# No Learner Left Behind: On the Complexity of Teaching Multiple Learners Simultaneously

**Xiaojin Zhu**
University of Wisconsin–Madison
jerryzhu@cs.wisc.edu

**Ji Liu**
University of Rochester
ji.liu.uwisc@gmail.com

**Manuel Lopes**
Instituto Superior Técnico / INESC-ID
manuel.lopes@tecnico.ulisboa.pt

## Abstract

We present a theoretical study of machine teaching in the setting where the teacher must use the same training set to teach multiple learners. This problem is a theoretical abstraction of the real-world classroom setting in which the teacher delivers the same lecture to academically diverse students. We define a minimax teaching criterion to guarantee the performance of the worst learner in the class. We prove that the teaching dimension increases with class diversity. For the classes of conjugate Bayesian learners and linear regression learners, respectively, we exhibit corresponding minimax teaching set. We then propose a method to enhance teaching by partitioning the class into sections. We present cases where the optimal partition minimizes aggregate teaching dimension while maintaining the guarantee of performance on all learners. Interestingly, we show personalized education (one learner per section) is not necessarily the optimal partition. Our results generalize machine teaching to multiple learners and offer insight on how to teach large classes. [1]

Machine learning from independent and identically-distributed (*i.i.d.*) training data is well-understood [Valiant, 1984; Vapnik, 1995]. In contrast, machine teaching studies the minimum training set to teach a target concept to a learner when a teacher chooses (possibly non-*i.i.d.*) training items. The minimum training set size is known as the *teaching dimension* (TD) [Goldman and Kearns, 1995; Shinohara and Miyano, 1991]. Machine teaching is of great theoretical interest as an inverse problem to machine learning [Liu *et al.*, 2016; Zhu, 2015; Doliwa *et al.*, 2014; Zilles *et al.*, 2011; Balbach and Zeugmann, 2009; Angluin, 2004; Angluin and Krikis, 1997; Goldman and Mathias, 1996; Mathias, 1997; Balbach and Zeugmann, 2006; Balbach, 2008; Kobayashi and Shinohara, 2009; Angluin and Krikis, 2003; Rivest and Yin, 1995; Ben-David and Eiron, 1998]. It also has application in education [Patil *et al.*, 2014; Singla *et al.*, 2014; Lindsey *et al.*, 2013; Cakmak and Thomaz, 2011], security [Alfeld *et al.*, 2017; 2016; Mei and Zhu, 2015], and in interactive machine learning [Suh *et al.*, 2016].

However, prior work in machine teaching is limited to teaching a *single* learner. Motivated by the most prevalent teaching paradigm in reality where a teacher gives a classroom lecture to academically-diverse students, we take machine teaching one step further: what is the teaching dimension (and the corresponding optimal training set) if the same training set must be applied to a class of different learners, such that *all* learners succeed?

We propose a new definition of *class teaching dimension* to answer this question. Our theoretical results support some well-known anecdotes: a teacher must spend more effort in order to teach a more diverse class; and teaching to the average student as a representative of the whole class [Clement *et al.*, 2016; Lee and Brunskill, 2012] can be justified under certain assumptions.

To enhance teaching, we further propose *optimal class partitioning*, where each partition of learners may have its own optimal training set. We show that the optimal partition can lead to a reduction in *aggregate* teaching dimension while still ensuring learner success. Interestingly, the optimal partition is not necessarily fully personalized education (each learner in its own partition) but rather encourages appropriate sharing.

## 1 Limitation of Classic Machine Teaching

In this section we review machine teaching and point out its limitations. We consider the problem of teaching a supervised learner, where the input space is $\mathbb{X}$ (e.g. $\mathbb{R}^d$) and the output space is $\mathbb{Y}$ (e.g. $\{0, 1\}$ for classification or $\mathbb{R}$ for regression). Consider the set of (not necessarily *i.i.d.*) training sets of all sizes $\mathbb{D} = \cup_{n=0}^{\infty}(\mathbb{X} \times \mathbb{Y})^n$. Let $\Theta$ be the model space. A learning algorithm $A$ is a function $\mathbb{D} \mapsto 2^{\Theta}$, i.e. $A$ takes a training set $D \in \mathbb{D}$ and maps it to a set of models. For example, the classic version-space learner $A_{vs}$ maps a $D$ to $A_{vs}(D) = \{\theta \in \Theta : \theta \text{ consistent with } D\}$. As another example, a soft-margin SVM learner maps $D$ to the singleton set $A_{svm}(D) = \{\hat{\theta}_D\}$, where $\hat{\theta}_D$ is the regularized empirical risk minimizer under the hinge loss.

**Definition 1 (Teaching Dimension)** *The TD of a target model $\theta^* \in \Theta$ for learner $A$ is the size of the smallest training*

[1] Supported by NSF IIS-1623605, DGE-1545481, IIS-0953219, CCF-1423237, CMMI-1561512, Fundação para a Ciência e a Tecnologia (FCT) UID/CEC/50021/2013, and EU FP7-ICT 610878. We thank Chuck Kalish for discussions on education.

