Iterative Machine Teaching

Weiyang Liu [1] **Bo Dai** [1] **Ahmad Humayun** [1] **Charlene Tay** [2] **Chen Yu** [2]
Linda B. Smith [2] **James M. Rehg** [1] **Le Song** [1]

Abstract

In this paper, we consider the problem of machine teaching, the inverse problem of machine learning. Different from traditional machine teaching which views the learners as batch algorithms, we study a new paradigm where the learner uses an iterative algorithm and a teacher can feed examples sequentially and intelligently based on the current performance of the learner. We show that the teaching complexity in the iterative case is very different from that in the batch case. Instead of constructing a minimal training set for learners, our iterative machine teaching focuses on achieving fast convergence in the learner model. Depending on the level of information the teacher has from the learner model, we design teaching algorithms which can provably reduce the number of teaching examples and achieve faster convergence than learning without teachers. We also validate our theoretical findings with extensive experiments on different data distribution and real image datasets.

1. Introduction

Machine teaching is the problem of constructing an optimal (usually minimal) dataset according to a target concept such that a student model can learn the target concept based on this dataset. Recently, there is a surge of interests in machine teaching which has found diverse applications in model compression (Bucila et al., 2006; Han et al., 2015; Ba & Caruana, 2014; Romero et al., 2014), transfer learning (Pan & Yang, 2010) and cyber-security problems (Alfeld et al., 2016; 2017; Mei & Zhu, 2015). Furthermore, machine teaching is also closely related to other subjects of interests, such as curriculum learning (Bengio et al., 2009) and knowledge distilation (Hinton et al., 2015).

[1]Georgia Institute of Technology [2]Indiana University. Correspondence to: Weiyang Liu <wyliu@gatech.edu>, Le Song <lsong@cc.gatech.edu>.

Proceedings of the 34^{th} International Conference on Machine Learning, Sydney, Australia, PMLR 70, 2017.

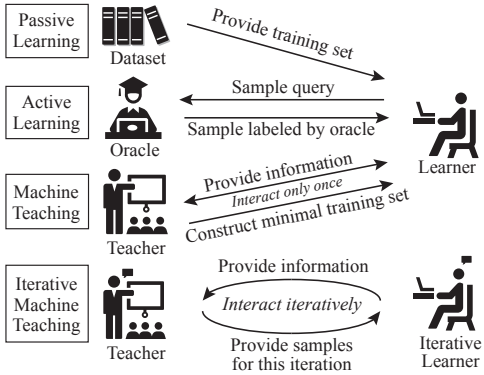


Figure 1. Comparison between iterative machine teaching and the other learning paradigms.

In the traditional machine learning paradigm, a teacher will typically construct a batch set of examples, and provide them to a learning algorithm in one shot; then the learning algorithm will work on this batch dataset trying to learn the target concept. Thus, many research work under this topic try to construct the smallest such dataset, or characterize the size of of such dataset, called the teaching dimension of the student model (Zhu, 2013; 2015). There are also many seminal theory work on analyzing the teaching dimension of different models (Shinohara & Miyano, 1991; Goldman & Kearns, 1995; Doliwa et al., 2014; Liu et al., 2016).

However, in many real world applications, the student model is typically updated via an iterative algorithm, and we get the opportunity to observe the performance of the student model as we feed examples to it. For instance,

- In model compression where we want to transfer a target "teacher model" to a destination "student model", we can constantly observe student model's prediction on current training points. Intuitively, such observations will allow us to get a better estimate where the student model is and pick examples more intelligently to better guide the student model to convergence.
- In cyber-security setting where an attack wants to mislead a recommendation system that learns online, the attacker can constantly generate fake clicks and observe the system's response. Intuitively, such feedback will allow the attacker to figure out the state of the learning system, and design better strategy to mislead the system.

From the aspects of both faster model compression and bet-

ter avoiding hacker attack, we seek to understand some fundamental questions, such as, *what is the sequence of examples that teacher should feed to the student in each iteration in order to achieve fast convergence? And how many such examples or such sequential steps are needed?*

In this paper, we will focus on this new paradigm, called ***iterative machine teaching***, which extends traditional machine teaching from batch setting to iterative setting. In this new setting, the teacher model can communicate with and influence the student model in multiple rounds, but the student model remains passive. More specifically, in each round, the teacher model can observe (potentially different levels of) information about the students to intelligently choose one example, and the student model runs a fixed iterative algorithm using this chosen example.

Furthermore, the smallest number of examples (or rounds) the teacher needs to construct in order for the student to efficiently learn a target model is called the ***iterative teaching dimension*** of the student algorithm. Notice that in this new paradigm, we shift from describing the complexity of a model to the complexity of an algorithm. Therefore, for the same student model, such as logistic regression, the iterative teaching dimension for a teacher model can be different depending on the student's learning algorithms, such as gradient descent versus conjugate gradient descent. In some sense, the teacher in this new setting is becoming active, but not the student. In Fig. 1, we summarize the differences of iterative machine teaching from traditional machine teaching, active learning and passive learning.

Besides introducing the new paradigm, we also propose three iterative teaching algorithms, called omniscient teacher, surrogate teacher and imitation teacher, based on the level of information about the student that the teacher has access to. Furthermore we provide partial theoretical analysis for these algorithms under different example construction schemes. Our analysis shows that under suitable conditions, iterative teachers can always perform better than passive teacher, and achieve exponential improvements. Our analysis also identifies two crucial properties, namely teaching monotonicity and teacher capability, which play critical roles in achieving fast iterative teaching.

To corroborate our theoretical findings, we also conduct extensive experiments on both synthetic data and real image data. In both cases, the experimental results verify our theoretical findings and the effectiveness of our proposed iterative teaching algorithms.

2. Related Work

Machine teaching. Machine teaching problem is to find an optimal training set given a student model and a target. (Zhu, 2015) proposes a general teaching framework. (Zhu, 2013) considers Bayesian learner in exponential family and expresses the machine teaching as an optimization problem over teaching examples that balance the future loss of the learner and the effort of the teacher. (Liu et al., 2016) provides the teaching dimension of several linear learners. The framework has been applied to security (Mei & Zhu, 2015), human computer interaction (Meek et al., 2016) and education (Khan et al., 2011). (Johns et al., 2015) further extends machine teaching to interactive settings. However, these work ignores the fact that a student model is typically learned by an iterative algorithm, and we usually care more about how fast the student can learn from the teacher.

Interactive Machine Learning. (Cakmak & Thomaz, 2014) consider the scenario of a human training an agent to perform a classification task by showing examples. They study how to improve human teacher by giving teaching guidance. (Singla et al., 2014) consider the crowdsourcing problem and propose a sequential teaching algorithm that can teach crowd worker to better classify the query. Both work consider a very different setting where the learner (i.e. human learner) is not iterative and does not have a particular optimization algorithm.

Active learning. Active learning enables a learner to interactively query the oracle to obtain the desired outputs at new samples. Machine teaching is different from active learning in the sense that active learners explore the optimal parameters by itself rather than being guided by the teacher. Therefore they have different sample complexity (Balcan et al., 2010; Zhu, 2013).

Curriculum learning. Curriculum learning (Bengio et al., 2009) is a general training strategy that encourages to input training examples from easy ones to difficult ones. Very interestingly, our iterative teacher model suggests similar training strategy in our experiments.

3. Iterative Machine Teaching

The proposed iterative machine teaching is a general concept, and the paper considers the following settings:

Student's Asset. In general, the asset of a student (learner) includes the initial parameter w_0, loss function, optimization algorithm, representation (feature), model, learning rate η_t over time (and initial η_0) and the trackability of the parameter w^t. The ideal case is that a teacher has access to all of them and can track the parameters and learning rate, while the worst case is that a teacher knows nothing. How practical the teaching is depends on how much the prior knowledge and trackability that a teacher has.

Representation. The teacher represents an example as (x, y) while the student represents the same example as $(\widetilde{x}, \widetilde{y})$ (typically $y = \widetilde{y}$). The representation $x \in \mathcal{X}$ and $\widetilde{x} \in \widetilde{\mathcal{X}}$ can be different but deterministically related. We assume there exists $\widetilde{x} = \mathcal{G}(x)$ for an unknown invertible mapping $\mathcal{G}$.

Model. The teacher uses a linear model $y = \langle v, x \rangle$ with pa-

rameter v^* (w^* for student's space) that is taught to the student. The student also uses a linear model $\widetilde{y}{=}\langle w, \widetilde{x}\rangle$ with parameter w, i.e., $\widetilde{y}{=}\langle w, \mathcal{G}(x)\rangle{=}f(x)$ in general. w and v do not necessarily lie in the same space, but for omniscient teacher, they are equivalent and interchangeably used.

Teaching protocol. In general, the teacher can only communicate with the student via examples. In this paper, the teacher provides one example x^t in one iteration, where t denotes the t-th iteration. The goal of the teacher is to provide examples in each iteration such that the student parameter w converge to its optimum w^* as fast as possible.

Loss function. The teacher and student share the same loss function. We assume this is a convex loss function $\ell(f(x), y)$, and the best model is usually found by minimizing the expected loss below:

$$w^* = \arg\min_w \ \mathbb{E}_{(x,y)} \left[\ell(\langle w, x\rangle, y)\right]. \tag{1}$$

where the sampling distribution $(x, y) \sim \mathbb{P}(x, y)$. Without loss of generality, we only consider typical loss functions, such as square loss $\frac{1}{2}(\langle w, x\rangle - y)^2$, logistic loss $\log(1 + \exp(-y\langle w, x\rangle))$ and hinge loss $\max(1 - y\langle w, x\rangle, 0)$.

Algorithm. The student uses the stochastic gradient descent to optimize the model. The iterative update is

$$w^{t+1} = w^t - \eta_t \frac{\partial \ell(\langle w, x\rangle, y)}{\partial w}. \tag{2}$$

Without teacher's guiding, the student can be viewed as being guided by a random teacher who randomly feed an example to the student in each iteration.

4. Teaching by an Omniscient Teacher

An omniscient teacher has access to the student's feature space, model, loss function and optimization algorithm. In specific, omniscient teacher's (x, y) and student's $(\tilde{x}, \tilde{y})$ share the same representation space, and teacher's optimal model v^* is also the same as student's optimal model w^*.

4.1. Intuition and teaching algorithm

In order to gain intuition on how to make the student model converge faster, we will start with looking into the difference between the current student parameter and the teacher parameter w^* during each iteration:

$$\begin{aligned} \left\|w^{t+1} - w^*\right\|_2^2 &= \left\|w^t - \eta_t \frac{\partial \ell(\langle w, x\rangle, y)}{\partial w} - w^*\right\|_2^2 \\ &= \left\|w^t - w^*\right\|_2^2 + \eta_t^2 \underbrace{\left\|\frac{\partial \ell(\langle w^t, x\rangle, y)}{\partial w^t}\right\|_2^2}_{T_1(x,y|w^t):\text{Difficulty of an example } (x,\,y)} \\ &\quad - 2\eta_t \underbrace{\left\langle w^t - w^*, \frac{\partial \ell(\langle w^t, x\rangle, y)}{\partial w^t}\right\rangle}_{T_2(x,y|w^t):\text{Usefulness of an example } (x,\,y)} \end{aligned} \tag{3}$$

Based on the decomposition of the parameter error, the teacher aims to choose a particular example (x, y) such that $\|w^{t+1} - w^*\|_2^2$ is most reduced compared to $\|w^t - w^*\|_2^2$ from the last iteration. Thus the general strategy for the teacher is to choose an example (x, y), such that $\eta_t^2 T_1 - 2\eta_t T_2$ is minimized in the t-th iteration:

$$\operatorname*{argmin}_{x\in\mathcal{X}, y\in\mathcal{Y}} \eta_t^2 T_1(x, y|w^t) - 2\eta_t T_2(x, y|w^t). \tag{4}$$

The teaching algorithm of omniscient teacher is summarized in Alg.1. The smallest value of $\eta_t^2 T_1 - 2\eta_t T_2$ is $-\|w^t - w^*\|_2^2$. If the teacher achieves this, it means that we have reached the teaching goal after this iteration. However, it usually cannot be done in just one iteration, because of the limitation of teacher's capability to provide examples. T_1 and T_2 have some nice intuitive interpretations:

Difficulty of an example. T_1 quantifies the difficulty level of an example. This interpretation for different loss functions becomes especially clear when the data lives on the surface of a sphere, *i.e.*, $\|x\| = 1$. For instance,

- For linear regression, $T_1 = (\langle w, x\rangle - y)^2$. The larger the norm of gradient is, the more difficult the example is.
- For logistic regression, we have $T_1 = \|\frac{1}{1+\exp(y\langle w,x\rangle)}\|_2^2$. We know that $\frac{1}{1+\exp(y\langle w,x\rangle)}$ is the probability of predicting the wrong label. The larger the number is, the more difficult the example is.
- For support vector machines, we have $T_1 = \frac{1}{2}(\text{sign}(1 - y\langle w, x\rangle) + 1)$. Different from above losses, the hinge loss has a threshold to identify the difficulty of examples. While the example is difficult enough, it will produce 1. Otherwise it is 0.

