Iterative Classroom Teaching

Teresa Yeo
LIONS, EPFL
teresa.yeo@epfl.ch

Parameswaran Kamalaruban
LIONS, EPFL
kamalaruban.parameswaran@epfl.ch

Adish Singla
MPI-SWS
adishs@mpi-sws.org

Arpit Merchant
MPI-SWS
arpitdm@mpi-sws.org

Thibault Asselborn
CHILI Lab, EPFL
thibault.asselborn@epfl.ch

Louis Faucon
CHILI Lab, EPFL
louis.faucon@epfl.ch

Pierre Dillenbourg
CHILI Lab, EPFL
pierre.dillenbourg@epfl.ch

Volkan Cevher
LIONS, EPFL
volkan.cevher@epfl.ch

Abstract

We consider the machine teaching problem in a classroom-like setting wherein the teacher has to deliver the same examples to a diverse group of students. Their diversity stems from differences in their initial internal states as well as their learning rates. We prove that a teacher with full knowledge about the learning dynamics of the students can teach a target concept to the entire classroom using $\mathcal{O}\left(\min\{d, N\}\log\frac{1}{\epsilon}\right)$ examples, where d is the ambient dimension of the problem, N is the number of learners, and ϵ is the accuracy parameter. We show the robustness of our teaching strategy when the teacher has limited knowledge of the learners' internal dynamics as provided by a noisy oracle. Further, we study the trade-off between the learners' workload and the teacher's cost in teaching the target concept. Our experiments validate our theoretical results and suggest that appropriately partitioning the classroom into homogenous groups provides a balance between these two objectives.

Introduction

Machine teaching considers the inverse problem of machine learning. Given a learning model and a target, the teacher aims to find an optimal set of training examples for the learner (Zhu et al. 2018; Liu et al. 2017). Machine teaching provides a rigorous formalism for various real-world applications such as personalized education and intelligent tutoring systems (Rafferty et al. 2016; Patil et al. 2014), imitation learning (Cakmak and Lopes 2012; Haug, Tschiatschek, and Singla 2018), program synthesis (Mayer, Hamza, and Kuncak 2017), adversarial machine learning (Mei and Zhu 2015), and human-in-the-loop systems (Singla et al. 2014; Singla et al. 2013).[1]

Individual teaching   Most of the research in this domain thus far, has focused on teaching a single student in the batch setting. Here, the teacher constructs an optimal training set (e.g., of minimum size) for a fixed learning model and a target concept and gives it to the student in a single interaction (Goldman and Kearns 1995; Zilles et al. 2011; Zhu 2013; Doliwa et al. 2014). Recently, there has been interest in studying the interactive setting (Liu et al. 2017; Zhu et al. 2018; Chen et al. 2018; Hunziker et al. 2018), wherein the teacher

focuses on finding an optimal sequence of examples to meet the needs of the student under consideration, which is, in fact, the natural expectation in a personalized teaching environment (Koedinger et al. 1997). (Liu et al. 2017) introduced the iterative machine teaching setting wherein the teacher has full knowledge of the internal state of the student at every time step using which she designs the subsequent optimal example. They show that such an "omniscient" teacher can help a single student approximately learn the target concept using $\mathcal{O}\left(\log\frac{1}{\epsilon}\right)$ training examples (where ϵ is the accuracy parameter) as compared to $\mathcal{O}\left(\frac{1}{\epsilon}\right)$ examples chosen randomly by the stochastic gradient descent (SGD) teacher.

Classroom teaching   In real-world classrooms, the teacher is restricted to providing the same examples to a large class of academically-diverse students. Customizing a teaching strategy for a specific student may not guarantee optimal performance of the entire class. Alternatively, teachers may constitute a partitioning of the students so as to maximize intra-group homogeneity while balancing the orchestration costs of managing parallel activities. (Zhu, Liu, and Lopes 2017) propose methods for explicitly constructing a minimal training set for teaching a class of batch learners based on a minimax teaching criterion. They also study optimal class partitioning based on prior distributions of the learners. However, they do not consider an interactive teaching setting.

Overview of our Approach

In this paper, we study the problem of designing optimal teaching examples for a classroom of iterative learners. We refer to this new paradigm as *iterative classroom teaching* (CT). We focus on online projected gradient descent learners under squared loss function. The learning dynamics comprise of the learning rates and the initial states which are different for different students. At each time step, the teacher constructs the next training example based on information regarding the students' learning dynamics. We focus on the following teaching objectives motivated by real-world classroom settings, where at the end of the teaching process:

(i) all learners in the class converge to the target model (*cf.* Eq. (1)),

(ii) the class on average converges to the target model (*cf.* Eq. (2)).

[1]http://teaching-machines.cc/nips2017/
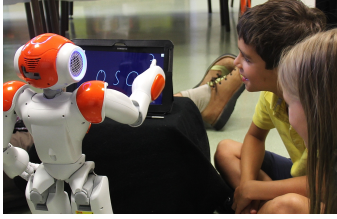
(a) A Nao robot writing on a digital tablet.
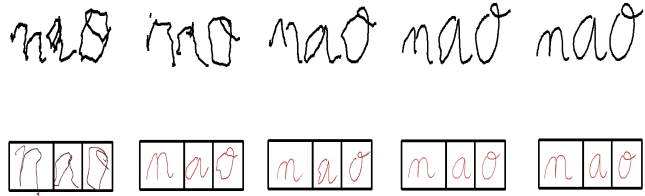
(b) Example of five interactions for writing the word "nao". The top row shows the writing of the robot, the bottom row shows the child's writing.

Figure 1: The robot writes iteratively adapted to the handwriting profile of the child; if the child's handwriting is shaky, the robot too writes with a shaky handwriting. In correcting the robot's handwriting, the child works towards remediating theirs.

Contributions We first consider that setting wherein at all times, the teacher has complete knowledge of the learning rates, loss function, and full observability of the internal states of all the students in the class. A naive approach here would be to apply (Liu et al. 2017)'s omniscient teaching strategy individually for each student in the class. This would require $\mathcal{O}\left(N \log \frac{1}{\epsilon}\right)$ teaching examples, where N is the number of students in the classroom. We present a teaching strategy that can achieve a convergence rate of $\mathcal{O}\left(k \log \frac{1}{\epsilon}\right)$, where k is the rank of the subspace in which the students of the classroom lie (i.e. $k \leq \min \{d, N\}$, where d is the ambient dimension of the problem). We also prove the robustness of our algorithm in noisy and incomplete information settings. We then explore the idea of partitioning the classroom into smaller groups of homogenous students based on either their learning ability or prior knowledge. We also validate our theoretical results on a simulated classroom of learners and demonstrate their practical applications to the task of teaching how to classify between butterflies and moths. Further, we show the applicability of our teaching strategy to the task of teaching children how to write (*cf.* Figure 1).

The Model

In this section, we consider a stylized model to derive a solid understanding for the dynamics of the learners. This simplicity of our model will then allow us to gain insights into classroom partitioning (i.e., how to create classrooms), and then explore the key trade-offs between the learners' workload as well as the teacher's orchestration costs. By orchestration costs, we mean the number of examples the teacher needs to teach the class.

Notation Define $\{a_i\}_{i=1}^N := \{a_1, \ldots, a_N\}$ as a set of N elements and $[N] := \{1, \ldots, N\}$ as the index set. For a given matrix A, denote $\lambda_i(A)$ and $e_i(A)$ to be the i-th largest eigenvalue of A and the corresponding eigenvector respectively. $\|\cdot\|$ denotes the Euclidean norm unless otherwise specified. The projection operation on a set $\mathcal{W}$ for any element y is defined as follows:

$$\texttt{Proj}_{\mathcal{W}}(y) := \arg\min_{x \in \mathcal{W}} \|x - y\|^2$$

Parameters In *synthesis-based teaching* (Liu et al. 2017), $\mathcal{X} = \{x \in \mathbb{R}^d, \|x\| \leq D_{\mathcal{X}}\}$ represents the feature space and the label set is given by $\mathcal{Y} = \mathbb{R}$ (for regression) or $\{1, 2, \ldots, m\}$ (for classification). A training example is denoted by $(x, y) \in \mathcal{X} \times \mathcal{Y}$. Further, we define the feasible hypothesis space by $\mathcal{W} = \{w \in \mathbb{R}^d, \|w\| \leq D_{\mathcal{W}}\}$, and denote the target hypothesis by w^*.

Classroom The classroom consists of N students. Each student $j \in [N]$ has two internal parameters: i) the learning rate (at time t) represented by η_j^t, and ii) the initial internal state given by $w_j^0 \in \mathcal{W}$. At each time step t, the classroom receives a labelled training example $(x^t, y^t) \in \mathcal{X} \times \mathcal{Y}$ and each student j performs a projected gradient descent step as follows:

$$w_j^{t+1} = \texttt{Proj}_{\mathcal{W}}\left(w_j^t - \eta_j^t \frac{\partial \ell\left(\pi(w_j^t, x^t), y^t\right)}{\partial w_j^t}\right),$$

where ℓ is the loss function and $\pi(w_j^t, x^t)$ is the student's label for example x^t. We restrict our analysis to the linear regression case where $\pi(w_j^t, x^t) = \langle w_j^t, x^t \rangle$ and $\ell\left(\langle w_j^t, x^t \rangle, y^t\right) = \frac{1}{2}\left(\langle w_j^t, x^t \rangle - y^t\right)^2$.

Teacher The teacher, over a series of iterations, interacts with the students in the classroom and guides them towards the target hypothesis by choosing "helpful" training examples. The choice of the training example depends on how much information she has about the students' learning dynamics.

- *Observability*: This represents the information that the teacher possesses about the internal state of each student. We study two cases: i) when the teacher knows the exact value $\{w_j^t\}_{j=1}^N$, and ii) when the teacher has a noisy estimate denoted by $\{\tilde{w}_j^t\}_{j=1}^N$ at any time t.

- *Knowledge*: This represents the information that the teacher has regarding the learning rates of each student. We consider two cases: i) when the learning rate of each student is constant and known to the teacher, and ii) when each student draws a value for the learning rate from a normal distribution at every time step, while the teacher only has access to the past values.

Teaching objective In the abstract machine teaching setting, the objective corresponds to approximately training a predictor. Given an accuracy value ϵ as input, at time T we say that a student $j \in [N]$ has approximately learnt the target concept w^* when $\left\| w_j^T - w^* \right\| \leq \epsilon$. In the strict sense, the teacher's goal may be to ensure that every student in the classroom converges to the target as quickly as possible, i.e.,

$$\left\| w_j^T - w^* \right\| \leq \epsilon, \forall j \in [N]. \tag{1}$$

The teacher's goal in the average case however, is to ensure that the classroom as a whole converges to w^* in a minimum number of interactions. More formally, the aim is to find the smallest value T such that the following condition holds:

$$\frac{1}{N} \sum_{i=1}^{N} \left\| w_j^T - w^* \right\|^2 \leq \epsilon. \tag{2}$$

Classroom Teaching

We study the omniscient and synthesis-based teacher, equivalent to the one considered in (Liu et al. 2017), but for the iterative classroom teaching problem under the squared loss given by $\ell\left(\langle w, x \rangle, y\right) := \frac{1}{2}\left(\langle w, x \rangle - y\right)^2$. Here the teacher has full knowledge of the target concept w^*, learning rates $\{\eta_j\}_{j=1}^N$ (assumed constant), and internal states $\{w_j^t\}_{j=1}^N$ of all the students in the classroom.

Teaching protocol At every time step t, the teacher uses all the information she has to choose a training example $x^t \in \mathcal{X}$ and the corresponding label $y^t = \langle w^*, x^t \rangle \in \mathcal{Y}$ (for linear regression). The idea is to pick the example which minimizes the average distance between the students' internal states and the target hypothesis at every time step. Formally,

$$x^t = \underset{x \in \mathcal{X}}{\arg\min} \frac{1}{N} \sum_{j=1}^{N} \left\| w_j^t - \eta_j \frac{\partial \ell\left(\langle w_j^t, x \rangle, \langle w^*, x \rangle\right)}{\partial w_j^t} - w^* \right\|^2.$$

Note that constructing the optimal example at time t, is in general a non-convex optimization problem. For the squared loss function, we present a closed-form solution below.

Example construction Define, $\hat{w}_j^t := w_j^t - w^*$, for $j \in [N]$. Then the teacher constructs the common example for the whole classroom at time t as follows:

1. the feature vector $x^t = \gamma_t \hat{x}^t$ such that

   (a) the magnitude γ_t:

   $$\gamma_t \leq D_{\mathcal{X}}, \text{ and } 2 - \eta_j \gamma_t^2 \geq 0, \forall j \in [N]. \tag{3}$$

   (b) the direction $\hat{x}^t$ (with $\|\hat{x}^t\| = 1$):

   $$\hat{x}^t := \underset{x:\|x\|=1}{\arg\max} \, x^\top W^t x = e_1\left(W^t\right), \tag{4}$$

   where

   $$\alpha_j^t := \eta_j \gamma_t^2 \left(2 - \eta_j \gamma_t^2\right) \tag{5}$$

   $$W^t := \frac{1}{N} \sum_{j=1}^{N} \alpha_j^t \hat{w}_j^t \left(\hat{w}_j^t\right)^\top. \tag{6}$$

2. the label $y^t = \langle w^*, x^t \rangle$.