*set such that $A$ learns $\theta^*$:*

$$TD_A(\theta^*) = \min_{D \in \mathbb{D}} |D|, \quad s.t.\ A(D) = \{\theta^*\}.$$

Here $|D|$ is the cardinality of $D$. Note this is exact teaching since $A$ must exclude all other models in $\Theta$. If $A = A_{vs}$ we recover the classic TD [Goldman and Kearns, 1995]; If $A =$ linear learners we recover the TD in [Liu *et al.*, 2016]. TD is known for monomials, hypercubes, monotone disjunctive normal forms, linear regression, logistic regression, SVMs, etc. TD has also been extended to cooperative teacher/learner pairs [Zilles *et al.*, 2011; Balbach, 2008].

To illustrate,[2] consider a student $A_\lambda$ who runs ridge regression with regularization parameter $\lambda > 0$ on training data $D = (X, Y)$ with $X \in \mathbb{R}^{n \times d}$, $Y \in \mathbb{R}^n$:

$$\begin{aligned} A_\lambda(D) &= \text{argmin}_\theta \frac{1}{2}\|X\theta - Y\|^2 + \frac{\lambda}{2}\|\theta\|^2 \\ &= (X^\top X + \lambda I)^{-1} X^\top Y. \end{aligned} \quad (1)$$

For any target model $\theta^* \in \mathbb{R}^d$, $\theta^* \neq 0$, the teacher can construct a single training item with $X = \theta^{*\top}$, $Y = \lambda + \|\theta^*\|^2$ so that $A_\lambda(D) = \{\theta^*\}$. Therefore, $TD_{A_\lambda}(\theta^*) = 1, \forall \theta^* \neq 0$ [Liu *et al.*, 2016].

Now consider a class of two students $A_\lambda$ and $A_{\lambda'}$, both running ridge regression but with $\lambda \neq \lambda'$. Critically, *we assume that the teacher has to deliver the same lecture $D$ to all students in the class.* It is easy to see that there is no finite $D$ that can make both students learn exactly $\theta^*$. Teaching dimension is not well-defined for a classroom setting.

## 2 Approximately Teaching a Single Student

Our plan to extend teaching dimension to a classroom setting has two steps. In this section we allow teaching to be approximate on a single student. In the next section we handle multiple students.

As the previous example shows, with multiple students but the same lecture it is generally impossible to achieve exact teaching for everyone. The teacher has to allow a certain degree of failure in some students' learned model.

**Definition 2 (Student failure)** *Given a target model $\theta^*$, the student failure is a function $\rho : 2^\Theta \mapsto \mathbb{R}_+$ with the property $\rho(\{\theta^*\}) = \min_{S \subseteq \Theta} \rho(S)$.*

**Example 1** *For a ridge regression learner* (1)*, one may define $\rho(A(D)) = \|A(D) - \theta^*\|^2$. We slightly abused the notation to use $A(D)$ for the element in the singleton set.*

**Example 2** *Exact teaching can be recovered by defining $\rho(A(D)) = 1_{[A(D) \neq \{\theta^*\}]}$.*

We say that a student $A$ $\epsilon$-approximately learns from lecture $D$ if $\rho(A(D)) \leq \rho(\{\theta^*\}) + \epsilon$. This allows us to extend teaching dimension to approximate teaching:

**Definition 3 (Approximate Teaching Dimension)** *The approximate teaching dimension is the smallest training set size for the learner to $\epsilon$-approximately learn $\theta^*$. Formally,*

$$TD_{A,\epsilon}(\theta^*) = \min_{D \in \mathbb{D}} |D|, \quad s.t.\ \rho(A(D)) \leq \rho(\{\theta^*\}) + \epsilon.$$

[2] We will use anthropomorphic terms in the paper by calling $A$ a student, $D$ a lecture, and so on; we use "diversity" to mean academic diversity, i.e. differences in the learning algorithm $A$.

**Proposition 1** *The more the teacher tolerates student failure, the easier (i.e. requiring less training items) it is to teach:*

1. $\forall \epsilon_1 \geq \epsilon_2, TD_{A,\epsilon_1}(\cdot) \leq TD_{A,\epsilon_2}(\cdot)$.
2. $\forall \epsilon > 0, TD_{A,\epsilon}(\cdot) \leq TD_A(\cdot)$.

Our definition of approximate TD bears a resemblance to the $k$-optimal teaching set in [Kobayashi and Shinohara, 2009], and can be viewed as its generalization to arbitrary learners and student failure functions. As is the case with classic TD [Goldman and Kearns, 1995], approximate TD is well-defined for all learners $A : \mathbb{D} \mapsto 2^\Theta$ but may not admit a closed-form expression. For concreteness, throughout the paper we will present two case studies: teaching conjugate Bayesian learners and ridge regression learners. We do not suggest that real-world students can be reduced to such simple learners. Nonetheless, the case studies will serve to illuminate the nuances of classroom teaching.

