Optimal Teaching for Online Perceptrons

Xuezhou Zhang, Hrag Gorune Ohannessian, Ayon Sen, Scott Alfeld, Xiaojin Zhu

Department of Computer Sciences, University of Wisconsin–Madison

Optimal Teaching Problem

- **Student**: A machine learner $\mathcal{A}$. In this work, we focus on **sequential** learners.
- **Teacher**: A person who knows a target model θ^*, and wants to teach it to the student $\mathcal{A}$ by creating a training set $\mathcal{D} \subseteq \mathcal{X} \times \mathcal{Y}$, where $\mathcal{X}$ is the input space and $\mathcal{Y}$ is the label space.
- **Constructive Setting**: In this work, we allow $\mathcal{X}$ to be the whole $\mathbb{R}^d$. This is called the **constructive** teaching setting, as opposed to the **pool-based** teaching setting, where $\mathcal{X}$ is a finite subset of $\mathbb{R}^d$.
- **Goal**: Find the 'best' training set.
- **General Optimization Formulation**:

$$\min_{\mathcal{D}} \quad \text{loss}(\mathcal{A}(\mathcal{D}), \theta^*) + \text{effort}(\mathcal{D})$$

Alternatively,

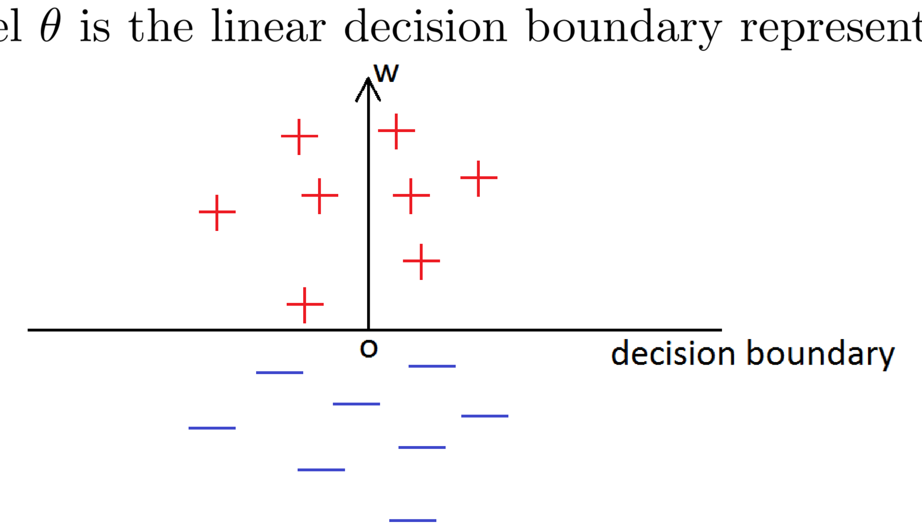
$$\begin{aligned} \min_{\mathcal{D}} \quad & \text{effort}(\mathcal{D}), \\ \text{s.t.} \quad & \text{loss}(\mathcal{A}(\mathcal{D}), \theta^*) \le \epsilon \end{aligned}$$

Online Perceptron

Algorithm 1 Online Perceptron

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

- **Homogeneous**: Linear decision boundary through the origin.
- **General Setting**: Allow non-zero $\mathbf{w}_0$ and arbitrary learning rate η.
- **Formulation**: In this work, the machine learner $\mathcal{A}$ is the perceptron, and model θ is the linear decision boundary represented by the parameter $\mathbf{w}$.