Algorithm 1 CT: Classroom teaching algorithm

Input: target $w^* \in \mathcal{W}$; students' learning rates $\{\eta_j\}_{j=1}^N$
Goal: accuracy ϵ
Initialize $t = 0$
Observe $\{w_j^0\}_{j=1}^N$
while $\frac{1}{N} \sum_{j=1}^{N} \left\| w_j^t - w^* \right\|^2 > \epsilon$ **do**
  Observe $\{w_j^t\}_{j=1}^N$
  Choose γ_t s.t. $\gamma_t \leq D_{\mathcal{X}}$, and $2 - \eta_j \gamma_t^2 \geq 0, \forall j \in [N]$
  Construct W^t given by Eq. (6)
  Pick example $x^t = \gamma_t \cdot e_1\left(W^t\right)$ and $y^t = \langle w^*, x^t \rangle$
  Provide the labeled example (x^t, y^t) to the classroom
  Students' update: $\forall j \in [N]$,

$$w_j^{t+1} \leftarrow \texttt{Proj}_{\mathcal{W}}\left(w_j^t - \eta_j\left(\langle w_j^t, x^t \rangle - y^t\right) x^t\right).$$

  $t \leftarrow t + 1$
end while

Algorithm 1 puts together our omniscient classroom teaching strategy. Theorem 1 provides the number of examples required to teach the target concept.[2]

Theorem 1. *Consider the teaching strategy given in Algorithm 1. Let $k := \max_t \{\text{rank}\left(W^t\right)\}$, where W^t is given by Eq. (6). Define $\alpha_j := \min_t \alpha_j^t$, $\alpha_{\min} := \min_{t,j} \alpha_j^t$, and $\alpha_{\max} := \max_{t,j} \alpha_j^t$, where α_j^t is given by Eq. (5). Then after $t = \mathcal{O}\left(\left(\log \frac{1}{1 - \frac{\alpha_{\min}}{k}}\right)^{-1} \log \frac{1}{\epsilon}\right)$ rounds, we have $\frac{1}{N} \sum_{j=1}^{N} \left\| w_j^t - w^* \right\|^2 \leq \epsilon$. Furthermore, after $t = \mathcal{O}\left(\max\left\{\left(\log \frac{1}{1 - \alpha_j}\right)^{-1} \log \frac{1}{\epsilon}, \left(\log \frac{1}{1 - \frac{\alpha_{\min}}{k}}\right)^{-1} \log \frac{1}{\epsilon}\right\}\right)$ rounds, we have $\left\| w_j^t - w^* \right\|^2 \leq \epsilon, \forall j \in [N]$.*

Remark 1. *By using the fact that $\exp(-2x) \leq \log(1 - x)$, for $2x \leq 1.59$, we can show that for $k \geq \frac{2\alpha_{\min}}{1.59}$,*

$$\mathcal{O}\left(\left(\log \frac{1}{1 - \frac{\alpha_{\min}}{k}}\right)^{-1} \log \frac{1}{\epsilon}\right) \approx \mathcal{O}\left(\frac{k}{\alpha_{\min}} \log \frac{1}{\epsilon}\right).$$

Based on Theorem 1 and Remark 1, note that the teaching strategy given in Algorithm 1 converges in $\mathcal{O}\left(k \log \frac{1}{\epsilon}\right)$ samples, where $k = \max_t \{\text{rank}\left(W^t\right)\} \leq \min\{d, N\}$. This is in fact a significant improvement (especially when $N \gg d$) over the sample complexity $\mathcal{O}\left(N \log \frac{1}{\epsilon}\right)$ of a teaching strategy which constructs personalized training examples for each student in the classroom.

Choice of magnitude γ_t We consider the following two choices of γ_t:

1. static $\gamma_t = \min\left\{\frac{1}{\max_{j \in [N]} \sqrt{\eta_j}}, D_{\mathcal{X}}\right\}$: This ensures that the classroom can converge without being partitioned into

---

[2]Proofs are given in the extended version of this paper (Yeo et al. 2018).

small groups of students. However, the value $\alpha_{\min}$ becomes small and as a result, the sample complexity increases.

2. dynamic $\gamma_t = \min\left\{\sqrt{\frac{\sum_{j=1}^{N}\eta_j\left\|w_j^t-w^*\right\|^2}{\sum_{j=1}^{N}\eta_j^2\left\|w_j^t-w^*\right\|^2}},D_{\mathcal{X}}\right\}$: This provides an optimal constant for the sample complexity, but requires that for effective teaching the classroom is partitioned appropriately. This value of γ_t is obtained by maximizing the term $\sum_{j=1}^{N}\eta_j\gamma_t^2\left(2-\eta_j\gamma_t^2\right)\left\|w_j^t-w^*\right\|^2$.

Natural partitioning based on learning rates In order to satisfy the requirements given in Eq. (3), for every student $j \in [N]$, we require (for dynamic γ_t):

$$
\begin{aligned}
\eta_j\gamma_t^2 &\leq \eta_{\max}\frac{\sum_{j=1}^{N}\eta_j\left\|w_j^t-w^*\right\|^2}{\sum_{j=1}^{N}\eta_j^2\left\|w_j^t-w^*\right\|^2}\\
&\leq \eta_{\max}\frac{\sum_{j=1}^{N}\eta_j\left\|w_j^t-w^*\right\|^2}{\eta_{\min}\sum_{j=1}^{N}\eta_j\left\|w_j^t-w^*\right\|^2} \leq 2, \quad (7)
\end{aligned}
$$

where $\eta_{\max} = \max_j \eta_j$, and $\eta_{\min} = \min_j \eta_j$. That is, if $\eta_{\max} \leq 2\eta_{\min}$, we can safely use the above optimal γ_t. This observation also suggests a natural partitioning of the classroom: $\{[\eta_{\min}, 2\eta_{\min}), [2\eta_{\min}, 4\eta_{\min}), \ldots, [2^m\eta_{\min}, 2\eta_{\max})\}$, where $m = \left\lfloor\log_2\frac{\eta_{\max}}{\eta_{\min}}\right\rfloor$.

Robust Classroom Teaching

In this section, we study the robustness of our teaching strategy in cases when the teacher can access the current state of the classroom only through a noisy oracle, or when the learning rates of the students vary with time.

Noise in w_j^t's

Here, we consider the setting where the teacher cannot directly observe the students' internal states $\left\{w_j^t\right\}_{j=1}^{N},\forall t$ but has full knowledge of students' learning rates $\left\{\eta_j\right\}_{j=1}^{N}$. Define $\alpha_{\min} := \min_{t,j}\alpha_j^t$, and $\alpha_{\text{avg}} := \max_t\frac{1}{N}\sum_{j=1}^{N}\alpha_j^t$, where α_j^t is given by Eq. (5). At every time step t, the teacher only observes a noisy estimate of w_j^t (for each $j \in [N]$) given by

$$
\tilde{w}_j^t := w_j^t + \delta_j^t, \quad (8)
$$

where δ_j^t is a random noise vector such that $\left\|\delta_j^t\right\| \leq \frac{\epsilon}{4\left(\frac{\alpha_{\text{avg}}}{\alpha_{\min}}d+1\right)D_{\mathcal{W}}}$. Then the teacher constructs the example as follows:

$$
\begin{aligned}
\hat{x}^t &:= \underset{x:\|x\|=1}{\arg\max}\, x^\top\left\{\frac{1}{N}\sum_{j=1}^{N}\alpha_j^t\hat{w}_j^t\left(\hat{w}_j^t\right)^\top\right\}x\\
x^t &:= \gamma_t\hat{x}^t \text{ and } y^t = \langle w^*, x^t\rangle, \quad (9)
\end{aligned}
$$

where $\hat{w}_j^t := \tilde{w}_j^t - w^*$, and γ_t satisfies the condition given in Eq. (3). The following theorem shows that even under this noisy observation setting Eq. (8), with the example construction strategy described in Eq. (9), the teacher can teach the classroom with linear convergence.

Theorem 2. *Consider the noisy observation setting given by Eq. (8). Let* $k := \max_t\{\text{rank}(W^t)\}$ *where* $W^t = \frac{1}{N}\sum_{j=1}^{N}\alpha_j^t\left(\tilde{w}_j^t-w^*\right)\left(\tilde{w}_j^t-w^*\right)^\top$. *Then for the robust teaching strategy given by Eq. (9), after* $t = \mathcal{O}\left(\left(\log\frac{1}{1-\frac{\alpha_{\min}}{k}}\right)^{-1}\log\frac{1}{\epsilon}\right)$ *rounds, we have* $\frac{1}{N}\sum_{j=1}^{N}\left\|w_j^t-w^*\right\|^2 \leq \epsilon$.

Noise in η_j^t

Here, we consider a classroom of online projected gradient descent learners with learning rates $\left\{\eta_j^t\right\}_{j=1}^{N}$, where $\eta_j^t \sim \mathcal{N}(\eta_j, \sigma)$. We assume that the teacher knows σ (which is constant across all the students) and $\{\eta_j\}_{j=1}^{N}$, but doesn't know $\left\{\eta_j^t\right\}_{j=1}^{N}$. Further, we assume that the teacher has full observability of $\left\{w_j^t\right\}_{j=1}^{N}$. At time t, the teacher has access to the history $H_t := \left(\left\{w_j^s\right\}_{s=1}^{t}, \left\{\eta_j^s\right\}_{s=1}^{t-1} : \forall j \in [N]\right)$. Then the teacher constructs the example as follows (depending only on H_t):

$$
\begin{aligned}
\hat{x}^t &:= \underset{x:\|x\|=1}{\arg\max}\, x^\top\bar{W}^t x = e_1\left(\bar{W}^t\right)\\
x^t &:= \gamma_t\hat{x}^t \text{ and } y^t = \langle w^*, x^t\rangle, \quad (10)
\end{aligned}
$$

where

$$
\gamma_t^2 \leq \frac{2\eta_j}{\sigma^2+\eta_j^2}, \forall j \in [N] \quad (11)
$$

$$
\bar{\eta}_j^t := \frac{1}{t-1}\sum_{s=1}^{t-1}\eta_j^s \quad (12)
$$

$$
\bar{\alpha}_j^t := 2\gamma_t^2\bar{\eta}_j^t - \gamma_t^4\left(\frac{t-2}{t-1}\sigma^2+\left(\bar{\eta}_j^t\right)^2\right) \quad (13)
$$

$$
\hat{w}_j^t := w_j^t - w^* \text{ and} \quad (14)
$$

$$
\bar{W}^t := \frac{1}{N}\sum_{j=1}^{N}\bar{\alpha}_j^t\hat{w}_j^t\left(\hat{w}_j^t\right)^\top. \quad (15)
$$

The following theorem shows that, in this setting, the teacher can teach the classroom in expectation with linear convergence.

Theorem 3. *Let* $k := \max_t\left\{\text{rank}\left(\bar{W}^t\right)\right\}$ *where* $\bar{W}^t$ *is given by Eq.* (15). *Define* $\bar{\alpha}_{\min} := \min_{t,j}\bar{\alpha}_j^t$, *and* $\beta_{\min} := \min_{j,t}\frac{\alpha_j^t}{\bar{\alpha}_j^t}$, *where* $\alpha_j^t := 2\gamma_t^2\eta_j - \gamma_t^4\left(\sigma^2+\eta_j^2\right)$ *and* $\bar{\alpha}_j^t$ *given by Eq.* (13). *Then for the teaching strategy given by Eq.* (10), *after* $t = \mathcal{O}\left(\left(\log\frac{1}{1-\frac{\beta_{\min}\bar{\alpha}_{\min}}{k}}\right)^{-1}\log\frac{1}{\epsilon}\right)$ *rounds, we have* $\mathbb{E}\left[\frac{1}{N}\sum_{j=1}^{N}\left\|w_j^t-w^*\right\|^2\right] \leq \epsilon$.

Classroom Partitioning

Individual teaching can be very expensive due to the effort required in producing personalized education resources. At

(a) Individual teaching (IT)   (b) Classroom teaching (CT)   (c) Classroom teaching with partitions (CTwP)
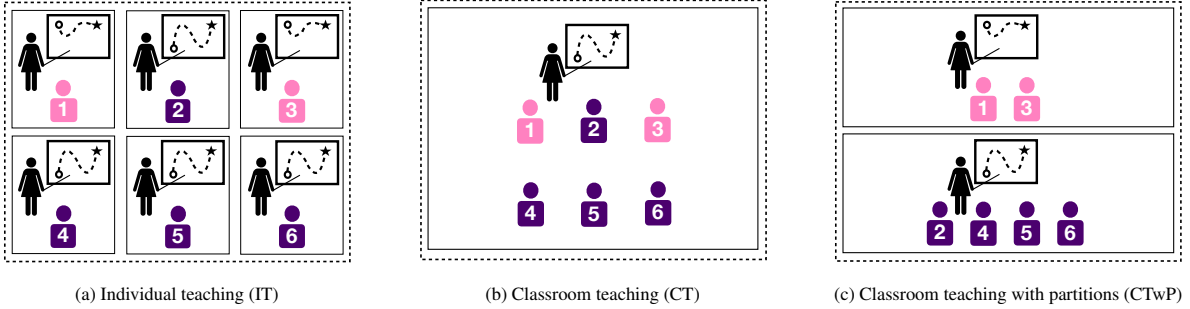
Figure 2: Comparisons between individual teaching (IT) and classroom teaching (CT and CTwP) paradigms.

the same time, classroom teaching increases the students' workload substantially because it requires catering to the needs of academically diverse learners. We overcome these pitfalls by partitioning the given classroom of N students into K groups such that the orchestration cost of the teacher and the workload of students is balanced. Figure 2 illustrates these three different teaching paradigms.

Let $T(K)$ be the total number of examples required by the teacher to teach all the groups. Let $S(K)$ be the average number of examples needed by a student to converge to the target. We study the total cost defined as:

$$\text{cost}(K) := T(K) + \lambda \cdot S(K),$$

where λ quantifies the trade-off factor, and its value is application dependent. In particular, for any given λ, we are interested in that value K that minimizes $\text{cost}(K)$. For example, when $\lambda = \infty$, the focus is on the student workload; thus the optimal teaching strategy is individual teaching, *i.e.*, $K = N$. Likewise, when $\lambda = 0$, the focus is on the orchestration cost; thus the optimal teaching strategy is classroom teaching without partitioning, i.e., $K = 1$. In this paper, we explore two homogeneous partitioning strategies: (a) based on learning rates of the students $\{\eta_j\}_{j=1}^N$, (b) based on prior knowledge of the students $\{w_j^0\}_{j=1}^N$.

Experiments

Teaching Linear Models with Synthetic Data

We first examine the performance of our teaching algorithms on simulated learners.

Setup We evaluate the following algorithms: (i) classroom teaching (CT) - the teacher gives an example to the entire class at each iteration, (ii) CT with optimal partitioning (CTwP-Opt) - the class is partitioned as defined in Section , (iii) CT with random partitioning (CTwP-Rand) - the class is randomly assigned to groups, and (iv) individual teaching (IT) - the teacher gives a tailored example to each student. An algorithm is said to converge when $\frac{1}{N}\sum_i \|w_i^t - w^*\|_2^2 \leq \epsilon$. We set the number of learners $N = 300$ and accuracy parameter $\epsilon = 0.1$.

Average error and robustness of CT We first consider the noise free classroom setting with $d = 25$, learning rates between $[0.05, 0.25]$, and $D_{\mathcal{X}} = 2$. The plot of the error over time is shown in Figure 3a, together with the performance

of four selected learners. Our algorithm exhibits linear convergence, as per Theorem 1. The slower the learners and the further away they are from w^*, the longer they take to converge. Figure 3d shows how convergence is affected as the noise level, δ, increases in the robust classroom teaching setting as described in Section . Although the number of iterations required for convergence increases, it is still significantly lower than the noise-free IT.

