Unlearn What You Have Learned: Adaptive Crowd Teaching with Exponentially Decayed Memory Learners

Yao Zhou
Arizona State University
Tempe, Arizona
yzhou174@asu.edu

Arun Reddy Nelakurthi
Arizona State University
Tempe, Arizona
arunreddy@asu.edu

Jingrui He
Arizona State University
Tempe, Arizona
jingrui.he@asu.edu

ABSTRACT

With the increasing demand for large amount of labeled data, crowdsourcing has been used in many large-scale data mining applications. However, most existing works in crowdsourcing mainly focus on label inference and incentive design. In this paper, we address a different problem of adaptive crowd teaching, which is a sub-area of machine teaching in the context of crowdsourcing. Compared with machines, human beings are extremely good at learning a specific target concept (e.g., classifying the images into given categories) and they can also easily transfer the learned concepts into similar learning tasks. Therefore, a more effective way of utilizing crowdsourcing is by supervising the crowd to label in the form of teaching. In order to perform the teaching and expertise estimation simultaneously, we propose an adaptive teaching framework named JEDI to construct the personalized optimal teaching set for the crowdsourcing workers. In JEDI teaching, the teacher assumes that each learner has an exponentially decayed memory. Furthermore, it ensures comprehensiveness in the learning process by carefully balancing teaching diversity and learner's accurate learning in terms of teaching usefulness. Finally, we validate the effectiveness and efficacy of JEDI teaching in comparison with the state-of-the-art techniques on multiple data sets with both synthetic learners and real crowdsourcing workers.

CCS CONCEPTS

• **Information systems** → **Crowdsourcing**; • **Applied computing** → *Interactive learning environments*; • **Computing methodologies** → Learning from demonstrations;

KEYWORDS

Crowd Teaching, Exponentially Decayed Memory, Human Learner

ACM Reference Format:
Yao Zhou, Arun Reddy Nelakurthi, and Jingrui He. 2018. Unlearn What You Have Learned: Adaptive Crowd Teaching with Exponentially Decayed Memory Learners. In *KDD 2018: The 24th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining, August 19–23, 2018, London, United Kingdom.* ACM, New York, NY, USA, 10 pages. https://doi.org/10.1145/3219819.3219952

Permission to make digital or hard copies of all or part of this work for personal or classroom use is granted without fee provided that copies are not made or distributed for profit or commercial advantage and that copies bear this notice and the full citation on the first page. Copyrights for components of this work owned by others than ACM must be honored. Abstracting with credit is permitted. To copy otherwise, or republish, to post on servers or to redistribute to lists, requires prior specific permission and/or a fee. Request permissions from permissions@acm.org.
KDD 2018, August 19–23, 2018, London, United Kingdom
© 2018 Association for Computing Machinery.
ACM ISBN 978-1-4503-5552-0/18/08…$15.00
https://doi.org/10.1145/3219819.3219952

1 INTRODUCTION

In many real-world applications, the performance of the learning models usually depends on the quality and the amount of labeled training examples. With the increasing attention on the large-scale data mining problems, the demand for large amount of labeled data also grows at an unprecedented scale. One of the most popular means of collecting the labeled data is through crowdsourcing platforms, such as Amazon Mechanical Turk, Crowdflower, etc. With the help of these crowdsourcing services, where the data is outsourced and labeled by a group of mostly unskilled online workers, the researchers and organizations are able to obtain large amount of label information within a short period of time at a low cost. However, the labels provided by these workers are often of low-quality due to the lack of expertise and lack of incentives, etc. In recent years, several works [12, 23–26] have been proposed to model and to estimate the expertise of the workers, and these approaches tend to improve the collective labeling quality by downweighting the votes from the weak annotators and trusting the experts. Another branch of crowdsourcing research [17, 18] focuses on the design of incentives that could motivate the workers to convey their knowledge more accurately by coupling it with a well-designed compensation mechanism. Despite the success of these works, they all omitted one important fact: human beings are extremely good at learning a specific target concept (e.g., classifying the images into given categories) and they can easily transfer the learned concepts into similar learning tasks especially when they have grasped certain prior knowledge regarding the original learning concept. Based on the above insightful observations, it is commonly assumed that a more effective way of utilizing crowdsourcing is by supervising the crowd to label in the form of teaching [8, 19].

The crowdsourcing workers usually have a variety of expertise. Therefore, teaching them a certain concept and estimating their labeling abilities at the same time is a challenging problem in general. From the context of teaching, there is an emerging research direction named machine teaching [29] which is the inverse problem of machine learning. Given the learners, the learning algorithm, and the target concept, machine teaching is concerned with a teacher who wants the learner to learn this target concept as fast as possible. Usually, the main principle of machine teaching is to improve the efficacy of the teachers either by minimizing the teaching effort (i.e., the teaching dimension [11, 27], which is defined as the cardinality of the optimal teaching set), or by maximizing the converging speed [13] (i.e., the number of the teaching iterations to reach teaching optimum). In this work, we focus on the problem of adaptive crowd teaching, which is a sub-area of machine teaching in the context of crowdsourcing. In crowd teaching, the learners are the crowdsourcing workers and the teacher is the machine that guides

the teaching procedure. This is similar to the computer tutoring system, where the teacher teaches by demonstrating the typical examples with answers to the students, and the teacher's goal is to help the students have good performance in similar tasks after tutoring. Very few works [8, 19] have been conducted to solve this problem, however, none of them have considered the human memory decay during learning, which has been shown to strongly affect real human learner's categorization decisions [4, 14, 16].

We propose an adaptive teaching paradigm based on the assumption that the learners have exponentially decayed memories [14]. Within our proposed paradigm, the teacher can gradually construct a personalized optimal teaching sequence for each learner by recommending a teaching example and querying the response from the learner interactively with multiple teaching iterations. Moreover, our teaching strategy ensures the teaching sequence diversity to help the learner develop more comprehensive knowledge on the learning task, and guarantees the teaching sequence usefulness to increase the learner's learning accuracy. To be specific, the main contributions of our work are summarized as follows:

- **Formulation**: We formulate the crowd teaching as a pool-based searching problem which performs teaching and expertise estimation simultaneously. The key idea of this teaching framework is to impose the trade-off between the principles of maximizing teaching usefulness and teaching diversity.
- **Models**: Each learner is assumed to be a gradient descent based model with exponentially decayed memory, and the teacher is formulated to minimize the discrepancy between the learner's current concept and the target concept. We also provide theoretical analyses to study the quality of the teaching examples.
- **Experiments**: We have provided a visualization of our teaching framework on a two-dimensional toy data set, and an exhaustive comparison on one synthetic data set and two real-world data sets using simulated learners. Furthermore, we have conducted a teaching experiment on real human learners and compared our results with state-of-the-art techniques with promising results.
- **Demonstration**: We have built a web-based teaching interface[1] for real human learners. This interface includes all three modules of teaching: memory length estimation, interactive teaching, and performance evaluation.

The rest of this paper is organized as follows. Section 2 briefly reviews some related work. In Section 3, we formally introduce the model of the learner and the model of the teacher, followed by the discussions of the algorithm and the analysis of the teaching performance. The adaptive teaching using harmonic function is proposed in Section 4. The experimental results are presented in Section 5 and we conclude this paper in Section 6.

2 RELATED WORK

In this section, we start by reviewing the research in machine teaching followed with its recent advances. Next, we will review the crowdsourcing works that have some overlap with machine teaching. In the end, we also introduce several other works closely related to human learning and teaching.

2.1 Machine Teaching

The inverse problem of machine learning is named as machine teaching, which typically assumes that there is a teacher who knows both the target concept and the learning algorithm used by a learner. Then, the teacher wants to teach the target concept to the learner by constructing an optimal teaching set of examples. One classic definition of this "optimal" is teaching dimension [5] which is referred to as the cardinality of the teaching set. Finding the optimal teaching set which strictly minimizes the teaching dimension is a difficult problem to solve in general. Thus, a relaxed formulation [29] of machine teaching has been proposed as an optimization problem that minimizes both the teaching risk and the teaching cost. In recent years, there has been a wide range of applications related to machine teaching, e.g., crowdsourcing [8, 19], educational tutoring, and data poisoning [15, 20], etc. In the meantime, several theoretical works have studied various aspects of machine teaching such as iterative machine teaching [13], recursive teaching dimension [2], and teaching dimension of linear learners [11], etc. Our work extends the study of machine teaching into the domain of crowdsourcing, and we studied the crowd teaching problem both theoretically and empirically.