Interestingly, the difficulty level is not related to the teacher w^*, but is based on the current parameters of the learner w^t. From another perspective, the difficulty level can also be interpreted as the information that an example carries. Essentially, a difficult example is usually more informative. In such sense, our difficulty level has similar interpretation to curriculum learning, but with different expression.

Usefulness of an example. T_2 quantifies the usefulness of an example. Concretely, T_2 is the correlation between discrepancy $w^t - w^*$ and the information (difficulty) of an example. If the information of the example has large correlation with the discrepancy, it means that this example is very useful in this teaching iteration.

Trade-off. Eq.(4) aims to minimize the difficulty level T_1 and maximize the usefulness T_2. In other word, the teacher always prefers easy but useful examples. When the learning rate is large, T_1 term plays a more important role. When learning rate is small, T_2 term plays a more important role. This suggests that initially the teacher should choose easier examples to feed into the student model, and later on the teacher should choose examples to focus more on reducing the discrepancy between $w^t - w^*$. Such examples are very likely the difficult ones. Even if the learning rate is fixed, the gradient $\nabla_w \ell$ is usually large for a convex loss

function at the beginning, so reducing the difficulty level (choosing easy examples) is more important. While near the optimum, the gradient $\nabla_w \ell$ is usually small, so T_2 becomes more important. It is also likely to choose difficult examples. It has nice connection with curriculum learning (easy example first and difficult later) and boosting (gradually focus on difficult examples).

4.2. Teaching monotonicity and universal speedup

Can the omniscient teacher always do better than a teacher who feed random examples to the student (in terms of convergence)? In this section, we identify generic conditions under which we can guarantee that the iterative teaching algorithm always perform better than random teacher.

Definition 1 (Teaching Volume) *For a specific loss function ℓ, we first define a teaching volume function $TV(w)$ with model parameter w as*

$$TV(w) = \max_{x\in\mathcal{X}, y\in\mathcal{Y}} \{-\eta_t^2 T_1(x,y|w) + 2\eta_t T_2(x,y|w)\} \quad (5)$$

Theorem 2 (Teaching Monotonicity) *Given a training set $\mathcal{X}$ and a loss function ℓ, if the inequality*

$$\|w_1 - w^*\|^2 - TV(w_1) \le \|w_2 - w^*\|^2 - TV(w_2) \quad (6)$$

holds for any w_1, w_2 that satisfy $\|w_1 - w^*\|^2 \le \|w_2 - w^*\|^2$, then with the same parameter initialization and learning rate, the omniscient teacher can always converge not slower than random teacher.

The teaching volume represents the teacher's teaching effort in this iteration, so $\|w^t - w^*\|^2 - TV(w^t)$ characterizes the remaining teaching effort needed to achieve the teaching goal after iteration t. Theorem 2 says that for a loss function and a training set, if the remaining teaching effort is monotonically decreasing while the model parameter gets closer to the optimum, we can guarantee that the omniscient teacher can always converge not slower than random teacher. It is a sufficient condition for loss functions to achieve faster convergence than SGD. For example, the square loss satisfies the condition with certain training set:

Proposition 3 *The square loss satisfies the teaching monotonicity condition given the training set $\{x | \|x\| \le R\}$.*

4.3. Teaching capability and exponential speedup

The theorem in previous subsection insures that under certain conditions the omniscient teacher can always lead to faster convergence for the student model, but can there be exponential speedup? To this end, we introduce further assumptions of the "richness" of teaching examples, which we call teaching capability. We start from the ideal case, *i.e.*, the synthesis-based omniscient teacher with hyperspherical feature space, and then, extend to real cases with the restrictions on teacher's knowledge domain, sampling scheme, and student information. We present specific teaching strategies in terms of teaching capability (strong to weak): synthesis, combination and (rescalable) pool.

Synthesis-based teaching. In synthesis-based teaching, the teacher can provide any samples from

$$\mathcal{X} = \{x \in \mathbb{R}^d, \|x\| \le R\}$$
$$\mathcal{Y} = \mathbb{R} \text{ (Regression) or } \{-1, 1\} \text{ (Classification).}$$

Theorem 4 (Exponential Synthesis-based Teaching) *For a synthesis-based omniscient teacher and a student with fixed learning rate $\eta \neq 0$, if the loss function $\ell(\cdot,\cdot)$ satisfies that for any $w \in \mathbb{R}^d$, there exists $\gamma \neq 0$, $|\gamma| \le \frac{R}{\|w - w^*\|}$ such that while $\hat{x} = \gamma(w - w^*)$ and $\hat{y} \in \mathcal{Y}$, we have*

$$0 < \gamma \nabla_{\langle w, \hat{x}\rangle} \ell(\langle w, \hat{x}\rangle, \hat{y}) \le \frac{1}{\eta},$$

then the student can learn an ϵ-approximation of w^* with $\mathcal{O}(C_1^{\gamma,\eta} \log \frac{1}{\epsilon})$ samples. We call such loss function $\ell(\cdot,\cdot)$ exponentially teachable in synthesis-based teaching.

The constant is $C_1^{\gamma,\eta} = (\log \frac{1}{1-\eta\nu(\gamma)})^{-1}$ in which $\nu(\gamma) := \min_{w,y} \gamma \nabla_{\langle w, \hat{x}\rangle} \ell(\langle w, \hat{x}\rangle, y) > 0$. $\nu(\gamma)$ is related to the convergence speed. Note that the sample complexity serves as the iterative teaching dimension corresponding to this particular teacher, student, algorithm and training data.

The sample complexity in iterative teaching is *deterministic*, different from the high probability bounds of traditional sample complexity with random *i.i.d.*samples or actively required samples. This is because the teacher provides the samples deterministically without noise in every iteration.

The radius R for $\mathcal{X}$, which can be interpreted as the knowledge domain of the teacher, will affect the sample complexity by constraining the valid values of γ, and thus $C_1^{\gamma,\eta}$. For example, for absolute loss, if R is large, such that $\frac{1}{\eta} \le \frac{R}{\|w^0 - w^*\|}$, γ can be set to $\frac{1}{\eta}$ and the $\nu(\gamma)$ will be $\frac{1}{\eta}$ in this case. Therefore, we have $C_1^{\gamma,\eta} = 0$, which means the student can learn with only one example (one iteration). However, if $\frac{1}{\eta} > \frac{R}{\|w^0 - w^*\|}$, we have $C_1^{\gamma,\eta} > 0$, and the student can converge exponentially. The similar phenomenon appears in the square loss, hinge loss, and logistic loss. Refer to Appendix A for details.

The exponential synthesis-based teaching is closely related to Lipschitz smoothness and strong convexity of loss functions in the sense that the two regularities provide positive lower and upper bound for $\gamma \nabla_{\langle w, x\rangle} \ell(\langle w, x\rangle, y)$.

Proposition 5 *The Lipschitz smooth and strongly convex loss functions are exponentially teachable in synthesis-based teaching.*

The exponential synthesis-based teachability is a weaker condition compared to the strong convexity and Lipschitz smoothness. We can show that besides the Lipschitz smooth and strongly convex loss, there are some other loss functions, which are not strongly convex, but still are exponentially teachable in synthesis-based scenario, *e.g.*, the hinge loss and logistic loss. Proofs are in Appendix A.

Combination-based teaching. In this scenario, the teacher

Algorithm 1 The omniscient teacher

1: Randomly initialize the student and teacher parameter w^0;
2: Set $t=1$ and the maximal iteration number T;
3: **while** w^t has not converged or $t<T$ **do**
4: Solve the optimization (e.g., pool-based teaching):

$$(x^t,y^t)=\underset{x\in\mathcal{X},y\in\mathcal{Y}}{\operatorname{argmin}}\ \eta_t^2\left\|\frac{\partial\ell\left(\left\langle w^{t-1},x\right\rangle,y\right)}{\partial w^{t-1}}\right\|^2 - 2\eta_t\left\langle w^{t-1}-w^*,\frac{\partial\ell\left(\left\langle w^{t-1},x\right\rangle,y\right)}{\partial w^{t-1}}\right\rangle$$

5: Use the selected example (x^t,y^t) to perform the update:

$$w^t=w^{t-1}-\eta_t\frac{\partial\ell\left(\left\langle w^{t-1},x^t\right\rangle,y^t\right)}{\partial w^{t-1}}.$$

6: $t\leftarrow t+1$
7: **end while**

can provide examples from ($\alpha_i\in\mathbb{R}$)

$$\mathcal{X}=\left\{x|\|x\|\le R, x=\sum_{i=1}^m\alpha_i x_i, x_i\in\mathcal{D}\right\}, \mathcal{D}=\{x_1,\dots,x_m\}$$

$\mathcal{Y}=\mathbb{R}$ (Regression) or $\{-1,1\}$ (Classification)

Corollary 6 *For a combination-based omniscient teacher and a student with fixed learning rate $\eta\neq 0$ and initialization w^0, if the loss function is exponentially synthesis-based teachable and $w^0-w^*\in span\,(\mathcal{D})$, the student can learn an ϵ-approximation of w^* with $\mathcal{O}\left(C_1^{\gamma,\eta}\log\frac{1}{\epsilon}\right)$ samples.*

Although the knowledge pool of teacher is more restricted compared to the synthesis-based scenario, with teacher's extra work to combine samples, the teacher can behave the same as the most knowledgable synthesis-based teacher.

Rescalable pool-based teaching. This scenario is further restricted in both knowledge pool and the effort to prepare samples. The teacher can provide examples from $\mathcal{X}\times\mathcal{Y}$:

$\mathcal{X}=\{x|\|x\|\le R, x=\gamma x_i, x_i\in\mathcal{D},\gamma\in\mathbb{R}\}, \mathcal{D}=\{x_1,\dots\}$
$\mathcal{Y}=\mathbb{R}$ (Regression) or $\{-1,1\}$ (Classification)

In such scenario, we cannot get arbitrary direction rather than the samples from the candidate pool. Therefore, to achieve the exponential improvement, the candidate pool should contain rich enough directions. To characterize the richness in finite case, we define the *pool volume* as

Definition 7 (Pool Volume) *Given the training example pool $\mathcal{X}\in\mathbb{R}^d$, the volume of $\mathcal{X}$ is defined as*

$$\mathcal{V}(\mathcal{X}):=\min_{w\in span(\mathcal{D})}\max_{x\in\mathcal{X}}\frac{\langle w,x\rangle}{\|w\|^2}.$$

Obviously, for the candidate pool of the synthesis-based teacher, we have $\mathcal{V}(\mathcal{X})=1$. In general, for finite candidate pool, the pool volume is $0<\mathcal{V}(\mathcal{X})<1$.