### 2.1 Case Study 1: Conjugate Bayesian Learner

The teacher wants to teach the mean $\mu^* \in \mathbb{R}$ of a Gaussian distribution $N(\mu^*, \sigma^2)$ to a student by constructing a training set $D = \{x_1, \ldots, x_n\}$. The teacher knows the following: the student $A_{\mu_0}$ is a Bayesian learner with a conjugate prior $\mu \sim N(\mu_0, \sigma_0^2)$ on the mean; $A_{\mu_0}$ knows the variance $\sigma^2$; Given $D$, $A_{\mu_0}$ performs standard Bayesian update on the mean:

$$A_{\mu_0}(D) = N\left(\frac{\sigma_0^2 \sum x_i + \sigma^2 \mu_0}{\sigma_0^2 n + \sigma^2}, \frac{1}{\frac{n}{\sigma^2} + \frac{1}{\sigma_0^2}}\right). \quad (2)$$

The teacher measures student failure by the squared distance between the posterior mean and $\mu^*$, disregarding posterior variance:

$$\rho(A_{\mu_0}(D)) = \left(\frac{\sigma_0^2 \sum x_i + \sigma^2 \mu_0}{\sigma_0^2 n + \sigma^2} - \mu^*\right)^2. \quad (3)$$

**Proposition 2** $TD_{A_{\mu_0},\epsilon}(\mu^*) = \begin{cases} 0, & \|\mu_0 - \mu^*\|^2 \leq \epsilon \\ 1, & \textit{otherwise.} \end{cases}$

**Proof:** If $\|\mu_0 - \mu^*\|^2 \leq \epsilon$ then there is no need to teach. Otherwise, the teacher can use a single training item:

$$x_1 = \mu^* + \frac{\sigma^2}{\sigma_0^2}(\mu^* - \mu_0) \quad (4)$$

to make $\rho(A_{\mu_0}(\{x_1\})) = 0$ by plugging it to (3). Note one teaching item is enough regardless of how far $\mu_0$ is from $\mu^*$. This case study can be generalized to exponential family learners [Zhu, 2013]. ■

### 2.2 Case Study 2: Ridge Regression Learner

For simplicity we discuss the 1D case. The teacher wants to teach the slope $\theta^* \in \mathbb{R}$ of a linear function $y = \theta^* x$ to a student by constructing a training set $D = \{(x_1, y_1), \ldots, (x_n, y_n)\}$. We further assume all training items are bounded: $|x| \leq 1$. The teacher knows the following: the student performs ridge regression as in (1) with regularization parameter $\lambda$. Given $D$, the student estimates the parameter as

$$A_\lambda(D) = \frac{\sum_{i=1}^n x_i y_i}{\lambda + \sum_{i=1}^n x_i^2}. \quad (5)$$

The teacher measures student failure by

$$\rho(A_\lambda(D), \theta^*) = \left( \frac{\sum_{i=1}^n x_i y_i}{\lambda + \sum_{i=1}^n x_i^2} - \theta^* \right)^2. \quad (6)$$

**Proposition 3** $TD_{A_\lambda,\epsilon}(\theta^*) = \begin{cases} 0, & \|\theta^*\|^2 \le \epsilon \\ 1, & \textit{otherwise} \end{cases}$

**Proof:** If $\|\theta^*\|^2 \le \epsilon$ there is no need to teach, since without teaching $A_\lambda(\emptyset) = \{0\}$ and this already satisfies approximate teaching. Otherwise, define

$$s \equiv \sum_{i=1}^n x_i^2, \; p \equiv \sum_{i=1}^n x_i y_i. \quad (7)$$

We have a constraint $0 \le s \le n$ due to the norm constraint on $x$. The teaching problem is

$$\min_{n,s,p} n \quad \text{s.t.} \left( \frac{p}{\lambda + s} - \theta^* \right)^2 \le \epsilon, \quad \text{and} \quad 0 \le s \le n. \quad (8)$$

By inspection, one solution is $n = 1, \; x = 1, \; y = \theta^*(1 + \lambda)$. With this training set, $A_\lambda(\{x, y\}) = \theta^*$ and student failure $\rho(A_\lambda(\{x, y\})) = 0$. ■

Both case studies will continue.

## 3 Teaching an Academically Diverse Class

We are now ready for the main problem of the paper. There are multiple (possibly infinite) students in a class, each with a different learning algorithm $A$. We denote the set of students by $\mathbb{A}$. The teacher knows all students $\mathbb{A}$, and must use the *same* lecture $D$ to teach them. As shown earlier, exact teaching is impossible in general. The teacher's goal is then to use as few training items as possible to make sure that no student failure is too severe – i.e. to leave no learner behind.

Formally we propose a minimax formulation: The teacher seeks the minimum lecture $D$ for all students, such that the worst student's failure is less than $\epsilon$:

$$\min_{D \in \mathbb{D}} \quad |D| \quad (9)$$

$$\text{s.t.} \quad \sup_{A \in \mathbb{A}} \rho(A(D)) \le \rho(\{\theta^*\}) + \epsilon. \quad (10)$$

**Definition 4 (Class Teaching Dimension)** *The class teaching dimension $TD_{\mathbb{A},\epsilon}(\theta^*)$ for all students $A$ in class $\mathbb{A}$ to $\epsilon$-approximately learn a target model $\theta^* \in \Theta$ w.r.t. student failure function $\rho$ is the minimum objective of* (9).

