Optimal Teaching for Online Perceptrons

Xuezhou Zhang **Hrag Gorune Ohannessian** **Ayon Sen** **Scott Alfeld** **Xiaojin Zhu**
Department of Computer Sciences, University of Wisconsin–Madison
{zhangxz1123, gorune, ayonsn, salfeld, jerryzhu}@cs.wisc.edu

Abstract

Consider a teacher designing a good lecture for students, or a hacker drafting a poisonous text input against Tay the chatterbot. Both cases can be formulated as a task of constructing special training data, such that a known learning algorithm taking the constructed data will arrive at a prespecified target model. This task is known as optimal teaching, which has a wide range of applications in education, psychology, computer security, program synthesis, etc.. Prior analysis of optimal teaching focused exclusively on batch learners. However, a theoretical understand of optimal teaching for online (sequential) learners is also important for many applications. This paper presents the first study of optimal teaching for an online learner, specifically the perceptron. We show how to construct the shortest input sequence for a perceptron, and prove that the sequence has length one when the teacher knows everything about the perceptron, or length three when the teacher does not know the initial weight vector of the perceptron.

1 Introduction

In some applications, one has a target model in mind and knows the learning algorithm used by a machine learner. One wants to *construct* a special training set to make the learner learn the target model. One also wants the training set to be of minimal size. This constructive task is known as optimal teaching [26, 17, 12, 24, 28, 7, 2, 3, 13, 18, 6, 5, 16, 4, 22, 9, 11]. The so-constructed training set has applications in education [21, 27, 15], computer security [8, 20, 19, 1], interactive machine learning [25, 10], program synthesis [14], and so on. However, prior work has restricted its attention to batch learners. Many applications involve a sequential learner instead. In this paper, we present the first theoretical study of optimal teaching to a sequential learner, namely online perceptron.

The (homogeneous) online perceptron is a classic online learning algorithm [23]. It is designed for binary classification with the hypothesis class of homogeneous decision boundaries, namely, $\mathcal{H} = \{\mathbf{x} \mapsto \text{sign}(\langle \mathbf{w}, \mathbf{x} \rangle), \mathbf{w} \in \mathbb{R}^d\}$, where $\mathbf{x} \in \mathcal{X} = \mathbb{R}^d$, the input space, and $\text{sign}(\langle \mathbf{w}, \mathbf{x} \rangle) \in \mathcal{Y} = \{-1, 1\}$, the output space. Here, the sign function is defined as $\text{sign}(z) = +1$ if $z > 0$ and -1 otherwise. The online perceptron updates the weight vector $\mathbf{w}$, which is a vector normal to its currently learned decision boundary. Therefore, two hypotheses are equivalent if and only if their corresponding weight vectors satisfy $\mathbf{w} = c\mathbf{w}'$, for some $c > 0$. The exact algorithm is presented in Algorithm 1:

Algorithm 1 Online Perceptron

1: Learning parameters: Initial weight vector $\mathbf{w_0} \in \mathbb{R}^d$, learning rate $\eta > 0$.
2: **for** $t = 1 \dots$ **do**
3: receive $\mathbf{x_t}$
4: predict $\hat{y}_t = \text{sign}(\langle \mathbf{w_{t-1}}, \mathbf{x_t} \rangle)$
5: receive y_t
6: $\mathbf{w}_t \leftarrow \mathbf{w}_{t-1} + \mathbb{1}_{(y_t \langle \mathbf{x}_t, \mathbf{w}_{t-1} \rangle \leq 0)} \eta y_t \mathbf{x}_t$

30th Conference on Neural Information Processing Systems (NIPS 2016), Barcelona, Spain.

Note that in this paper, we allow the perceptron learner to have a possibly non-zero initial parameter $\mathbf{w}_0$, and arbitrary step size η. These two variables capture the notion of 'background knowledge' and 'learning rate' and fully specify the perceptron learner.

2 Optimal Teaching with Full Knowledge of the Perceptrons

In this section, we study the optimal teaching problem of online perceptron in the idealistic teaching setting, where the teacher has full knowledge of the learner. In particular, the teacher knows that the learner is an online perceptron with initial weight vector $\mathbf{w}_0$ and learning rate η. The teacher's goal is to teach a target decision boundary represented by an equivalence class of target parameters, $\{c\mathbf{w}^* : c > 0\}$ for some $\mathbf{w}^* \neq 0$, to the student. To achieve this teaching goal, the teacher is allowed to construct a teaching sequence S using any training items that are consistent with the target decision boundary. In other word, the teacher can use any $\mathbf{x} \in \mathcal{X}$ to construct the training sequence, but the corresponding label y has to satisfy $y = \text{sign}(\langle \mathbf{w}^*, \mathbf{x} \rangle)$. We will show that in this teaching setting, the teacher is able to teach the exact target decision boundary to the learner, which gives rise to the following defintion and theorem.