2.2 Crowdsourcing

Crowdsourcing is a special sourcing model in which pieces of micro-tasks are distributed to a pool of online workers. It has become a popular research topic in the recent decades because of its widely commercial and academic adoptions in related areas. One of the fundamental problems of interest is how to properly guide the online workers and teach them the correct labeling concept given the fact that those hired workers are usually non-experts. Based on the learning and teaching styles [3] that students progress towards concept understanding, human learners can be categorized as either the sequential learners (who learn things in continual steps) or global learners (who learn things in large jumps, holistically). Inspired by this pioneer work, recently, several teaching models have been proposed: the gradient descent model proposed in [13] studied the teaching paradigm for sequential learners and their study conducted on human toddlers has demonstrated the effectiveness of iterative machine teaching; the non-stationary hypothesis transition model proposed in [19] assumes crowdsourcing workers are global learners and their learned concepts are randomly switched based on observed workers' feedback; the expected error reduction model proposed in [8] learns to present the most informative teaching images to the students by using an online estimation of their current knowledge. Compared with the former approaches, our work explicitly models the human learner with an exponentially decayed memory which is suitable for the human short-term memory concept learning [6]. Meanwhile, our teaching paradigm is an adaptive crowd teaching framework that ensures both the usefulness and the diversity of the teaching examples.

[1] A demo of this teaching interface is available at: JEDI-Web-Demo. The latest source code is available at: JEDI-Crowd-Teaching.

2.3 Other Related Work

Besides the existing works on machine teaching and crowdsourcing, the proposed work in this paper is also closely related to many other research subjects such as active learning [21] and curriculum learning [1]. The learner in active learning can query the label of an example from the oracle; however, the teaching example in machine teaching is recommended by the teacher. Curriculum learning, which is inspired by the learning process of humans and animals, suggests an easy-to-complex teaching strategy. The empirical results conducted on human subjects in [9] have indicated that human teachers tend to follow the curriculum learning principle. In curriculum learning, samples in the teaching sequence are selected merely based on the example difficulty. However, as a comparison, self-paced learning with diversity [7] which also favors example diversity has shown its superior performance on various learning tasks such as detection and classification.

3 THE CROWD TEACHING FRAMEWORK

In this paper, we denote $\mathcal{X} \subset \mathbb{R}^m$ as the m-dimensional feature representations of all examples (e.g., images or documents) and $\mathcal{Y}$ as the collection of labels. The teacher has access to a labeled subset $\Phi \subset \mathcal{X} \times \mathcal{Y}$, which is named as the teaching set[2] of the teaching task. For binary concept learning, $\mathbf{x} \in \mathcal{X}$ is the feature representation of one example, and $y \in \{-1, +1\}$ is its corresponding binary class label. We assume the teacher knows the target concept $\mathbf{w}_* \in \mathbb{R}^m$ and the learning model (e.g., logistic regression) of each learner. The teacher wants to teach the target concept to the learner using a personalized teaching set which is constructed by interacting with the learner for multiple teaching iterations. To be specific, each teaching iteration (e.g., the t-th iteration) includes the following three major steps:

- First, the teacher estimates the current concept $\mathbf{w}_{t-1}$ grasped by the learner and recommends a new teaching example $(\mathbf{x}_t, y_t)$ to the learner.
- Next, the teacher will show the recommended teaching example (without revealing its true label y_t), and ask the learner to provide its label estimation $\tilde{y}_t$.
- At last, the teacher reveals the true label y_t to the learner, and the learner will perform the learning use $(\mathbf{x}_t, y_t)$.

3.1 Model of the Learner

To begin with, we assume that the learners to be taught are active learners who are seeking for improvement and aim to become the experts of the given task. Therefore, we do not take the spammers or adversaries into consideration under this teaching setting.

Now, we formally introduce the model of the learner, whose assets include its *initial concept* $\mathbf{w}_0$, *learning loss* $\mathcal{L}(\cdot, \cdot)$, *learning procedure*, and *learning rate* η_t. After the t-th teaching iteration, the learner applies a linear model, i.e., $\mathbf{w}_t^T \mathbf{x}$, to predict using its learned concept $\mathbf{w}_t$. Similar to the learning model proposed in [13], we also assume that the learner uses a gradient descent learning procedure. However, based on the fact that the real human learner's categorization decisions are guided by a small set of examples retrieved from memory at the time of decision [4, 16], and the retrievability of memory is usually approximated with an exponential curve [14], we further assume that each learner has an exponentially decayed retrievability for the learned concept in terms of the order of the teaching examples, i.e.,:

$$
\begin{aligned}
\mathbf{v}_1 &= \beta \mathbf{v}_0 + \frac{\partial \mathcal{L}(\mathbf{w}_0^T \mathbf{x}_1, y_1)}{\partial \mathbf{w}_0} \\
\mathbf{v}_2 &= \beta \mathbf{v}_1 + \frac{\partial \mathcal{L}(\mathbf{w}_1^T \mathbf{x}_2, y_2)}{\partial \mathbf{w}_1} \\
&= \beta^2 \mathbf{v}_0 + \left[\beta \frac{\partial \mathcal{L}(\mathbf{w}_0^T \mathbf{x}_1, y_1)}{\partial \mathbf{w}_0} + \frac{\partial \mathcal{L}(\mathbf{w}_1^T \mathbf{x}_2, y_2)}{\partial \mathbf{w}_1} \right] \\
&\cdots \\
\mathbf{v}_t &= \beta^t \mathbf{v}_0 + \sum_{s=1}^{t} \beta^{t-s} \frac{\partial \mathcal{L}(\mathbf{w}_{s-1}^T \mathbf{x}_s, y_s)}{\partial \mathbf{w}_{s-1}}
\end{aligned}
\tag{1}
$$

where $\beta \in (0, 1)$ is the personalized memory decay rate. Various learners can have different memory lengths, and this personalized memory length is parameterized by β. The learners with large β can actually retrieve more information from their memory. The concept momentum $\mathbf{v}_t$ is defined as the linear combination of its previous concept momentum $\mathbf{v}_{t-1}$ and the gradient of the learner's loss $\frac{\partial \mathcal{L}(\mathbf{w}_{t-1}^T \mathbf{x}_t, y_t)}{\partial \mathbf{w}_{t-1}}$. The initial momentum $\mathbf{v}_0$ is usually set to $\mathbf{0}$ in practice. With the properly chosen learning rate η_t, the learner uses the gradient descent learning procedure to improve their concept in an iterative way:

$$
\mathbf{w}_t \leftarrow \mathbf{w}_{t-1} - \eta_t \mathbf{v}_t \tag{2}
$$

Similar to stochastic gradient descent (SGD) with momentum, the learner will update his/her concept $\mathbf{w}_t$ towards the target concept $\mathbf{w}_*$ along the direction of the negative concept momentum $-\mathbf{v}_t$, which is the linear combination of the negative gradients of the learning losses with exponentially decayed weights. Intuitively, the concept learned by a human learner depends on a sequence of teaching examples. The latest example will contribute more (has larger weights) towards learning than the earlier ones.

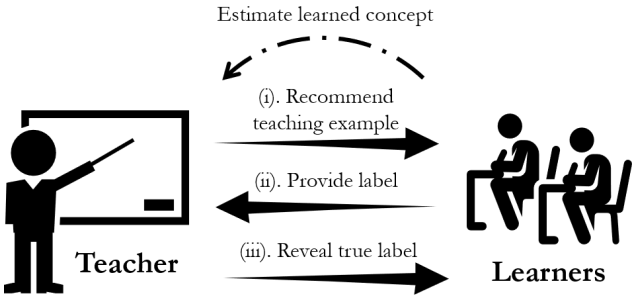

Figure 1: Illustration of JEDI (one teaching iteration)

3.2 Model of the Teacher

Initially, we assume that the teacher has access to the learner's current concept $\mathbf{w}_t$, learning loss, learning procedure, etc., and the teacher intends to guide the learner towards the target concept $\mathbf{w}_*$.