It follows that teaching a more academically diverse classroom requires no less effort.

**Proposition 4** $\forall \mathbb{A} \subseteq \mathbb{A}', \forall \epsilon \ge 0, TD_{\mathbb{A},\epsilon}(\cdot) \le TD_{\mathbb{A}',\epsilon}(\cdot)$.

Also the more class failure the teacher tolerates, the easier it is to teach:

**Proposition 5** $\forall \mathbb{A}, \forall \epsilon_1 \ge \epsilon_2, TD_{\mathbb{A},\epsilon_1}(\cdot) \le TD_{\mathbb{A},\epsilon_2}(\cdot)$.

For general learners the class TD can be difficult to compute, since $\min_D$ is over a combinatorial space of training sets and $\sup_A$ involves highly nonlinear functions. We turn our attention to the two case studies, where we are able to derive a closed-form expression for $TD_{\mathbb{A},\epsilon}$. They serve to quantitatively illustrate Proposition 4, namely teaching becomes harder if the class is more inhomogeneous.

### 3.1 Teaching a Class in Case Study 1

We continue the story in case study 1. Now instead of a single student $A_{\mu_0}$, there is a class of (possibly infinite) students $\mathbb{A}$. All students know that the target distribution has the form $N(\cdot, \sigma^2)$. They know $\sigma^2$ and must infer the mean from the teacher's training data $D = \{x_1, \ldots, x_n\}$. All students are conjugate Bayesian learners. Each student $A_{\mu_0^i}$ has a different prior on the mean $\mu \sim N(\mu_0^i, \sigma_0^2)$, where the prior mean $\mu_0^i$ varies across the students but all students have the same prior variance $\sigma_0^2$. We can identify the class $\mathbb{A} = \{\mu_0^i : A_{\mu_0^i} \text{ is a student}\}$ by the set of prior means. Define the lower and upper value of prior means as

$$\mu_0^L = \inf \mathbb{A}, \; \mu_0^U = \sup \mathbb{A}. \quad (11)$$

The teacher knows all the above information, and must design a good lecture $D$ to $\epsilon$-approximately teach all students the true mean $\mu^*$. The teacher uses the same student failure function as in (3). We now show that the class TD is a monotonic function of $\mu_0^U - \mu_0^L$, namely the spread of students.

**Lemma 1** *Let $a > 0$ and $b_1 \le b_2$. We have*

$$\frac{1}{4}(b_2 - b_1)^2 = \min_{x \in \mathbb{R}} \left\{ f(x) := \max\{(ax - b_1)^2, (ax - b_2)^2\} \right\}$$

*and the optimal $x^*$ satisfies $x^* = \frac{b_1 + b_2}{2a}$.*

We omit proof for lemmas due to space, while sketching out proof outlines for theorems.

**Theorem 1** *The class TD of conjugate Bayesian learners is*

$$TD_{\mathbb{A},\epsilon}(\mu^*) = \begin{cases} 0, & \textit{if } \sup_{\mu_0 \in \mathbb{A}} \|\mu_0 - \mu^*\|^2 \le \epsilon \\ \left\lceil \max\left\{ \left( \frac{\mu_0^U - \mu_0^L}{2\sqrt{\epsilon}} - 1 \right) \frac{\sigma^2}{\sigma_0^2}, 1 \right\} \right\rceil, & \textit{o.w.} \end{cases} \quad (12)$$

**Proof:** If $\sup_{\mu_0 \in \mathbb{A}} \|\mu_0 - \mu^*\|^2 \le \epsilon$ then there is no need to teach, and $TD_{\mathbb{A},\epsilon}(\mu^*) = 0$. Otherwise, $TD_{\mathbb{A},\epsilon}(\mu^*) \ge 1$. The training set $D$ can be reduced to its sufficient statistics $\sum x_i$ and $n$. We may write the class TD problem (9) as

$$\min_{n, \sum x_i} n \quad \text{s.t.} \sup_{\mu_0 \in \mathbb{A}} \rho(A_{\mu_0}(D)) \le \epsilon. \quad (13)$$

Plugging in the posterior mean for $A_{\mu_0}(D)$ in (2), the constraint involves a quadratic form in $\mu_0$:

$$\sup_{\mu_0 \in \mathbb{A}} \left( \frac{\sigma^2}{\sigma_0^2 n + \sigma^2} \mu_0 + \frac{\sigma_0^2 \sum x_i}{\sigma_0^2 n + \sigma^2} - \mu^* \right)^2 \le \epsilon. \quad (14)$$

Since the quadratic form is convex, the supremum occurs at the boundary of $\mathbb{A}$. This allows us to replace the supremum over possibly infinite students with a maximum between the two extreme students $A_{\mu_0^L}, A_{\mu_0^U}$:

$$\max_{\mu_0 \in \{\mu_0^L, \mu_0^U\}} \left( \frac{\sigma^2}{\sigma_0^2 n + \sigma^2} \mu_0 + \frac{\sigma_0^2 \sum x_i}{\sigma_0^2 n + \sigma^2} - \mu^* \right)^2 \le \epsilon. \quad (15)$$

In other words, one of these two extreme students will have the worst student failure in the class.