Convergence for classroom with diverse η We study the effect of partitioning by η on the performance of the algorithms described. The diversity of the classroom varies from 0 (where all learners in the classroom have $\eta = 0.1$) to 0.5 (where for all learners $\eta \in [0.1, 0.6]$ chosen randomly), and so on. Figure 3b and Figure 3c depict the number of iterations and number of examples needed by the teacher and students respectively to achieve convergence. As expected, IT performs best, and CTwP-Opt consistently outperforms CT. For a class with low diversity, partitioning is costly. However as diversity increases, partitioning is beneficial from the teachers' perspective. Note that the dip at a diversity of 0.15 for both plots is due to the value of $D_{\mathcal{X}}$. For the static γ_t, with $D_{\mathcal{X}} = 2$, all learners with rates less than 0.25 will be negatively affected. As the minimum value of η is 0.1, at zero diversity, all learners are affected the most. As diversity increases to 0.15, all learners are affected but to a lesser degree. Figure 3e shows how the optimal algorithm, the one that minimizes cost, changes with λ and diversity of η. When diversity is low and there is a low trade-off factor on the students' workload, CT performs best. At high values, IT has the lowest cost. CTwP-Opt falls between these two regimes.

Convergence for classroom with diverse w^0 Next, we study partitioning based on prior knowledge. We generate each cluster from a Gaussian distribution centered on a point along different axes. At diversity 1, all 300 learners are centered on a point on one axis, whereas at diversity 2, 150 learners are centered on one axis and the other 150 on another. Thus at 10, we have 30 learners around a point at each of the 10 axes. Each cluster represents one partition. Although the convergence plots from the teacher and students' perspective are not presented, they exhibit the same behaviour as partitioning by η. Figure 3f shows the cost trade off plot in 10 dimensions as the number of clusters of w^0 increase. The results are the same as with η partitioning and CTwP-Opt outperforms in most regimes.

(a) Error plot for CT; average error of the classroom and the error of four selected learners.

(b) Total iterations needed for convergence from the teacher's perspective.

(c) Total iterations per student needed for convergence from the students' perspective.

(d) Total iterations needed for convergence in noisy w_t case as the noise, δ, increases.

(e) λ: Trade-off between teacher's and students' cost with increasing η diversity.

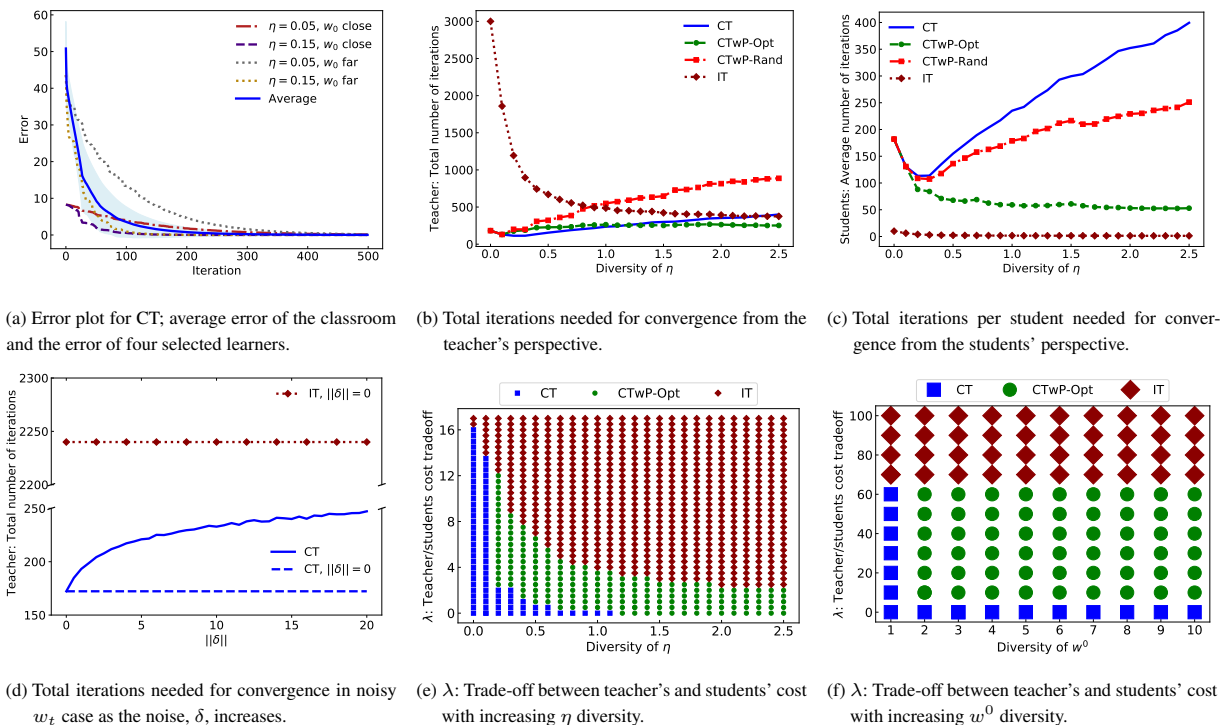(f) λ: Trade-off between teacher's and students' cost with increasing w^0 diversity.

Figure 3: (3a) and (3d) show the convergence results for the noise-free and noisy settings. CT is robust and exhibits linear convergence. (3b), (3c) and (3e) show the convergence results and trade-off for a classroom with diverse η. (3f) shows the trade-off for a classroom with diverse w^0.

Teaching How to Classify Butterflies and Moths

We now demonstrate the performance of our teaching algorithms on a binary image classification task for identifying insect species, a prototypical task in crowdsourcing applications and an important component in citizen science projects such as eBird (Sullivan et al. 2009).

Images and Euclidean embedding We use a collection of 160 images (40 each) of four species of insects, namely (a) Caterpillar Moth (cmoth), (b) Tiger Moth (tmoth), (c) Ringlet Butterfly (rbfly), and (d) Peacock Butterfly (pbfly), to form the teaching set $\mathcal{X}$. Given an image, the task is to classify if it is a butterfly or a moth. However, we need a Euclidean embedding of these images so that they can be used by a teaching algorithm. Based on the data collected by (Singla et al. 2014), we obtained binary labels (whether a given image is a butterfly or not) for $\mathcal{X}$ from a set of 67 workers from Amazon Mechanical Turk. Using this annotation data, the Bayesian inference algorithm of (Welinder et al. 2010) allows us to obtain an embedding, shown in Figure 4a, along with the target w^* (the best fitted linear hypothesis for $\mathcal{X}$).

Learners' hypotheses The process described above to obtain the embedding in Figure 4a simultaneously generates an embedding of each of the 67 annotators as linear hypotheses in the same 2D space. Termed as "schools of thought" by (Welinder et al. 2010), these hypotheses capture various real-world idiosyncrasies in the AMT workers' annotation behavior. For our experiments, we identified four types of learners' hypotheses; those who (i) *misclassify tmoth as but-*

terfly (P1), (ii) *misclassify rbfly as moth (P2)*, (iii) *misclassify pbfly as moth (P3)* and (iv) *misclassify tmoth and cmoth as butterflies*. Figure 4b shows an embedding of three distinct hypotheses each of the four types of learners.

Creating the classroom We denote the hypotheses described above as initial states w^0 of the learners/students. Due to sparsity of data, we create a supersample of size 60 for the four types of learners by adding a small noise. We set the classroom size $N = 60$. The diversity of the class, defined by the number of different types of learners present, varies from 1 to 4. Thus, diversity of 1 refers to the case when all 60 learners are of same type (randomly picked from P1, P2, P3, or P4), and diversity of 4 refers to the case when there are 15 learners of each type. We set a constant learning rate of $\eta = 0.05$ for all students.

Teaching and performance metrics We study the performance of CT, CTwP-Rand, and IT teachers. We also examine the CTwP-Opt teacher that partitions the learners of the class based on their types. All teachers are assumed to have complete information about the learners at all times. We set accuracy parameter $\epsilon = 0.2$ and the classroom is said to have converged when $\frac{1}{N} \sum_i \|w_i^t - w^*\|_2^2 \leq \epsilon$.

Teaching examples Figure 4c consists of 5 rows of 20 thumbnail images each, depicting the training examples chosen in an actual run of the experiment when the diversity of the classroom is 4. The first row corresponds to the images chosen by CT. For instance, in iteration 1, CT chooses a tmoth example. While this example is most helpful for learners in

(a) Dataset of images $\mathcal{X}$ and target w^*

(b) Initial w^0 of 4 types of learners

(c) Teaching examples, visualized twice every 10 iterations.

(d) Total iterations needed for convergence from the teacher's perspective

(e) Total iterations needed for convergence from the student's perspective

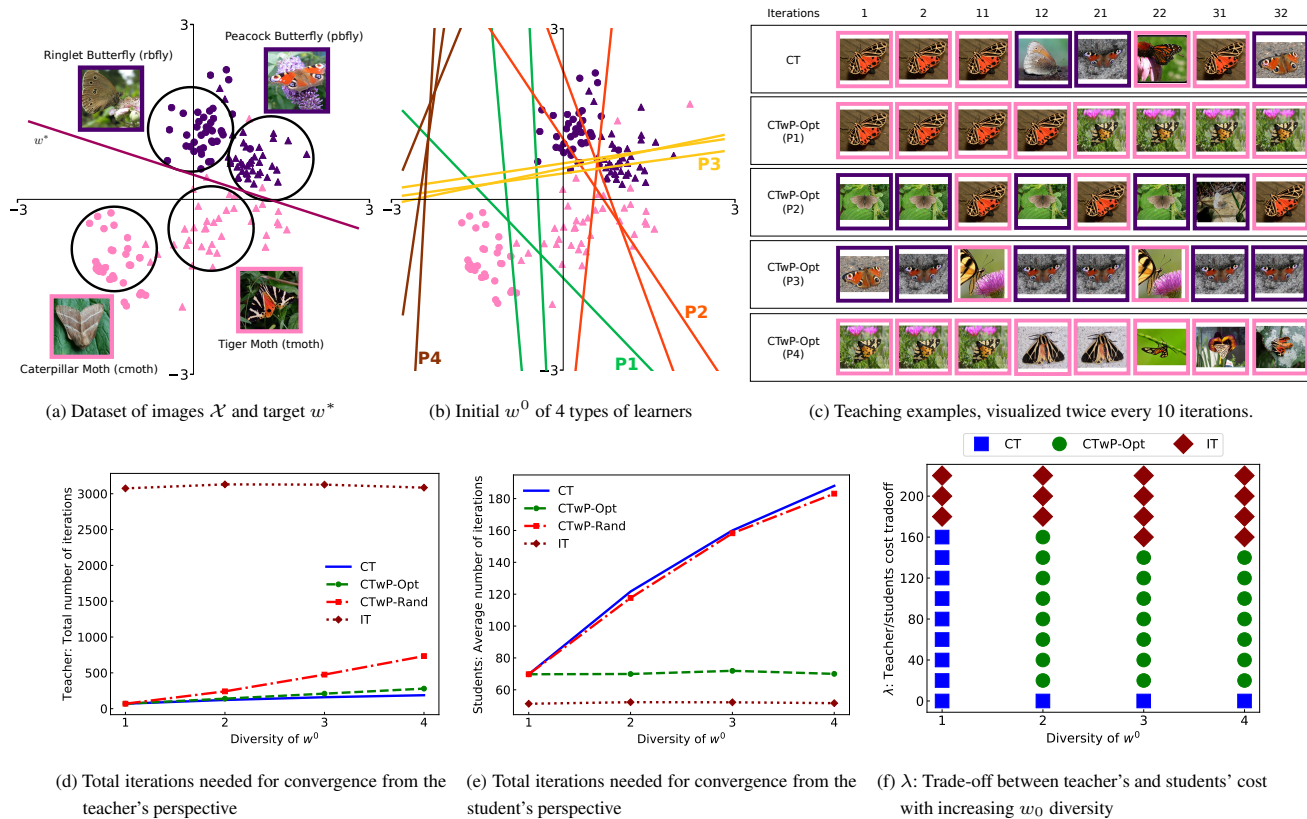(f) λ: Trade-off between teacher's and students' cost with increasing w_0 diversity

Figure 4: (4a) shows a low-dimensional embedding of the dataset and the target concept. (4b) shows an embedding of the initial states of three learners of each of the 4 types. (4c) are training examples selected by CT and CTwP-Opt teachers when the class has diversity 4. (4d) and (4e) show the number of iterations required to achieve ϵ-convergence from the teacher and student perspectives. (4f) shows how the optimal algorithm changes as we vary the trade-off parameter λ, and diversity of the class.

P1 (confusing tmoths as butterflies), however, learners in P2 and P3 would have benefited more from seeing examples of butterflies. This increases the workload for the learners. The next four rows in Figure 4c correspond to the images chosen by CTwP-Opt when teaching partitions P1, P2, P3, and P4 respectively—these thumbnails show the personalized effect given the homogeneity of these partitions. For instance, for P1, the CTwP-Opt focuses on choosing tmoth examples thereby allowing these learners to converge faster while ensuring that the cost for learners in other partitions does not increase.

Convergence Figure 4d compares the performances of the teachers in terms of the total number of iterations required for convergence. CT performs optimally because every example chosen is provided to the entire class; CTwP-Opt requires only a few examples more, given the homogeneity of the partition and the partitions being of equal size. IT constructs individually tailored examples for each learner in the class. Thus the combined number of iterations is much higher in comparison.

Teacher/students cost trade-off On the other hand, Figure 4e depicts the average number of examples required by each learner to achieve convergence as a function of diversity. This represents the learning cost from the students' persective. IT

performs best because the teacher chooses personalized examples for each learner. CTwP-Opt performs considerably better than CT. This happens because partitioning groups together learners of the same type. Figure 4f represents optimal algorithm given the diversity of the class and the trade-off factor λ as defined in Section . As diversity increases, CTwP-Opt outperforms the other teachers in terms of the total cost.

Teaching How to Write

Despite formal training, between 5% to 25% of children struggle to acquire handwriting skills. Being unable to write legibly and rapidly limits a child's ability to simultaneously handle other tasks such as grammar and composition which may lead to general learning difficulties (Feder and Majnemer 2007; Christensen 2009). (Johal et al. 2016) and (Chase et al. 2009) adopt an approach where the child plays the role of the "teacher" and an agent a "learner" that needs help. This method of learning by teaching boosts a child's self esteem and increases their commitment to the task as they are given the role of the one who "knows and teaches" (Rohrbeck et al. 2003; Chase et al. 2009). In our experiments, a robot iteratively proposes a handwriting adapted to the handwriting profile of the child, that they try to correct (*cf.* Figure 1b). We now demonstrate the performance of our algorithm in choosing this sequence of examples.
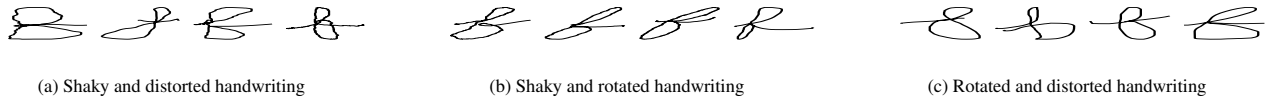
(a) Shaky and distorted handwriting   (b) Shaky and rotated handwriting   (c) Rotated and distorted handwriting

Figure 5: (5a) to (5c) shows samples of children's handwriting where two of the three defined features are poor and the third is good.