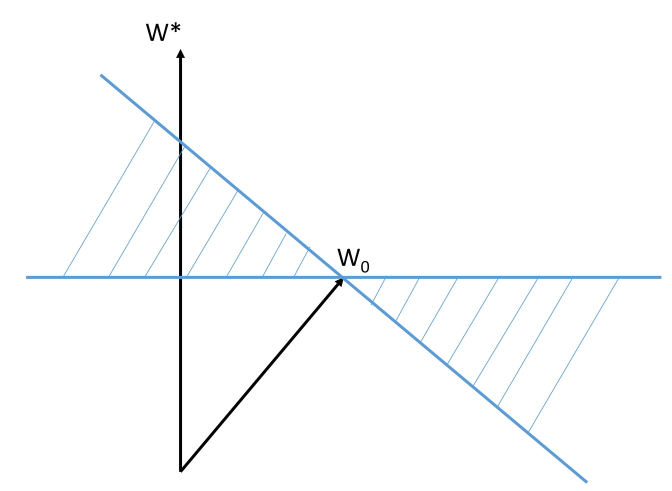


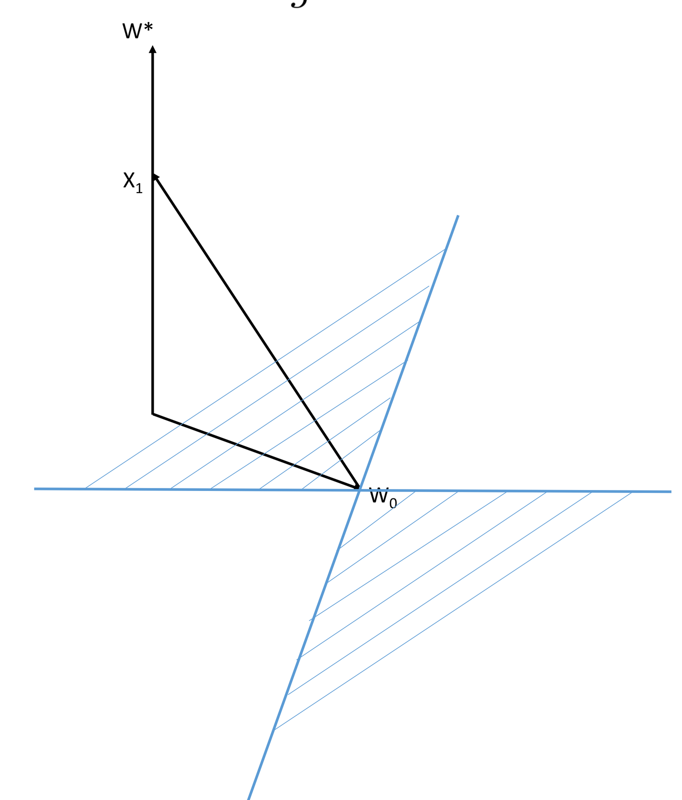
Teaching with Full Knowledge of Perceptron

Definition. *The* ***Exact Teaching dimension*** *of perceptron is defined as*

$$\begin{aligned} \arg\min_{\mathcal{D}} \quad & |\mathcal{D}|, \\ s.t. \quad & \mathcal{A}(\mathcal{D}) = \mathbf{w}^* \end{aligned}$$

Theorem. *For any target parameter $\mathbf{w}^*$, a perceptron with any initial weight $\mathbf{w}_0$ and learning rate η has exact teaching dimension 1.*





Case 1: $\langle \mathbf{w}_0, \mathbf{w}^* \rangle > 0$.

Choose $c \in \left(\frac{\langle \mathbf{w}_0, \mathbf{w}^* \rangle}{\|\mathbf{w}^*\|^2}, \frac{\|\mathbf{w}_0\|^2}{\langle \mathbf{w}_0, \mathbf{w}^* \rangle} \right]$,

and $y_1 \in \{-1, +1\}$,

let $\mathbf{x_1} = \frac{c\mathbf{w}^* - \mathbf{w_0}}{\eta y_1}$.

Case 2: $\langle \mathbf{w}_0, \mathbf{w}^* \rangle \le 0$.