[2]The definition of teaching set in this paper is the same as in [8, 19]. However, in the concept of teaching dimension [11], the definition of teaching set is different from ours.

Notice that in real-world teaching, the teacher generally does not have direct access to a learner's current concept. The alternative of estimating a learner's concept will be introduced in Section 4. Thus, the objective of teaching is proposed as follows:

$$\min ||\mathbf{w}_t - \mathbf{w}_*||_2^2 \tag{3}$$

This objective is designed to minimize the discrepancy between the target concept $\mathbf{w}_*$ and the learner's current concept $\mathbf{w}_t$ after t rounds of teaching. The objective can be decomposed into three parts by substituting Eq. (2) into it:

$$\begin{aligned} O(\mathbf{x}_t, y_t) &= \|\mathbf{w}_t - \mathbf{w}_*\|_2^2 \\ &= \|\mathbf{w}_{t-1} - \mathbf{w}_*\|_2^2 + \eta_t^2 \underbrace{\left\| \sum_{s=1}^{t} \beta^{t-s} \frac{\partial \mathcal{L}(\mathbf{w}_{s-1}^T \mathbf{x}_s, y_s)}{\partial \mathbf{w}_{s-1}} \right\|_2^2}_{T_1\text{: Diversity of the teaching sequence}} \\ &\quad \underbrace{-2\eta_t \left\langle \mathbf{w}_{t-1} - \mathbf{w}_*, \sum_{s=1}^{t} \beta^{t-s} \frac{\partial \mathcal{L}(\mathbf{w}_{s-1}^T \mathbf{x}_s, y_s)}{\partial \mathbf{w}_{s-1}} \right\rangle}_{T_2\text{: Usefulness of the teaching sequence}} \end{aligned} \tag{4}$$

The first part is the discrepancy between $\mathbf{w}_*$ and learner's previous concept $\mathbf{w}_{t-1}$, and the second part T_1 essentially measures the diversity of the teaching sequence. The third part T_2 measures the usefulness of the teaching sequence and the intuitive explanations of T_1 and T_2 will be clear later.

Meanwhile, we assume that the teacher has an infinite memory of the teaching sequence of examples $\mathcal{D}_t = \{(\mathbf{x}_1, y_1), \ldots, (\mathbf{x}_t, y_t)\}$, as well as the corresponding estimate of the concept sequence from the learner $\mathcal{W}_t = \{\mathbf{w}_0, \mathbf{w}_1, \ldots, \mathbf{w}_t\}$.

Diversity of the teaching sequence.

In order to simplify T_1, we further decompose it into two intermediate terms:

$$\begin{aligned} T_1 = &\sum_{s=1}^{t} \beta^{2(t-s)} \left\| \frac{\partial \mathcal{L}(\mathbf{w}_{s-1}^T \mathbf{x}_s, y_s)}{\partial \mathbf{w}_{s-1}} \right\|_2^2 \\ &+ \sum_{s=1}^{t} \sum_{r \neq s}^{t} \beta^{2t-s-r} \left\langle \frac{\partial \mathcal{L}(\mathbf{w}_{s-1}^T \mathbf{x}_s, y_s)}{\partial \mathbf{w}_{s-1}}, \frac{\partial \mathcal{L}(\mathbf{w}_{r-1}^T \mathbf{x}_r, y_r)}{\partial \mathbf{w}_{r-1}} \right\rangle \end{aligned} \tag{5}$$

The selection of the learning loss can be flexible. For the task of teaching a classification concept, we utilize the logistic loss, which is convex and smooth, to illustrate the key idea, i.e., $\log(1 + \exp(-y\mathbf{w}^T\mathbf{x}))$, although the proposed framework can be extended to other loss functions. Easily, we can have the gradient norm of each teaching example as $\left(\frac{-y}{1+\exp(y\mathbf{w}^T\mathbf{x})}\right)^2 \|\mathbf{x}\|_2^2$, which has the interpretation of *example difficulty* when all the example feature $\mathbf{x}$ lies on a hypersphere (e.g., L2-normalized bag-of-words features in document classification). In that case, $\|\mathbf{x}\|_2 = 1$ and the first term of T_1 becomes the sum of squares of the probability of incorrect predictions.

Our goal of teaching is to recommend the next teaching example $(\mathbf{x}_t, y_t)$, therefore, these observed gradients (with indices $s = 1, \ldots, t-1$) are not relevant in this teaching optimization sub-problem of minimizing T_1. If we substitute the gradient of logistic loss into the objective, it is straightforward to get the following equivalent optimization sub-problem:

$$\begin{aligned} &\min_{(\mathbf{x}_t, y_t)} \; T_1 \\ \Leftrightarrow &\min_{(\mathbf{x}_t, y_t)} \left(\frac{y_t}{1 + \exp(y_t \mathbf{w}_{t-1}^T \mathbf{x}_t)} \right)^2 \|\mathbf{x}_t\|_2^2 \\ &+ \sum_{s=1}^{t-1} 2\beta^{t-s} \left[\frac{y_s y_t}{\left(1 + \exp(y_s \mathbf{w}_{s-1}^T \mathbf{x}_s)\right)\left(1 + \exp(y_t \mathbf{w}_{t-1}^T \mathbf{x}_t)\right)} \right] \mathbf{x}_s^T \mathbf{x}_t \end{aligned} \tag{6}$$

Usefulness of the teaching sequence.

T_2 part in the objective serves as the measurement of the usefulness of the whole teaching sequence. Specifically, it is the weighted sum of all inner products between $\mathbf{w}_{t-1} - \mathbf{w}_*$ and the gradients of the teaching sequence examples. It means that the entire teaching sequence $\mathcal{D}_t$ will contribute to maximizing the convergence of the teaching. The larger value the inner product has, the more useful this teaching sequence is. However, similar to the T_1 minimization sub-problem, only $(\mathbf{x}_t, y_t)$ relevant terms matter for the purpose of maximizing T_2:

$$\begin{aligned} &\max_{(\mathbf{x}_t, y_t)} \; T_2 \\ \Leftrightarrow &\max_{(\mathbf{x}_t, y_t)} \sum_{s=1}^{t-1} \beta^{t-s} \left\langle \mathbf{w}_{t-1} - \mathbf{w}_*, \frac{\partial \mathcal{L}(\mathbf{w}_{s-1}^T \mathbf{x}_s, y_s)}{\partial \mathbf{w}_{s-1}} \right\rangle \\ &+ \left\langle \mathbf{w}_{t-1} - \mathbf{w}_*, \frac{\partial \mathcal{L}(\mathbf{w}_{t-1}^T \mathbf{x}_t, y_t)}{\partial \mathbf{w}_{t-1}} \right\rangle \\ \Leftrightarrow &\max_{(\mathbf{x}_t, y_t)} \left\langle \mathbf{w}_{t-1} - \mathbf{w}_*, \frac{-y_t \mathbf{x}_t}{1 + \exp(y_t \mathbf{w}_{t-1}^T \mathbf{x}_t)} \right\rangle \end{aligned} \tag{7}$$

Trade-off between diversity and usefulness.

For simplicity, we denote $f_t := \frac{1}{1+\exp(y_t \mathbf{w}_{t-1}^T \mathbf{x}_t)}$ and $f_s := \frac{1}{1+\exp(y_s \mathbf{w}_{s-1}^T \mathbf{x}_s)}$, where $s = 1, \ldots, t-1$. Then, the overall teaching problem becomes:

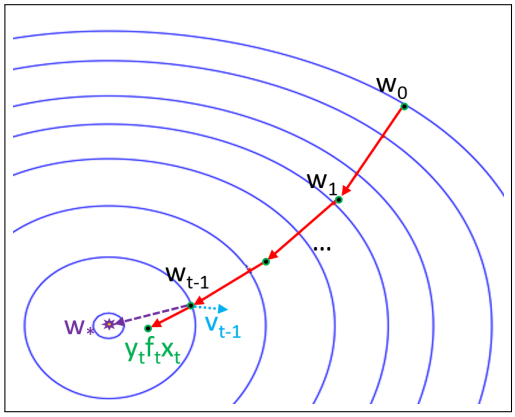

Figure 2: Trade-off between diversity and usefulness

$$
\begin{aligned}
&\min_{(\mathbf{x}_t, y_t)} O(\mathbf{x}_t, y_t) \\
\Leftrightarrow &\min_{(\mathbf{x}_t, y_t)} \eta_t^2 \left[\left\langle y_t f_t \mathbf{x}_t, y_t f_t \mathbf{x}_t \right\rangle + 2 \sum_{s=1}^{t-1} \beta^{t-s} \left\langle y_s f_s \mathbf{x}_s, y_t f_t \mathbf{x}_t \right\rangle \right] \\
&- 2\eta_t \left\langle \mathbf{w}_{t-1} - \mathbf{w}_*, -y_t f_t \mathbf{x}_t \right\rangle \\
\Leftrightarrow &\min_{(\mathbf{x}_t, y_t)} \eta_t^2 \left\| y_t f_t \mathbf{x}_t - \mathbf{v}_{t-1} \right\|_2^2 - 2\eta_t \left\langle \mathbf{w}_* - \mathbf{w}_{t-1}, y_t f_t \mathbf{x}_t \right\rangle
\end{aligned} \tag{8}
$$

Prob. (8) aims to maximize the teaching diversity (T_1 part) and the teaching usefulness (T_2 part) at the same time. As illustrated in Figure 2, the teacher prefers the negative gradient $y_t f_t \mathbf{x}_t$ of next teaching example is similar to the concept momentum $\mathbf{v}_{t-1}$ and has large correlation with the target learning direction $\mathbf{w}_* - \mathbf{w}_{t-1}$. The learning rate η_t is usually set to a small value ($\eta_t < 1$) in optimization. Therefore, the teaching usefulness (T_2 part) dominates the teaching process. It is straightforward to see that when the $\beta = 0$, our objective is directly reduced to the no memory teaching framework proposed in [13].

In order to solve this teaching problem, we denote $\mathbf{a}_t := y_t f_t \mathbf{x}_t$. Then, the teaching objective can be further simplified as:

$$
\begin{aligned}
O(\mathbf{x}_t, y_t) &= \eta_t^2 \left[\left\| \mathbf{a}_t \right\|_2^2 - 2 \left\langle \mathbf{a}_t, \mathbf{v}_{t-1} + \frac{\mathbf{w}_* - \mathbf{w}_{t-1}}{\eta_t} \right\rangle \right] \\
&= \eta_t^2 \left\| \mathbf{a}_t - \left(\mathbf{v}_{t-1} + \frac{\mathbf{w}_* - \mathbf{w}_{t-1}}{\eta_t} \right) \right\|_2^2 \\
&\quad - \eta_t^2 \left\| \mathbf{v}_{t-1} + \frac{\mathbf{w}_* - \mathbf{w}_{t-1}}{\eta_t} \right\|_2^2
\end{aligned} \tag{9}
$$

The concept momentum $\mathbf{v}_{t-1}$ is the weighted sum of gradients of the teaching sequence $\mathcal{D}_{t-1}$ using exponentially decayed weights.

Based on this new objective, we propose the teaching algorithm **JEDI** (Ad**J**ustable **E**xponential **D**ecay Memory **I**nteractive Crowd Teaching). The JEDI teaching algorithm with omniscient teacher (having access to learner's concept sequence $\mathcal{W}_t$) is shown in **Algorithm** 1. It is given the learner's memory decay rate, initial concept, target concept, learning rate, teaching set as input, and will output the personalized teaching sequence. JEDI works as follows. We first initialize the teaching iterator $t = 1$ and initial momentum $\mathbf{v}_0 = \mathbf{0}$. Then, in each iteration, the teacher searches through the teaching set Φ and finds the example $(\mathbf{x}_t, y_t)$ that minimizes the objective function in Eqn. (10), where $f = \frac{1}{1+\exp(y \mathbf{w}_{t-1}^T \mathbf{x})}$ is the probability of incorrect prediction of example $(\mathbf{x}, y)$ in the teaching set Φ. Next, the learner performs the labeling on $\mathbf{x}_t$ using its current concept $\mathbf{w}_{t-1}$. Then, the label y_t is revealed by the teacher and the learner performs learning using Eqn. (2). This interactive teaching will continue until the stopping criteria is satisfied.

In exponential weighted average, the number of examples being used is usually approximated [10] as $\frac{1}{1-\beta}$. Therefore, we can assume that there exists a memory window size (i.e,. how many examples or corresponding concept gradients the learner can memorize) for each learner, and it can be approximated as $\frac{1}{1-\beta}$. In the following, we use the two-example teaching scenario (e.g., memory decay rate is as low as $\beta \approx 0.5$, or it is the second iteration of teaching $t = 2$) as a running example, where the learner can memorize a teaching sequence of size 2. Notice that the analysis and conclusions can be extended to other values of β as well. In the two-example teaching scenario, the trade-off between diversity and usefulness will lead to further insights with the help of the following definitions and theorem.

Algorithm 1 JEDI with omniscient teacher

1: **Input:** Learner's memory decay rate β, initial concept $\mathbf{w}_0$, target concept $\mathbf{w}_*$, initial learning rate η_0, teaching set Φ, MaxIter.
2: **Initialization:**

$$\mathbf{v}_0 \leftarrow \mathbf{0}$$
$$t \leftarrow 1$$

3: **Repeat:**
4: (i). Among all examples $(\mathbf{x}, y)$ in teaching set Φ and their probabilities of incorrect prediction f, the teacher recommends example $(\mathbf{x}_t, y_t)$ to the learner by solving:

$$(\mathbf{x}_t, y_t) = \arg\min_{(\mathbf{x}, y) \in \Phi} \left\| y f \mathbf{x} - \left(\mathbf{v}_{t-1} + \frac{\mathbf{w}_* - \mathbf{w}_{t-1}}{\eta_t} \right) \right\|_2^2 \tag{10}$$

5: (ii). Learner performs the labeling.
6: (iii). Learner performs learning after teacher reveals y_t.
7: (iv). t ← t + 1
8: **Until** converged **or** $t >$ MaxIter
9: **Output:** The teaching sequence $\mathcal{D}_t$

Definition 3.1. Given the previous teaching example $(\mathbf{x}_{t-1}, y_{t-1})$, if the teacher recommends a new teaching example $(\mathbf{x}_t, y_t)$ which has different label $y_t \neq y_{t-1}$, this teaching action is named **Exploration**. If the new teaching example has the same label $y_t = y_{t-1}$, this teaching action is named **Exploitation**.

Definition 3.2. Given the previous teaching example $(\mathbf{x}_{t-1}, y_{t-1})$, its negative gradient $y_{t-1} f_{t-1} \mathbf{x}_{t-1}$ and its optimal teaching direction $\mathbf{w}_* - \mathbf{w}_{t-1}$ has an angle $\theta \in [0, \pi]$. Then, this teaching example is **not useful** towards teaching optimal $\mathbf{w}_*$ if the angle satisfies $\theta \geq \frac{\pi}{2}$.

Theorem 3.3. ***(Exploration vs. Exploitation)*** *For two-example teaching, if the previous teaching example $(\mathbf{x}_{t-1}, y_{t-1})$ is not useful towards teaching optimal, the teacher will recommend large diversity teaching example $(\mathbf{x}_t, y_t)$ for exploitation, i.e., $y_t = y_{t-1}$, or recommend highly similar teaching example $(\mathbf{x}_t, y_t)$ for exploration, i.e., $y_t \neq y_{t-1}$.*

Proof. Let $\mathbf{a}_{t-1} := \beta y_{t-1} f_{t-1} \mathbf{x}_{t-1}$, then the teaching objective becomes:

$$
\begin{aligned}
O(\mathbf{x}_t, y_t) &= \eta_t^2 \left\| \mathbf{a}_t + \left(\mathbf{a}_{t-1} - \frac{\mathbf{w}_* - \mathbf{w}_{t-1}}{\eta_t} \right) \right\|_2^2 \\
&\quad - \eta_t^2 \left\| \mathbf{a}_{t-1} - \frac{\mathbf{w}_* - \mathbf{w}_{t-1}}{\eta_t} \right\|_2^2
\end{aligned}
$$

The minimum of the objective is guaranteed to be reached when the next teaching example $\mathbf{x}_t$ is selected as follows:

$$\mathbf{x}_t = -\beta \frac{y_{t-1}}{y_t} \frac{f_{t-1}}{f_t} \mathbf{x}_{t-1} + \frac{1}{y_t f_t} \frac{\mathbf{w}_* - \mathbf{w}_{t-1}}{\eta_t}$$