(a) Teaching examples for shaky and rotated handwriting   (b) Teaching examples for distorted and shaky handwriting   (c) Teaching examples for distorted and rotated handwriting   (d) Teaching examples for distorted, rotated, shaky handwriting
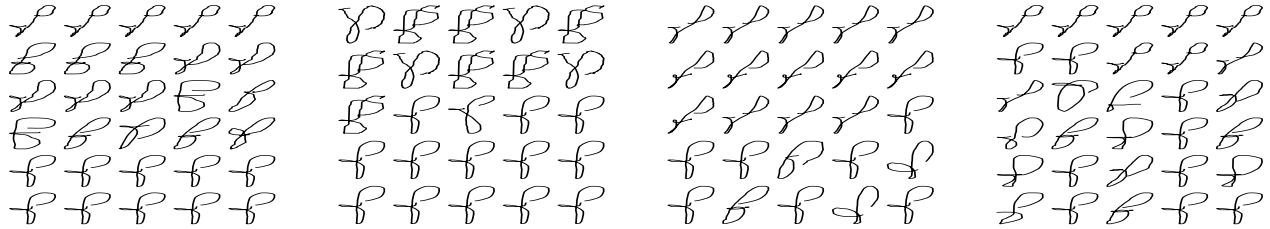
Figure 6: (6a) to (6d) shows the sequence of examples, visualized every other iteration, chosen by our algorithm for different initial hypothesis of the children.

Generating handwriting dynamics A LSTM is used to learn and generate handwriting dynamics (Graves 2013). It is trained on children's handwriting data collected from 1014 children from 14 schools.[3] In the extended version of this paper (Yeo et al. 2018), we showed that the pool of samples has to be rich enough for teaching to be effective.[4] As our generative model outputs a distribution, we can sample from it to get a diverse set of teaching examples. We analyze our results for a cursive "f", similar results apply for the other letters.

Handwriting features Concise Evaluation Scale (BHK) (Hamstra-Bletz, DeBie, and Den Brinker 1987) is a standard handwriting test used in Europe to evaluate children's handwriting quality. We adopt features such as (i) distortion, (ii) rotation, (iii) shakiness, and label each generated sample with a score for each of these features.

Creating the classroom Given a child's handwriting sample, we estimate their initial hypothesis, w_0 by how well each of the above features have been written, in a similar fashion to the scoring of samples. As most children fair poorly in two out of the three features, we selected and partitioned them according to the following three types of handwriting characteristics, substantial (i) shakiness and rotation, (ii) distortion and shakiness, and (iii) distortion and rotation. Original samples of each are shown in Figures 5a to 5c. We set a constant learning rate for all learners.

Teaching examples Figures 6a to 6c shows the training examples chosen by CTwP-Opt and Figure 6d by our CT algorithm. Each of the synthesized handwriting samples are labelled as good or bad, based on the average of their nor-

malized scores. We then run a classification algorithm on each partition and the entire class. This returns a sequence of examples that the robot would propose, for the children to correct. For children with handwriting that is shaky and rotated but not distorted, the sequence of examples chosen by our algorithm shows examples that are not distorted but progressively smoother and upright. Similarly, for children with handwriting that is distorted and shaky, the sequence of examples shown is upright with decreasing distortion and shakiness. We did not show the convergence plots as they have similar characteristics as those from the previous experiments.

Conclusion

We studied the problem of constructing an optimal teaching sequence for a classroom of online gradient descent learners. In general, this problem is non-convex, but for the squared loss, we presented and analyzed a teaching strategy with linear convergence. We achieved a sample complexity of $\mathcal{O}\left(\min\{d, N\}\log\frac{1}{\epsilon}\right)$, which is a significant improvement over $\mathcal{O}\left(N\log\frac{1}{\epsilon}\right)$ samples as required by the individual teaching strategy. We also showed that a homogeneous grouping of learners allows us to achieve a good trade-off between the learners' workload and the teacher's orchestration cost. Further, we compared the individual teaching (IT), classroom teaching (CT), and classroom teaching with partitioning (CTwP): we showed that a homogeneous grouping of learners (based on learning ability or prior knowledge) allows us to achieve a good trade-off between the learners' workload and the teacher's orchestration cost. The sequence of examples returned by our experiments are interpretable and they clearly demonstrate a significant potential in automation for robotics.

---

[3]The model has 3 layers, 300 hidden units and outputs a 20 component bivariate Gaussian mixture and a Bernoulli variable indicating the end of the letter. Each child was asked to write, in cursive, the 26 letters of the alphabet and the 10 digits on a tablet.

[4]The attained result is for the squared loss function, however, the analysis holds for other loss function.

References

[Cakmak and Lopes 2012] Cakmak, M., and Lopes, M. 2012. Algorithmic and human teaching of sequential decision tasks. In *AAAI*.

[Chase et al. 2009] Chase, C. C.; Chin, D. B.; Oppezzo, M. A.; and Schwartz, D. L. 2009. Teachable agents and the protégé effect: Increasing the effort towards learning. *Journal of Science Education and Technology* 18(4):334–352.

[Chen et al. 2018] Chen, Y.; Singla, A.; Mac Aodha, O.; Perona, P.; and Yue, Y. 2018. Understanding the role of adaptivity in machine teaching: The case of version space learners. In *NIPS*.

[Christensen 2009] Christensen, C. A. 2009. The critical role handwriting plays in the ability to produce high-quality written text. *The SAGE handbook of writing development* 284–299.

[Doliwa et al. 2014] Doliwa, T.; Fan, G.; Simon, H. U.; and Zilles, S. 2014. Recursive teaching dimension, vc-dimension and sample compression. *Journal of Machine Learning Research* 15(1):3107–3131.

[Feder and Majnemer 2007] Feder, K. P., and Majnemer, A. 2007. Handwriting development, competency, and intervention. *Developmental Medicine & Child Neurology* 49(4):312–317.

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

[Graves 2013] Graves, A. 2013. Generating sequences with recurrent neural networks. *arXiv preprint arXiv:1308.0850*.

[Hamstra-Bletz, DeBie, and Den Brinker 1987] Hamstra-Bletz, L.; DeBie, J.; and Den Brinker, B. 1987. Concise evaluation scale for children's handwriting. *Lisse: Swets* 1.

[Haug, Tschiatschek, and Singla 2018] Haug, L.; Tschiatschek, S.; and Singla, A. 2018. Teaching inverse reinforcement learners via features and demonstrations. In *NIPS*.

[Hunziker et al. 2018] Hunziker, A.; Chen, Y.; Mac Aodha, O.; Gomez-Rodriguez, M.; Krause, A.; Perona, P.; Yue, Y.; and Singla, A. 2018. Teaching multiple concepts to a forgetful learner. *CoRR* abs/1805.08322.

[Johal et al. 2016] Johal, W.; Jacq, A.; Paiva, A.; and Dillenbourg, P. 2016. Child-robot spatial arrangement in a learning by teaching activity. In *Robot and Human Interactive Communication (RO-MAN), 2016 25th IEEE International Symposium on*, 533–538. IEEE.

[Koedinger et al. 1997] Koedinger, K. R.; Anderson, J. R.; Hadley, W. H.; and Mark, M. A. 1997. Intelligent tutoring goes to school in the big city. *International Journal of Artificial Intelligence in Education (IJAIED)* 8:30–43.

[Liu et al. 2017] Liu, W.; Dai, B.; Humayun, A.; Tay, C.; Yu, C.; Smith, L. B.; Rehg, J. M.; and Song, L. 2017. Iterative machine teaching. In *ICML*, 2149–2158.

[Mayer, Hamza, and Kuncak 2017] Mayer, M.; Hamza, J.; and Kuncak, V. 2017. Proactive synthesis of recursive tree-to-string functions from examples (artifact). In *DARTS-Dagstuhl Artifacts Series*, volume 3.

[Mei and Zhu 2015] Mei, S., and Zhu, X. 2015. Using machine teaching to identify optimal training-set attacks on machine learners. In *AAAI*, 2871–2877.

[Patil et al. 2014] Patil, K. R.; Zhu, X.; Kopeć, Ł.; and Love, B. C. 2014. Optimal teaching for limited-capacity human learners. In *NIPS*, 2465–2473.

[Rafferty et al. 2016] Rafferty, A. N.; Brunskill, E.; Griffiths, T. L.; and Shafto, P. 2016. Faster teaching via pomdp planning. *Cognitive science* 40(6):1290–1332.

[Rohrbeck et al. 2003] Rohrbeck, C. A.; Ginsburg-Block, M. D.; Fantuzzo, J. W.; and Miller, T. R. 2003. Peer-assisted learning interventions with elementary school students: A meta-analytic review.

[Singla et al. 2013] Singla, A.; Bogunovic, I.; Bartók, G.; Karbasi, A.; and Krause, A. 2013. On actively teaching the crowd to classify. In *NIPS Workshop on Data Driven Education*.

[Singla et al. 2014] Singla, A.; Bogunovic, I.; Bartók, G.; Karbasi, A.; and Krause, A. 2014. Near-optimally teaching the crowd to classify. In *ICML*, 154–162.

[Sullivan et al. 2009] Sullivan, B. L.; Wood, C. L.; Iliff, M. J.; Bonney, R. E.; Fink, D.; and Kelling, S. 2009. ebird: A citizen-based bird observation network in the biological sciences. *Biological Conservation* 142(10):2282–2292.

[Welinder et al. 2010] Welinder, P.; Branson, S.; Perona, P.; and Belongie, S. J. 2010. The multidimensional wisdom of crowds. In *NIPS*, 2424–2432.

[Yeo et al. 2018] Yeo, T.; Kamalaruban, P.; Singla, A.; Merchant, A.; Asselborn, T.; Faucon, L.; Dillenbourg, P.; and Cevher, V. 2018. Iterative classroom teaching. *CoRR* abs/1811.03537.

[Zhu et al. 2018] Zhu, X.; Singla, A.; Zilles, S.; and Rafferty, A. N. 2018. An overview of machine teaching. *CoRR* abs/1801.05927.

[Zhu, Liu, and Lopes 2017] Zhu, X.; Liu, J.; and Lopes, M. 2017. No learner left behind: On the complexity of teaching multiple learners simultaneously. In *IJCAI*, 3588–3594.

[Zhu 2013] Zhu, X. 2013. Machine teaching for bayesian learners in the exponential family. In *NIPS*, 1905–1913.

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

Additional Robust Teaching Settings

Noise in W^t

Assume that the teacher only receives the noisy version of W^t given by

$$\tilde{W}^t := W^t + \delta^t, \tag{16}$$

where δ^t is some random noise matrix such that $\lambda_1\left(\delta^t\right) \leq \frac{\alpha_{\min}\epsilon}{2(d-1)}$. Then the teacher constructs the example as follows:

$$\hat{x}^t := \underset{x:\|x\|=1}{\arg\max}\, x^\top \tilde{W}^t x = e_1\left(\tilde{W}^t\right)$$
$$x^t := \gamma_t \hat{x}^t \text{ and } y^t = \langle w^*, x^t \rangle, \tag{17}$$

where γ_t satisfies the condition given in (3). In this setting also, the classroom teaching is possible with linear convergence, as shown in the following theorem.

Theorem 4. *Consider the noisy observation setting given by* (16). *Let* $k := \max_t\left\{\mathrm{rank}\left(\tilde{W}^t\right)\right\}$ *where* $\tilde{W}^t$ *is given by* (16). *Define* $\alpha_{\min} := \min_{t,j}\alpha_j^t$, *where* $\alpha_j^t = \eta_j \gamma_t^2\left(2 - \eta_j \gamma_t^2\right)$. *Then for the robust teaching strategy given by* (17), *after* $t = \mathcal{O}\left(\left(\log\frac{1}{1-\frac{\alpha_{\min}}{k}}\right)^{-1}\log\frac{1}{\epsilon}\right)$ *rounds, we have* $\frac{1}{N}\sum_{i=1}^N\|w_i^t - w^*\|^2 \leq \epsilon$.

SGLD Learners

Here we consider a classroom of *Stochastic Gradient Langevin Dynamics* (SGLD) learners. For a given example $(x, y) \in \mathcal{X} \times \mathcal{Y}$ at time t, the update rule of the student $i \in [N]$ is given by

$$w_i^{t+1} = \mathtt{Proj}_{\mathcal{W}}\left(w_i^t - \eta_i\frac{\partial\ell\left(\langle w_i^t, x\rangle, y\right)}{\partial w_i^t} + \sqrt{2\eta_j\beta^{-1}}\xi_j^t\right), \tag{18}$$

where $\xi_j^t \sim \mathcal{N}(0, I)$ is a standard Gaussian random vector in $\mathbb{R}^d$, and $\beta > 0$ is the inverse temperature parameter. We assume that the teacher has full observability of $\{w_i^t\}_{i=1}^N$, and has full knowledge of students' learning rates $\{\eta_i\}_{i=1}^N$, but doesn't know β. Then the teacher constructs the example as follows (depending on $H_t := \left(\{w_j^s\}_{s=1}^t : \forall j \in [N]\right)$):

$$\hat{x}^t := \underset{x:\|x\|=1}{\arg\max}\, x^\top W^t x = e_1\left(W^t\right)$$
$$x^t := \gamma_t \hat{x}^t \text{ and } y^t = \langle w^*, x^t \rangle, \tag{19}$$

where

$$\gamma_t^2 \leq \frac{2}{\eta_j}, \forall j \in [N] \tag{20}$$
$$\alpha_j^t := \eta_j \gamma_t^2\left(2 - \eta_j \gamma_t^2\right) \tag{21}$$
$$\hat{w}_j^t := w_j^t - w^* \text{ and} \tag{22}$$
$$W^t := \frac{1}{N}\sum_{j=1}^N \alpha_j^t \hat{w}_j^t\left(\hat{w}_j^t\right)^\top. \tag{23}$$

The following theorem shows that, in this setting, the teacher can teach the classroom in expectation with linear convergence.

Theorem 5. *Consider the classroom model given by* (18). *Let* $k := \max_t\left\{\mathrm{rank}\left(W^t\right)\right\}$ *where* W^t *is given by* (23). *Define* $\alpha_{\min} := \min_{t,j}\alpha_j^t$, *and* $\eta_{\mathrm{avg}} := \frac{1}{N}\sum_{j=1}^N \eta_j$, *where* α_j^t *given by* (21). *Then for the teaching strategy given by* (19) *and for* $\beta^{-1} \leq \frac{\alpha_{\min}}{4\eta_{\mathrm{avg}}d^2}\epsilon$, *after* $t = \mathcal{O}\left(\left(\log\frac{1}{1-\frac{\alpha_{\min}}{k}}\right)^{-1}\log\frac{1}{\epsilon}\right)$ *rounds, we have* $\mathbb{E}\left[\frac{1}{N}\sum_{i=1}^N\|w_i^t - w^*\|^2\right] \leq \epsilon$.