We now fix $n$ and treat $\sum x_i$ as a free variable in $\mathbb{R}$. To satisfy the constraint as much as possible, the optimal $\sum x_i$

should minimize the LHS of (15). From Lemma 1, we have $\frac{\sigma_0^2 \sum x_i}{\sigma_0^2 n+\sigma^2} - \mu^* = \frac{\sigma^2}{\sigma_0^2 n+\sigma^2} \frac{-\mu_0^L-\mu_0^U}{2}$. This leads to the solution

$$\sum x_i = \frac{\sigma_0^2 n + \sigma^2}{\sigma_0^2} \mu^* - \frac{\sigma^2}{\sigma_0^2} \frac{\mu_0^L + \mu_0^U}{2}. \quad (16)$$

Plugging it back in (15), we have the following optimization problem:

$$\min_n n \quad \text{s.t.} \left( \frac{\mu_0^U - \mu_0^L}{2} \cdot \frac{\sigma^2}{\sigma_0^2 n + \sigma^2} \right)^2 \leq \epsilon. \quad (17)$$

The constraint is

$$n \geq \left( \frac{\mu_0^U - \mu_0^L}{2\sqrt{\epsilon}} - 1 \right) \frac{\sigma^2}{\sigma_0^2}. \quad (18)$$

Recall that we must have $n \geq 1$ to teach, and $n$ must be an integer. Taken together, the minimum $n$ is the class TD in (12). An actual lecture $D$ can be easily constructed from $n$ and $\sum x_i$. ■

Theorem 1 can be interpreted as follows. The class TD of those conjugate Bayesian learners depends on class spread $\mu_0^U - \mu_0^L$, but not on how many students there are in the class. When the spread is sufficiently small, it is possible to teach the whole class successfully with a single training item. The threshold for this spread is

$$\mu_0^U - \mu_0^L = 2\sqrt{\epsilon} \left( \frac{\sigma_0^2}{\sigma^2} + 1 \right). \quad (19)$$

It is easy to see that teaching is easier as the failure tolerance $\epsilon$ increases. Teaching is also easy if the students' prior variance $\sigma_0^2$ is large: in that case the students are less "stubborn" around their prior means and can be easily convinced by new evidence in data $D$. But when the class spread is sufficiently large, $TD_{\mathbb{A},\epsilon}$ is a linear function of the spread (up to rounding). The teacher has to use more training items to teach as the class spread increases.

### 3.2 Teaching a Class in Case Study 2

To extend case study 2 to multiple students, let the (possibly infinite) students all be ridge regression learners as in (1) but with different regularization weights $\lambda$. Let $\Lambda$ be the set of regularization weights, which is equivalent to the class $\mathbb{A}$. The teacher knows $\Lambda$ and must design a lecture $D = \{(x_1, y_1) \dots (x_n, y_n)\}$ to $\epsilon$-approximately teach all students in the class. Again, the teacher uses the student failure function (6).

Recall the definition of $s, p$ in (7). The class teaching dimension problem (9) for this class can be stated as an extension of problem (8) to multiple learners:

$$\min_{n,s,p} n \quad \text{s.t.} \sup_{\lambda \in \Lambda} \left( \frac{p}{\lambda + s} - \theta^* \right)^2 \leq \epsilon, \quad 0 \leq s \leq n. \quad (20)$$

Define

$$\lambda^L = \inf \Lambda, \quad \lambda^U = \sup \Lambda. \quad (21)$$

The class TD of ridge regression learners depends on these two quantities, but in a more complex manner:

**Lemma 2** *The student failure function* $\rho(A_\lambda(D), \theta^*) = \left( \frac{p}{\lambda+s} - \theta^* \right)^2$ *is quasiconvex in* $\lambda$.

**Theorem 2** *The class TD of ridge regression learners is*

$$TD_{\Lambda,\epsilon}(\theta^*) = \begin{cases} 0, & \text{if } \|\theta^*\|^2 \leq \epsilon \\ \left\lceil \max\left\{ \frac{\lambda^U - \lambda^L}{2\sqrt{\epsilon}} |\theta^*| - \frac{\lambda^L + \lambda^U}{2}, 1 \right\} \right\rceil, & \text{o.w.} \end{cases} \quad (22)$$

**Proof:** If $\|\theta^*\|^2 \leq \epsilon$ the teacher can let $D = \emptyset$. The argmin in ridge regression (1) is then determined by the regularizer. This leads to $A_\lambda(\emptyset) = 0, \forall \lambda \in \Lambda$, and 0 is within $\epsilon$ of the target $\theta^*$ by definition. Therefore, in this case the class TD is zero.