Definition 1. *Let $\mathbb{S}$ be the space of finite training sequences. Define an online perceptron with initial weight $\mathbf{w}_0$ and learning rate η as a function $\mathcal{A}^{\eta}_{\mathbf{w}_0} : \mathbb{S} \to \mathbb{R}^d$, which takes a sequence of input $S \in \mathbb{S}$ and produces a weight vector $\mathbf{w} \in \mathbb{R}^d$. Then, the **Exact Teaching Dimension** of $\mathcal{A}^{\eta}_{\mathbf{w}_0}$ with target decision boundary $\{c\mathbf{w}^*\}$ is*

$$TD_{exact}(\mathbf{w}^*, \mathcal{A}^{\eta}_{\mathbf{w}_0}) = \min \left\{ |S| : \mathcal{A}^{\eta}_{\mathbf{w}_0}(S) = c\mathbf{w}^* \text{ for some } c > 0 \right\}. \tag{1}$$

Theorem 1. *Given any target decision boundary $\{c\mathbf{w}^* : c > 0\}$ for $\mathbf{w}^* \neq 0$, an online perceptron $\mathcal{A}^{\eta}_{\mathbf{w}_0}$ with initial parameter $\mathbf{w}_0 \notin \{c\mathbf{w}^*\}$ and learning rate η has an exact teaching dimension $TD_{exact}(\mathbf{w}^*, \mathcal{A}^{\eta}_{\mathbf{w}_0}) = 1$.*

Proof. We prove the teaching dimension by exhibiting the teaching sequence in two separate cases.
Case 1: $\langle \mathbf{w}_0, \mathbf{w}^* \rangle > 0$. In this case, $(\mathbf{x_1}, y_1)$ can be constructed as follows:

$$\text{Choose any } c \in \left(\frac{\langle \mathbf{w}_0, \mathbf{w}^* \rangle}{\|\mathbf{w}^*\|^2}, \frac{\|\mathbf{w}_0\|^2}{\langle \mathbf{w}_0, \mathbf{w}^* \rangle} \right], \text{ and } y_1 \in \{-1, +1\}, \text{ let } \mathbf{x_1} = \frac{c\mathbf{w}^* - \mathbf{w}_0}{\eta y_1}. \tag{2}$$

Case 2: $\langle \mathbf{w}_0, \mathbf{w}^* \rangle \leq 0$. In this case, $(\mathbf{x_1}, y_1)$ can be constructed as follows:

$$\text{Choose any } c \in (0, \infty), \text{ and } y_1 \in \{-1, +1\}, \text{ let } \mathbf{x_1} = \frac{c\mathbf{w}^* - \mathbf{w}_0}{\eta y_1}. \tag{3}$$

We now verify that our constructions successfully teach the target decision boundary. In order for our construction to be valid, it has to satisfy three conditions: it has to be consistent with the target parameter $\mathbf{w}^*$, it has to make the perceptron update its weight vector, and it has to update $\mathbf{w}_0$ to exactly $c\mathbf{w}^*$, for some $c > 0$. Written mathematically, $(\mathbf{x_1}, y_1)$ needs to satisfy the following:

$$\begin{cases} y_1 = \text{sign}(\langle \mathbf{w}^*, \mathbf{x}_1 \rangle) & (4) \\ y_1 \langle \mathbf{w}_0, \mathbf{x_1} \rangle \leq 0 & (5) \\ c\mathbf{w}^* = \mathbf{w}_0 + \eta y_1 \mathbf{x_1}, \text{ for some } c > 0 & (6) \end{cases}$$

