of active learning. *Theoretical computer science*, 412(19): 1767–1781, 2011.

Sanjoy Dasgupta, Daniel Hsu, Stefanos Poulis, and Xiaojin Zhu. Teaching a black-box learner. In *International Conference on Machine Learning*, pages 1547–1555. PMLR, 2019.

François Denis. Pac learning from positive statistical queries. In *International Conference on Algorithmic Learning Theory*, pages 112–126. Springer, 1998.

Thorsten Doliwa, Gaojian Fan, Hans Ulrich Simon, and Sandra Zilles. Recursive teaching dimension, vc-dimension and sample compression. *The Journal of Machine Learning Research*, 15(1):3107–3131, 2014.

Linton C Freeman. *Elementary applied statistics: for students in behavioral science*. New York: Wiley, 1965.

Yarin Gal, Riashat Islam, and Zoubin Ghahramani. Deep bayesian active learning with image data. In *International Conference on Machine Learning*, pages 1183–1192. PMLR, 2017.

Mingfei Gao, Zizhao Zhang, Guo Yu, Sercan Ö Arık, Larry S Davis, and Tomas Pfister. Consistency-based semi-supervised active learning: Towards minimizing labeling cost. In *European Conference on Computer Vision*, pages 510–526. Springer, 2020.

Ran Gilad-Bachrach, Amir Navot, and Naftali Tishby. Query by committee made real. In *Advances in neural information processing systems*, pages 443–450, 2006.

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

Daniel Golovin and Andreas Krause. Adaptive submodularity: A new approach to active learning and stochastic optimization. In *COLT*, pages 333–345, 2010.

Daniel Golovin, Andreas Krause, and Debajyoti Ray. Near-optimal bayesian active learning with noisy observations. In *Proceedings of the 23rd International Conference on Neural Information Processing Systems-Volume 1*, pages 766–774, 2010.

Alon Gonen, Sivan Sabato, and Shai Shalev-Shwartz. Efficient active learning of halfspaces: an aggressive approach. In *International Conference on Machine Learning*, pages 480–488. PMLR, 2013.

Agrim Gupta, Silvio Savarese, Surya Ganguli, and Li Fei-Fei. Embodied intelligence via learning and evolution. *Nature communications*, 12(1):5721, 2021.

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

Steve Hanneke. Teaching dimension and the complexity of active learning. In *International Conference on Computational Learning Theory*, pages 66–81. Springer, 2007b.

Steve Hanneke. *Theoretical foundations of active learning*. Carnegie Mellon University, 2009.

Steve Hanneke. Activized learning: Transforming passive to active with improved label complexity. *The Journal of Machine Learning Research*, 13(1):1469–1587, 2012.

Steve Hanneke et al. Theory of disagreement-based active learning. *Foundations and Trends® in Machine Learning*, 7(2-3):131–309, 2014.

Miriam Huijser and Jan C van Gemert. Active decision boundary annotation with deep generative models. In *Proceedings of the IEEE international conference on computer vision*, pages 5286–5295, 2017.

Matti Kääriäinen, Tuomo Malinen, and Tapio Elomaa. Selective rademacher penalization and reduced error pruning of decision trees. *The Journal of Machine Learning Research*, 5:1107–1126, 2004.

Michael Kampffmeyer, Arnt-Borre Salberg, and Robert Jenssen. Semantic segmentation of small objects and modeling of uncertainty in urban remote sensing images using deep convolutional neural networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition workshops*, pages 1–9, 2016.

Andreas Kirsch, Joost Van Amersfoort, and Yarin Gal. Batchbald: Efficient and diverse batch acquisition for deep bayesian active learning. *Advances in neural information processing systems*, 32:7026–7037, 2019.

Akshay Krishnamurthy, Alekh Agarwal, Tzu-Kuo Huang, Hal Daumé III, and John Langford. Active learning for cost-sensitive classification. In *International Conference on Machine Learning*, pages 1915–1924. PMLR, 2017.

Ji Liu, Xiaojin Zhu, and Hrag Ohannessian. The teaching dimension of linear learners. In *International Conference on Machine Learning*, pages 117–126. PMLR, 2016.

Weiyang Liu, Bo Dai, Ahmad Humayun, Charlene Tay, Chen Yu, Linda B Smith, James M Rehg, and Le Song. Iterative machine teaching. In *International Conference on Machine Learning*, pages 2149–2158. PMLR, 2017.

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

Tambet Matiisen, Avital Oliver, Taco Cohen, and John Schulman. Teacher–student curriculum learning. *IEEE transactions on neural networks and learning systems*, 31(9): 3732–3740, 2019.

Zhong Meng, Jinyu Li, Yifan Gong, and Biing-Hwang Juang. Adversarial teacher-student learning for unsupervised domain adaptation. In *2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 5949–5953. IEEE, 2018.

Zhong Meng, Jinyu Li, Yashesh Gaur, and Yifan Gong. Domain adaptation via teacher-student learning for end-to-end speech recognition. In *2019 IEEE Automatic Speech Recognition and Understanding Workshop (ASRU)*, pages 268–275. IEEE, 2019.

Dang Nguyen, Sunil Gupta, Kien Do, and Svetha Venkatesh. Black-box few-shot knowledge distillation. In *European Conference on Computer Vision*, pages 196–211. Springer, 2022.