Theorem 8 *For a rescalable pool-based omniscient teacher and a student with fixed learning rate $\eta\neq 0$ and initialization w^0, if for any $w\in\mathbb{R}^d$, $w\not\perp w^*$ and $w^0-w^*\in span\,(\mathcal{D})$, there exists $\{x,y\}\in\mathcal{X}\times\mathcal{Y}$ and γ such that while $\hat{x}=\frac{\gamma\|w-w^*\|}{\|x\|}x, \hat{y}=y$, we have*

$$0<\gamma\nabla_{\langle w,\hat{x}\rangle}\ell\left(\langle w,\hat{x}\rangle,\hat{y}\right)<\frac{2\mathcal{V}(\mathcal{X})}{\eta},$$

then the student can learn an ϵ-approximation of w^* with $\mathcal{O}(C_2^{\eta,\gamma,\mathcal{V}(\mathcal{X})}\log\frac{1}{\epsilon})$ samples. We say such loss function is exponentially teachable in rescalable pool-based teaching.

The pool volume plays a vital role in pool-based teaching. It not only affects the existence of γ and $\{\hat{x},\hat{y}\}$ to satisfy the conditions, but also changes the convergence rate. While $\mathcal{V}(\mathcal{X})$ increases, $C_2^{\eta,\gamma,\mathcal{V}(\mathcal{X})}$ will decrease, yielding smaller sample complexity. With $\mathcal{V}(\mathcal{X})<1$, the rescalable pool-based teaching requires more samples than the synthesis-based teaching. As $\mathcal{V}(\mathcal{X})$ increases to 1, the candidate pool becomes $\left\{x\in\mathbb{R}^d,\|x\|\le R\right\}$ and $C_2^{\eta,\gamma,\mathcal{V}(\mathcal{X})}$ approaches to $C_1^{\gamma,\eta}$. Then the convergence speed of rescalable pool-based teaching approaches to the synthesis/combination-based teaching.

5. Teaching by a less informative teacher

To make the teacher model useful in practice, we further design two less informative teacher model that requires less and less information from the student.

5.1. The surrogate teacher

Suppose we can only query the function output from the learned $\langle w^t,x\rangle$, but we can not directly access w^t. How can we choose the example? In this case we propose to make use of the the convexity of the loss function. That is

$$\left\langle w^t-w^*,\frac{\partial\ell(\langle w^t,x\rangle,y)}{\partial w^t}x\right\rangle\geqslant\ell(\langle w^t,x\rangle,y)-\ell(\langle w^*,x\rangle,y). \tag{7}$$

Taking the pool-based teaching as an example, we can instead optimize the following surrogate loss function:

$$(x^t,y^t)=\underset{\{x,y\}\in\mathcal{X}}{\operatorname{argmin}}\ \eta_t^2\left\|\frac{\partial\ell(\langle w^t,x\rangle,y)}{\partial w^t}\right\|_2^2 - 2\eta_t\left(\ell(\langle w^t,x\rangle,y)-\ell(\langle w^*,x\rangle,y)\right) \tag{8}$$

by replacing $\left\langle w^t-w^*,\frac{\partial\ell(\langle w^t,x\rangle,y)}{\partial w^t}\right\rangle$ with its lower bound. The advantage of this approach is that the teacher only need to query the learner for the function output $\langle w^t,x\rangle$ to choose the example, without the need to access the learner parameter w^t directly. Furthermore, after noticing that in this formulation, the teacher makes prediction via inner products, we find that the surrogate teacher can also be applied to the scenario where the teacher and the student use different feature spaces by further replacing $(\ell(\langle w^t,x\rangle,y)-\ell(\langle w^*,x\rangle,y))$ with $(\ell(\langle w^t,x\rangle,y)-\ell(\langle v^*,\widetilde{x}\rangle,y))$. With this modification, we can provide examples without using information about w^*. The performance of the surrogate teacher largely depends on the tightness of such convexity lower bound.

Algorithm 2 The imitation teacher

1: Randomly initialize the student parameter w^0 and the teacher parameter v^0; Randomly select a training sample (x^0, y^0);
2: Set $t = 1$ and the maximal iteration number T;
3: **while** w^t has not converged or $t < T$ **do**
4: Perform the update:

$$v^t = v^{t-1} - \eta_v \left(\left\langle v^{t-1}, x^{t-1} \right\rangle - \left\langle w^t, x^{t-1} \right\rangle \right) x^{t-1}.$$

5: Solve the optimization (e.g., pool-based teaching):

$$(x^t, y^t) = \underset{x \in \mathcal{X}, y \in \mathcal{Y}}{\operatorname{argmin}} \ \eta_t^2 \left\| \frac{\partial \ell \left(\left\langle w^t, x \right\rangle, y \right)}{\partial v^t} \right\|^2 - 2\eta_t \left\langle v^t - v^*, \frac{\partial \ell \left(\left\langle v^t, x \right\rangle, y \right)}{\partial v^t} \right\rangle.$$

6: Provide the selected example (x^t, y^t) for the student to perform the update ;

$$w^{t+1} = w^t - \eta_t \frac{\partial \ell \left(\left\langle w^t, x \right\rangle, y \right)}{\partial w}.$$

7: $t \leftarrow t + 1$
8: **end while**

5.2. The imitation teacher

When the teacher and the student have different feature spaces, this teaching setting will be much closer to practice than all the previous settings and also more challenging. To this end, we present an imitation teacher who learns to imitate the inner product output $\langle w^t, x \rangle$ of the student model and simultaneously choose examples in teacher's own feature space. The teacher can possibly use active learning to imitate the student's $\langle w^t, x \rangle$. In this imitation, the student model stays unchanged and the teacher model could update itself via multiple queries to the student (input an example and see the inner product output of the student). We present a more simple and straightforward imitation teacher (Alg. 2) which works in a way similar to stochastic mirror descent (Nemirovski et al., 2009; Hall & Willett, 2013). In specific, the teacher first learns to approximate the student's $\langle w^t, x \rangle$ with the following iterative update:

$$v^{t+1} = v^t - \eta_v \left(\left\langle v^t, x \right\rangle - \left\langle w^t, x \right\rangle \right) x \tag{9}$$

where η_v is the learning rate for the update. Then we use v^{t+1} to perform the example synthesis or selection in teacher's own feature space. We summarize this simple yet effective imitation teacher model in Alg. 2.

6. Discussion

Optimality of the teacher model. For arbitary loss function, the optimal teacher model for a student model should find the training example sequence to achieve the fastest possible convergence. Exhaustively finding such example sequence is computational impossible. For example, there are n^T possible training sequences (T is the iteration number) for n-size pool-based teaching. As a results, we need to make use of the properties of loss function to design the teacher model. The proposed teacher models are not necessarily optimal, but they are good enough under some conditions for loss function, student model and training data.

Theoretical aspects of the teacher model. The theoretical study of the teacher model includes finding the conditions for the loss function and training data such that the teacher model is optimal, or achieves provable faster convergence rate, or provably converges faster than the random teacher. We desire these conditions to be sufficient and necessary, but sometimes sufficient conditions suffice in practice. For different student models, the theoretical analysis may be different and we merely consider stochastic gradient learner here. There are still lots of optimization algorithms that can be considered. Besides, our teacher models are not necessarily the best, so it is also important to come up with better teacher models with provable guarantees. Although our paper mainly focuses on the fixed learning rate, our results are still applicable for the dynamic learning rate. However, the teacher should be more powerful in synthesizing or choosing examples (R should be larger than fixed learning rate case). In human teaching, it actually makes sense because while teaching a student who learns knowledge with dynamic speed, the teacher should be more powerful so that the student consistently learn fast.

Practical aspects of the teacher model. In practice, we usually want the teacher model to be less and less informative to the student model, scalable to large datasets, efficient to compute. How to make the teacher model scalable, efficient and less informative remains open challenges.

7. Experiments

7.1. Experimental details

Performance metric. We use three metric to evaluate the convergence performance: objective value w.r.t. the training set, difference between w^t and w^* ($\|w^t - w^*\|_2$), and the classification accuracy on testing set.

Parameters and setup. Detailed experimental setup is given in Appendix B. We mostly evaluate the practical pool-based teaching (without rescaling). We evaluate the different teaching strategies in Appendix C, and give more experiments on spherical data (Appendix E) and infant egocentric visual data (Appendix F). For fairness, learning rates for all methods are the same.

7.2. Teaching linear models on Gaussian data

This experiment explores the convergence of three typical linear models: ridge regression (RR), logistic regression (LR) and support vector machine (SVM) on Gaussian data. Note that SGD on selected set is to run SGD on the union of all samples selected by the omniscient teacher. For the scenario of different feature spaces, we use a random orthogonal projection matrix to generate the teacher's feature space from student's. All teachers use pool-based teaching strategy. For fair comparisons, we use the same random initialization and the same learning rate.

Teaching in the same feature space. The results in Fig.

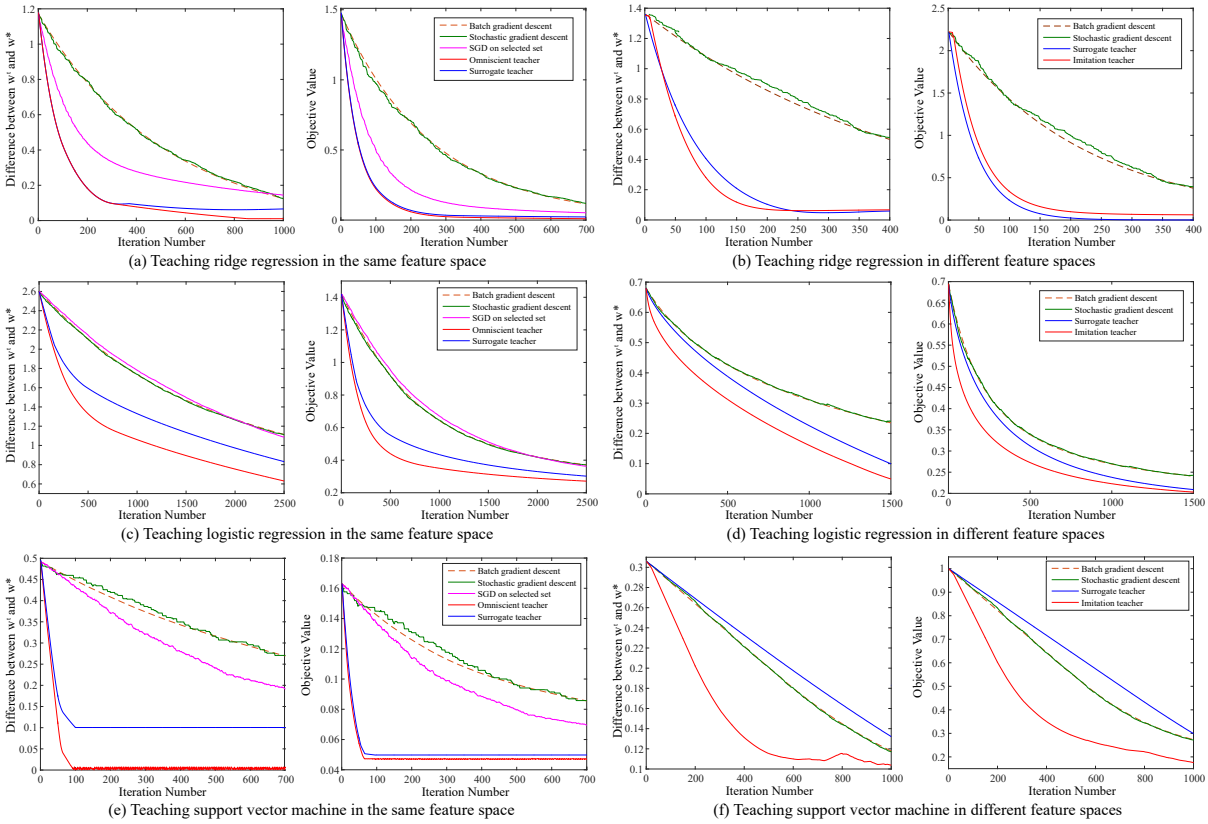


Figure 2. Convergence results on Gaussian distributed data.