Using $\mathbf{a}_{t-1}$ as the reference, the optimal teaching direction vector can be decomposed as $\mathbf{w}_* - \mathbf{w}_{t-1} = (\mathbf{w}_* - \mathbf{w}_{t-1})_{\parallel} + (\mathbf{w}_* - \mathbf{w}_{t-1})_{\perp}$

in $\mathbf{a}_{t-1}$'s parallel direction and perpendicular direction. If the previous teaching example is not useful (i.e., $\theta \geq \frac{\pi}{2}$), without loss of generality, we can assume $(\mathbf{w}_* - \mathbf{w}_{t-1})_{\parallel} = \alpha \mathbf{a}_{t-1}$, where $\alpha \leq 0$ is obviously satisfied. Then, we have:

$$\mathbf{x}_t = -\gamma_+ \frac{f_{t-1}}{f_t} \mathbf{x}_{t-1} + \xi_t \qquad (\Leftarrow \text{Exploitation})$$

$$\mathbf{x}_t = \gamma_+ \frac{f_{t-1}}{f_t} \mathbf{x}_{t-1} + \xi_t \qquad (\Leftarrow \text{Exploration})$$

where $\gamma_+ = (1 - \frac{\alpha}{\eta_t})\beta$ is a positive scalar and $\xi_t = \frac{1}{y_t f_t} \frac{(\mathbf{w}_* - \mathbf{w}_{t-1})_{\perp}}{\eta_t}$ is the teaching perturbation. If the previous teaching example $\mathbf{x}_{t-1}$ is not useful, then the teacher will prefer the next teaching example $\mathbf{x}_t$ to be very different from the previous one for exploitation (intra-class teaching) or to be similar with the previous one for exploration (inter-class teaching). □

The teaching action choice between exploration and exploitation is very clear especially when the previous teaching example is most useless (i.e., $\theta = \pi$), under which scenario the recommended teaching example has zero teaching perturbation $(\mathbf{w}_* - \mathbf{w}_{t-1})_{\perp} = 0$. The magnitude of the teaching perturbation is positively correlated with the usefulness of the previous teaching example since $(\mathbf{w}_* - \mathbf{w}_{t-1})_{\perp} \propto \sin(\theta)$ and $\theta \geq \frac{\pi}{2}$. Therefore, if the previous teaching example is less useful (θ becomes larger), the perturbation will become smaller, and the teacher has less uncertainty to decide whether the next teaching recommendation should be an exploitation action or an exploration action.

Proposition 3.4. *For the examples that live on a hypersphere, if the previous teaching example $(\boldsymbol{x}_{t-1}, y_{t-1})$ is most useless ($\theta = \pi$) towards teaching optimal and the learning rate satisfies $\eta_t \geq \frac{\alpha\beta}{\beta - 1}$, then the teacher recommended example $(\boldsymbol{x}_t, y_t)$ is guaranteed to have better labeling quality than $(\boldsymbol{x}_{t-1}, y_{t-1})$, i.e., the learner can correctly label example $\boldsymbol{x}_t$ with higher probability than labeling example $\boldsymbol{x}_{t-1}$.*

Proof. We have $f_t = (\frac{\alpha}{\eta_t} - 1)\beta \frac{y_{t-1}}{y_t} \frac{<\mathbf{x}_{t-1}, \mathbf{x}_t>}{<\mathbf{x}_t, \mathbf{x}_t>} f_{t-1}$ from Theorem 3.3. For hyperspherical feature space, $\left|\frac{<\mathbf{x}_{t-1}, \mathbf{x}_t>}{<\mathbf{x}_t, \mathbf{x}_t>}\right| \leq 1$ and no matter if the teaching action is exploration or exploitation, the coefficient of f_{t-1} is always smaller than 1. Therefore, f_t (probability of incorrectly labeling $\mathbf{x}_t$) is smaller than f_{t-1} (probability of incorrectly labeling $\mathbf{x}_{t-1}$). □

For the teaching scenarios with multiple teaching examples (e.g., β is large), the above theoretical analyses are also applicable by treating the previous teaching sequence $\mathcal{D}_{t-1}$ as one pseudo teaching example with its decayed negative gradient as $\mathbf{v}_{t-1}$.

4 ADAPTIVELY TEACHING THE HUMAN LEARNERS

In this section, we first discuss the challenges for teaching the real human learners. Then, we present the methodology which can estimate the human learner's current concept using the harmonic function. In the end, we formally present the algorithm JEDI teaching with harmonic function estimation.

4.1 Teaching in the Real World

All examples help teaching. After the teacher reveals the true label of the recommended teaching example, the human learner can improve the concept learning either by verify the correctness of his/her labels or by gaining information from the mistakes he/she made.

Repeated teaching examples. Memories are so volatile that human learners have to be provided with repeated examples to strengthen the learned concept. Due to this reason, the teaching sequence $\mathcal{D}_t$ selected from the teaching set Φ should have repeated examples especially when these examples are incorrectly labeled or the learner's memory window size is small.

Pool-based teaching. Similar to the pool-based active learning, in many real-world teaching tasks, the synthetically generated teaching examples that meet the global optimum of JEDI objective are not valid real-world examples (e.g., images, documents). Thus, a pool-based search is a more realistic alternative. In other words, the JEDI teacher will search for the best teaching examples in the teaching set Φ instead of the whole feature and label space.

Teacher has no access to learner's concept. To address this challenge, notice that by utilizing the first-order convexity of the learning loss, we can have:

$$\left\langle \mathbf{w}_{t-1} - \mathbf{w}_*, \frac{\partial \mathcal{L}(\mathbf{w}_{t-1}^T \mathbf{x}_t, y_t)}{\partial \mathbf{w}_{t-1}} \right\rangle \geq \mathcal{L}(\mathbf{w}_{t-1}^T \mathbf{x}_t, y_t) - \mathcal{L}(\mathbf{w}_*^T \mathbf{x}_t, y_t) \tag{11}$$

Then, minimizing T_2 can be relaxed to the problem of optimizing its lower bound. This relaxation enables the teacher to query the learner's prediction $\text{sign}(\mathbf{w}^T \mathbf{x})$ instead of requiring access to his/her concept $\mathbf{w}$ directly (which is impossible for real human learners). The effectiveness of this relaxation depends on the tightness of the lower bound. Therefore, the smaller $\|\mathbf{w}_{t-1} - \mathbf{w}_*\|$ is, the tighter the bound is. In other words, this relaxed problem is gradually becoming a reliable approximation of the original problem with more and more teaching iterations.

4.2 Concept Estimation using Harmonic Function

In the teaching phase, for every observed teaching example (with indices $s = 1, \ldots, t-1$), the teacher has access to the features $\mathbf{x}_s$ and the learner provided label $\tilde{y}_s$. However, the teacher still needs the learner's probability of incorrect prediction $f = \frac{1}{1+\exp(y \mathbf{w}_{t-1}^T \mathbf{x})}$ on every example $(\mathbf{x}, y)$ in Φ to start teaching. One naive way of estimating f is by using learner provided labels to train a supervised classification model, and predict the unlabeled ones with this classifier to get f. However, due to the limited number of labeled examples, a semi-supervised model [22, 28] should be more effective than supervised models. One alternative to estimating f is by using graph-based semi-supervised learning method proposed in [28]. Given the teaching sequence $\mathcal{D}_{t-1}$, for every unlabeled example, we can estimate its probability of labels using semi-supervised Gaussian random fields and harmonic functions:

$$F_u = (D_{uu} - A_{uu})^{-1} A_{ul} F_l \tag{12}$$

In the above formulation, A is the affinity matrix of all examples and D is a diagonal matrix (with $D_{ii} = \sum_{j=1} A_{ij}$). Matrix A can be reordered and split into four blocks as: $A = \begin{bmatrix} A_{ll} & A_{lu} \\ A_{ul} & A_{uu} \end{bmatrix}$ and similar block split operation is applied on D as well. $F_l \in \{0,1\}^{|\mathcal{D}_{t-1}|\times 2}$ is the label matrix associated with learner provided labels, where each element is set to 1 if the corresponding label has been provided by the learner and 0 otherwise. Following this convention, the affinity matrix can be constructed as follow:

$$A_{ij} = \exp\left(-\sum_{d=1}^{m} \frac{(\mathbf{x}_{id} - \mathbf{x}_{jd})^2}{\sigma_d^2}\right) \tag{13}$$