Choose $c \in (0, \infty)$,

and $y_1 \in \{-1, +1\}$,

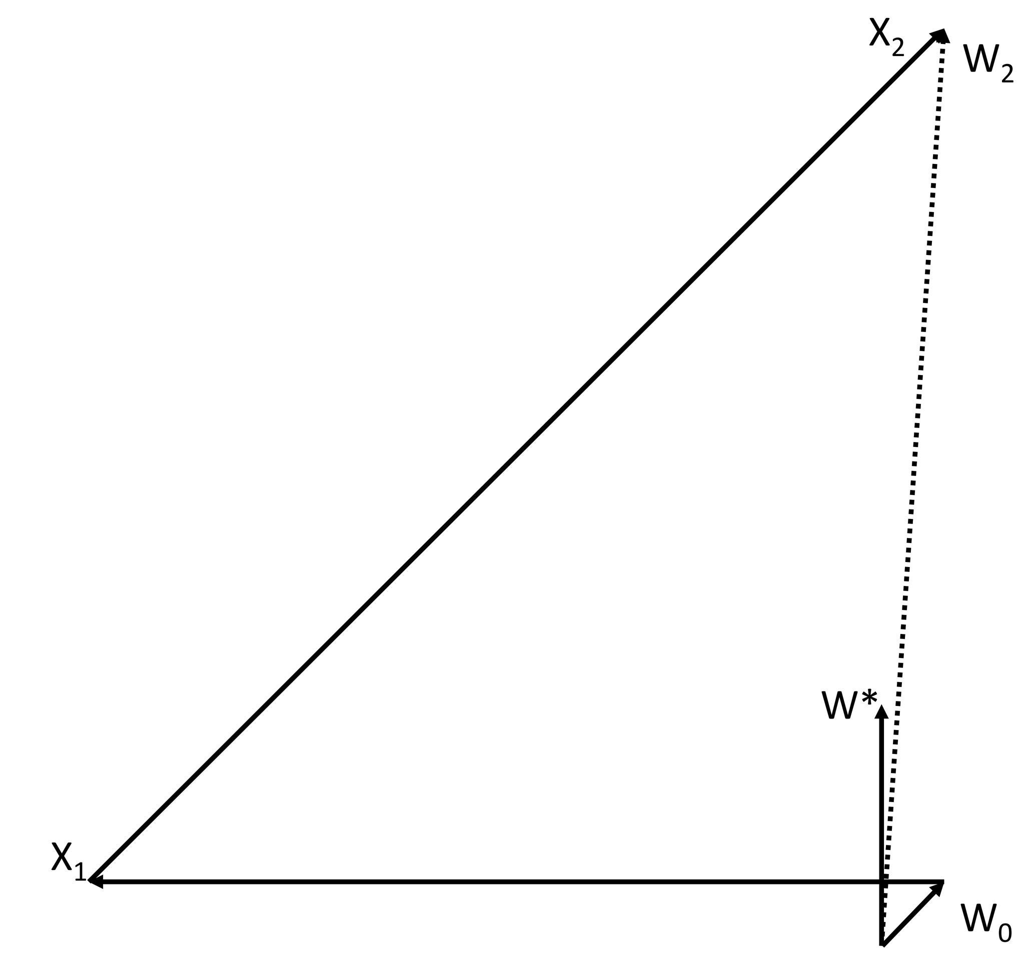
let $\mathbf{x_1} = \frac{c\mathbf{w}^* - \mathbf{w_0}}{\eta y_1}$.

Approximate Teaching with Unknown W_0

Definition. *The* ***ϵ-Approximate Teaching dimension*** *of perceptron is defined as*

$$\begin{aligned} \arg\min_{\mathcal{D}} \quad & |\mathcal{D}|, \\ s.t. \quad & \frac{\langle \mathcal{A}(\mathcal{D}), \mathbf{w}^* \rangle}{\|\mathcal{A}(\mathcal{D})\| \|\mathbf{w}^*\|} \ge 1 - \epsilon \end{aligned}$$

Theorem. *For any target parameter $\mathbf{w}^*$ and precision ϵ, a perceptron with unknown initial weight $\mathbf{w}_0$ and known learning rate η has ϵ-approximate teaching dimension 3.*



- Pick $\mathbf{x}_1 \in \mathbb{R}^d$, s.t. $\|\mathbf{x}_1\| = \frac{2b}{\epsilon\eta}$ and $\langle \mathbf{w}^*, \mathbf{x}_1 \rangle = 0$.
- Let $\mathbf{x}_1' = -\mathbf{x}_1$.
- At iteration 1, feed $\mathbf{x}_1$ to the learner. If $\mathbf{x}_1$ triggers an update, let $\mathbf{x}_2 = \frac{2b}{\epsilon\eta}\mathbf{w}^* + \mathbf{x}_1$ and feed $\mathbf{x}_2$ to the learner.
- Else if $\mathbf{x}_1$ does not trigger an update, let $\mathbf{x}_2 = \mathbf{x}_1'$ and $\mathbf{x}_3 = \frac{2b}{\epsilon\eta}\mathbf{w}^* + \mathbf{x}_2$. Feed $\mathbf{x}_2$, $\mathbf{x}_3$ to the learner.

Discussion and Future Work

- How do **constraints** on the (e.g. norm of the) training items affect teaching dimension?
- How do teachers handle other **uncertainties**, e.g. an unknown learning rate η?
- We can already see from the construction above, with the existence of uncertainty, each step of the teaching sequence depend on the student's feedback in previous steps. This **interactive** nature of teaching is worth studying.
- **Vision**: A '**cooperative**' learning setting, where the student tries to learn the target model, while the teacher is learning to teach.
- Extend optimal teaching problem to other **sequential learners** such as stochastic gradient descent (SGD).

Contact

Xuezhou Zhang,
Department of Computer Sciences
University of Wisconsin–Madison
Email: zhangxz1123@cs.wisc.edu