Otherwise, we must have $TD_{\Lambda,\epsilon} \geq 1$. By Lemma 2 the supremum in constraint (20) happens at one of $\lambda^L, \lambda^U$ due to quasi-convexity. We may replace the supremum over the whole $\Lambda$ by a maximum over those two values:

$$\max \left\{ \left( \frac{p}{\lambda^L + s} - \theta^* \right)^2, \left( \frac{p}{\lambda^U + s} - \theta^* \right)^2 \right\} \leq \epsilon. \quad (23)$$

Fixing $n, s$ and applying Lemma 1 again, the optimal $p$ denoted by $p^*(s)$ must happen at $\frac{p^*(s)}{\lambda^L+s} - \theta^* = -\frac{p^*(s)}{\lambda^U+s} + \theta^*$. This leads to $p^*(s) = 2\theta^* \left( \frac{1}{\lambda^L+s} + \frac{1}{\lambda^U+s} \right)^{-1}$. Plugging back this optimal value of $p$ the teaching problem becomes

$$\min_{n,s} n \quad \text{s.t.} \left( \frac{\lambda^U - \lambda^L}{\lambda^L + \lambda^U + 2s} \theta^* \right)^2 \leq \epsilon, \quad \text{and } 0 \leq s \leq n. \quad (24)$$

By inspection the (real-valued) solution is

$$n = s = \frac{\lambda^U - \lambda^L}{2\sqrt{\epsilon}} |\theta^*| - \frac{\lambda^L + \lambda^U}{2}. \quad (25)$$

The class TD is obtained by $n \geq 1$ and rounding. An optimal training set can be constructed from $n, s, p$. ■

Again, the class TD depends on the class spread $\lambda^U - \lambda^L$, not the number of students. As the class spread becomes wider, the class TD increases at a rate of $(\lambda^U - \lambda^L)/(2\sqrt{\epsilon})$.

## 4 Optimal Class Partitioning

The previous section highlights Proposition 4, namely it is hard to teach an academically diverse class. The root of the problem is the classroom setting, where the teacher must use the same lecture $D$ on all students. We now consider a setting where the teacher is allowed to partition the class into sections, and different sections may use different lectures. This is a familiar tradeoff in reality: Parents want smaller class sizes, but schools need to control expenses. One extreme partition is to have each student in their own section, i.e., personalized education. But this may not be optimal in the teacher's effort to deliver lectures: presumable very similar students should share the same lecture. We now show that there is a sweet spot, such that the aggregate training set size is minimized while guaranteeing no learner left behind.

We formulate the optimal class partition problem as follows. For a class $\mathbb{A}$ we use $\pi = \pi(\mathbb{A})$ to denote a partition,

i.e. a set of sections: $\emptyset \notin \pi$; $\cup_{\mathbb{S}\in\pi}\mathbb{S} = \mathbb{A}$; $\forall \mathbb{S}, \mathbb{S}' \in \pi, \mathbb{S} \neq \mathbb{S}' \Rightarrow \mathbb{S} \cap \mathbb{S}' = \emptyset$. Each section $\mathbb{S} \in \pi$ is now a smaller class. To $\epsilon$-approximately teach all students in $\mathbb{S}$, by definition the teacher needs $TD_{\mathbb{S},\epsilon}(\theta^*)$ training items. Define the aggregate training set size under $\pi$ as $n(\pi) \equiv \sum_{\mathbb{S}\in\pi} TD_{\mathbb{S},\epsilon}(\theta^*)$. This implies that the teaching effort is the accumulative teaching time over all sections, where each training item takes unit time. Note $n(\pi)$ is not the number of unique training items across all sections.

**Definition 5** *The partition teaching dimension is defined as*

$$TD_{\pi^*,\epsilon}(\theta^*) = \min_{\pi} n(\pi) \tag{26}$$

*where the optimal class partition $\pi^*$ is the one that achieves the minimum.*

Our two case studies, continued below, demonstrates that there are nontrivial class partition sweet spots.

### 4.1 The Optimal Class Partition in Case Study 1

Recall the class $\mathbb{A}$ consists of students with different prior distributions $N(\mu_0, \sigma_0^2)$, where $\mu_0$ is bounded between $\mu_0^L$ and $\mu_0^U$. We first note that, unlike the class TD which only depends on the spread $\mu_0^U - \mu_0^L$, the optimal partition also depends on the topology of $\mathbb{A}$. For example, when all students form two degenerate (elements overlapping) clusters at the two extrema $\mu_0^L$ and $\mu_0^U$, the teacher can partition the class into two sections by the clusters. Each section would require at most one training item per Theorem 1, leading to a partition TD of at most two. In contrast, the partition TD is very different when the class is dense in the interval $\mathbb{A} = [\mu_0^L, \mu_0^U]$, which is our focus below.

Let $\mathbf{cl}(\cdot)$ be the closure of a set (for example, $\mathbf{cl}([0,1)) = [0,1]$), and $\mathbf{conv}(\cdot)$ be the convex hull of a set.