Notice that any $(\mathbf{x}_1, y_1)$ in the form $\mathbf{x_1} = \frac{c\mathbf{w}^* - \mathbf{w}_0}{\eta y_1}$ satisfies condition (6), so it remains to show that our construction satisfy condition (4) and (5), and we again look at each of the two cases.
Case 1: $\langle \mathbf{w}_0, \mathbf{w}^* \rangle > 0$. First notice that the interval $\left(\frac{\langle \mathbf{w}_0, \mathbf{w}^* \rangle}{\|\mathbf{w}^*\|^2}, \frac{\|\mathbf{w}_0\|^2}{\langle \mathbf{w}_0, \mathbf{w}^* \rangle} \right]$ will not be empty for any $\mathbf{w}_0 \neq c\mathbf{w}^*$ for some $c > 0$, as a result of the Cauchy-Schwarz inequality:

$$\frac{\|\mathbf{w}_0\|^2}{\langle \mathbf{w}_0, \mathbf{w}^* \rangle} - \frac{\langle \mathbf{w}_0, \mathbf{w}^* \rangle}{\|\mathbf{w}^*\|^2} = \frac{\|\mathbf{w}_0\|^2 \|\mathbf{w}^*\|^2 - \langle \mathbf{w}_0, \mathbf{w}^* \rangle^2}{\langle \mathbf{w}_0, \mathbf{w}^* \rangle \|\mathbf{w}^*\|^2} > 0. \tag{7}$$

Let $(\mathbf{x}_1, y_1)$ be constructed according to (2), then

$$\text{sign}(\langle \mathbf{w}^*, \mathbf{x}_1 \rangle) = y_1 \text{sign}\left(\frac{c\|\mathbf{w}^*\|^2 - \langle \mathbf{w}_0, \mathbf{w}^* \rangle}{\eta} \right) = y_1, \text{ as } c > \frac{\langle \mathbf{w}_0, \mathbf{w}^* \rangle}{\|\mathbf{w}^*\|^2}, \tag{8}$$

so condition (4) is satisfied, and now we show condition (5) is also satisfied:

$$y_1\langle \mathbf{w}_0, \mathbf{x}_1\rangle = \frac{c\langle \mathbf{w}_0, \mathbf{w}^*\rangle - \|\mathbf{w}_0\|^2}{\eta} \leq 0, \text{ as } \langle \mathbf{w}_0, \mathbf{w}^*\rangle > 0 \text{ and } c \leq \frac{\|\mathbf{w}_0\|^2}{\langle \mathbf{w}_0, \mathbf{w}^*\rangle}. \tag{9}$$

Therefore, construction (2) is valid.
Case 2: $\langle \mathbf{w}_0, \mathbf{w}^*\rangle \leq 0$. Let $(\mathbf{x}_1, y_1)$ be constructed according to (3), similarly we have

$$\text{sign}(\langle \mathbf{w}^*, \mathbf{x}_1\rangle) = y_1 \text{sign}\left(\frac{c\|\mathbf{w}^*\|^2 - \langle \mathbf{w}_0, \mathbf{w}^*\rangle}{\eta}\right) = y_1, \text{ as } c > 0 \text{ and } \langle \mathbf{w}_0, \mathbf{w}^*\rangle \leq 0 \tag{10}$$

and

$$y_1\langle \mathbf{w}_0, \mathbf{x}_1\rangle = \frac{c\langle \mathbf{w}_0, \mathbf{w}^*\rangle - \|\mathbf{w}_0\|^2}{\eta} \leq 0, \text{ as } \langle \mathbf{w}_0, \mathbf{w}^*\rangle \leq 0 \text{ and } c \geq 0 \tag{11}$$

Therefore, construction (3) is valid.
Moreover, notice that the optimality of our construction is obvious, as we have constructed a teaching sequence of length 1, and clearly for $\mathbf{w}_0 \notin \{c\mathbf{w}^*\}$, an empty teaching sequence does not update the initial weight vector $\mathbf{w}_0$, and so does not achieve the teaching goal. This completes the proof of theorem 1. □

3 Approximate Teaching with Unknown $\mathbf{w}_0$

In this section, we look into a different teaching setting, where the teacher does not have full knowledge of the learner. In particular, the teacher knows that the student is a perceptron learner with learning rate η but does not know the exact form of its initial weight vector $\mathbf{w}_0$. Instead, she knows that there is an upper bound on the norm of the weight vector: $\|\mathbf{w}_0\| \leq b$. To compensate for this uncertainty, the perceptron reports back to the teacher whether it performs an update or not each time she feeds the perceptron with a training item $(\mathbf{x}_t, y_t)$. Recall that the perceptron performs an update if $y_t\langle \mathbf{x}_t, \mathbf{w}_{t-1}\rangle \leq 0$. Similar to the previous setting, the teacher can only use training labels that are consistent with the target model. In this teaching setting,it is not clear that teaching the *exact* target model is possible, so a natural question is how to define a realizable teaching goal. In this section, we show that what we can achieve is an approximate teaching goal such that teaching is considered successful if the learner arrives at a model close enough to the target model. First, we define our teaching goal in this teaching setting by introducing a notion called teaching strategy.