It should be noticed that the teaching examples could be repeatedly recommended by the JEDI teacher, and this is different from the crowd teaching model of [8], which also uses the harmonic function but only allows each example to be recommended once. Therefore, before applying the harmonic solution, we only keep the unique examples that have the latest labels provided by the learner in the teaching sequence. Meanwhile, in order to guarantee all examples in the teaching set Φ could be recommended for next round of teaching, affinity matrix A are padded using extra nodes and edges constructed from the existing teaching sequence $\mathcal{D}_{t-1}$. After applying the harmonic solution, the labeling probability estimation of every example $\mathbf{x}$ in Φ corresponds to a row (whose entries are p and $1-p$) of matrix F_u:

$$\begin{aligned} P(y=1|\mathbf{x}, \mathcal{D}_{t-1}) = p &= \frac{1}{1+\exp(-\mathbf{w}^T\mathbf{x})} \\ P(y=-1|\mathbf{x}, \mathcal{D}_{t-1}) = 1-p &= \frac{1}{1+\exp(\mathbf{w}^T\mathbf{x})} \end{aligned} \tag{14}$$

To calculate concept momentum $\mathbf{v}_{t-1}$ in T_1, which utilizes the probability of incorrect prediction f_s of teaching example $(\mathbf{x}_s, y_s)$ where $s = 1, \ldots, t-1$, the estimated labeling probabilities are used together with the teacher revealed ground truth label y_s. They are calculated as:

$$f_s := \frac{1}{1+\exp(y_s \mathbf{w}_{s-1}^T \mathbf{x}_s)} = (1-p_s)^{\frac{y_s+1}{2}} p_s^{\frac{1-y_s}{2}} \tag{15}$$

where p_s is the harmonic probability estimate of teaching example $\mathbf{x}_s$. Similarly, in order to calculate T_2 term, the estimated labeling probabilities are jointly used with ground truth label y_t as:

$$\frac{1}{f_{-t}} := 1+\exp(-y_t \mathbf{w}_{t-1}^T \mathbf{x}_t) = \left(\frac{1}{p_t}\right)^{\frac{y_t+1}{2}} \left(\frac{1}{1-p_t}\right)^{\frac{1-y_t}{2}} \tag{16}$$

where p_t is the harmonic probability estimate of teaching example $\mathbf{x}_t$.

4.3 Teaching Algorithm

The details of the JEDI algorithm using harmonic function estimation are provided in **Algorithm** 2. It is given the learner's memory decay rate, target concept, learning rate, teaching set as input, and will output the personalized teaching sequence. It works as follows. We first initialize the iterator $t = 1$ and initial momentum $\mathbf{v}_0 = \mathbf{0}$. Then, in each teaching iteration, the JEDI teacher estimates the probability of incorrect labeling using harmonic function Eqn. (12). Next, the JEDI teacher searches through the teaching set Φ and finds the example $(\mathbf{x}_t, y_t)$ that minimizes the objective function in Eqn. (17) which uses the f_s (where $s = 1, \ldots, t-1$), and $\frac{1}{f_{-t}}$. Next, the learner performs the labeling on $\mathbf{x}_t$ and the JEDI teacher updates affinity matrix A, diagonal matrix D, and label matrix F_l using the methods described in Section 4.2. At last, the teacher reveals the true label y_t and the learner performs learning. The JEDI teaching with harmonic function estimation will stop when the maximum number of iterations has been reached.

Algorithm 2 JEDI with harmonic function estimation

1: **Input:** Learner's memory decay rate β, target concept $\mathbf{w}_*$, initial learning rate η_0, teaching set Φ, affinity matrix A, diagonal matrix D, MaxIter.
2: **Initialization:**

$$\mathbf{v}_0 \leftarrow \mathbf{0}$$
$$t \leftarrow 1$$

3: **Repeat:**
4: (i). Teacher estimates F_u using Eq. (12) and calculates f_s and $\frac{1}{f_{-t}}$ using Eq. (15) and Eq. (16).
5: (ii). Teacher recommends example $(\mathbf{x}_t, y_t)$ to the learner:

$$(\mathbf{x}_t, y_t) = \arg\min_{(\mathbf{x},y)\in\Phi} \eta_t^2 \|yf\mathbf{x} - \mathbf{v}_{t-1}\|_2^2 - 2\eta_t \log \frac{1+\exp(-y\mathbf{w}_{t-1}^T\mathbf{x})}{1+\exp(-y\mathbf{w}_*^T\mathbf{x})} \tag{17}$$

6: (iii). Learner performs the labeling and then teacher updates A, D, and F_l.
7: (iv). Learner performs learning after teacher reveals y_t.
8: (v). t ← t + 1
9: **Until** $t >$ MaxIter
10: **Output:** The teaching sequence $\mathcal{D}_t$

Data set	# Examples (Teach)	# Examples (Evaluate)	# Features
10D-Guassian	400	1600	10
Comp. vs. Sci	375	1500	150
Rec. vs. Talk	369	1475	150

Table 1: Statistics of the three data sets with synthetic learners.

5 EXPERIMENTS

In this section, we first conduct the experiments on a toy data set to illustrate the trade-off between diversity and usefulness using JEDI with omniscient teacher. Then, we evaluate the convergence and the performance of JEDI with harmonic function estimation on three data sets using synthetically generated learners. At last, we evaluate the effectiveness of JEDI teaching on two real-world data sets by hiring and teaching a group of crowdsourcing workers.

5.1 Toy Data Set Visualization

In order to visualize the selected examples of the teaching sequence, we apply three different teaching methods: SGD, Iterative Machine Teaching (IMT) [13], and JEDI (omniscient teacher) on a 2D Gaussian mixture data set. This data set is draw from two Gaussian

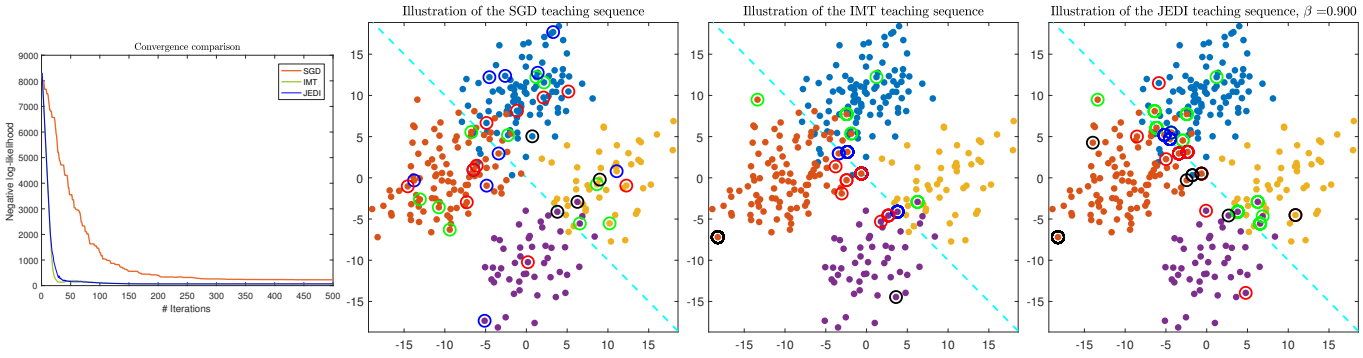

Figure 3: Left: Comparison of convergence. Right: The unique examples in the teaching sequences of SGD, IMT, and JEDI. Selected examples: Green (iter. 1-100), Blue (iter. 101-200), Red (iter. 201-300), and Black (iter. 301-500). Cyan dashed line: optimal classification hyperplane.

mixture distributions (one positive class and one negative class):

$$p_+(\mathbf{x}) = \frac{2}{3}\mathcal{N}(\mathbf{x}|\mu_1, \Sigma_1) + \frac{1}{3}\mathcal{N}(\mathbf{x}|\mu_2, \Sigma_2)$$
$$p_-(\mathbf{x}) = \frac{2}{3}\mathcal{N}(\mathbf{x}|\mu_3, \Sigma_1) + \frac{1}{3}\mathcal{N}(\mathbf{x}|\mu_4, \Sigma_2) \quad (18)$$

where the parameters are $\mu_1 = (0, 8)$, $\mu_2 = (8, 0)$, $\mu_3 = (-8, 0)$, $\mu_4 = (0, -8)$, $\Sigma_1 = [12\ 6; 6\ 12]$ and $\Sigma_2 = [10\ 5; 5\ 10]$. The number of examples in each class is 150. To guarantee a fair comparison, the target concepts $\mathbf{w}_*$ are the same for IMT and JEDI (SGD has no target concept), and the initial concept $\mathbf{w}_0$, the first teaching example $(\mathbf{x}_1, y_1)$, the step size $\eta = 0.03$ are set to be identical for all three methods.