**Lemma 3** *Given any $\mathbb{A} \subset \mathbb{R}$, we have the class TD*

$$TD_{\mathbb{A},\epsilon}(\cdot) = TD_{\mathbf{conv}(\mathbb{A}),\epsilon}(\cdot) = TD_{\mathbf{cl}(\mathbf{conv}(\mathbb{A})),\epsilon}(\cdot).$$

**Lemma 4** *For a bounded $\mathbb{A} \subset \mathbb{R}$, one optimal partition $\pi^*$ where $\mathbb{A} = \cup_{i=1}^{TD_{\pi^*,\epsilon}(\cdot))}\mathbb{A}_i$ as in* (26) *must satisfy*

1. *All $\mathbf{conv}(\mathbb{A}_i)$'s are disjoint;*
2. *Any $\mathbb{A}_i$ can be taught by 1 item, that is, the class $TD_{\mathbb{A}_i,\epsilon}(\cdot) = 1$.*

**Theorem 3** *For $\mathbb{A} = [\mu_0^L, \mu_0^U]$ satisfying $\sup_{\mu\in\mathbb{A}} \|\mu - \mu^*\|^2 > \epsilon$, an optimal partitioning $\pi^*$ is to break $\mathbb{A}$ into intervals each of length defined by the RHS of* (19)*, where each section can be taught with a single teaching item. The corresponding partition teaching dimension is*

$$TD_{\pi^*,\epsilon}(\mu^*) = \left\lceil \frac{\mu_0^U - \mu_0^L}{2\sqrt{\epsilon}\left(\frac{\sigma_0^2}{\sigma^2} + 1\right)} \right\rceil. \tag{27}$$

**Proof:** From Lemma 4, we know that one optimal partition should be in the form of $\mathbb{A} = \cup_{i=1}^{TD_{\pi^*,\epsilon}}\mathbb{A}_i$ with $\mathbb{A}_i = [\mu^{i-1}, \mu^i]$ where $\mu^0 = \mu_0^L$ and $\mu^{TD_{\pi^*,\epsilon}} = \mu_0^U$. Given such partition structure, we can apply the greedy procedure to find one optimal partition: start from $\mu^0$ and maximize $\mu^1$ for $\mathbb{A}_1$ given $TD_{\mathbb{A}_1,\epsilon} = 1$, then maximize $\mu^2$ for $\mathbb{A}_2$ given $TD_{\mathbb{A}_2,\epsilon} = 1$, and so on until that $\mu^i$ exceeds $\mu_0^U$. From Theorem 1, the largest range of students $\mu^i - \mu^{i-1}$ that one training item can teach satisfies:

$$\left\lceil \max\left\{ \left( \frac{\mu^i - \mu^{i-1}}{2\sqrt{\epsilon}} - 1 \right) \frac{\sigma^2}{\sigma_0^2}, 1 \right\} \right\rceil = 1.$$

(Note that the "0" teaching case is excluded automatically because $\sup_{\mu\in\mathbb{A}} \|\mu - \mu^*\|^2 > \epsilon$). It implies that the largest range is $\mu^i - \mu^{i-1} \leq 2\sqrt{\epsilon}\left(\frac{\sigma_0^2}{\sigma^2} + 1\right)$, precisely (19). Thus the minimal number of sections is given by (27). ■

Comparing the partition TD in (27) to the whole class TD in (12), we see that class partition is especially helpful when $\sigma_0^2 \to 0$, i.e. when the students are stubborn. Stubborn students blow up the whole class TD, but having them in different sections prevents the blow up due to the added 1 in the denominator.

### 4.2 The Optimal Class Partition in Case Study 2

For simplicity, we consider a class consisting of students dense in the interval $\mathbb{A} \equiv \Lambda = [\lambda^L, \lambda^U]$. If $|\theta^*| \leq \sqrt{\epsilon}$ then trivially $TD_{\pi,\epsilon}(\theta^*) = 0$. We only consider the nontrivial case $|\theta^*| > \sqrt{\epsilon}$ below.

**Theorem 4** *For a class $\mathbb{A} = [\lambda^L, \lambda^U]$ and $|\theta^*| > \sqrt{\epsilon}$, let $a = 2\left(\frac{|\theta^*|}{\sqrt{\epsilon}} - 1\right)^{-1}$. The partition teaching dimension is*

$$TD_{\pi^*,\epsilon}(\theta^*) = \left\lceil \log_{(1+a)}\left(\frac{\lambda^U + 1}{\lambda^L + 1}\right) \right\rceil. \tag{28}$$

**Proof:** (sketch) It can be shown that each section should be an interval of students; Moreover, if one interval of students can be taught by $n$ training items, then it can be partitioned into two intervals such that one can be taught by $n-1$ items, and the other can be taught by 1 item. These imply that an optimal class partition divides $\Lambda$ at $\lambda^L = \lambda_0, \lambda_1, \ldots, \lambda_{n-1} < \lambda^U, \lambda_n \geq \lambda^U$, such that each section (except the last one) is maximal while still can be $\epsilon$-approximately taught with one training item. Then $\pi^* = \{[\lambda_{i-1}, \lambda_i] : i = 1 \ldots n\}$ and $TD_{\pi^*,\epsilon}(\theta^*) = n$.