Definition 2. *Let $\mathbb{S}$ be the space of finite training sequences. Let $\mathcal{A}^\eta$ be an online perceptron with unknown initial weight vector and known learning rate η. Define an **ϵ-approximate teaching strategy** for a target decision boundary $\{c\mathbf{w}^*\}$ and learning algorithm $\mathcal{A}^\eta$ as a function $t_\epsilon : \mathbb{R}^d \to \mathbb{S}$, such that for all $\mathbf{w}_0 \in \mathbb{R}^d$,*

$$1 - \frac{\langle \mathcal{A}^\eta_{\mathbf{w}_0}(t_\epsilon(\mathbf{w}_0)), \mathbf{w}^*\rangle}{\|\mathcal{A}^\eta_{\mathbf{w}_0}(t_\epsilon(\mathbf{w}_0))\|\|\mathbf{w}^*\|} \leq \epsilon \tag{12}$$

Here, the quantity $1 - \frac{\langle \mathcal{A}^\eta_{\mathbf{w}_0}(t_\epsilon(\mathbf{w}_0)), \mathbf{w}^*\rangle}{\|\mathcal{A}^\eta_{\mathbf{w}_0}(t_\epsilon(\mathbf{w}_0))\|\|\mathbf{w}^*\|}$ is a measure of the angular (cosine) difference between the taught parameter $\mathcal{A}^\eta_{\mathbf{w}_0}(t_\epsilon(\mathbf{w}_0))$ and the target parameter $\mathbf{w}^*$. Now we define the notion of teaching dimension in this teaching setting, and compute it for online perceptrons.

Definition 3. *Let $\mathbb{T}_\epsilon$ be the space of ϵ-approximate teaching strategies for a target decision boundary $\{c\mathbf{w}^*\}$ and an online perceptron $\mathcal{A}^\eta$ with unknown initial weight vector and known learning rate η. Then, the **ϵ-Approximate Teaching Dimension** of $\mathcal{A}^\eta$ with target decision boundary $\{c\mathbf{w}^*\}$ is*

$$TD_\epsilon(\mathbf{w}^*, \mathcal{A}^\eta) = \min_{t \in \mathbb{T}_\epsilon} \max_{\mathbf{w}_0 \in \mathbb{R}^d} |t(\mathbf{w}_0)| \tag{13}$$

Theorem 2. *For any $\epsilon > 0$ and target decision boundary $\{c\mathbf{w}^* : c > 0 \text{ and } \mathbf{w}^* \neq 0\}$, an online perceptron $\mathcal{A}^\eta$ with learning rate η and an unknown initial parameter $\mathbf{w}_0 \notin \{c\mathbf{w}^*\}$ with a known upper bound $\|\mathbf{w}_0\| \leq b$ has an ϵ-approximate teaching dimension $TD_\epsilon(\mathbf{w}^*, \mathcal{A}^\eta) = 3$.*

Proof. We first demonstrate a teaching strategy that will generate a teaching sequence of size at most 3. Without loss of generality, assume $\|\mathbf{w}^*\| = 1$. First, pick any $\mathbf{x}_1 \in \mathbb{R}^d$, s.t. $\|\mathbf{x}_1\| = \frac{2b}{\epsilon\eta}$ and $\langle \mathbf{w}^*, \mathbf{x}_1\rangle = 0$. Note that such $\mathbf{x}_1$ always exists and is not unique. Then, let $\mathbf{x}'_1 = -\mathbf{x}_1$. Recall that $\text{sign}(0) = -1$. Correspondingly, we have $y_1 = \text{sign}(\langle \mathbf{w}^*, \mathbf{x}_1\rangle) = -1$ and $y'_1 = \text{sign}(\langle \mathbf{w}^*, \mathbf{x}'_1\rangle) =$