Narges Norouzi, Snigdha Chaturvedi, and Matthew Rutledge. Lessons learned from teaching machine learning and natural language processing to high school students. In *Proceedings of the AAAI conference on artificial intelligence*, volume 34, pages 13397–13403, 2020.

Tribhuvanesh Orekondy, Bernt Schiele, and Mario Fritz. Knockoff nets: Stealing functionality of black-box models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 4954–4963, 2019.

Robert Pinsler, Jonathan Gordon, Eric Nalisnick, and José Miguel Hernández-Lobato. Bayesian batch active learning as sparse subset approximation. *Advances in neural information processing systems*, 32:6359–6370, 2019.

Antti Rasmus, Mathias Berglund, Mikko Honkala, Harri Valpola, and Tapani Raiko. Semi-supervised learning with ladder networks. *Advances in Neural Information Processing Systems*, 28:3546–3554, 2015.

Nicholas Roy and Andrew McCallum. Toward optimal active learning through monte carlo estimation of error reduction. *ICML, Williamstown*, 2:441–448, 2001.

Sebastian Ruder. An overview of gradient descent optimization algorithms. *arXiv preprint arXiv:1609.04747*, 2016.

Burr Settles. Active learning literature survey. 2009.

Claude Elwood Shannon. A mathematical theory of communication. *ACM SIGMOBILE mobile computing and communications review*, 5(1):3–55, 2001.

Yawar Siddiqui, Julien Valentin, and Matthias Nießner. Viewal: Active learning with viewpoint entropy for semantic segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 9433–9443, 2020.

Patrice Y Simard, Saleema Amershi, David M Chickering, Alicia Edelman Pelton, Soroush Ghorashi, Christopher Meek, Gonzalo Ramos, Jina Suh, Johan Verwey, Mo Wang, et al. Machine teaching: A new paradigm for building machine learning systems. *arXiv e-prints*, pages arXiv–1707, 2017.

Samarth Sinha, Sayna Ebrahimi, and Trevor Darrell. Variational adversarial active learning. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 5972–5981, 2019.

Emil Talpes, Debjit Das Sarma, Ganesh Venkataramanan, Peter Bannon, Bill McGee, Benjamin Floering, Ankit Jalote, Christopher Hsiong, Sahil Arora, Atchyuth Gorti, et al. Compute solution for tesla's full self-driving computer. *IEEE Micro*, 40(2):25–35, 2020.

Simon Tong and Daphne Koller. Support vector machine active learning with applications to text classification. *Journal of machine learning research*, 2(Nov):45–66, 2001.

Pei Wang, Kabir Nagrecha, and Nuno Vasconcelos. Gradient-based algorithms for machine teaching. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 1387–1396, 2021.

Shuo Wang, Yuexiang Li, Kai Ma, Ruhui Ma, Haibing Guan, and Yefeng Zheng. Dual adversarial network for deep active learning. In *European Conference on Computer Vision*, pages 680–696. Springer, 2020a.

Xionghui Wang, Jian-Fang Hu, Jianhuang Lai, Jianguo Zhang, and Wei-Shi Zheng. Progressive teacher-student learning for early action prediction. In *Conference on Computer Vision and Pattern Recognition 2019*, pages 3551–3560. IEEE, 2020b.

Zi Wang. Zero-shot knowledge distillation from a decision-based black-box model. In *International conference on machine learning*, pages 10675–10685. PMLR, 2021.

Manfred KK Warmuth, Gunnar Rätsch, Michael Mathieson, Jun Liao, and Christian Lemmen. Active learning in the drug discovery process. *Advances in Neural information processing systems*, 14, 2001.

Weng Kee Wong. Comparing robust properties of a, d, e and g-optimal designs. *Computational statistics & data analysis*, 18(4):441–448, 1994.

Songbai Yan and Chicheng Zhang. Revisiting perceptron: efficient and label-optimal learning of halfspaces. In *Proceedings of the 31st International Conference on Neural Information Processing Systems*, pages 1056–1066, 2017.

Chen Zhang, Xiaofeng Cao, Weiyang Liu, Ivor Tsang, and James Kwok. Nonparametric iterative machine teaching. *ICML 2023*, 2023a.

Chen Zhang, Xiaofeng Cao, Weiyang Liu, Ivor Tsang, and James Kwok. Nonparametric teaching for multiple learners. *NeurIPS 2023*, 2023b.

Chicheng Zhang and Kamalika Chaudhuri. Beyond disagreement-based agnostic active learning. *Advances in Neural Information Processing Systems*, 27:442–450, 2014.

Xiaojin Zhu. Machine teaching: An inverse problem to machine learning and an approach toward optimal education. In *Proceedings of the AAAI conference on artificial intelligence*, volume 29, 2015.

Xiaojin Zhu, Ji Liu, and Manuel Lopes. No learner left behind: On the complexity of teaching multiple learners simultaneously. In *IJCAI*, pages 3588–3594, 2017.

Xiaojin Zhu, Adish Singla, Sandra Zilles, and Anna N Rafferty. An overview of machine teaching. *arXiv preprint arXiv:1801.05927*, 2018.