Proofs

Theorem 1. *Consider the teaching strategy given in Algorithm 1. Let* $k := \max_t\left\{\mathrm{rank}\left(W^t\right)\right\}$, *where* W^t *is given by Eq.* (6). *Define* $\alpha_j := \min_t\alpha_j^t$, $\alpha_{\min} := \min_{t,j}\alpha_j^t$, *and* $\alpha_{\max} := \max_{t,j}\alpha_j^t$, *where* α_j^t *is given by Eq.* (5). *Then after* $t = \mathcal{O}\left(\left(\log\frac{1}{1-\frac{\alpha_{\min}}{k}}\right)^{-1}\log\frac{1}{\epsilon}\right)$ *rounds, we have* $\frac{1}{N}\sum_{j=1}^N\|w_j^t - w^*\|^2 \leq \epsilon$. *Furthermore, after* $t = \mathcal{O}\left(\max\left\{\left(\log\frac{1}{1-\alpha_j}\right)^{-1}\log\frac{1}{\epsilon}, \left(\log\frac{1}{1-\frac{\alpha_{\min}}{k}}\right)^{-1}\log\frac{1}{\epsilon}\right\}\right)$ *rounds, we have* $\|w_j^t - w^*\|^2 \leq \epsilon, \forall j \in [N]$.

Proof. Let $G\left(w;x,y\right) = \frac{\partial \ell(\langle w,x\rangle,y)}{\partial w}$. For the student $j \in [N]$ with the update rule $w_j^{t+1} \leftarrow \texttt{Proj}_{\mathcal{W}}\left(w_j^t - \eta_j G\left(w_j^t;x,y\right)\right)$ and any input example $(x^t, y^t) \in \mathcal{X} \times \mathcal{Y}$ (with $y^t = \langle w^*, x^t\rangle$) we have

$$
\begin{aligned}
\left\|w_j^{t+1} - w^*\right\|^2 &\overset{(i)}{\leq} \left\|w_j^t - \eta_j G\left(w_j^t;x^t,y^t\right) - w^*\right\|^2 \\
&= \left\|w_j^t - w^*\right\|^2 + \eta_j^2 \left\|G\left(w_j^t;x^t,y^t\right)\right\|^2 - 2\eta_j \left\langle w_j^t - w^*, G\left(w_j^t;x^t,y^t\right)\right\rangle \\
&\overset{(ii)}{=} \left\|w_j^t - w^*\right\|^2 + \eta_j^2 \left(\langle w_j^t, x^t\rangle - y^t\right)^2 \left\|x^t\right\|^2 - 2\eta_j \left(\langle w_j^t, x^t\rangle - y^t\right)\left\langle w_j^t - w^*, x^t\right\rangle \\
&= \left\|w_j^t - w^*\right\|^2 + \eta_j \langle w_j^t - w^*, x^t\rangle^2 \left(\eta_j \left\|x^t\right\|^2 - 2\right),
\end{aligned} \tag{24}
$$

where (i) is by the property of projection, and (ii) is due to the fact that $G\left(w;x,y\right) = \left(\langle w,x\rangle - y\right)\cdot x$ for the squared loss function. Then for the example construction strategy described in Algorithm 1, we have

$$
\begin{aligned}
\frac{1}{N}\sum_{j=1}^N \left\|w_j^{t+1} - w^*\right\|^2 &\leq \frac{1}{N}\sum_{j=1}^N \left\|w_j^t - w^*\right\|^2 + \frac{1}{N}\sum_{j=1}^N \eta_j \langle w_j^t - w^*, x^t\rangle^2 \left(\eta_j \left\|x^t\right\|^2 - 2\right) \\
&\overset{(i)}{=} \frac{1}{N}\sum_{j=1}^N \left\|w_j^t - w^*\right\|^2 + \frac{1}{N}\sum_{j=1}^N \eta_j \gamma_t^2 \langle w_j^t - w^*, \hat{x}^t\rangle^2 \left(\eta_j \gamma_t^2 - 2\right) \\
&\overset{(ii)}{=} \frac{1}{N}\sum_{j=1}^N \left\|w_j^t - w^*\right\|^2 - \frac{1}{N}\sum_{j=1}^N \alpha_j^t \langle w_j^t - w^*, \hat{x}^t\rangle^2 \\
&\overset{(iii)}{=} \frac{1}{N}\sum_{j=1}^N \left\|w_j^t - w^*\right\|^2 - \frac{1}{N}\sum_{j=1}^N \alpha_j^t \langle \hat{w}_j^t, \hat{x}^t\rangle^2 \\
&= \frac{1}{N}\sum_{j=1}^N \left\|w_j^t - w^*\right\|^2 - \frac{1}{N}\sum_{j=1}^N \alpha_j^t \left(\hat{x}^t\right)^\top \hat{w}_j^t \left(\hat{w}_j^t\right)^\top \hat{x}^t \\
&\overset{(iv)}{=} \frac{1}{N}\sum_{j=1}^N \left\|w_j^t - w^*\right\|^2 - \left(\hat{x}^t\right)^\top W^t \hat{x}^t,
\end{aligned} \tag{25}
$$

where (i) is due to the fact that $x^t = \gamma_t \hat{x}^t$ with $\|\hat{x}^t\| = 1$, (ii) is due to the fact that $\left(\eta_j \gamma_t^2 - 2\right) < 0, \forall j \in [N]$, (iii) is by the definition of $\hat{w}_j^t$, and (iv) is by the definition of W^t. Since $\hat{x}^t$ is the first principal component of W^t *i.e.* eigenvector corresponding to the largest eigenvalue of W^t, we have

$$
\begin{aligned}
\left(\hat{x}^t\right)^\top W^t \hat{x}^t &= \lambda_1\left(W^t\right) \\
&= \frac{\lambda_1\left(W^t\right)}{\sum_{j=1}^d \lambda_j\left(W^t\right)} \cdot \mathrm{tr}\left(W^t\right) \\
&= \frac{\lambda_1\left(W^t\right)}{\sum_{j=1}^d \lambda_j\left(W^t\right)} \cdot \frac{1}{N}\sum_{j=1}^N \alpha_j^t \left\|\hat{w}_j^t\right\|^2 \\
&\geq \frac{1}{k} \cdot \frac{1}{N}\sum_{j=1}^N \alpha_j^t \left\|\hat{w}_j^t\right\|^2 \\
&= \frac{1}{k} \cdot \frac{1}{N}\sum_{j=1}^N \alpha_j^t \left\|w_j^t - w^*\right\|^2 \\
&\geq \frac{\alpha_{\min}}{k} \cdot \frac{1}{N}\sum_{j=1}^N \left\|w_j^t - w^*\right\|^2,
\end{aligned} \tag{26}
$$

where $\alpha_{\min} := \min_{t,j} \alpha_j^t$, and $k := \max_t \{\mathrm{rank}\left(W^t\right)\}$. From (25) and (26), we get

$$
\frac{1}{N}\sum_{j=1}^N \left\|w_j^{t+1} - w^*\right\|^2 \leq \left(1 - \frac{\alpha_{\min}}{k}\right)\frac{1}{N}\sum_{j=1}^N \left\|w_j^t - w^*\right\|^2. \tag{27}
$$

That is after
$$t + 1 \geq \left(\log \frac{1}{1 - \frac{\alpha_{\min}}{k}} \right)^{-1} \log \frac{\frac{1}{N} \sum_{i=1}^{N} \left\| w_i^0 - w^* \right\|^2}{\epsilon}$$

iterations we get $\frac{1}{N} \sum_{i=1}^{N} \left\| w_i^{t+1} - w^* \right\|^2 \leq \epsilon$. This completes the proof of the first part of the theorem.

For any student $j \in [N]$, and the example constructed in Algorithm 1, we have

$$
\begin{aligned}
\left\| w_j^{t+1} - w^* \right\|^2 &\leq \left\| w_j^t - w^* \right\|^2 + \eta_j \langle w_j^t - w^*, x^t \rangle^2 \left(\eta_j \left\| x^t \right\|^2 - 2 \right) \\
&= \left\| w_j^t - w^* \right\|^2 + \eta_j \gamma_t^2 \langle w_j^t - w^*, \hat{x}^t \rangle^2 \left(\eta_j \gamma_t^2 - 2 \right) \\
&= \left\| w_j^t - w^* \right\|^2 - \alpha_j^t \langle w_j^t - w^*, \hat{x}^t \rangle^2 \\
&= \left\| w_j^t - w^* \right\|^2 - \alpha_j^t \langle \hat{w}_j^t, \hat{x}^t \rangle^2 \\
&= \left\| w_j^t - w^* \right\|^2 - \left(\hat{x}^t \right)^\top W_j^t \hat{x}^t \\
&= \left\| w_j^t - w^* \right\|^2 - \left(\hat{x}^t \right)^\top \left(W^t - \delta_j^t \right) \hat{x}^t \\
&\leq \left\| w_j^t - w^* \right\|^2 - \left(\hat{x}^t \right)^\top W^t \hat{x}^t + \max_{x : \|x\| = 1} x^\top \delta_j^t x,
\end{aligned}
\tag{28}
$$

where $W_j^t := \alpha_j^t \hat{w}_j^t \left(\hat{w}_j^t \right)^\top$ and $\delta_j^t := W^t - W_j^t$. Consider

$$
\begin{aligned}
\max_{x : \|x\| = 1} x^\top \delta_j^t x &= \lambda_1 \left(\delta_j^t \right) \\
&\leq \operatorname{tr} \left(\delta_j^t \right) \\
&= \operatorname{tr} \left(W^t \right) - \operatorname{tr} \left(W_j^t \right) \\
&= \frac{1}{N} \sum_{i=1}^{N} \alpha_i^t \left\| \hat{w}_i^t \right\|^2 - \alpha_j^t \left\| \hat{w}_j^t \right\|^2 \\
&= \frac{1}{N} \sum_{i=1}^{N} \alpha_i^t \left\| w_i^t - w^* \right\|^2 - \alpha_j^t \left\| w_j^t - w^* \right\|^2 \\
&= \alpha_{\max} \cdot \frac{1}{N} \sum_{i=1}^{N} \left\| w_i^t - w^* \right\|^2 - \alpha_j \left\| w_j^t - w^* \right\|^2,
\end{aligned}
\tag{29}
$$

where $\alpha_{\max} := \max_{t,i} \alpha_i^t$ and $\alpha_j := \min_t \alpha_j^t$. Thus from (26), (27), (28), and (29), we get

$$
\begin{aligned}
\left\| w_j^{t+1} - w^* \right\|^2 &\leq \left\| w_j^t - w^* \right\|^2 - \left(\hat{x}^t \right)^\top W^t \hat{x}^t + \max_{x : \|x\| = 1} x^\top \delta_j^t x \\
&\leq \left\| w_j^t - w^* \right\|^2 - \frac{\alpha_{\min}}{k} \cdot \frac{1}{N} \sum_{i=1}^{N} \left\| w_i^t - w^* \right\|^2 + \alpha_{\max} \cdot \frac{1}{N} \sum_{i=1}^{N} \left\| w_i^t - w^* \right\|^2 - \alpha_j \left\| w_j^t - w^* \right\|^2 \\
&= (1 - \alpha_j) \left\| w_j^t - w^* \right\|^2 + \left(\alpha_{\max} - \frac{\alpha_{\min}}{k} \right) \Delta_t \\
&\leq (1 - \alpha_j) \left\| w_j^t - w^* \right\|^2 + \left(\alpha_{\max} - \frac{\alpha_{\min}}{k} \right) \left(1 - \frac{\alpha_{\min}}{k} \right)^t \Delta_0 \\
&\leq (1 - \alpha_j) \left\{ (1 - \alpha_j) \left\| w_j^{t-1} - w^* \right\|^2 + \left(\alpha_{\max} - \frac{\alpha_{\min}}{k} \right) \left(1 - \frac{\alpha_{\min}}{k} \right)^{t-1} \Delta_0 \right\} \\
&\quad + \left(\alpha_{\max} - \frac{\alpha_{\min}}{k} \right) \left(1 - \frac{\alpha_{\min}}{k} \right)^t \Delta_0 \\
&\leq (1 - \alpha_j)^{t+1} \left\| w_j^0 - w^* \right\|^2 + \left(\alpha_{\max} - \frac{\alpha_{\min}}{k} \right) \Delta_0 \sum_{s=0}^{t} (1 - \alpha_j)^s \left(1 - \frac{\alpha_{\min}}{k} \right)^{t-s},
\end{aligned}
$$

where $\Delta_t := \frac{1}{N} \sum_{i=1}^{N} \left\| w_i^t - w^* \right\|^2$. Since $1 - \alpha_j \leq 1 - \frac{\alpha_{\min}}{k}$, we have

$$\sum_{s=0}^{t} (1 - \alpha_j)^s \left(1 - \frac{\alpha_{\min}}{k} \right)^{t-s} = \left(1 - \frac{\alpha_{\min}}{k} \right)^t \sum_{s=0}^{t} \left(\frac{1 - \alpha_j}{1 - \frac{\alpha_{\min}}{k}} \right)^s$$

$$\leq \left(1 - \frac{\alpha_{\min}}{k}\right)^t \sum_{s=0}^{\infty} \left(\frac{1-\alpha_j}{1 - \frac{\alpha_{\min}}{k}}\right)^s$$

$$= \left(1 - \frac{\alpha_{\min}}{k}\right)^t \frac{1}{1 - \frac{1-\alpha_j}{1 - \frac{\alpha_{\min}}{k}}}$$

$$= \frac{1}{\alpha_j - \frac{\alpha_{\min}}{k}} \left(1 - \frac{\alpha_{\min}}{k}\right)^{t+1}.$$

Thus we get

$$\left\|w_j^{t+1} - w^*\right\|^2 \ \leq \ (1-\alpha_j)^{t+1} \left\|w_j^0 - w^*\right\|^2 + \frac{\alpha_{\max}}{\alpha_j - \frac{\alpha_{\min}}{k}} \Delta_0 \left(1 - \frac{\alpha_{\min}}{k}\right)^{t+1}$$

$$= \ (1-\alpha_j)^{t+1} \left\|w_j^0 - w^*\right\|^2 + C_j \Delta_0 \left(1 - \frac{\alpha_{\min}}{k}\right)^{t+1}$$

$$\leq \ \frac{\epsilon}{2} + \frac{\epsilon}{2} = \epsilon,$$

where $C_j := \frac{\alpha_{\max}}{\alpha_j - \frac{\alpha_{\min}}{k}}$. That is we get $\left\|w_j^{t+1} - w^*\right\|^2 \leq \epsilon$ after

$$t_j + 1 = \max \left\{ \left(\log \frac{1}{1-\alpha_j}\right)^{-1} \log \frac{2\left\|w_j^0 - w^*\right\|^2}{\epsilon}, \ \left(\log \frac{1}{1 - \frac{\alpha_{\min}}{k}}\right)^{-1} \log \frac{2C_j \Delta^0}{\epsilon} \right\}$$

iterations. This completes the proof. $\qquad \square$

Theorem 2. *Consider the noisy observation setting given by Eq. (8). Let $k := \max_t \{\mathrm{rank}\,(W^t)\}$ where $W^t = \frac{1}{N}\sum_{j=1}^N \alpha_j^t \left(\tilde{w}_j^t - w^*\right)\left(\tilde{w}_j^t - w^*\right)^\top$. Then for the robust teaching strategy given by Eq. (9), after $t = \mathcal{O}\left(\left(\log \frac{1}{1 - \frac{\alpha_{\min}}{k}}\right)^{-1} \log \frac{1}{\epsilon}\right)$ rounds, we have $\frac{1}{N}\sum_{j=1}^N \left\|w_j^t - w^*\right\|^2 \leq \epsilon$.*