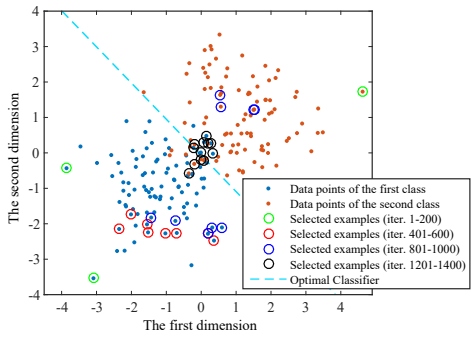


Figure 3. The examples selected by omniscient teacher for logistic regression on 2D binary-class Gaussian data.

2 show that the learner can converge much faster using the example provided by the teacher, showing the effectiveness of our teaching models. As expected, we find that the omniscient teacher consistently achieves faster convergence than the surrogate teacher who has no access to w. It is because the omniscient teacher always has more information about the learner. More interestingly, our guiding algorithms also consistently outperform SGD on the selected set, showing that the order of inputting training samples matters.

Teaching in different feature spaces. It is a more practical scenario that teacher and student use different feature spaces. While the omniscient teacher model is no longer applicable here, we teach the student model using the surrogate teacher and the imitation teacher. While the feature spaces are totally different, it can be expected that there will be a mismatch gap between the teacher model parameter and the student model parameter. Even in such a challenging scenario, the experimental results show that our teacher model still outperforms the conventional SGD and batch GD in most cases. One can observe that the surrogate teacher performs poorly in the SVM, which may be caused by the tightness of the approximated lower bound of the T_2 term. Compared to the surrogate teacher, the imitation teacher is more stable and consistently improves the convergence in all three linear models.

7.3. Teaching Linear Classifiers on MNIST Dataset

We further evaluate our teacher models on MNIST dataset. We use 24D random features to classify the digits (0/1, 3/5 as examples). We generate the teacher's features using a random projection matrix from the original 24D student's features. Note that, omniscient teacher and surrogate teacher (same space) assume the teacher uses the student's feature space, while surrogate teacher (different space) and imitation teacher assume the teacher uses its own space. From Fig. 4, one can observe that all these teacher model produces significant convergence speedup. We can see that the omniscient teacher converges fastest as expected. Interestingly, our imitation teacher achieves very similar con-

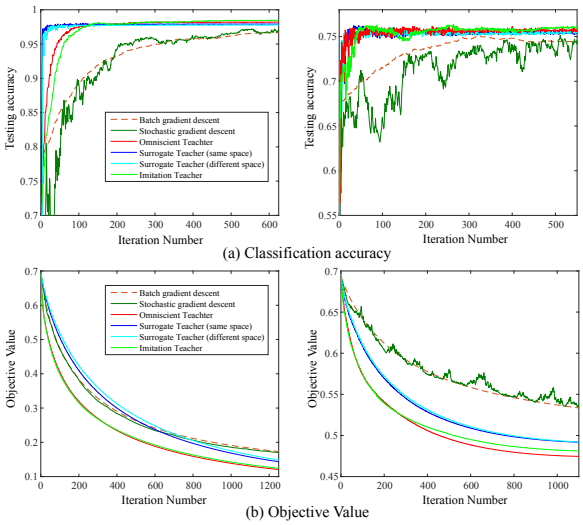


Figure 4. Teaching logistic regression on MNIST dataset. Left column: 0/1 classification. Right column: 3/5 classification

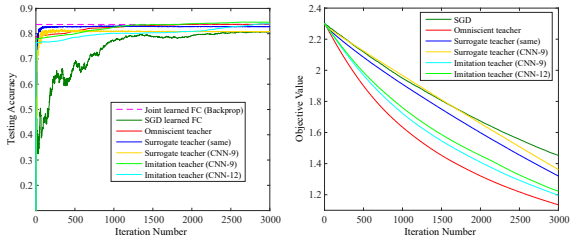


Figure 5. Teaching fully connected layers of CNNs on CIFAR-10. Left: testing accuracy. Right: training objective value.

vergence speedup to the omniscient teacher under the condition that the teacher does not know the student's feature space. In Fig.6, we also show some examples of teacher's selected digit images (0/1 as examples) and find that the teacher tends to select easy example at the beginning and gradually shift the focus to difficult examples. This also has the intrinsic connections with the curriculum learning.

7.4. Teaching Fully Connected Layers in CNNs

We extend our teacher models from binary classification to multi-class classification. The teacher models are used to teach the final fully connected (FC) layers in convolutional neural network on CIFAR-10. We first train three baseline CNNs (6/9/12 convolution layers, detailed configuration is in Appendix B) on CIFAR-10 without data augmentation and obtain the 83.5%, 86.1%, 87.2% accuracy. First, we applied the omniscient teacher and the surrogate teacher to the CNN-6 student using the optimal FC layer from the joint backprop training. It is essentially to teach the FC layer in the same feature space. Second, we applied the surrogate teacher and the imitation teacher to the CNN-6 student using the parameters of optimal FC layers from CNN-9 and CNN-12. It is to teach the FC layer in different feature spaces. More interestingly, this different feature space may not necessarily have an invertible one-to-one mapping, but we could still observe convergence speedup using our teacher models. From Fig. 5, we can see that all the teacher models produces very fast convergence in terms of testing accuracy. Our teacher models can even produces better testing accuracy than the backprop-learned FC layer. For objective value, the omniscient teacher shows the largest convergence speedup, and the imitation teacher performs slightly worse but still much better than the SGD.

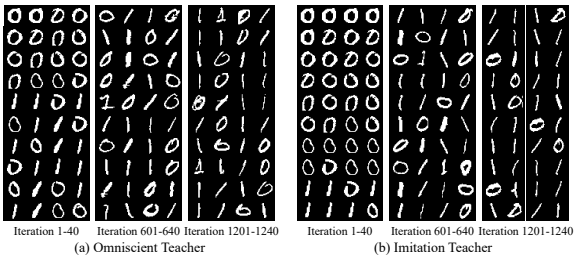


Figure 6. Some selected training examples on MNIST.

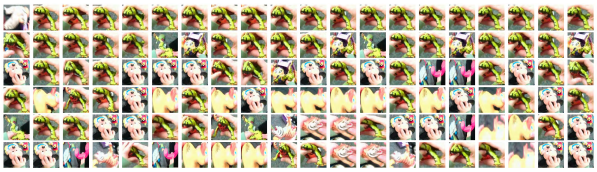


Figure 7. Selected training examples by the omniscient teacher on ego-centric data of infants. (The examples are visualized every 100 iteration, with left-to-right and top-to-bottom ordering)

7.5. Teaching on ego-centric visual data of infants

Using our teaching model, we analyze cropped object instances obtained from ego-centric video of an infant playing with toys (Yurovsky et al., 2013). Full detailed settings and results are in Appendix F. The results in Fig. 7 demonstrate a strong qualitative agreement between the training examples selected by the omniscient teacher and the order of examples received by a child in a naturalistic play environment. In both cases, the learner experiences extended bouts of viewing the same object. In contrast, the standard SGD learner receives random inputs. Our convergence results demonstrate that the learner converges significantly faster when receiving similar inputs to the child. Previous works have documented the unique temporal structure of the image examples that a child receives during object play (Bambach et al., 2016; Pereira et al., 2014). We believe these are the first results demonstrating that similar orderings can be obtained via a machine teaching approach.

8. Concluding Remarks

The paper proposes an iterative machine teaching framework. We elaborate the settings of the framework, and then study two important properties: teaching monotonicity and teaching capability. Based on the framework, we propose three teacher models for gradient learners, and give theoretical analysis for the learner to provably achieve fast convergence. Our theoretical findings are verified by experiments.

Acknowledgement

We would like to sincerely thank all the reviewers and Prof. Xiaojin Zhu for the valuable suggestions to improve the paper, Dan Yurovsky and Charlotte Wozniak for their help in collecting the dataset of children's visual inputs during object learning, and Qian Shao for help with the annotations. This project was supported in part by NSF IIS-1218749, NIH BIGDATA 1R01GM108341, NSF CAREER IIS-1350983, NSF IIS-1639792 EAGER, ONR N00014-15-1-2340, NSF Awards (BCS-1524565, BCS-1523982, and IIS-1320348) Nvidia and Intel. In addition, this work was partially supported by the Indiana University Areas of Emergent Research initiative in Learning: Brains, Machines, Children.

References

Alfeld, Scott, Zhu, Xiaojin, and Barford, Paul. Data poisoning attacks against autoregressive models. In *AAAI*, pp. 1452–1458, 2016.

Alfeld, Scott, Zhu, Xiaojin, and Barford, Paul. Explicit defense actions against test-set attacks. In *AAAI*, 2017.

Ba, Jimmy and Caruana, Rich. Do deep nets really need to be deep? In *Advances in neural information processing systems*, pp. 2654–2662, 2014.

Balcan, Maria-Florina, Hanneke, Steve, and Vaughan, Jennifer Wortman. The true sample complexity of active learning. *Machine learning*, 80(2-3):111–139, 2010.

Bambach, Sven, Crandall, David J, Smith, Linda B, and Yu, Chen. Active Viewing in Toddlers Facilitates Visual Object Learning: An Egocentric Vision Approach. *Proceedings of the 38th Annual Meeting of the Cognitive Science Society*, 2016.

Bengio, Yoshua, Louradour, Jerome, Collobert, Ronan, and Weston, Jason. Curriculum learning. In *ICML*, 2009.

Bucila, Cristian, Caruana, Rich, and Niculescu-Mizil, Alexandru. Model compression. In *Proceedings of the 12th ACM SIGKDD international conference on Knowledge discovery and data mining*, pp. 535–541. ACM, 2006.

Cakmak, Maya and Thomaz, Andrea L. Eliciting good teaching from humans for machine learners. *Artificial Intelligence*, 217:198–215, 2014.

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

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

Hall, Eric C and Willett, Rebecca M. Online optimization in dynamic environments. *arXiv preprint arXiv:1307.5944*, 2013.

Han, Song, Mao, Huizi, and Dally, William J. Deep compression: Compressing deep neural networks with pruning, trained quantization and huffman coding. *arXiv preprint arXiv:1510.00149*, 2015.

Hinton, Geoffrey, Vinyals, Oriol, and Dean, Jeff. Distilling the knowledge in a neural network. *arXiv preprint arXiv:1503.02531*, 2015.

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

Khan, Faisal, Mutlu, Bilge, and Zhu, Xiaojin. How do humans teach: On curriculum learning and teaching dimension. In *NIPS*, 2011.

Liu, Ji, Zhu, Xiaojin, and Ohannessian, H Gorune. The teaching dimension of linear learners. In *ICML*, 2016.

Meek, Christopher, Simard, Patrice, and Zhu, Xiaojin. Analysis of a design pattern for teaching with features and labels. *arXiv preprint arXiv:1611.05950*, 2016.

Mei, Shike and Zhu, Xiaojin. Using machine teaching to identify optimal training-set attacks on machine learners. In *AAAI*, 2015.

Nemirovski, Arkadi, Juditsky, Anatoli, Lan, Guanghui, and Shapiro, Alexander. Robust stochastic approximation approach to stochastic programming. *SIAM Journal on optimization*, 19(4):1574–1609, 2009.

Pan, Sinno Jialin and Yang, Qiang. A survey on transfer learning. *IEEE Transactions on knowledge and data engineering*, 22(10):1345–1359, 2010.

Pereira, Alfredo F, Smith, Linda B, and Yu, Chen. A Bottom-up View of Toddler Word Learning. *Psychonomic bulletin & review*, 21(1):178–185, 2014.

Romero, Adriana, Ballas, Nicolas, Kahou, Samira Ebrahimi, Chassang, Antoine, Gatta, Carlo, and Bengio, Yoshua. Fitnets: Hints for thin deep nets. *arXiv preprint arXiv:1412.6550*, 2014.