The numbers of unique teaching examples of SGD, IMT, and JEDI are 182, 16, and 27 respectively. As we expected, SGD almost selects half of all examples for teaching. Thus, we visualize the selected examples of SGD with a stride size of 5. From the visualization results, there are several interesting observations. First, the convergence rate of IMT and JEDI are comparable and both have better convergence than SGD. Second, the selected teaching examples of JEDI are much diverse than those of IMT, yet the convergence speed of JEDI is guaranteed. As we can see, in the first 100 iterations (when teaching converges very fast), the unique teaching examples of IMT mainly focus on the upper left two data distributions, but the teaching examples of JEDI are more diversely distributed in all data distributions. The advantages of JEDI is significant especially when the examples are drawn from a mixture distribution because JEDI objective considers both the usefulness and the diversity of the teaching examples. Third, the JEDI selected examples are symmetrically scattered over the optimal classification hyperplane and have more appearances on the data distribution boundaries which reflects our theoretical analysis regarding the exploration and exploitation actions.

5.2 Adaptive Teaching with Synthetic Learners

The teacher in this group of experiments doesn't have access to the learners' concepts. Thus, teachers of Random Teaching (RT), IMT, and JEDI can only estimate a learner's concept using harmonic function. We evaluates these methods on three data sets, as shown in Table 1, which include a 10-dimensional Gaussian data set and two text data sets from 20 Newsgroups. The learners are randomly generated as a vector that has the same length as the example features. After performing the learning, a random Gaussian noise will be added to the learner's concept vector to simulate the learning uncertainty. Below are the settings of the data sets:

- **10D-Gaussian**: The means are $\mu_1 = (-0.6, \ldots, -0.6)$, $\mu_2 = (0.6, \ldots, 0.6)$ and the diagonal of its covariance matrix has random values between 1 to 10. Initial step size is set to $\eta_0 = 0.03$ and it is gradually decreased as $\eta_t = \frac{20}{20+t}\eta_0$ where t is the teaching iterations.
- **Comp.vs.Sci** and **Rec.vs.Talk**: We use their largest classification tasks for each of them. The extracted features are TF-IDF. Initial step size is set to $\eta_0 = 0.03$ and it is gradually decreased as $\eta_t = \frac{200}{200+t}\eta_0$.

All three data sets are randomly split into 20% as the teaching set (has true labels for the teacher) and 80% as the evaluation set. For the JEDI teacher, we have five different synthetic learners with $\beta \in \{0.368, 0.5, 0.75, 0.875, 0.999\}$ and these values represent for learner with memory window size of $\{1, 2, 4, 8, \text{Inf}\}$. To guarantee a fair comparison, all learners have the same initial concept $\mathbf{w}_0$, first teaching example $(\mathbf{x}_1, y_1)$, and target concept $\mathbf{w}_*$ (except RT, which doesn't have target concept). As we can see from the experiment results in Figure 4 and Table 2, the IMT and JEDI have consistently better convergence speed than RT. We also observe that the JEDI learners with exponential decay memories usually outperforms the no memory learners of IMT in terms of either the convergence speed or the evaluation accuracy. Meanwhile, as we expected, the JEDI learners with larger β has slower convergence speed, increasing number of unique teaching examples, and possibly better performance in the evaluation.

5.3 Adaptive Teaching with Real Human Learners

The teaching experiments with real human learners are designed for crowdsourcing workers to learn the concept of labeling different animals images [26] based on the animal breed. We utilize two categories of images (Cat and Canidae) and the label of each image is either *domestic* or *wild*. Following the same convention of [26],

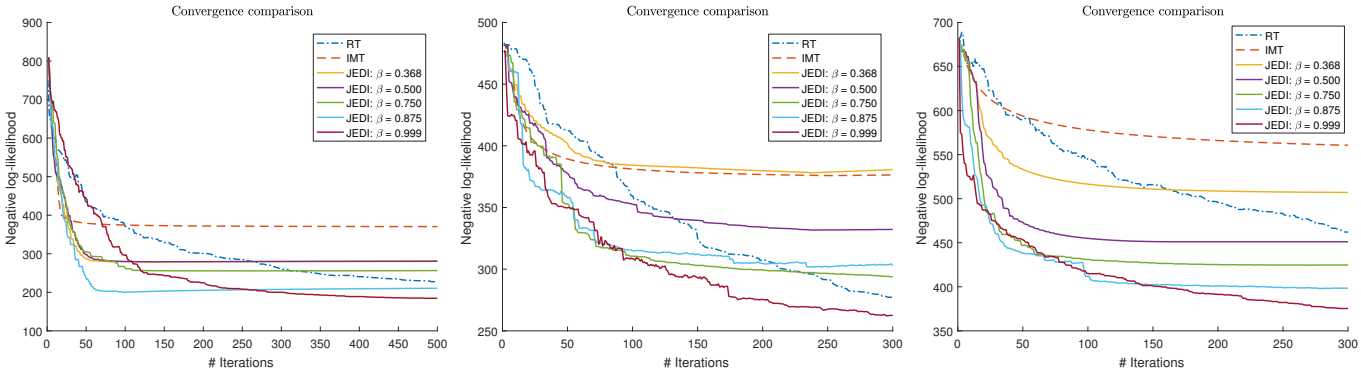

Figure 4: Convergence plots of data sets: 10D-Gaussian (left), Comp.vs.Sci (middle), Rec.vs.Talk (right)

	10D-Guassian		Comp. vs. Sci		Rec. vs. Talk	
	# Unique Teaching Examples	*Evaluation Accuracy*	*# Unique Teaching Examples*	*Evaluation Accuracy*	*# Unique Teaching Examples*	*Evaluation Accuracy*
RT	283	0.7625	210	0.7247	208	0.6041
IMT	13	0.7075	9	0.6887	5	0.5607
JEDI: β = 0.368	16	0.7575	14	0.6827	12	0.5736
JEDI: β = 0.500	18	0.7575	16	0.6927	17	0.6102
JEDI: β = 0.750	30	0.7613	27	0.7207	23	0.6292
JEDI: β = 0.875	53	0.7775	45	0.7007	34	0.6617
JEDI: β = 0.999	197	0.7825	64	0.7307	50	0.6915

Table 2: Results of three data sets with synthetic learners

each image is represented by the top 110 TF-IDF features using bag-of-visual-words extracted from a three-level image pyramid. The experiment is designed to have three modules: *memory length estimation, interactive teaching*, and *performance evaluation.*

Memory length estimation: For each crowdsourcing learner, his/her memory decay rate β is not available to the JEDI teacher beforehand. Thus, we propose to use the images sorting task to estimate each learner's β. This sorting task shows increasing number (from 2 to 9) of randomly ordered images to the learner for a few seconds (from 3s to 10s), and then ask the learner to recover the correct order of these images after random shuffling. Each learner's memory decay rate is ad-hocly estimated as $\beta = 1 - \frac{1}{\overline{n}}$ where $\overline{n}$ is the mean of the maximum number of ordered images that this learner can recover. There are three memory length estimation trials, we drop the one with smallest memory length and take the mean of the remaining two.