Proof. From (24), we have

$$\left\|w_j^{t+1} - w^*\right\|^2 \ = \ \left\|w_j^t - w^*\right\|^2 + \eta_j \langle w_j^t - w^*, x^t\rangle^2 \left(\eta_j \left\|x^t\right\|^2 - 2\right).$$

Then for the example construction strategy described, we have

$$\left\|w_j^{t+1} - w^*\right\|^2 \ \leq \ \left\|w_j^t - w^*\right\|^2 + \eta_j \gamma_t^2 \langle w_j^t - w^*, \hat{x}^t\rangle^2 \left(\eta_j \gamma_t^2 - 2\right)$$

$$= \ \left\|w_j^t - w^*\right\|^2 - \eta_j \gamma_t^2 \left(2 - \eta_j \gamma_t^2\right) \langle \tilde{w}_j^t - \delta - w^*, \hat{x}^t\rangle^2$$

$$= \ \left\|w_j^t - w^*\right\|^2 - \eta_j \gamma_t^2 \left(2 - \eta_j \gamma_t^2\right) \left\{ \langle \tilde{w}_j^t - w^*, \hat{x}^t\rangle^2 - 2\langle \tilde{w}_j^t - w^*, \hat{x}^t\rangle \langle \delta, \hat{x}^t\rangle + \langle \delta, \hat{x}^t\rangle^2 \right\}$$

$$\leq \ \left\|w_j^t - w^*\right\|^2 - \eta_j \gamma_t^2 \left(2 - \eta_j \gamma_t^2\right) \langle \tilde{w}_j^t - w^*, \hat{x}^t\rangle^2 + 2\eta_j \gamma_t^2 \left(2 - \eta_j \gamma_t^2\right) \langle \tilde{w}_j^t - w^*, \hat{x}^t\rangle \langle \delta, \hat{x}^t\rangle$$

$$\leq \ \left\|w_j^t - w^*\right\|^2 - \alpha_j^t \langle \hat{w}_j^t, \hat{x}^t\rangle^2 + 2\alpha_j^t \left\|\tilde{w}_j^t - w^*\right\| \left\|\hat{x}^t\right\| \left\|\delta\right\| \left\|\hat{x}^t\right\|$$

$$\leq \ \left\|w_j^t - w^*\right\|^2 - \alpha_j^t \langle \hat{w}_j^t, \hat{x}^t\rangle^2 + 2\alpha_j^t D_{\mathcal{W}} \left\|\delta\right\|.$$

Since

$$\left\|\tilde{w}_j^t - w^*\right\|^2 \ = \ \left\|w_j^t + \delta - w^*\right\|^2$$

$$= \ \left\|w_j^t - w^*\right\|^2 + \left\|\delta\right\|^2 - 2\langle w_j^t - w^*, \delta\rangle$$

$$\geq \ \left\|w_j^t - w^*\right\|^2 + \left\|\delta\right\|^2 - 2\left\|w_j^t - w^*\right\| \left\|\delta\right\|$$

$$\geq \ \left\|w_j^t - w^*\right\|^2 - 2D_{\mathcal{W}} \left\|\delta\right\|,$$

we have

$$\frac{1}{N}\sum_{j=1}^N \left\|w_j^{t+1} - w^*\right\|^2 \ \leq \ \frac{1}{N}\sum_{j=1}^N \left\|w_j^t - w^*\right\|^2 - \frac{1}{N}\sum_{j=1}^N \alpha_j^t \langle \hat{w}_j^t, \hat{x}^t\rangle^2 + 2\frac{1}{N}\sum_{j=1}^N \alpha_j^t D_{\mathcal{W}} \left\|\delta\right\|$$

$$\leq \frac{1}{N} \sum_{j=1}^{N} \left\| w_j^t - w^* \right\|^2 - \frac{\alpha_{\min}}{k} \frac{1}{N} \sum_{j=1}^{N} \left\| \tilde{w}_j^t - w^* \right\|^2 + 2\alpha_{\mathrm{avg}} D_{\mathcal{W}} \left\| \delta \right\|$$

$$\leq \frac{1}{N} \sum_{j=1}^{N} \left\| w_j^t - w^* \right\|^2 - \frac{\alpha_{\min}}{k} \frac{1}{N} \sum_{j=1}^{N} \left(\left\| w_j^t - w^* \right\|^2 - 2D_{\mathcal{W}} \left\| \delta \right\| \right) + 2\alpha_{\mathrm{avg}} D_{\mathcal{W}} \left\| \delta \right\|$$

$$= \left(1 - \frac{\alpha_{\min}}{k} \right) \frac{1}{N} \sum_{j=1}^{N} \left\| w_j^t - w^* \right\|^2 + 2 \left(\frac{\alpha_{\min}}{k} + \alpha_{\mathrm{avg}} \right) D_{\mathcal{W}} \left\| \delta \right\|$$

$$\leq \left(1 - \frac{\alpha_{\min}}{k} \right)^{t+1} \frac{1}{N} \sum_{j=1}^{N} \left\| w_j^0 - w^* \right\|^2 + 2 \left(\frac{\alpha_{\min}}{k} + \alpha_{\mathrm{avg}} \right) D_{\mathcal{W}} \left\| \delta \right\| \sum_{s=0}^{t} \left(1 - \frac{\alpha_{\min}}{k} \right)^{s}$$

$$\leq \left(1 - \frac{\alpha_{\min}}{k} \right)^{t+1} \frac{1}{N} \sum_{j=1}^{N} \left\| w_j^0 - w^* \right\|^2 + 2 \left(\frac{\alpha_{\min}}{k} + \alpha_{\mathrm{avg}} \right) D_{\mathcal{W}} \left\| \delta \right\| \frac{1}{1 - \left(1 - \frac{\alpha_{\min}}{k} \right)}$$

$$= \left(1 - \frac{\alpha_{\min}}{k} \right)^{t+1} \frac{1}{N} \sum_{j=1}^{N} \left\| w_j^0 - w^* \right\|^2 + 2 \left(\frac{k\alpha_{\mathrm{avg}}}{\alpha_{\min}} + 1 \right) D_{\mathcal{W}} \left\| \delta \right\|$$

$$\leq \frac{\epsilon}{2} + \frac{\epsilon}{2} = \epsilon,$$

where $\alpha_{\min} := \min_{t,j} \alpha_j^t$, and $\alpha_{\mathrm{avg}} := \max_t \frac{1}{N} \sum_{j=1}^{N} \alpha_j^t$. Thus for

$$t + 1 \geq \left(\log \frac{1}{1 - \frac{\alpha_{\min}}{k}} \right)^{-1} \log \frac{\frac{2}{N} \sum_{i=1}^{N} \left\| w_i^0 - w^* \right\|^2}{\epsilon}$$

and $\|\delta\| \leq \frac{\epsilon}{4 \left(\frac{k\alpha_{\mathrm{avg}}}{\alpha_{\min}} + 1 \right) D_{\mathcal{W}}}$, we have $\frac{1}{N} \sum_{i=1}^{N} \left\| w_i^{t+1} - w^* \right\|^2 \leq \epsilon$. $\qquad \square$

Theorem 3. *Let* $k := \max_t \left\{ \mathrm{rank} \left(\bar{W}^t \right) \right\}$ *where* $\bar{W}^t := \frac{1}{N} \sum_{j=1}^{N} \bar{\alpha}_j^t \hat{w}_j^t \left(\hat{w}_j^t \right)^{\top}$. *Define* $\bar{\alpha}_{\min} := \min_{t,j} \bar{\alpha}_j^t$, *and* $\beta_{\min} := \min_{j,t} \frac{\alpha_j^t}{\bar{\alpha}_j^t}$, *where* $\alpha_j^t := 2\gamma_t^2 \eta_j - \gamma_t^4 \left(\sigma^2 + \eta_j^2 \right)$ *and* $\bar{\alpha}_j^t$ *given by Eq. (13). Then for the teaching strategy given by Eq. (10), after*
$t = \mathcal{O} \left(\left(\log \frac{1}{1 - \frac{\beta_{\min} \bar{\alpha}_{\min}}{k}} \right)^{-1} \log \frac{1}{\epsilon} \right)$ *rounds, we have* $\mathbb{E} \left[\frac{1}{N} \sum_{j=1}^{N} \left\| w_j^t - w^* \right\|^2 \right] \leq \epsilon$.

Proof. For the student $j \in [N]$, with update rule $w_j^{t+1} \leftarrow \mathrm{Proj}_{\mathcal{W}} \left(w_j^t - \eta_j^t G \left(w_j^t; x, y \right) \right)$, where $\eta_j^t \sim \mathcal{N} \left(\eta_j, \sigma \right)$, and for any example (x^t, y^t), from (24), we have

$$\left\| w_j^{t+1} - w^* \right\|^2 \leq \left\| w_j^t - w^* \right\|^2 + \eta_j^t \langle w_j^t - w^*, x^t \rangle^2 \left(\eta_j^t \left\| x^t \right\|^2 - 2 \right).$$

Let the history up to time t be $H_t := \left(\{ w_j^s \}_{s=1}^{t}, \{ \eta_j^s \}_{s=1}^{t-1} : \forall j \in [N] \right)$, and define $\mathbb{E}_t [\cdot] := \mathbb{E} [\cdot \mid H_t]$. Suppose the teacher constructs the example $x^t = \gamma_t \hat{x}^t$ (with $\|\hat{x}^t\| = 1$) based on the history H_t. Then given H_t, only η_j^t and w_j^{t+1} are random variables in the above equation *i.e.* we have

$$\mathbb{E}_t \left[\left\| w_j^{t+1} - w^* \right\|^2 \right] \leq \left\| w_j^t - w^* \right\|^2 + \mathbb{E}_t \left[\eta_j^t \gamma_t^2 \left(\eta_j^t \gamma_t^2 - 2 \right) \right] \langle w_j^t - w^*, \hat{x}^t \rangle^2$$

$$= \left\| w_j^t - w^* \right\|^2 + \left\{ \gamma_t^4 \left(\sigma^2 + \eta_j^2 \right) - 2\gamma_t^2 \eta_j \right\} \langle w_j^t - w^*, \hat{x}^t \rangle^2$$

$$= \left\| w_j^t - w^* \right\|^2 - \alpha_j^t \langle w_j^t - w^*, \hat{x}^t \rangle^2,$$

where $\alpha_j^t := 2\gamma_t^2 \eta_j - \gamma_t^4 \left(\sigma^2 + \eta_j^2 \right)$ and $\gamma_t^2 \leq \frac{2\eta_j}{\sigma^2 + \eta_j^2}, \forall j \in [N]$. Thus for the classroom of students, we have

$$\mathbb{E}_t \left[\frac{1}{N} \sum_{j=1}^{N} \left\| w_j^{t+1} - w^* \right\|^2 \right] \leq \frac{1}{N} \sum_{j=1}^{N} \left\| w_j^t - w^* \right\|^2 - \frac{1}{N} \sum_{j=1}^{N} \alpha_j^t \langle w_j^t - w^*, \hat{x}^t \rangle^2$$

$$= \frac{1}{N} \sum_{j=1}^{N} \left\| w_j^t - w^* \right\|^2 - \frac{1}{N} \sum_{j=1}^{N} \alpha_j^t \langle \hat{w}_j^t, \hat{x}^t \rangle^2$$

$$= \frac{1}{N} \sum_{j=1}^{N} \|w_j^t - w^*\|^2 - \frac{1}{N} \sum_{j=1}^{N} \alpha_j^t \left(\hat{x}^t\right)^\top \hat{w}_j^t \left(\hat{w}_j^t\right)^\top \hat{x}^t$$

$$= \frac{1}{N} \sum_{j=1}^{N} \|w_j^t - w^*\|^2 - \left(\hat{x}^t\right)^\top W^t \hat{x}^t, \tag{30}$$

where $\hat{w}_j^t := \left(w_j^t - w^*\right)$ and $W^t := \frac{1}{N} \sum_{j=1}^{N} \alpha_j^t \hat{w}_j^t \left(\hat{w}_j^t\right)^\top$. The teacher constructs the example $\hat{x}^t$ as follows (depending only on H_t):

$$\hat{x}^t := \underset{x:\|x\|=1}{\arg\max} \, x^\top \left\{ \frac{1}{N} \sum_{j=1}^{N} \bar{\alpha}_j^t \hat{w}_j^t \left(\hat{w}_j^t\right)^\top \right\} x$$

$$= \underset{x:\|x\|=1}{\arg\max} \, x^\top \bar{W}^t x = e_1\left(\bar{W}^t\right),$$

where $\bar{\alpha}_j^t := 2\gamma_t^2 \bar{\eta}_j^t - \gamma_t^4 \left(\frac{t-2}{t-1}\sigma^2 + \left(\bar{\eta}_j^t\right)^2\right)$, $\bar{\eta}_j^t := \frac{1}{t-1}\sum_{s=1}^{t-1}\eta_j^s$ and $\bar{W}^t := \frac{1}{N}\sum_{j=1}^{N}\bar{\alpha}_j^t \hat{w}_j^t \left(\hat{w}_j^t\right)^\top$. Note that $\mathbb{E}\left[\bar{\alpha}_j^t\right] = \alpha_j^t$. For this example, from (26), we have

$$\left(\hat{x}^t\right)^\top \bar{W}^t \hat{x}^t \geq \frac{\bar{\alpha}_{\min}}{k} \cdot \frac{1}{N} \sum_{j=1}^{N} \|w_j^t - w^*\|^2,$$

where $\bar{\alpha}_{\min} := \min_{t,j} \bar{\alpha}_j^t$. Since $\hat{w}_j^t \left(\hat{w}_j^t\right)^\top$ is a positive semidefinite matrix, we have

$$\left(\hat{x}^t\right)^\top W^t \hat{x}^t = \left(\hat{x}^t\right)^\top \left\{ \frac{1}{N} \sum_{j=1}^{N} \alpha_j^t \hat{w}_j^t \left(\hat{w}_j^t\right)^\top \right\} \hat{x}^t$$

$$\geq \min_{j,t} \frac{\alpha_j^t}{\bar{\alpha}_j^t} \left(\hat{x}^t\right)^\top \left\{ \frac{1}{N} \sum_{j=1}^{N} \bar{\alpha}_j^t \hat{w}_j^t \left(\hat{w}_j^t\right)^\top \right\} \hat{x}^t$$

$$= \beta_{\min} \left(\hat{x}^t\right)^\top \bar{W}^t \hat{x}^t$$

$$\geq \frac{\beta_{\min}\bar{\alpha}_{\min}}{k} \cdot \frac{1}{N} \sum_{j=1}^{N} \|w_j^t - w^*\|^2,$$

where $\beta_{\min} := \min_{j,t} \frac{\alpha_j^t}{\bar{\alpha}_j^t}$. By using the above inequality in (30), we get

$$\mathbb{E}_t \left[\frac{1}{N} \sum_{j=1}^{N} \|w_j^{t+1} - w^*\|^2 \right] \leq \left(1 - \frac{\beta_{\min}\bar{\alpha}_{\min}}{k}\right) \frac{1}{N} \sum_{j=1}^{N} \|w_j^t - w^*\|^2.$$