Shinohara, Ayumi and Miyano, Satoru. Teachability in computational learning. *New Generation Computing*, 8 (4):337–347, 1991.

Singla, Adish, Bogunovic, Ilija, Bartok, Gabor, Karbasi, Amin, and Krause, Andreas. Near-optimally teaching the crowd to classify. In *ICML*, pp. 154–162, 2014.

Yurovsky, Daniel, Smith, Linda B, and Yu, Chen. Statistical Word Learning at Scale: The Baby's View is Better Developmental Science. *Developmental Science*, 16(6): 959–966, 2013.

Zhu, Xiaojin. Machine teaching for bayesian learners in the exponential family. In *NIPS*, 2013.

Zhu, Xiaojin. Machine teaching: An inverse problem to machine learning and an approach toward optimal education. In *AAAI*, pp. 4083–4087, 2015.

Appendix

A. Details of the Proof

Proof of Theorem 2 We assume the optimization starts with an initialized weights w^0. t is denoted as the iteration index. Let w_g^t and w_s^t be the model parameter updated by our omniscient teacher and SGD, respectively. We first consider the case where $t = 1$. For SGD, the first gradient update w_s^1 is

$$w_s^1 = w^0 - \eta_t \frac{\partial \ell(\langle w^0, x_s \rangle, y_s)}{\partial w^0}. \tag{10}$$

Then we compute the difference between w_s^1 and w^*:

$$\begin{aligned} \left\|w_s^1 - w^*\right\|_2^2 &= \left\| w^0 - \eta_t \frac{\partial \ell(\langle w^0, x \rangle, y)}{\partial w^0} - w^* \right\|_2^2 \\ &= \left\|w^0 - w^*\right\|_2^2 + \eta_t^2 \left\| \frac{\partial \ell(\langle w^0, x \rangle, y)}{\partial w^0} \right\|_2^2 - 2\eta_t \left\langle w^0 - w^*, \frac{\partial \ell(\langle w^0, x \rangle, y)}{\partial w^0} \right\rangle \end{aligned} \tag{11}$$

Because the omniscient teacher is to minimize last two term, so we are guaranteed to have

$$\left\|w_g^1 - w^*\right\|_2^2 \leq \left\|w_s^1 - w^*\right\|_2^2. \tag{12}$$

So with the same initialization $w_g^0 = w_s^0$, $\left\|w_g^1 - w^*\right\|_2^2 \leq \left\|w_s^1 - w^*\right\|_2^2$ is always true. Then we consider the case where $t = k, k \geq 1$. We first compute the difference between w_g^{k+1} and w^*:

$$\begin{aligned} \left\|w_g^{k+1} - w^*\right\|_2^2 &= \left\| w_g^k - \eta_t \frac{\partial \ell(\langle w_g^k, x \rangle, y)}{\partial w^{k+1}} - w^* \right\|_2^2 \\ &= \left\|w_g^k - w^*\right\|_2^2 + \min_{\{x,y\}} \left\{ \eta_t^2 \left\| \frac{\partial \ell(\langle w_g^k, x \rangle, y)}{\partial w_g^k} \right\|_2^2 - 2\eta_t \left\langle w_g^k - w^*, \frac{\partial \ell(\langle w_g^k, x \rangle, y)}{\partial w_g^k} \right\rangle \right\} \\ &= \left\|w_g^k - w^*\right\|_2^2 + \eta_t^2 \left\| \frac{\partial \ell(\langle w_g^k, x_*^k \rangle, y_*^k)}{\partial w_g^k} \right\|_2^2 - 2\eta_t \left\langle w_g^k - w^*, \frac{\partial \ell(\langle w_g^k, x_*^k \rangle, y_*^k)}{\partial w_g^k} \right\rangle \\ &= \left\|w_g^k - w^*\right\|_2^2 - TV(w_g^k) \end{aligned} \tag{13}$$

where x_*^k, y_*^k is the sample selected by the omniscient teacher in the k-th iteration. Using the given conditions, we can bound the difference between w_s^{k+1} and w^* from below:

$$\begin{aligned} \left\|w_s^{k+1} - w^*\right\|_2^2 &= \left\| w_s^k - \eta_t \frac{\partial \ell(\langle w_s^k, x^s \rangle, y^s)}{\partial w_s^k} - w^* \right\|_2^2 \\ &= \left\|w_s^k - w^*\right\|_2^2 + \eta_t^2 \left\| \frac{\partial \ell(\langle w_s^k, x_s^k \rangle, y_s^k)}{\partial w_s^k} \right\|_2^2 - 2\eta_t \left\langle w_s^k - w^*, \frac{\partial \ell(\langle w_s^k, x_s^k \rangle, y_s^k)}{\partial w_s^k} \right\rangle \\ &\geq \left\|w_s^k - w^*\right\|_2^2 - TV(w_s^k) \end{aligned} \tag{14}$$

where x_s^k, y_s^k is the sample selected by the random teacher in the k-th iteration. Comparing Eq. 13 and Eq. 14 and using the condition in the theorem, the following inequality always holds under the condition $\left\|w_g^k - w^*\right\|_2^2 \leq \left\|w_s^k - w^*\right\|_2^2$:

$$\left\|w_s^{k+1} - w^*\right\|_2^2 = \left\|w_s^k - w^*\right\|_2^2 - TV(w_s^k) \geq \left\|w_g^k - w^*\right\|_2^2 - TV(w_g^k) = \left\|w_g^{k+1} - w^*\right\|_2^2. \tag{15}$$

Further because we already know that $\left\|w_g^1 - w^*\right\|_2^2 \leq \left\|w_s^1 - w^*\right\|_2^2$, using induction we can conclude that $\left\|w_g^t - w^*\right\|_2^2$ will be always not larger than $\left\|w_s^t - w^*\right\|_2^2$ (t can be any iteration). Therefore, in each iteration the omniscient teacher can always converge not slower than random teacher (SGD). ■

Proof of Proposition 3 Consider the square loss $\ell(\langle w,x\rangle,y)=(\langle w,x\rangle-y)^2$, we have $\frac{\partial\ell(\langle w,x\rangle,y)}{\partial w}=2(\langle w,x\rangle-y)x$. Suppose we are given two initializations w_1,w_2 satisfying $\|w_1-w_*\|_2^2\le\|w_2-w_*\|_2^2$. For square loss, we first write out

$$\begin{aligned}&\|w_1-w^*\|^2-TV(w_1)=\|w_1-w^*\|^2+\min_{x\in\mathcal{X},y\in\mathcal{Y}}\{\eta_t^2T_1(x,y|w_1)-2\eta_tT_2(x,y|w_1)\}\\&=\|w_1-w^*\|^2+\min_{\{x,y\}}\left\{\eta_t^2\left\|\frac{\partial\ell(\langle w_1,x\rangle,y)}{\partial w_1}\right\|_2^2-2\eta_t\left\langle w_1-w^*,\frac{\partial\ell(\langle w_1,x^*\rangle,y^*)}{\partial w_1}\right\rangle\right\}\\&=\|w_1-w^*\|^2+\begin{cases}2(\frac{R}{\|w_1-w^*\|})^2\|w_1-w^*\|^2(w_1-w^*),&\text{if }\frac{R}{\|w_1-w^*\|}<\frac{1}{\eta_t}\\-\|w_1-w^*\|^2,&\text{if }\frac{R}{\|w_1-w^*\|}\ge\frac{1}{\eta_t}\end{cases}\end{aligned}\tag{16}$$

Similarly for w_2, we have

$$\begin{aligned}&\|w_2-w^*\|^2-TV(w_2)\\&=\|w_2-w^*\|^2+\begin{cases}2(\frac{R}{\|w_2-w^*\|})^2\|w_2-w^*\|^2(w_2-w^*),&\text{if }\frac{R}{\|w_2-w^*\|}<\frac{1}{\eta_t}\\-\|w_2-w^*\|^2,&\text{if }\frac{R}{\|w_2-w^*\|}\ge\frac{1}{\eta_t}\end{cases}\end{aligned}\tag{17}$$

There will be three scenarios to consider: (1) $R\eta_t\le\|w_1-w^*\|\le\|w_2-w^*\|$; (2) $\|w_1-w^*\|\le R\eta_t\le\|w_2-w^*\|$; (3) $\|w_1-w^*\|\le\|w_2-w^*\|\le R\eta_t$. It is easy to verify that under all three scenarios, we have

$$\|w_1-w^*\|^2-TV(w_1)\le\quad\|w_2-w^*\|^2-TV(w_2)\tag{18}$$

■

To simplify notations, we denote $\beta_{(\langle w,x\rangle,y)}=\nabla_{\langle w,x\rangle}\ell(\langle w,x\rangle,y)$ for a loss function $\ell(\cdot,\cdot)$ in the following proof. For omniscient teacher, $(\hat{x},\hat{y})$ denotes a specific construction of (x,y). Notice that $(\tilde{x},\tilde{y})$ will not be used in omniscient teacher case to avoid ambiguity, since the student and the teacher use the same representation space.

Proof of Theorem 4 At t-step, the omniscient teacher selects the samples via optimization

$$\min_{x\in\mathcal{X},y\in\mathcal{Y}}\eta^2\|\nabla_{w^t}\ell\left(\left\langle w^t,x\right\rangle,y\right)\|^2-2\eta\left\langle w^t-w^*,\nabla_{w^t}\ell\left(\left\langle w^t,x\right\rangle,y\right)\right\rangle.$$

We denote $\hat{x}=\gamma(w^t-w^*)$ and $\hat{y}\in\mathcal{Y}$, since $\gamma(w-w^*)\in\mathcal{X}$, we have

$$\min_{x\in\mathcal{X},y\in\mathcal{Y}}\eta^2\|\nabla_{w^t}\ell\left(\left\langle w^t,x\right\rangle,y\right)\|^2-2\eta\left\langle w^t-w^*,\nabla_{w^t}\ell\left(\left\langle w^t,x\right\rangle,y\right)\right\rangle\tag{19}$$

$$\le\quad\left(\eta^2\beta_{(\langle w^t,\hat{x}\rangle,\hat{y})}^2\gamma^2-2\eta\beta_{(\langle w^t,\hat{x}\rangle,\hat{y})}\gamma\right)\|w^t-w^*\|_2^2.\tag{20}$$

Plug Eq. (19) into the recursion Eq. (3), we have

$$\begin{aligned}\left\|w^{t+1}-w^*\right\|_2^2&=\min_{x\in\mathcal{X},y\in\mathcal{Y}}\left\|w^t-\eta\frac{\partial\ell(\langle w,x\rangle,y)}{\partial w}-w^*\right\|_2^2\\&=\left\|w^t-w^*\right\|_2^2+\min_{x\in\mathcal{X},y\in\mathcal{Y}}\eta^2\left\|\frac{\partial\ell(\langle w^t,x\rangle,y)}{\partial w^t}\right\|_2^2-2\eta\left\langle w^t-w^*,\frac{\partial\ell(\langle w^t,x\rangle,y)}{\partial w^t}\right\rangle\\&\le\left(1+\eta^2\beta_{(\langle w^t,\hat{x}\rangle,\hat{y})}^2\gamma^2-2\eta\beta_{(\langle w^t,\hat{x}\rangle,\hat{y})}\gamma\right)\|w^t-w^*\|_2^2=\left(1-\eta\beta_{(\langle w^t,\gamma(w^t-w^*)\rangle,\hat{y})}\gamma\right)^2\|w^t-w^*\|_2^2.\end{aligned}\tag{21}$$

First we let $\nu(\gamma)=\min_{w,y}\gamma\nabla_{\langle w,\gamma(w-w^*)\rangle}\ell(\langle w,\gamma(w-w^*)\rangle,y)$. Then we have the condition $0<\nu(\gamma)\le\gamma\beta_{(\langle w,\gamma(w-w^*)\rangle,\hat{y})}\le\frac{1}{\eta}<\infty$ for any w,y, so we can obtain

$$0\le1-\gamma\eta\beta_{(\langle w,\gamma(w-w^*)\rangle,\hat{y})}\le1-\eta\nu(\gamma),$$

after simplifying $\nu(\gamma)$ to ν, we therefore have the following inequality from Eq. (21):

$$\left\|w^{t+1}-w^*\right\|_2^2\le(1-\eta\nu)^2\left\|w^t-w^*\right\|_2^2,$$

Thus we can have the exponential convergence:

$$\left\|w^t-w^*\right\|_2\le(1-\eta\nu)^t\left\|w^0-w^*\right\|_2,$$

in other words, the student needs $\left(\log\frac{1}{1-\eta\nu}\right)^{-1}\log\frac{\|w^0-w^*\|}{\epsilon}$ samples to achieve an ϵ-approximation of w^*.