-1. Since $y_1\langle \mathbf{w}_0, \mathbf{x}_1\rangle + y_1'\langle \mathbf{w}_0, \mathbf{x}_1'\rangle = 0$, at least one of the two terms is less than or equal to 0, implying at least one of them will make the perceptron update.
With $(\mathbf{x}_1, y_1)$ and $(\mathbf{x}_1', y_1')$ defined, the teaching strategy generates a teaching sequence as follows:
At iteration 1, feed $(\mathbf{x}_1, y_1)$ to the learner.
Case 1: If $(\mathbf{x}_1, y_1)$ triggers an update, then we know $\langle \mathbf{x}_1, \mathbf{w}_0\rangle \geq 0$, and $\mathbf{w}_1 = \mathbf{w}_0 + \eta y_1 \mathbf{x}_1 = \mathbf{w}_0 - \eta \mathbf{x}_1$.
At iteration 2, let $\mathbf{x}_2 = \frac{2b}{\epsilon\eta}\mathbf{w}^* + \mathbf{x}_1$ and $y_2 = 1$. Feed $(\mathbf{x}_2, y_2)$ to the learner. Note that $(\mathbf{x}_2, y_2)$ is consistent with $\mathbf{w}^*$, as $\text{sign}(\langle \mathbf{x}_2, \mathbf{w}^*\rangle) = \text{sign}\left(\frac{2b}{\epsilon\eta}\|\mathbf{w}^*\|^2 + \langle \mathbf{x}_1, \mathbf{w}^*\rangle\right) = \text{sign}\left(\frac{2b}{\epsilon\eta}\|\mathbf{w}^*\|^2\right) = y_2$.
We also have

$$y_2 \langle \mathbf{x}_2, \mathbf{w}_1\rangle = \left\langle \frac{2b}{\epsilon\eta}\mathbf{w}^* + \mathbf{x}_1, \mathbf{w}_0 - \eta\mathbf{x}_1 \right\rangle = \left\langle \frac{2b}{\epsilon\eta}\mathbf{w}^* + \mathbf{x}_1, \mathbf{w}_0 \right\rangle - \eta\|\mathbf{x}_1\|^2 \tag{14}$$

$$\leq \left(\frac{2b}{\epsilon\eta}\|\mathbf{w}^*\| + \|\mathbf{x}_1\|\right)\|\mathbf{w}_0\| - \eta\|\mathbf{x}_1\|^2 \leq \frac{4b^2}{\epsilon\eta} - \frac{4b^2}{\epsilon^2\eta} < 0, \text{ for } \epsilon < 1 \tag{15}$$

in which (15) is a result of triangle inequality and Cauchy-Schwarz inequality. Therefore, $(\mathbf{x}_2, y_2)$ also triggers an update. As a result, $\mathbf{w}_2 = \mathbf{w}_1 + \eta y_2 \mathbf{x}_2 = \mathbf{w}_0 + \frac{2b}{\epsilon}\mathbf{w}^*$, and

$$\frac{\langle \mathbf{w}_2, \mathbf{w}^*\rangle}{\|\mathbf{w}_2\|\|\mathbf{w}^*\|} = \frac{\frac{2b}{\epsilon}\|\mathbf{w}^*\|^2 + \langle \mathbf{w}_0, \mathbf{w}^*\rangle}{\|\mathbf{w}_2\|\|\mathbf{w}^*\|} \geq \frac{\frac{2b}{\epsilon} - b}{\frac{2b}{\epsilon} + b} = 1 - \frac{2b}{2b + b\epsilon}\epsilon \geq 1 - \epsilon. \tag{16}$$

Therefore, $\mathcal{S} = (\mathbf{x}_1, y_1), (\mathbf{x}_2, y_2)$ achieves the teaching goal.
Case 2: If $(\mathbf{x}_1, y_1)$ does not trigger an update, then let $\mathbf{x}_2 = \mathbf{x}_1'$ and $y_2 = y_1'$, and $\mathbf{x}_3 = \frac{2b}{\epsilon\eta}\mathbf{w}^* + \mathbf{x}_2$ and $y_3 = 1$. Arguments similar to those above show that $\mathbf{x}_2$ and $\mathbf{x}_3$ both trigger an update, and $\mathbf{w}_3$ satisfies the ϵ constraint. In this case, $\mathcal{S} = (\mathbf{x}_1, y_1), (\mathbf{x}_2, y_2), (\mathbf{x}_3, y_3)$ achieves the teaching goal.