Then by the law of total expectation and the above recurrence relationship, we have

$$\mathbb{E} \left[\frac{1}{N} \sum_{j=1}^{N} \|w_j^{t+1} - w^*\|^2 \right] = \mathbb{E}_0 \left[\frac{1}{N} \sum_{j=1}^{N} \|w_j^{t+1} - w^*\|^2 \right]$$

$$= \mathbb{E}_0 \mathbb{E}_1 \cdots \mathbb{E}_{t-1} \mathbb{E}_t \left[\frac{1}{N} \sum_{j=1}^{N} \|w_j^{t+1} - w^*\|^2 \right]$$

$$\leq \left(1 - \frac{\beta_{\min}\bar{\alpha}_{\min}}{k}\right) \mathbb{E}_0 \mathbb{E}_1 \cdots \mathbb{E}_{t-1} \left[\frac{1}{N} \sum_{j=1}^{N} \|w_j^t - w^*\|^2 \right]$$

$$\leq \left(1 - \frac{\beta_{\min}\bar{\alpha}_{\min}}{k}\right)^{t+1} \frac{1}{N} \sum_{j=1}^{N} \|w_j^0 - w^*\|^2.$$

That is after

$$t+1 \geq \left(\log \frac{1}{1 - \frac{\beta_{\min}\bar{\alpha}_{\min}}{k}}\right)^{-1} \log \frac{\frac{1}{N}\sum_{j=1}^{N} \left\|w_j^0 - w^*\right\|^2}{\epsilon}$$

iterations we get $\mathbb{E}\left[\frac{1}{N}\sum_{j=1}^{N}\left\|w_j^{t+1} - w^*\right\|^2\right] \leq \epsilon$.

$\square$

Theorem 4. *Consider the noisy observation setting given by* (16). *Let* $k := \max_t \left\{\mathrm{rank}\left(\tilde{W}^t\right)\right\}$ *where* $\tilde{W}^t$ *is given by* (16). *Define* $\alpha_{\min} := \min_{t,j} \alpha_j^t$, *where* $\alpha_j^t = \eta_j \gamma_t^2 \left(2 - \eta_j \gamma_t^2\right)$. *Then for the robust teaching strategy given by* (17), *after* $t = \mathcal{O}\left(\left(\log \frac{1}{1 - \frac{\alpha_{\min}}{k}}\right)^{-1} \log \frac{1}{\epsilon}\right)$ *rounds, we have* $\frac{1}{N}\sum_{i=1}^{N}\left\|w_i^t - w^*\right\|^2 \leq \epsilon$.

Proof. From (25), we have

$$\frac{1}{N}\sum_{j=1}^{N}\left\|w_j^{t+1} - w^*\right\|^2 \leq \frac{1}{N}\sum_{j=1}^{N}\left\|w_j^t - w^*\right\|^2 - \left(\hat{x}^t\right)^\top W^t \hat{x}^t,$$

where $\alpha_j^t := \eta_j \gamma_t^2 \left(2 - \eta_j \gamma_t^2\right)$, $\hat{w}_j^t := w_j^t - w^*$ and $W^t := \frac{1}{N}\sum_{j=1}^{N} \alpha_j^t \hat{w}_j^t \left(\hat{w}_j^t\right)^\top$. Then for the example construction strategy described in section , we have

$$\frac{1}{N}\sum_{j=1}^{N}\left\|w_j^{t+1} - w^*\right\|^2 \leq \frac{1}{N}\sum_{j=1}^{N}\left\|w_j^t - w^*\right\|^2 - \left(\hat{x}^t\right)^\top W^t \hat{x}^t$$

$$= \frac{1}{N}\sum_{j=1}^{N}\left\|w_j^t - w^*\right\|^2 - \left(\hat{x}^t\right)^\top \left(\tilde{W}^t - \delta\right)\hat{x}^t$$

$$= \frac{1}{N}\sum_{j=1}^{N}\left\|w_j^t - w^*\right\|^2 - \left(\hat{x}^t\right)^\top \tilde{W}^t \hat{x}^t + \left(\hat{x}^t\right)^\top \delta \hat{x}^t$$

$$= \frac{1}{N}\sum_{j=1}^{N}\left\|w_j^t - w^*\right\|^2 - \lambda_1\left(\tilde{W}^t\right) + \left(\hat{x}^t\right)^\top \delta \hat{x}^t$$

$$= \frac{1}{N}\sum_{j=1}^{N}\left\|w_j^t - w^*\right\|^2 - \frac{\lambda_1\left(\tilde{W}^t\right)}{\sum_{j=1}^{d}\lambda_j\left(\tilde{W}^t\right)}\mathrm{tr}\left(\tilde{W}^t\right) + \left(\hat{x}^t\right)^\top \delta \hat{x}^t$$

$$\leq \frac{1}{N}\sum_{j=1}^{N}\left\|w_j^t - w^*\right\|^2 - \frac{1}{k}\mathrm{tr}\left(\tilde{W}^t\right) + \left(\hat{x}^t\right)^\top \delta \hat{x}^t$$

$$\leq \frac{1}{N}\sum_{j=1}^{N}\left\|w_j^t - w^*\right\|^2 - \frac{1}{k}\mathrm{tr}\left(W^t + \delta\right) + \lambda_1\left(\delta\right)$$

$$\leq \frac{1}{N}\sum_{j=1}^{N}\left\|w_j^t - w^*\right\|^2 - \frac{\alpha_{\min}}{k}\frac{1}{N}\sum_{j=1}^{N}\left\|w_j^t - w^*\right\|^2 + \lambda_1\left(\delta\right) - \frac{1}{k}\lambda_1\left(\delta\right)$$

$$= \left(1 - \frac{\alpha_{\min}}{k}\right)\frac{1}{N}\sum_{j=1}^{N}\left\|w_j^t - w^*\right\|^2 + \left(1 - \frac{1}{k}\right)\lambda_1\left(\delta\right)$$

$$\leq \left(1 - \frac{\alpha_{\min}}{k}\right)^{t+1}\frac{1}{N}\sum_{j=1}^{N}\left\|w_j^0 - w^*\right\|^2 + \left(1 - \frac{1}{k}\right)\lambda_1\left(\delta\right)\sum_{s=0}^{t}\left(1 - \frac{\alpha_{\min}}{k}\right)^s$$

$$\leq \left(1 - \frac{\alpha_{\min}}{k}\right)^{t+1}\frac{1}{N}\sum_{j=1}^{N}\left\|w_j^0 - w^*\right\|^2 + \left(1 - \frac{1}{k}\right)\lambda_1\left(\delta\right)\frac{1}{1 - \left(1 - \frac{\alpha_{\min}}{k}\right)}$$

$$= \left(1 - \frac{\alpha_{\min}}{k}\right)^{t+1} \frac{1}{N} \sum_{j=1}^{N} \left\|w_j^0 - w^*\right\|^2 + \frac{k-1}{\alpha_{\min}} \lambda_1(\delta)$$

$$\leq \frac{\epsilon}{2} + \frac{\epsilon}{2} = \epsilon,$$

where $\alpha_{\min} := \min_{t,j} \alpha_j^t$. Thus for

$$t + 1 \geq \left(\log \frac{1}{1 - \frac{\alpha_{\min}}{k}}\right)^{-1} \log \frac{\frac{2}{N} \sum_{i=1}^{N} \left\|w_i^0 - w^*\right\|^2}{\epsilon}$$

and $\lambda_1(\delta) \leq \frac{\alpha_{\min}\epsilon}{2(k-1)}$, we have $\frac{1}{N} \sum_{i=1}^{N} \left\|w_i^{t+1} - w^*\right\|^2 \leq \epsilon$. $\qquad \square$

Theorem 5. *Consider the classroom model given by* (18). *Let* $k := \max_t \{\mathrm{rank}(W^t)\}$ *where* W^t *is given by* (23). *Define* $\alpha_{\min} := \min_{t,j} \alpha_j^t$, *and* $\eta_{\mathrm{avg}} := \frac{1}{N} \sum_{j=1}^{N} \eta_j$, *where* α_j^t *given by* (21). *Then for the teaching strategy given by* (19) *and for* $\beta^{-1} \leq \frac{\alpha_{\min}}{4\eta_{\mathrm{avg}} d^2} \epsilon$, *after* $t = \mathcal{O}\left(\left(\log \frac{1}{1 - \frac{\alpha_{\min}}{k}}\right)^{-1} \log \frac{1}{\epsilon}\right)$ *rounds, we have* $\mathbb{E}\left[\frac{1}{N} \sum_{i=1}^{N} \left\|w_i^t - w^*\right\|^2\right] \leq \epsilon$.

Proof. For the student $j \in [N]$ with the update rule $w_j^{t+1} \leftarrow \mathtt{Proj}_{\mathcal{W}}\left(w_j^t - \eta_j G\left(w_j^t; x, y\right) + \sqrt{2\eta_j \beta^{-1}} \xi_j^t\right)$ (where $\xi_j^t \sim \mathcal{N}(0, I)$, and $\beta > 0$) and any input example $(x^t, y^t) \in \mathcal{X} \times \mathcal{Y}$ (with $y^t = \langle w^*, x^t \rangle$) we have

$$\left\|w_j^{t+1} - w^*\right\|^2 \overset{(i)}{\leq} \left\|w_j^t - \eta_j G\left(w_j^t; x^t, y^t\right) + \sqrt{2\eta_j \beta^{-1}} \xi_j^t - w^*\right\|^2$$

$$= \left\|w_j^t - w^*\right\|^2 + \left\|\sqrt{2\eta_j \beta^{-1}} \xi_j^t - \eta_j G\left(w_j^t; x^t, y^t\right)\right\|^2 + 2\left\langle w_j^t - w^*, \sqrt{2\eta_j \beta^{-1}} \xi_j^t - \eta_j G\left(w_j^t; x^t, y^t\right)\right\rangle$$

$$= \left\|w_j^t - w^*\right\|^2 + 2\eta_j \beta^{-1} \left\|\xi_j^t\right\|^2 + \eta_j^2 \left\|G\left(w_j^t; x^t, y^t\right)\right\|^2 - 2\left\langle \sqrt{2\eta_j \beta^{-1}} \xi_j^t, \eta_j G\left(w_j^t; x^t, y^t\right)\right\rangle$$

$$+ 2\left\langle w_j^t - w^*, \sqrt{2\eta_j \beta^{-1}} \xi_j^t\right\rangle - 2\left\langle w_j^t - w^*, \eta_j G\left(w_j^t; x^t, y^t\right)\right\rangle$$

$$\overset{(ii)}{=} \left\|w_j^t - w^*\right\|^2 + 2\eta_j \beta^{-1} \left\|\xi_j^t\right\|^2 + \eta_j^2 \left(\langle w_j^t, x^t \rangle - y^t\right)^2 \left\|x^t\right\|^2 - 2\eta_j \sqrt{2\eta_j \beta^{-1}} \left(\langle w_j^t, x^t \rangle - y^t\right) \langle \xi_j^t, x^t \rangle$$

$$+ 2\sqrt{2\eta_j \beta^{-1}} \langle w_j^t - w^*, \xi_j^t \rangle - 2\eta_j \left(\langle w_j^t, x^t \rangle - y^t\right) \langle w_j^t - w^*, x^t \rangle$$

$$= \left\|w_j^t - w^*\right\|^2 + \eta_j \langle w_j^t - w^*, x^t \rangle^2 \left(\eta_j \left\|x^t\right\|^2 - 2\right) + 2\eta_j \beta^{-1} \left\|\xi_j^t\right\|^2$$

$$- 2\eta_j \sqrt{2\eta_j \beta^{-1}} \langle w_j^t - w^*, x^t \rangle \langle \xi_j^t, x^t \rangle + 2\sqrt{2\eta_j \beta^{-1}} \langle w_j^t - w^*, \xi_j^t \rangle, \tag{31}$$

where (i) is by the property of projection, and (ii) is due to the fact that $G(w; x, y) = (\langle w, x \rangle - y) \cdot x$ for the squared loss function. Let the history up to time t be $H_t := \left(\{w_j^s\}_{s=1}^t : \forall j \in [N]\right)$, and define $\mathbb{E}_t[\cdot] := \mathbb{E}[\cdot \mid H_t]$. Suppose the teacher constructs the example $x^t = \gamma_t \hat{x}^t$ (with $\|\hat{x}^t\| = 1$) and $y^t = \langle w^*, x^t \rangle$ based on the history H_t. Then given H_t, only ξ_j^t and w_j^{t+1} are random variables in the above equation *i.e.* we have

$$\mathbb{E}_t\left[\left\|w_j^{t+1} - w^*\right\|^2\right] \leq \left\|w_j^t - w^*\right\|^2 + \eta_j \langle w_j^t - w^*, x^t \rangle^2 \left(\eta_j \left\|x^t\right\|^2 - 2\right) + 2\eta_j \beta^{-1} \mathbb{E}_t\left[\left\|\xi_j^t\right\|^2\right]$$

$$- 2\eta_j \sqrt{2\eta_j \beta^{-1}} \langle w_j^t - w^*, x^t \rangle \langle \mathbb{E}_t[\xi_j^t], x^t \rangle + 2\sqrt{2\eta_j \beta^{-1}} \langle w_j^t - w^*, \mathbb{E}_t[\xi_j^t] \rangle$$

$$\overset{(i)}{=} \left\|w_j^t - w^*\right\|^2 + \eta_j \langle w_j^t - w^*, x^t \rangle^2 \left(\eta_j \left\|x^t\right\|^2 - 2\right) + 2\eta_j \beta^{-1} d$$