■

Proof of Proposition 5 Because $\ell\left(\langle w,x\rangle,y\right)$ is ζ_1-strongly convex w.r.t. w, we have

$$\zeta_1\left(\ell\left(\langle w,x\rangle,y\right)-\min_w \ell\left(\langle w,x\rangle,y\right)\right)\leq\left\|\nabla_w\ell\left(\langle w,x\rangle,y\right)\right\|^2=\beta^2_{(\langle w,x\rangle,y)}\left\|x\right\|^2,\quad\forall\{x,y\}\in\mathcal{X}\times\mathcal{Y},$$

where $\mathcal{X}=\left\{x\in\mathbb{R}^d,\|x\|\leq R\right\}$. Using $\hat{x}=\gamma(w-w^*),\gamma\geq 0$, we have

$$\sqrt{\zeta_1\left(\ell\left(\langle w,\gamma(w-w^*)\rangle,y\right)-\min_w \ell\left(\langle w,\gamma(w-w^*)\rangle,y\right)\right)}\leq\beta_{(\langle w,\gamma(w-w^*)\rangle,y)}\gamma\|w-w^*\|.$$

We assume the loss function is always non-negative, i.e., $\ell\left(\langle w,x\rangle,y\right)\geq 0$. Therefore we have

$$\sqrt{\zeta_1\left(\ell\left(\langle w,\gamma(w-w^*)\rangle,y\right)\right)}\leq\beta_{(\langle w,\gamma(w-w^*)\rangle,y)}\gamma\|w-w^*\|.$$

Because $\ell\left(\langle w,x\rangle,y\right)$ is ζ-strongly convex w.r.t. w, it is also ζ_2-strongly convex w.r.t. $\langle w,x\rangle$. Then we perform Taylor expansion to $\ell\left(\langle w,\gamma(w-w^*)\rangle,y\right)$ w.r.t. $\langle w,x\rangle$ at the point $\langle w^*,x\rangle$ and obtain

$$\ell\left(\langle w,\gamma(w-w^*)\rangle,y\right)\geq\ell\left(\langle w,\gamma(w^*-w^*)\rangle,y\right)+\nabla_{\langle w,x\rangle}\ell\left(\langle w,\gamma(w^*-w^*)\rangle,y\right)(w-w^*)^Tx+\frac{\zeta_2}{2}\|(w-w^*)^Tx\|^2$$

which leads to

$$\ell\left(\langle w,\gamma(w-w^*)\rangle,y\right)\geq\frac{\zeta_2}{2}\gamma^2\|w-w^*\|^4$$

Combining pieces, we have

$$\sqrt{\frac{\zeta_1\zeta_2}{2}}\gamma\|w-w^*\|\leq\beta_{(\langle w,\gamma(w-w^*)\rangle,y)}\gamma.$$

Then if we set $\gamma=\min\left\{\sqrt{\frac{2}{\zeta_1\zeta_2}}\frac{1}{\|w-w^*\|\eta},\frac{R}{\|w-w^*\|}\right\}$, we can have $\frac{1}{\eta}\leq\beta_{(\langle w,\gamma(w-w^*)\rangle,y)}\gamma$. Because $\ell\left(\langle w,x\rangle,y\right)$ is Lipschitz smooth w.r.t. $\langle w,x\rangle$ with parameter L, we have

$$\left\|\beta_{(\langle w,x\rangle,y)}-\beta_{(\langle w^*,x\rangle,y)}\right\|\leq LR\left\|w-w^*\right\|$$

Because $\beta_{(\langle w^*,x\rangle,y)}=0$, we have the following inequality:

$$\left\|\beta_{(\langle w,x\rangle,y)}\right\|\leq LR\left\|w-w^*\right\|$$

If we multiply both side with γ, we can have

$$\beta_{(\langle w,x\rangle,y)}\gamma\leq LR\left\|w-w^*\right\|\gamma$$

By setting γ as $\frac{1}{LR\eta\|w-w^*\|}$, we arrive at $\beta_{(\langle w,x\rangle,y)}\gamma<\frac{1}{\eta}$. Combining pieces, as long as we set

$$\gamma=\min\left\{\sqrt{\frac{2}{\zeta_1\zeta_2}}\frac{1}{\eta\|w-w^*\|},\frac{R}{\|w-w^*\|},\frac{1}{LR\eta\left\|w-w^*\right\|}\right\},$$

then we can have

$$0<c\leq\beta_{(\langle w,\gamma\hat{x}\rangle,\hat{y})}\gamma\leq\frac{1}{\eta}.$$

where c is a non-zero positive constant. Therefore, we achieve the condition for the exponential synthesis-based teaching.

■

By the Proposition 5, the absolute loss and sqaure loss are exponentially teachable in synthesis-based case, and we can obtain γ by plugging into the general form. We will tighten the γ up by analyzing absolute loss and square loss separately. Besides that, we also show the commonly used loss functions for classification, *e.g.*, hinge loss and logistic loss, are also exponentially teachable in synthesis-based teaching if $\|w^*\|$ can be bounded.

Proposition 9 *Absolute loss is exponentially teachable in synthesis-based teaching.*

Proof To show one loss function is exponentially teachable in synthesis-based case, we just need to find the appropriate γ such that the learning intensity is bounded below and above, according to Theorem 4. For the absolute loss, *i.e.*,

$$\ell\left(\langle w,x\rangle,y\right)=\left|\langle w,x\rangle-y\right|,$$

its sub-gradient is

$$\nabla_w\ell(\langle w,x\rangle,y)=\operatorname{sign}(\langle w,x\rangle-y)x,$$

and thus, the learning intensity $\beta_{(\langle w,x\rangle,y)}=\operatorname{sign}\left(\langle w,x\rangle-y\right)$. For $w\neq w^*$, plugging $\hat{x}=\gamma\left(w-w^*\right)$ and

$\hat{y} = \langle w^*, \gamma (w - w^*) \rangle$ into the learning intensity, we have

$$\beta_{\gamma\langle w,\hat{x}\rangle,\hat{y}}\gamma = \text{sign}\left(\gamma^2 \langle w - w^*, w - w^* \rangle\right)\gamma = \gamma.$$

Recall that $\gamma \neq 0$, $|\gamma| \leq \frac{R}{\|w^t - w^*\|}$, $\forall t \in \mathbb{N}$, we have

$$\gamma \leq \min_{t\in\mathbb{N}} \frac{R}{\|w^t - w^*\|} := C.$$

Set $\gamma = \min\{C, \frac{1}{\eta}\}$, we have $\nu = \min\{C, \frac{1}{\eta}\}$. Therefore, we obtain the exponential decay. In fact, since the $\|w^t - w^*\|$ decreases in every step, we have $C = \frac{R}{\|w^0 - w^*\|}$. In following proof, we will follow the same argument to use this fact. ■

Proposition 10 *Square loss is exponentially teachable in synthesis-based teaching.*

Proof For square loss, *i.e.*,

$$\ell\left(\langle w, x\rangle, y\right) = \left(\langle w, x\rangle - y\right)^2,$$

its gradient is

$$\nabla_w \ell\left(\langle w, x\rangle, y\right) = 2\left(\langle w, x\rangle - y\right)x,$$

and thus, the learning intensity $\beta_{\langle w,x\rangle,y} = 2\left(\langle w, x\rangle - y\right)$. For $w \neq w^*$, plugging $\hat{x} = \gamma(w - w^*)$ and $\hat{y} = \langle w^*, \gamma(w - w^*)\rangle$ into the learning intensity, we have

$$\beta_{(\langle w,\hat{x}\rangle,\hat{y})}\gamma = 2\gamma^2 \|w - w^*\|^2.$$

Set $\gamma = \min\left\{\frac{1}{\sqrt{2\eta}\|w^t - w^*\|}, \frac{R}{\|w^t - w^*\|}\right\}$, we achieve the exponential teachable condition. ■

Proposition 11 *Hinge loss is exponentially teachable in synthesis-based teaching if* $\|w^*\| \leq 1$.

Proof For hinge loss, *i.e.*,

$$\ell\left(\langle w, x\rangle, y\right) = \max\left(1 - y\langle w, x\rangle, 0\right),$$

as long as $1 - y\langle w, x\rangle > 0$, its subgradient will be

$$\nabla_w \ell\left(\langle w, x\rangle, y\right) = -yx.$$

Denote $\hat{x} = \gamma(w - w^*)$, we have $\beta_{\langle w,\hat{x}\rangle,\hat{y}} = -\hat{y}$ where $\hat{y} \in \{-1, 1\}$. To satisfy the exponential teachable condition, we need to select $\hat{y}$ and γ such that

$$\begin{cases} 1 - \hat{y}\langle w, \hat{x}\rangle > 0 \\ 0 < -\hat{y}\gamma \leq \frac{1}{\eta} \\ |\gamma| \leq \frac{R}{\|w - w^*\|} \end{cases} \Rightarrow \begin{cases} \hat{y}\gamma\langle w, w - w^*\rangle < 1 \\ -\frac{1}{\eta} \leq \hat{y}\gamma < 0 \\ |\gamma| \leq \frac{R}{\|w - w^*\|} \end{cases} \Rightarrow \begin{cases} \langle w, w - w^*\rangle > -1 \\ -\frac{1}{\eta} \leq \hat{y}\gamma < 0 \\ |\gamma| \leq \frac{R}{\|w - w^*\|} \end{cases}.$$

If $\|w^*\| \leq 1$, we can show

$$\langle w, w^*\rangle \leq \|w\|\|w^*\| \leq \|w\| < 1 + \|w\|^2,$$

where the last inequality comes from the fact $1 + a^2 - a > 0$, and thus, we have $\langle w, w - w^*\rangle > -1$. Therefore, we select any configuration of $\hat{y}$ and γ satisfying

$$-\frac{1}{\eta} \leq \hat{y}\gamma < 0, \quad \text{and} \quad |\gamma| \leq \frac{R}{\|w - w^*\|}.$$

Particularly, we set $\hat{y} = -1$ and $\gamma = \min\left\{\frac{1}{\eta}, \frac{R}{\|w^0 - w^*\|}\right\}$.

■

Proposition 12 *Logistic loss is exponentially teachable in synthesis-based teaching if* $\|w^*\| \leq 1$.