Interactive teaching: In total, we hired 58 crowdsourcing workers (30 for cat data set, 28 for canidae data set). All human workers are graduate students who are hired from Arizona State University and have machine learning background. Besides RT, IMT, and JEDI, we also add the Expected Error Reduction (EER [8]) as a comparison teaching method which is specially designed for teaching real human learners. To guarantee a fair comparison, each learner will be assigned with one of these four teachers using round-robin scheduling. The numbers of teaching images of JEDI are 20, 30, or 40 if $\overline{n}$ falls into these ranges $[2, 4.5], (4.5, 6.5], (6.5, 9]$ respectively. The numbers of teaching images of RT, IMT, and EER are fixed as 30. It should be noticed that the ideal number of teaching examples for different teaching tasks could have a large variation due to the various learning abilities of the learners, different scales of the data set, etc. We have left the exploration of this specific setting to the future work. To deal with the "cold start" issue, the first five teaching examples are randomly selected from the teaching set. All workers know that they will be taught, and the incentive for workers to learn is by giving double payment if they have the top 20 percent labeling accuracies among all workers.

Performance evaluation: For each crowdsourcing worker, they are asked to label 100 images (50 domestic/50 wild) in the evaluation stage. The purpose of teaching is to let the human learner grasp the idea of this domestic/wild classification concept. Thus, we propose to use *teaching gain* as the evaluation metric which is defined as the labeling accuracy during evaluation minus the labeling accuracy (of these first seen teaching examples) during teaching. From the plots shown in Figure 5, we observe that the teachers with an explicit learner's model (e.g. IMT and JEDI) performs better than model-agnostic teachers and the JEDI teacher consistently performs the best over all teaching strategies. Interestingly, we also see that JEDI is the only teacher that has positive teaching gain on the cat data set. One possible explanation is that cat breed classification is a very difficult task and these crowdsourcing workers without given a properly selected teaching sequence could hardly grasp this labeling concept. On the other hand, the human learners have difficulties to visually differentiate the wild and domestic cats is

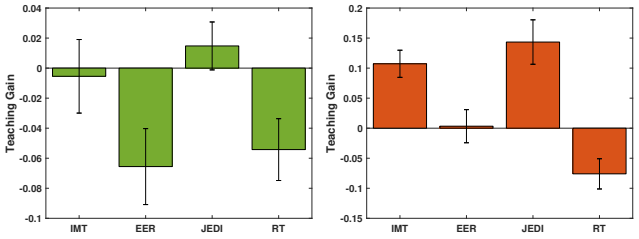

Figure 5: Teaching Gain of Cat (left) and Canidae (right)

because of that the subtle difference of their discriminative features could be easily forgotten by human learners. However, JEDI is the only model that could capture this memory loss by assuming that each learner's memory has an exponentially decayed rate.

6 CONCLUSION

In this paper, we study the problem of crowd teaching which applies the machine teaching paradigm into the crowdsourcing applications. The proposed JEDI teaching framework advances the state-of-the-art techniques in multiple dimensions in terms of human memory decay modeling, converging speed, teaching comprehensiveness and teaching accuracy, etc. The experimental results on several data sets with synthetic learners and crowdsourcing workers show the superiority of JEDI teaching. Future work could focus on multiple directions. Adapting to the multi-class teaching or finding an alternative method to perform concept estimation are two possible straightforward extensions. Furthermore, other promising explorations could be teaching the gray-box learners (using different loss functions or different learning procedures), teaching the black-box learners (model agnostic) and teaching with human interpretable explanations (e.g., area of interest on images or key phrases of documents).

ACKNOWLEDGMENTS

This work is supported by National Science Foundation under Grant No. IIS-1552654, Grant No. IIS-1813464 and Grant No. CNS-1629888, the U.S. Department of Homeland Security under Grant Award Number 2017-ST-061-QA0001, and an IBM Faculty Award. The views and conclusions are those of the authors and should not be interpreted as representing the official policies of the funding agencies or the government.

REFERENCES

[1] Yoshua Bengio, Jérôme Louradour, Ronan Collobert, and Jason Weston. 2009. Curriculum learning. In *ICML*. 41–48.

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

[3] Richard M. Felder and Linda K. Silverman. 1988. Learning and teaching styles in engineering education. *Engineering Education* (1988).

[4] Gyslain Giguère and Bradley C. Love. 2013. Limits in decision making arise from limits in memory retrieval. *Proceedings of the National Academy of Sciences* 110, 19 (2013), 7613–7618.

[5] Sally A. Goldman and Michael J. Kearns. 1995. On the Complexity of Teaching. *J. Comput. System Sci.* 50, 1 (1995), 20–31.

[6] Richard N.A. Henson. 1998. Short-Term Memory for Serial Order: The Start-End Model. *Cognitive Psychology* 36, 2 (1998), 73 – 137.

[7] Lu Jiang, Deyu Meng, Shoou-I Yu, Zhen-Zhong Lan, Shiguang Shan, and Alexander G. Hauptmann. 2014. Self-Paced Learning with Diversity. In *NIPS*. 2078–2086.

[8] Edward Johns, Oisin Mac Aodha, and Gabriel J. Brostow. 2015. Becoming the expert - interactive multi-class machine teaching. In *CVPR*. 2616–2624.

[9] Faisal Khan, Xiaojin Zhu, and Bilge Mutlu. 2011. How Do Humans Teach: On Curriculum Learning and Teaching Dimension. In *NIPS*. 1449–1457.

[10] Diederik P. Kingma and Jimmy Ba. 2014. Adam: A Method for Stochastic Optimization. *CoRR* abs/1412.6980 (2014).

[11] Ji Liu and Xiaojin Zhu. 2016. The Teaching Dimension of Linear Learners. *Journal of Machine Learning Research* 17 (2016), 162:1–162:25.

[12] Qiang Liu, Jian Peng, and Alexander T. Ihler. 2012. Variational Inference for Crowdsourcing. In *NIPS*. 701–709.

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

[14] Geoffrey R. Loftus. 1985. Evaluating Forgetting Curves. *Journal of Experimental Psychology: Learning, Memory, and Cognition* 11, 2 (1985), 397–406.

[15] Shike Mei and Xiaojin Zhu. 2015. Using Machine Teaching to Identify Optimal Training-Set Attacks on Machine Learners. In *AAAI*. 2871–2877.

[16] Kaustubh R. Patil, Xiaojin Zhu, Lukasz Kopec, and Bradley C. Love. 2014. Optimal Teaching for Limited-Capacity Human Learners. In *NIPS*. 2465–2473.

[17] Nihar B. Shah and Dengyong Zhou. 2016. No Oops, You Won't Do It Again: Mechanisms for Self-correction in Crowdsourcing. In *ICML*. 1–10.

[18] Nihar B. Shah, Dengyong Zhou, and Yuval Peres. 2015. Approval Voting and Incentives in Crowdsourcing. In *ICML*. 10–19.

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

[20] Huang Xiao, Battista Biggio, Gavin Brown, Giorgio Fumera, Claudia Eckert, and Fabio Roli. 2015. Is Feature Selection Secure against Training Data Poisoning?. In *ICML*. 1689–1698.

[21] Yan Yan, Rómer Rosales, Glenn Fung, and Jennifer G. Dy. 2011. Active Learning from Crowds. In *ICML*. 1161–1168.

[22] Dengyong Zhou, Olivier Bousquet, Thomas Navin Lal, Jason Weston, and Bernhard Schölkopf. 2003. Learning with Local and Global Consistency. In *NIPS*. 321–328.

[23] Dengyong Zhou, Qiang Liu, John C. Platt, Christopher Meek, and Nihar B. Shah. 2015. Regularized Minimax Conditional Entropy for Crowdsourcing. *CoRR* abs/1503.07240 (2015).

[24] Yao Zhou and Jingrui He. 2016. Crowdsourcing via Tensor Augmentation and Completion. In *IJCAI*. 2435–2441.

[25] Yao Zhou and Jingrui He. 2017. A Randomized Approach for Crowdsourcing in the Presence of Multiple Views. In *IEEE ICDM*. 685–694.

[26] Yao Zhou, Lei Ying, and Jingrui He. 2017. MultiC^2: an Optimization Framework for Learning from Task and Worker Dual Heterogeneity. In *SDM*. 579–587.

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

[28] Xiaojin Zhu, Zoubin Ghahramani, and John D. Lafferty. 2003. Semi-Supervised Learning Using Gaussian Fields and Harmonic Functions. In *ICML*. 912–919.

[29] Xiaojin Zhu, Adish Singla, Sandra Zilles, and Anna Rafferty. 2018. An Overview of Machine Teaching. *ArXiv* (Jan. 2018). arXiv:cs.LG/1801.05927