Now we are left to show that no teaching strategy can construct a teaching sequence of size two for all $\mathbf{w}_0$, and we show this by contradiction. Suppose that there exists such a teaching strategy, i.e. there exists $(\mathbf{x}_1, y_1)$, $(\mathbf{x}_2, y_2)$ and $(\mathbf{x}_2', y_2')$, such that $S_1 = (\mathbf{x}_1, y_1), (\mathbf{x}_2', y_2')$ achieves the teaching goal if the perceptron updates on $(\mathbf{x}_1, y_1)$, and $S_2 = (\mathbf{x}_1, y_1), (\mathbf{x}_2, y_2)$ achieves the teaching goal if the perceptron does not update on $(\mathbf{x}_1, y_1)$. In particular, consider S_2 in two possible cases.
If S_2 satisfies $\mathbf{x}_2 \neq -a\mathbf{x}_1$ for any $a > 0$. If $y_2 = y_1$, in the case that $\mathbf{w}_0 = y_1(\frac{\mathbf{x}_1}{\|\mathbf{x}_1\|} + \frac{\mathbf{x}_2}{\|\mathbf{x}_2\|})$, we have by Cauchy-Schwarz inequality,

$$y_1\langle \mathbf{x}_1, \mathbf{w}_0\rangle = \|\mathbf{x}_1\|\left(\frac{\|\mathbf{x}_1\|^2}{\|\mathbf{x}_1\|^2} + \frac{\langle \mathbf{x}_1, \mathbf{x}_2\rangle}{\|\mathbf{x}_1\|\|\mathbf{x}_2\|}\right) > 0, \text{ implying} \tag{17}$$

$$y_2\langle \mathbf{x}_2, \mathbf{w}_1\rangle = y_1\langle \mathbf{x}_2, \mathbf{w}_0\rangle = \|\mathbf{x}_2\|\left(\frac{\|\mathbf{x}_2\|^2}{\|\mathbf{x}_2\|^2} + \frac{\langle \mathbf{x}_1, \mathbf{x}_2\rangle}{\|\mathbf{x}_1\|\|\mathbf{x}_2\|}\right) > 0. \tag{18}$$

If $y_2 = -y_1$, in the case $\mathbf{w}_0 = y_1\left(\frac{\mathbf{x}_1}{\|\mathbf{x}_1\|} - \frac{\mathbf{x}_2}{\|\mathbf{x}_2\|}\right)$, similar calculations show that $y_1\langle \mathbf{x}_1, \mathbf{w}_0\rangle > 0$ and $y_2\langle \mathbf{x}_2, \mathbf{w}_1\rangle > 0$. Therefore, if $\mathbf{x}_2 \neq -a\mathbf{x}_1$ for any $a > 0$, S_2 does not guarantee to trigger an update for all $\mathbf{w}_0$, so S_2 fails to teach the target model. If instead $\mathbf{x}_2 = -a\mathbf{x}_1$, for some $a > 0$, still we will have $y_2\langle \mathbf{x}_2, \mathbf{w}_2\rangle = y_1\langle \mathbf{x}_1, \mathbf{w}_0\rangle > 0$, except if $\mathbf{x}_1$ and $\mathbf{x}_2$ both lie on the decision boundary, in which case $y_2 = y_1 = -1$. Then, suppose $\langle \mathbf{x}_1, \mathbf{w}^*\rangle = 0$ and $\mathbf{x}_2 = -a\mathbf{x}_1$, for some $a > 0$, consider the case $\langle \mathbf{w}_0, \mathbf{w}^*\rangle < 0$, since $\mathbf{x}_1$ does not trigger the update and $\mathbf{x}_2$ triggers the update, $\mathbf{w}_1 = \mathbf{w}_0$, and $\mathbf{w}_2 = \mathbf{w}_1 + \eta y_2 \mathbf{x}_2$, yet still we have $\langle \mathbf{w}_2, \mathbf{w}^*\rangle = \langle \mathbf{w}_0, \mathbf{w}^*\rangle < 0$, as $\langle \mathbf{x}_2, \mathbf{w}^*\rangle = 0$, so S_2 fails to teach the target model. Therefore, in either, S_2 fails to achieve the teaching goal for all $\mathbf{w}_0$. □

4 Discussions

There are many interesting avenues to continue exploring. For example, our construction in the unknown $\mathbf{w}_0$ setting is somewhat extreme, in the sense that the norms of the input vectors in our teaching sequences grow dramatically as ϵ decreases. This naturally gives rise to the question: How do constraints on the (e.g. norm of the) training items affect teaching dimension? Another important question is how do teachers handle other uncertainties, e.g. an unknown learning rate η. The definition of *teaching strategy* already captures the adaptive nature of optimal teaching in the situation of uncertainty. The teacher can only generate a teaching sequence 'on-site' that adapts to the student's feedbacks. Last but not least, one would also want to extend optimal teaching problem to other sequential learners such as stochastic gradient descent (SGD).