Proof For the logistic loss, *i.e.*,

$$\ell\left(\langle w, x\rangle, y\right) = \log\left(1 + \exp(-y\langle w, x\rangle)\right),$$

its gradient is

$$\nabla_w \ell\left(\langle w, x\rangle, y\right) = -\frac{yx}{1+\exp(y\langle w, x\rangle)}.$$

Denote $\hat{x} = \gamma(w - w^*)$, we have $\beta_{\langle w,\hat{x}\rangle,\hat{y}} = -\frac{\hat{y}}{1+\exp(\hat{y}\langle w,\hat{x}\rangle)}$ where $\hat{y} \in \{-1, 1\}$. To satisfy the exponential teachable condition, we need to select $\hat{y}$ and γ such that

$$\begin{cases} 0 < -\frac{\hat{y}\gamma}{1+\exp(\hat{y}\langle w,\hat{x}\rangle)} \leq \frac{1}{\eta} \\ |\gamma| \leq \frac{R}{\|w-w^*\|} \end{cases}.$$

Particularly, we set $\hat{y} = -1$, we can fix the γ by

$$0 < \frac{\gamma}{1+\exp(\gamma)} < \frac{\gamma}{1+\exp(\hat{y}\langle w,\hat{x}\rangle)} \leq \gamma \leq \frac{1}{\eta}, \quad \text{and} \quad |\gamma| \leq \frac{R}{\|w-w^*\|}.$$

The $\frac{\gamma}{1+\exp(\gamma)} < \frac{\gamma}{1+\exp(\hat{y}\langle w,\hat{x}\rangle)}$ is obtained by the monotonicity of $\exp(\cdot)$ and $\langle w, w - w^*\rangle > -1$ when $\|w^*\|$. Therefore, we can choose $\gamma = \min\left\{\frac{1}{\eta}, \frac{R}{\|w^0-w^*\|}\right\}$, and thus, the lower bound $\nu = \frac{\gamma}{1+\exp(\gamma)}$. ■

Proof of Corollary 6 In each update, given the training sample $x \in \text{span}(\mathcal{X})$, we have $w^{t+1} = w^t - \eta\beta_{\langle w,x\rangle,y}x$, therefore, the $\Delta_{t+1}w := w^{t+1} - w^0 \in \text{span}(\mathcal{X})$. If $w^0 - w^* \in \text{span}(\mathcal{X})$, $w^{t+1} - w^* \in \text{span}(\mathcal{X})$, which means by linear combination, we can construct $\hat{\gamma}\sum_{i=1}^n \alpha_i^t x_i = \gamma(w^t - w^*)$. With the condition that the loss function is exponentially synthesis-based teachable, we achieve the conclusion that the combination-based omniscient teacher will converge at least exponentially with the same rate to the synthesis-based teaching. ■

Proof of Theorem 8 The proof is similar to the synthesis-based case. However, we introduce the consideration of the effect of pool-based teaching. Specifically, we first obtain a virtual training sample in full space, and then, we generate the sample from the candidate pool to mimic the virtual sample.

With the condition $w^0 - v^* \in \text{span}(\mathcal{D})$, as we discussed in the proof of Corollary 6, in every iteration, $w^t - v^* \in \text{span}(\mathcal{D})$. Therefore, we only need to consider in the space of span $(\mathcal{D})$. Meanwhile, since the teacher can rescale the sample, without loss of generality, we assume if $x \in \mathcal{X}$, then $-x \in \mathcal{X}$ to make the rescaling is always positive.

At t-step, as the loss is exponentially synthesis-based teachable with γ, therefore, we have the virtually constructed sample $\{x_v, y_v\}$ where $x_v = \gamma(w^t - w^*)$ with γ satisfying the condition of exponentially teachable in synthesis-based settings, we first rescale the candidate pool $\mathcal{X}$ such that

$$\forall x \in \mathcal{X}, \gamma_x \|x\| = \|x_v\| = \gamma\left\|w^t - w^*\right\|.$$

We denote the rescaled candidate pool as $\mathcal{X}_t$, under the condition of rescalable pool-based teachability, there is a sample $\{\hat{x}, \hat{y}\} \in \mathcal{X} \times \mathcal{Y}$ with scale factor $\hat{\gamma}$ such that

$$\begin{aligned} \min_{(x,y)\in\mathcal{X}_t\times\mathcal{Y}} \quad & \eta^2\|\nabla_{w^t}\ell\left(\left\langle w^t, x\right\rangle, y\right)\|^2 - 2\eta\left\langle w^t - w^*, \nabla_{w^t}\ell\left(\left\langle w^t, x\right\rangle, y\right)\right\rangle \\ \leq \quad & \eta^2\beta^2_{\langle w^t,\hat{\gamma}\hat{x}\rangle,\hat{y}}\|\hat{x}\|^2 - 2\eta\beta_{\langle w^t,\hat{\gamma}\hat{x}\rangle,\hat{y}}\langle w^t - w^*, \hat{\gamma}\hat{x}\rangle. \end{aligned}$$

We decompose the $\hat{\gamma}\hat{x} = ax_v + x_{v\perp}$ with $a = \frac{\langle\hat{\gamma}\hat{x}, x_v\rangle}{\|x_v\|^2}$. and $x_{v\perp} = \hat{\gamma}\hat{x} - ax_v$. Then, we have

$$\begin{aligned} \min_{(x,y)\in\mathcal{X}_t\times\mathcal{Y}} \quad & \eta^2\|\nabla_{w^t}\ell\left(\left\langle w^t, x\right\rangle, y\right)\|^2 - 2\eta\left\langle w^t - w^*, \nabla_{w^t}\ell\left(\left\langle w^t, x\right\rangle, y\right)\right\rangle \\ \leq \quad & \eta^2\beta^2_{\langle w^t,\hat{\gamma}\hat{x}\rangle,\hat{y}}\|\hat{x}\|^2 - 2\eta\beta_{\langle w^t,\hat{\gamma}\hat{x}\rangle,\hat{y}}\langle w^t - w^*, \hat{\gamma}\hat{x}\rangle \\ = \quad & \eta^2\beta^2_{\langle w^t,\hat{\gamma}\hat{x}\rangle,\hat{y}}\gamma^2\|w - w^*\|^2 - 2\eta\beta_{\langle w^t,\hat{\gamma}\hat{x}\rangle,\hat{y}}\langle w^t - w^*, ax_v + x_{v\perp}\rangle \\ = \quad & \eta^2\beta^2_{\langle w^t,\hat{\gamma}\hat{x}\rangle,\hat{y}}\gamma^2\|w - w^*\|^2 - 2\eta\beta_{\langle w^t,\hat{\gamma}\hat{x}\rangle,\hat{y}}\gamma a\left\|w^t - w^*\right\|^2. \end{aligned}$$

Under the condition

$$0 < \gamma\beta_{\langle w,\gamma\frac{w-w^*}{\hat{x}}\rangle,\hat{y}} < \frac{2\mathcal{V}(\mathcal{X})}{\eta},$$

we denote $\nu(\gamma) = \min_{w,\hat{x}\in\mathcal{X},\hat{y}\in\mathcal{Y}}\gamma\beta_{\langle w,\gamma\frac{w-w^*}{\hat{x}}\rangle,\hat{y}} > 0$ and $\mu(\gamma) = \max_{w,\hat{x}\in\mathcal{X},\hat{y}\in\mathcal{Y}}\gamma\beta_{\langle w,\gamma\frac{w-w^*}{\hat{x}}\rangle,\hat{y}} < \frac{2\mathcal{V}(\mathcal{X})}{\eta}$.

we have the recursion

$$\left\|w^{t+1}-w^*\right\|_2^2 \leq r(\eta,\gamma)\left\|w^t-w^*\right\|_2^2,$$

with $r(\eta,\gamma,\mathcal{V}(\mathcal{X})) := \max\left\{1+\eta^2\mu\left(\gamma\right)^2-2\eta\mu\left(\gamma\right)\mathcal{V}(\mathcal{X}), 1+\eta^2\nu\left(\gamma\right)^2-2\eta\nu\left(\gamma\right)\mathcal{V}(\mathcal{X})\right\}$ and $0 \leq r(\eta,\gamma) < 1$. Therefore, the algorithm converges exponentially

$$\left\|w^t-w^*\right\|_2 \leq r\left(\eta,\gamma\right)^{t/2}\left\|w^0-w^*\right\|_2,$$

in other words, the student needs $2\left(\log\frac{1}{r(\eta,\gamma,\mathcal{V}(\mathcal{X}))}\right)^{-1}\log\frac{\|w^0-w^*\|}{\epsilon}$ samples to achieve an ϵ-approximation of w^*. For clearity, we define the constant term as $C_2^{\eta,\gamma,\mathcal{V}(\mathcal{X})} = 2\left(\log\frac{1}{r(\eta,\gamma,\mathcal{V}(\mathcal{X}))}\right)^{-1}$. ■

B. Detailed Experimental Setting

Layer	CNN-6	CNN-9	CNN-12
Conv1.x	[3×3, 16]×2	[3×3, 16]×3	[3×3, 16]×4
Pool1	2×2 Max, Stride 2		
Conv2.x	[3×3, 32]×2	[3×3, 32]×3	[3×3, 32]×4
Pool2	2×2 Max, Stride 2		
Conv3.x	[3×3, 64]×2	[3×3, 64]×3	[3×3, 64]×4
Pool3	2×2 Max, Stride 2		
FC1	32	32	32

Table 1. Our standard CNN architectures for CIFAR-10. Conv1.x, Conv2.x and Conv3.x denote convolution units that may contain multiple convolution layers. E.g., [3×3, 16]×3 denotes 3 cascaded convolution layers with 16 filters of size 3×3. The CNNs learning ends at 20K iterations with multi-step rate decay.

General Settings We have used three linear models in the experiments. In specific, the formulation of ridge regression (RR) is

$$\min_{w\in\mathbb{R}^d, b\in\mathbb{R}} \frac{1}{n}\sum_{i=1}^{n}\frac{1}{2}(w^T x_i + b - y_i)^2 + \frac{\lambda}{2}\|w\|^2$$

The formulation of logistic regression (LR) is

$$\min_{w\in\mathbb{R}^d, b\in\mathbb{R}} \frac{1}{n}\sum_{i=1}^{n}\log(1+\exp\{-y_i(w^T x_i + b)\}) + \frac{\lambda}{2}\|w\|^2$$

The formulation of support vector machine (SVM) is

$$\min_{w\in\mathbb{R}^d, b\in\mathbb{R}} \frac{1}{n}\sum_{i=1}^{n}\max(1 - y_i(w^T x_i + b), 0) + \frac{\lambda}{2}\|w\|^2$$

Comparison of different teaching strategies We use a linear regression model (ridge regression with $\lambda = 0$) for this experiment. We set R as 1 and uniformly generate 30 data points as our knowledge pool for the teacher. In this first case, we set the feature dimension as 2, while in the second case, feature dimension is 70. The learning rate is set as 0.0001 for pool-based teaching, same as BGD and SGD.

Experiments on Gaussian data Specifically, RR is run on training data $(\boldsymbol{x}_i, y)$ where each entry in $\boldsymbol{x}_i$ is Gaussian distributed and $y = \langle \boldsymbol{w}^*, \boldsymbol{x}_i\rangle + \epsilon$. LR and SVM are run on $\{\mathcal{X}_1, +1\}$ and $\{\mathcal{X}_2, -1\}$ where $\boldsymbol{x}_i \in \mathcal{X}_1$ is Gaussian distributed in each entry and $+1, -1$ are the labels. Specifically, we use the 10-dimension data that is Gaussian distributed with $(0.5, \cdots, 0.5)$ (label $+1$) and $(-0.5, \cdots, -0.5)$ (label -1) as mean and identity matrix as covariance matrix. We generate 1000 training data points for each class. Learning rate for the same feature space is 0.0001, while learning rate for different feature spaces are 0.00001. λ is set as 0.00005.

Experiments on uniform spherical data We first generate the training data that are uniformly distributed on a unit sphere $\|\boldsymbol{x}_i\|_2 = 1$. Then we set the data points on half of the sphere ($(0, \pi]$) as label $+1$ and the other half ($(\pi, 2\pi]$) as label -1. All the generated data points are 2D. For the scenario of different features, we use a random orthogonal projection matrix to generate the teacher's feature space from student's. Learning rate for the same feature space is 0.001, while learning rate for different feature spaces are 0.0001. λ is set as 0.00005.

Experiments on MNIST dataset We use 24D random features (projected by a random matrix $\mathbb{R}^{784\times 24}$) for the MNIST dataset. The learning rate for all the compared methods are 0.001. Note that, we generate the teacher's features using a random projection matrix ($\mathbb{R}^{24\times 24}$) from the original 24D student's features. λ is set as 0.00005.