$$= \left\|w_j^t - w^*\right\|^2 - \eta_j \gamma_t^2 \left(2 - \eta_j \gamma_t^2\right) \langle w_j^t - w^*, \hat{x}^t \rangle^2 + 2\eta_j \beta^{-1} d$$

$$\overset{(ii)}{=} \left\|w_j^t - w^*\right\|^2 - \alpha_j^t \langle w_j^t - w^*, \hat{x}^t \rangle^2 + 2\eta_j \beta^{-1} d,$$

where (i) is by the facts that $\mathbb{E}_t\left[\left\|\xi_j^t\right\|^2\right] = \mathrm{tr}(I) = d$ and $\mathbb{E}_t[\xi_j^t] = 0$, and (ii) is due to $\alpha_j^t := \eta_j \gamma_t^2 \left(2 - \eta_j \gamma_t^2\right)$ and $\gamma_t^2 \leq \frac{2}{\eta_j}, \forall j \in [N]$. Thus for the classroom of students, we have

$$\mathbb{E}_t\left[\frac{1}{N} \sum_{j=1}^{N} \left\|w_j^{t+1} - w^*\right\|^2\right] \leq \frac{1}{N} \sum_{j=1}^{N} \left\|w_j^t - w^*\right\|^2 - \frac{1}{N} \sum_{j=1}^{N} \alpha_j^t \langle w_j^t - w^*, \hat{x}^t \rangle^2 + \frac{2\beta^{-1} d}{N} \sum_{j=1}^{N} \eta_j$$

$$= \frac{1}{N} \sum_{j=1}^{N} \|w_j^t - w^*\|^2 - \frac{1}{N} \sum_{j=1}^{N} \alpha_j^t \langle \hat{w}_j^t, \hat{x}^t \rangle^2 + \frac{2\beta^{-1}d}{N} \sum_{j=1}^{N} \eta_j$$

$$= \frac{1}{N} \sum_{j=1}^{N} \|w_j^t - w^*\|^2 - \frac{1}{N} \sum_{j=1}^{N} \alpha_j^t \left(\hat{x}^t\right)^\top \hat{w}_j^t \left(\hat{w}_j^t\right)^\top \hat{x}^t + \frac{2\beta^{-1}d}{N} \sum_{j=1}^{N} \eta_j$$

$$= \frac{1}{N} \sum_{j=1}^{N} \|w_j^t - w^*\|^2 - \left(\hat{x}^t\right)^\top W^t \hat{x}^t + \frac{2\beta^{-1}d}{N} \sum_{j=1}^{N} \eta_j, \tag{32}$$

where $\hat{w}_j^t := \left(w_j^t - w^*\right)$ and $W^t := \frac{1}{N} \sum_{j=1}^{N} \alpha_j^t \hat{w}_j^t \left(\hat{w}_j^t\right)^\top$. The teacher constructs the example $\hat{x}^t$ as follows (depending only on H_t):

$$\hat{x}^t := \arg\max_{x:\|x\|=1} x^\top W^t x = e_1\left(W^t\right).$$

For this example, from (26), we have

$$\left(\hat{x}^t\right)^\top W^t \hat{x}^t \geq \frac{\alpha_{\min}}{k} \cdot \frac{1}{N} \sum_{j=1}^{N} \|w_j^t - w^*\|^2,$$

where $\alpha_{\min} := \min_{t,j} \alpha_j^t$. By using the above inequality in (32), we get

$$\mathbb{E}_t \left[\frac{1}{N} \sum_{j=1}^{N} \|w_j^{t+1} - w^*\|^2 \right] \leq \left(1 - \frac{\alpha_{\min}}{k}\right) \frac{1}{N} \sum_{j=1}^{N} \|w_j^t - w^*\|^2 + \frac{2\beta^{-1}d}{N} \sum_{j=1}^{N} \eta_j.$$

Then by the law of total expectation and the above recurrence relationship, we have

$$\mathbb{E} \left[\frac{1}{N} \sum_{j=1}^{N} \|w_j^{t+1} - w^*\|^2 \right] = \mathbb{E}_0 \left[\frac{1}{N} \sum_{j=1}^{N} \|w_j^{t+1} - w^*\|^2 \right]$$

$$= \mathbb{E}_0 \mathbb{E}_1 \cdots \mathbb{E}_{t-1} \mathbb{E}_t \left[\frac{1}{N} \sum_{j=1}^{N} \|w_j^{t+1} - w^*\|^2 \right]$$

$$\leq \left(1 - \frac{\alpha_{\min}}{k}\right) \mathbb{E}_0 \mathbb{E}_1 \cdots \mathbb{E}_{t-1} \left[\frac{1}{N} \sum_{j=1}^{N} \|w_j^t - w^*\|^2 \right] + \frac{2\beta^{-1}d}{N} \sum_{j=1}^{N} \eta_j$$

$$\leq \left(1 - \frac{\alpha_{\min}}{k}\right)^{t+1} \frac{1}{N} \sum_{j=1}^{N} \|w_j^0 - w^*\|^2 + \frac{2\beta^{-1}d}{N} \sum_{j=1}^{N} \eta_j \frac{1}{1 - \left(1 - \frac{\alpha_{\min}}{k}\right)}$$

$$= \left(1 - \frac{\alpha_{\min}}{k}\right)^{t+1} \frac{1}{N} \sum_{j=1}^{N} \|w_j^0 - w^*\|^2 + \frac{2\beta^{-1}k d \eta_{\text{avg}}}{\alpha_{\min}}$$

$$\leq \frac{\epsilon}{2} + \frac{\epsilon}{2} = \epsilon.$$

That is for $\beta^{-1} \leq \frac{\alpha_{\min}}{4\eta_{\text{avg}}kd}\epsilon$ and after

$$t + 1 \geq \left(\log \frac{1}{1 - \frac{\alpha_{\min}}{k}} \right)^{-1} \log \frac{\frac{2}{N} \sum_{j=1}^{N} \|w_j^0 - w^*\|^2}{\epsilon}$$

iterations we get $\mathbb{E}\left[\frac{1}{N} \sum_{j=1}^{N} \|w_j^{t+1} - w^*\|^2 \right] \leq \epsilon$. $\qquad \square$

Re-scalable Pool based Teaching under Squared Loss

Here we restrict the teacher to select examples only from

$$\mathcal{X} := \{x : \|x\| \leq R, x = \gamma x_i, x_i \in \mathcal{D}, \gamma \in \mathbb{R}\}$$

$$\mathcal{Y} := \mathbb{R} \text{ (regression) or } \{-1, 1\} \text{ (classification)},$$

where $\mathcal{D} := \{x_1, \ldots, x_m : \|x_i\| = 1, \forall i \in [m]\}$ is a pool of directions. For teaching to be effective, the pool should contain *rich enough* directions.

Single Learner

For the student $j \in [N]$ with the update rule $w_j^{t+1} \leftarrow \mathtt{Proj}_{\mathcal{W}} \left(w_j^t - \eta_j G \left(w_j^t; x, y \right) \right)$ and any input example $(x^t, y^t) \in \mathcal{X} \times \mathcal{Y}$ (with $x^t = \gamma_t \hat{x}^t$, $y^t = \langle w^*, x^t \rangle$, and $\|\hat{x}^t\| = 1$) we have

$$\left\| w_j^{t+1} - w^* \right\|^2 \leq \left\| w_j^t - w^* \right\|^2 - \eta_j \gamma_t^2 \left(2 - \eta_j \gamma_t^2 \right) \left\langle w_j^t - w^*, \hat{x}^t \right\rangle^2. \tag{33}$$

Given a pool of unit vector directions $\mathcal{D} := \{x_1, \ldots, x_m : \|x_i\| = 1, \forall i \in [m]\}$, the teacher constructs the example as follows

$$\hat{x}^t := \arg\max_{x \in \mathcal{D}} \left\langle w_j^t - w^*, x \right\rangle.$$

Let the optimal example of synthesis based teaching be $\hat{x}_{\text{syn}}^t = \frac{w_j^t - w^*}{\|w_j^t - w^*\|}$. Then for some $a_t, b_t \in \mathbb{R}$, we can decompose the example $\hat{x}^t$ as follows

$$\hat{x}^t = \frac{a_t}{\sqrt{a_t^2 + b_t^2}} \hat{x}_{\text{syn}}^t + \frac{b_t}{\sqrt{a_t^2 + b_t^2}} \left(\hat{x}_{\text{syn}}^t \right)_\perp.$$

If the pool is rich enough we would have $\left| \frac{a_t}{\sqrt{a_t^2 + b_t^2}} \right| \approx 1$. Consider

$$
\begin{aligned}
\left\langle w_j^t - w^*, \hat{x}^t \right\rangle^2 &= \left\langle w_j^t - w^*, \frac{a_t}{\sqrt{a_t^2 + b_t^2}} \hat{x}_{\text{syn}}^t + \frac{b_t}{\sqrt{a_t^2 + b_t^2}} \left(\hat{x}_{\text{syn}}^t \right)_\perp \right\rangle^2 \\
&= \left\langle w_j^t - w^*, \frac{a_t}{\sqrt{a_t^2 + b_t^2}} \hat{x}_{\text{syn}}^t \right\rangle^2 \\
&= \frac{a_t^2}{a_t^2 + b_t^2} \left\langle w_j^t - w^*, \hat{x}_{\text{syn}}^t \right\rangle^2 \\
&= \frac{a_t^2}{a_t^2 + b_t^2} \left\| w_j^t - w^* \right\|^2.
\end{aligned}
$$

Then by applying the above equality in (33), we get

$$
\begin{aligned}
\left\| w_j^{t+1} - w^* \right\|^2 &\leq \left\| w_j^t - w^* \right\|^2 - \eta_j \gamma_t^2 \left(2 - \eta_j \gamma_t^2 \right) \frac{a_t^2}{a_t^2 + b_t^2} \left\| w_j^t - w^* \right\|^2 \\
&\leq \left\| w_j^t - w^* \right\|^2 - \alpha_j^t \left\| w_j^t - w^* \right\|^2 \\
&\leq \left\| w_j^t - w^* \right\|^2 - \alpha_j \left\| w_j^t - w^* \right\|^2 \\
&\leq (1 - \alpha_j)^{t+1} \left\| w_j^0 - w^* \right\|^2,
\end{aligned}
$$

where $\alpha_j^t := \frac{a_t^2}{a_t^2 + b_t^2} \eta_j \gamma_t^2 \left(2 - \eta_j \gamma_t^2 \right)$, and $\alpha_j := \min_t \alpha_j^t$.

Classroom Setting

For the classroom, from (25), we have

$$\frac{1}{N} \sum_{j=1}^N \left\| w_j^{t+1} - w^* \right\|^2 \leq \frac{1}{N} \sum_{j=1}^N \left\| w_j^t - w^* \right\|^2 - \left(\hat{x}^t \right)^\top W^t \hat{x}^t. \tag{34}$$

Given a pool of unit vector directions $\mathcal{D} := \{x_1, \ldots, x_m : \|x_i\| = 1, \forall i \in [m]\}$, the teacher constructs the example as follows

$$\hat{x}^t := \arg\max_{x \in \mathcal{D}} x^\top W^t x.$$

Since W^t is a symmetric positive semidefinite matrix:

- it has orthogonal eigenvectors, i.e., $e_i \left(W^t \right) \perp e_j \left(W^t \right), \forall i, j \in [d] ; i \neq j$.
- the eigenvectors span $\mathbb{R}^d$ i.e. $\text{span} \left\{ e_i \left(W^t \right) : i \in [d] \right\} = \mathbb{R}^d$.
- for $i \neq j$: $e_i \left(W^t \right)^\top W^t e_j \left(W^t \right) = e_i \left(W^t \right)^\top \lambda_j \left(W^t \right) e_j \left(W^t \right) = 0$.

Let the optimal example of synthesis based teaching be $\hat{x}^t_{\text{syn}} = e_1(W^t)$. Then for some $a_{t,i} \in \mathbb{R}, i \in [d]$, we can decompose the example $\hat{x}^t$ as follows

$$\hat{x}^t = \frac{1}{\sqrt{\sum_{i=1}^d a_{t,i}^2}} \sum_{i=1}^d a_{t,i} e_i(W^t).$$

If the pool is rich enough we would have $\left| \frac{a_{t,1}}{\sqrt{\sum_{i=1}^d a_{t,i}^2}} \right| \approx 1$. Consider

$$
\begin{aligned}
\left(\hat{x}^t\right)^\top W^t \hat{x}^t &= \frac{1}{\sum_{i=1}^d a_{t,i}^2} \left(\sum_{i=1}^d a_{t,i} e_i(W^t)\right)^\top W^t \left(\sum_{i=1}^d a_{t,i} e_i(W^t)\right) \\
&= \frac{1}{\sum_{i=1}^d a_{t,i}^2} \left\{ \sum_{i=1}^d a_{t,i}^2 e_i(W^t)^\top W^t e_i(W^t) + \sum_{i \neq j} a_{t,i} a_{t,j} e_i(W^t)^\top W^t e_j(W^t) \right\} \\
&= \frac{1}{\sum_{i=1}^d a_{t,i}^2} \sum_{i=1}^d a_{t,i}^2 e_i(W^t)^\top W^t e_i(W^t) \\
&\geq \frac{a_{t,1}^2}{\sum_{i=1}^d a_{t,i}^2} e_1(W^t)^\top W^t e_1(W^t) \\
&= \frac{a_{t,1}^2}{\sum_{i=1}^d a_{t,i}^2} \left(\hat{x}^t_{\text{syn}}\right)^\top W^t \hat{x}^t_{\text{syn}} \\
&\geq \frac{a_{t,1}^2}{\sum_{i=1}^d a_{t,i}^2} \cdot \frac{\alpha_{\min}}{k} \cdot \frac{1}{N} \sum_{j=1}^N \left\| w_j^t - w^* \right\|^2,
\end{aligned}
$$

where last inequality is from (26). Then by applying the above inequality in (34), we get

$$
\begin{aligned}
\frac{1}{N} \sum_{j=1}^N \left\| w_j^{t+1} - w^* \right\|^2 &\leq \left(1 - \frac{a_{t,1}^2}{\sum_{i=1}^d a_{t,i}^2} \cdot \frac{\alpha_{\min}}{k}\right) \frac{1}{N} \sum_{j=1}^N \left\| w_j^t - w^* \right\|^2 \\
&\leq \left(1 - a_{\min} \cdot \frac{\alpha_{\min}}{k}\right) \frac{1}{N} \sum_{j=1}^N \left\| w_j^t - w^* \right\|^2 \\
&\leq \left(1 - a_{\min} \cdot \frac{\alpha_{\min}}{k}\right)^{t+1} \frac{1}{N} \sum_{j=1}^N \left\| w_j^0 - w^* \right\|^2,
\end{aligned}
\tag{35}
$$

where $a_{\min} := \min_t \frac{a_{t,1}^2}{\sum_{i=1}^d a_{t,i}^2}$.