References

[1] S. Alfeld, X. Zhu, and P. Barford. Data poisoning attacks against autoregressive models. *AAAI*, 2016.

[2] D. Angluin. Queries revisited. *Theoretical Computer Science*, 313(2):175–194, 2004.

[3] D. Angluin and M. Krikis. Teachers, learners and black boxes. *COLT*, 1997.

[4] D. Angluin and M. Krikis. Learning from different teachers. *Machine Learning*, 51(2):137–163, 2003.

[5] F. J. Balbach. Measuring teachability using variants of the teaching dimension. *Theor. Comput. Sci.*, 397(1-3):94–113, 2008.

[6] F. J. Balbach and T. Zeugmann. Teaching randomized learners. *COLT*, pages 229–243, 2006.

[7] F. J. Balbach and T. Zeugmann. Recent developments in algorithmic teaching. In *Proceedings of the 3rd International Conference on Language and Automata Theory and Applications*, pages 1–18, 2009.

[8] M. Barreno, B. Nelson, A. D. Joseph, and J. D. Tygar. The security of machine learning. *Machine Learning Journal*, 81(2):121–148, 2010.

[9] S. Ben-David and N. Eiron. Self-directed learning and its relation to the VC-dimension and to teacher-directed learning. *Machine Learning*, 33(1):87–104, 1998.

[10] M. Cakmak and A. Thomaz. Mixed-initiative active learning. *ICML Workshop on Combining Learning Strategies to Reduce Label Cost*, 2011.

[11] T. Doliwa, G. Fan, H. U. Simon, and S. Zilles. Recursive teaching dimension, VC-dimension and sample compression. *Journal of Machine Learning Research*, 15:3107–3131, 2014.

[12] S. Goldman and M. Kearns. On the complexity of teaching. *Journal of Computer and Systems Sciences*, 50(1):20–31, 1995.

[13] S. A. Goldman and H. D. Mathias. Teaching a smarter learner. *Journal of Computer and Systems Sciences*, 52(2):255–267, 1996.

[14] S. Jha and S. A. Seshia. A theory of formal synthesis via inductive learning. *CoRR*, 2015.

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

[16] H. Kobayashi and A. Shinohara. Complexity of teaching by a restricted number of examples. *COLT*, pages 293–302, 2009.

[17] Ji Liu, Xiaojin Zhu, and H. Gorune Ohannessian. The teaching dimension of linear learners. In *The 33rd International Conference on Machine Learning (ICML)*, 2016.

[18] H. David Mathias. A model of interactive teaching. *J. Comput. Syst. Sci.*, 54(3):487–501, 1997.

[19] S. Mei and X. Zhu. The security of latent Dirichlet allocation. *AISTATS*, 2015.

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

[21] K. Patil, X. Zhu, L. Kopec, and B. C. Love. Optimal teaching for limited-capacity human learners. *Advances in Neural Information Processing Systems (NIPS)*, 2014.

[22] R. L. Rivest and Y. L. Yin. Being taught can be faster than asking questions. *COLT*, 1995.

[23] F. Rosenblatt. The perceptron: A probabilistic model for information storage and organization in the brain. *Psychological Review*, 65(6):386–408, 1958.

[24] A. Shinohara and S. Miyano. Teachability in computational learning. *New Generation Computing*, 8(4):337—-348, 1991.

[25] J. Suh, X. Zhu, and S. Amershi. The label complexity of mixed-initiative classifier training. *International Conference on Machine Learning (ICML)*, 2016.

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

[27] Xiaojin Zhu. Machine teaching for bayesian learners in the exponential family. In *Advances in Neural Information Processing Systems (NIPS)*, pages 1905–1913, 2013.

[28] S. Zilles, S. Lange, R. Holte, and M. Zinkevich. Models of cooperative teaching and learning. *Journal of Machine Learning Research*, 12:349–384, 2011.