Experiments on CIFAR-10 dataset The learning rate for all the compared methods are 0.001. λ is set as 0.00005. The goal is to learn the $\mathbb{R}^{32\times 10}$ fully connected layer, which is also the classifiers for 10 classes. The three network we use in the experiments are shown as follows:

Experiments on infant ego-centric dataset We manually crop and label all the objects that the child is holding for this experiments. For feature extraction, we use VGG-16 network that is pre-trained on Imagenet dataset. Then we use PCA to reduce the 4096 dimension to 64 dimension. We train a multi-class logistic regression to classify the objects. Note that, the omniscient teacher is also applied to train the logistic regression model. The learning rate is set to 0.001 for both SGD and omniscient teacher.

C. Comparison of different teaching strategies

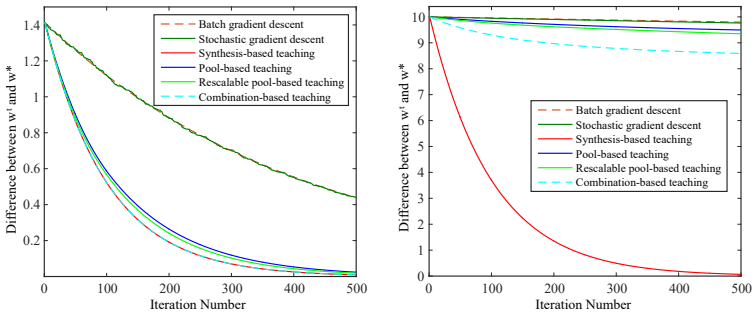


Figure 8. Comparison of different teaching strategies.

We first compare four different teaching strategies for the omniscient teacher. We consider two scenarios. One is that the dimension of feature space is smaller than the number of samples (the given features are sufficient to represent the entire feature), and the other is that the feature dimension is greater than the number of samples (the given features are not sufficient to represent the entire feature). In these two scenarios, we find that synthesis-based teaching usually works the best and always achieves exponential convergence. The combination-based teaching is exactly the same as the synthesis-based teaching in the first scenario, but it is much worse than synthesis in the second scenario. Rescalable pool-based teaching is also better than pool-based teaching. Empirically, the experiment verifies our theoretical findings: the more flexible the teaching strategy is, the more convergence gain we may obtain.

D. More experiments on MNIST dataset

We provide more experimental results on MNIST dataset. Fig. 9 shows the selected examples from 7/9 binary digit classification. The results further verify the teacher models tend to select easy examples at first and gradually shift their focuses to difficult examples, very much resembling the human learning. Fig. 10 shows the difference between the current model parameter and the optimal model parameter over iterations. It also shows that our teachers achieve faster convergence.

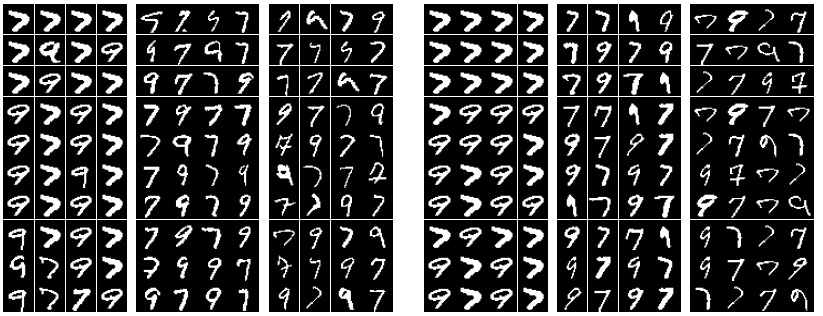


Figure 9. Selected training examples during iteration. (7/9 classification)

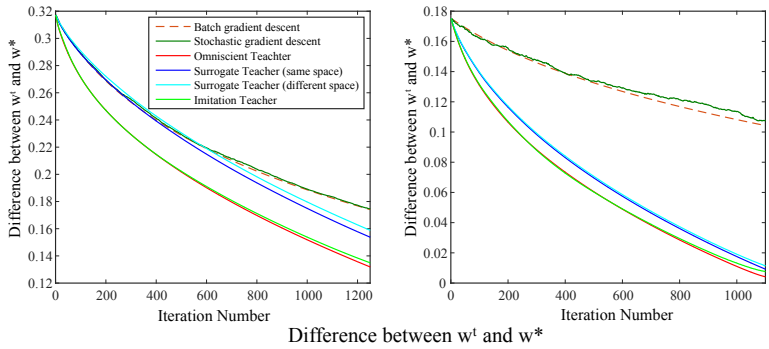


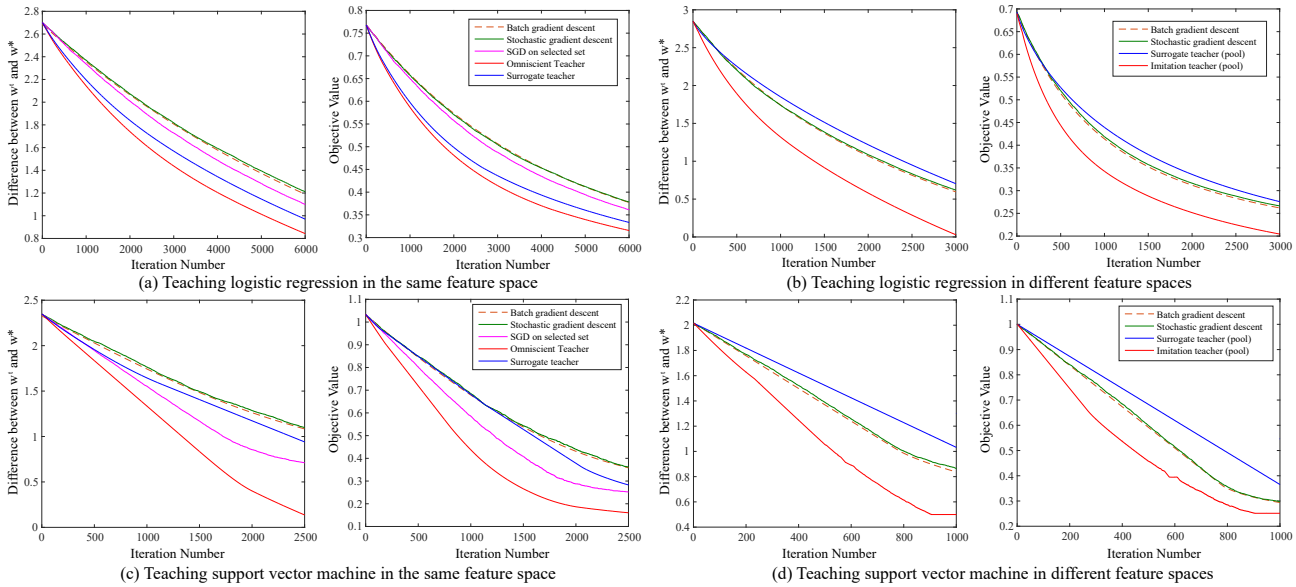
Figure 10. Teaching logistic regression on MNIST dataset. Left column: 0/1 classification. Right column: 3/5 classification

E. Teaching linear models on uniform spherical data

In this experiment, we use a different data distribution to further evaluate the teacher models. We will examine LR and SVM by classifying uniform spherical data.

Teaching in the same feature space. From Fig. 11, one can observe that the convergence is consistently improved while using omniscient teacher to provide examples to learners. We find that the significance of improvement is related to the training data distribution and loss function, as indicated by our theoretical results. The surrogate teacher produces less convergence gain in SVM, because the convexity lower bound becomes very loose in this case. Overall, omniscient teacher still presents strong teaching capability. More interestingly, we use simple SGD run on the sample set selected by the omniscient teacher and also get faster convergence, showing that the selected example set is better than the entire set in terms of convergence.

Teaching in different feature spaces. While the teacher and student use different feature spaces, one can observe from Fig. 11 that the surrogate teacher performs very poorly, even worse than the original SGD and BGD. The imitation teacher works much better and achieves consistent and significant convergence speedup, showing its superiority while the teacher and the student use different features.



(a) Teaching logistic regression in the same feature space

(b) Teaching logistic regression in different feature spaces

(c) Teaching support vector machine in the same feature space

(d) Teaching support vector machine in different feature spaces

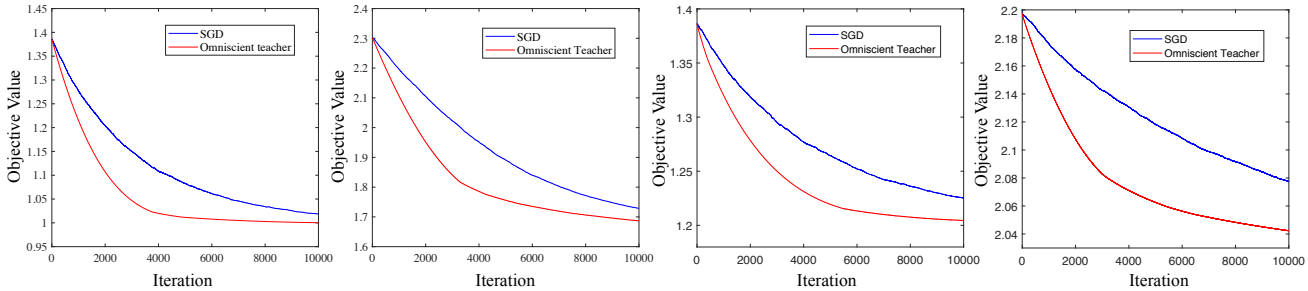
Figure 11. Convergence results on uniform spherical data.

F. Object learning experiment on children's ego-centric visual data

We experiment with a dataset capturing children and parents interacting with toys in a naturalistic setting (Yurovsky et al., 2013). These interactions are recorded for around 10.5 minutes with a camera worn low on the child's forehead. The head-camera's visual field was 90 degrees wide, providing a broad view of objects visible to the infant. The camera was attached to a headband that was tightened so that it did not move once set on the child. To calibrate the camera, the experimenter noted when the child focused on an object and adjusted the camera until the object was in the center of the image in the control monitor.

For our experiments, we selected interactions of 4 one year old infants. For each parent-child dyad, we annotated the bounding box location and category of the toy attended to by the infant at each frame. There are 10 objects in total: doll (34 frames), toy (53 frames), duck (335 frames), frog (2108 frames), helicopter (169 frames), horse (42 frames), mickey (472 frames), phone (394 frames), sheep (119 frames) and tiger (266 frames). We use a VGG-16 network that is pre-trained on Imagenet dataset as our feature extraction. We first extract the 4096D features from these images and then use PCA to reduce the dimension to 64D. Finally, we run our omniscient teacher on these ego-centric data.

One can observe from Fig. 12 that our omniscient teacher achieves faster convergence than the random teacher. Moreover, we give part of the selected training examples of random teacher and omniscient teacher in Fig. 14 and Fig. 15, respectively. We visualize the selected samples every 50 iterations from the first iteration to the 10000th iteration. Interestingly, we find that the training samples that are selected by the omniscient teacher consist of contiguous bouts of experience with the same object instance, unlike the random teacher. The adjacent samples are similar and the object changes in a smooth way. These inputs are qualitatively similar in their ordering to the actual visual experiences of infants in our study, as illustrated in Fig. 13. This can be seen as partial algorithmic confirmation of the desirable structural properties of children's natural learning environment, which emphasizes a smooth and continuous evolution of visual experience, in sharp contrast to random sample selection.



(a) Video of infant 1 (b) Video of infant 2 (c) Video of infant 3 (d) Video of infant 4

Figure 12. Convergence comparison on infant ego-centric visual data.

Figure 13. Training examples corresponding to the natural sequence of objects experienced by a single infant in our study.

Figure 14. Training examples selected by the random teacher (Stochastic gradient descent).

Figure 15. Training examples selected by the omniscient teacher.