To find out the values of $\lambda_i$ and $n$, we identify those maximal sections. Starting from $\lambda^L = \lambda_0$ and applying (22), the maximum $\lambda_1$ must satisfy $\frac{\lambda_1 - \lambda_0}{2\sqrt{\epsilon}}|\theta^*| - \frac{\lambda_0 + \lambda_1}{2} = 1$. This leads to $\lambda_1 = a + (1+a)\lambda_0$ and more generally $\lambda_t = a + (1+a)\lambda_{t-1}$. Note the intervals increase with $t$, which is different from the equal partitions in case study 1. Let $z_t = \lambda_t + 1$ then the above series becomes $z_t = (1+a)z_{t-1} = (1+a)^t z_0$, hence $\lambda_t = (1+a)^t(\lambda^L + 1) - 1$. The value $n$ is the smallest $t$ such that $\lambda_t \geq \lambda^U$, namely $t \geq \log_{(1+a)}\left(\frac{\lambda^U+1}{\lambda^L+1}\right)$. Rounding completes the proof. Compared to the class TD which is $O(\lambda^U - \lambda^L)$, the partition TD is always better. ■

### 4.3 A Simulation for Illustration

We illustrate in Figure 1 the benefit of teaching an optimally partitioned class with a simulation for case study 2. There are 1000 ridge regression students (1) with $\lambda^L = 10^{-1}, \lambda^U =$

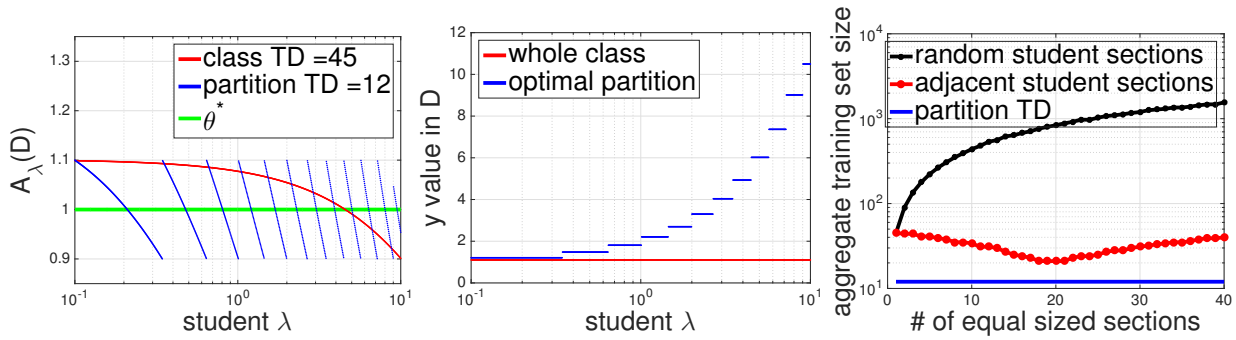


Figure 1: Demonstration of partition TD much smaller than class TD. (Left) model $A_\lambda(D)$ learned by student $\lambda$; (Center) Optimal training sets $D$; and (Right) Non-optimal class partitions. See text for details.

$10^1$ and the students are evenly spaced in $\log(\lambda)$. The students form the class $\mathbb{A}$ (or $\Lambda$), and they are the $x$-axis in Figure 1. The teacher wants to $\epsilon$-approximately teach the target model $\theta^* = 1$, with $\epsilon = 0.01$.

The left panel shows what model each student learned. The green curve simply marks the target model $\theta^*$. The red curve shows the effect of teaching the whole class $\mathbb{A}$ with the optimal "whole-class lecture" $D$ as defined in Theorem 2 proof. The class $TD_{\mathbb{A},0.01}(1) = 45$. The teacher needs 45 training items in $D$ to guarantee $\epsilon$-learning for everyone. Since we defined student failure as the squared difference (6), the students succeed if they learn a model $A_\lambda(D)$ within $\pm\sqrt{\epsilon} = \pm 0.1$ of $\theta^*$. This is precisely what the red curve shows. In contrast, the blue curves show the optimal partition as in Theorem 4 proof. Each blue segment represents a section that can be $\epsilon$-taught with one training item. Indeed each section of the learned models are within $\theta^* \pm 0.1$. There are 12 sections in the optimal partition: $TD_{\pi^*,0.01}(1) = 12$.

The center panel shows the lectures $D$. In all $D$, we have the construction $x = 1$ always, while $y$ and multiplicity differ. The curves should be read as follows: the value $y$ at a student $\lambda$ means that the student is sharing a lecture which contains the training point $(x = 1, y)$. The red curve gives $y = 1.101$ for the "whole class lecture," which means that $D$ consists of 45 identical instances of the point $(1, 1.101)$. The blue curve segments are for the 12 sections of the class, each section's $D$ has a single instance. For example, the right-most section has $D = \{(1, 10.496)\}$. Note how the teacher increases $y$ for sections with larger $\lambda$'s to account for regularization.

The right panel shows the aggregate training set size $n(\pi)$ needed to teach various non-optimal class partitions $\pi$. We simply divide the 1000 students into equal-sized sections, where the number of sections is on the $x$-axis. The red curve shows the case where students with adjacent $\lambda$ values are partitioned together. When the number of sections is 1, the aggregate training set size is 45 as we mentioned earlier (it is bad to teach a large class). When the number of sections is 1000 (not shown), that size is trivially 1000 (it is inefficient to perform fully personalized education). Note the minimum happens at around 20 – not 12 – sections. And the minimum aggregate size is 21, larger than the partition TD of 12 indicated by the blue line (simply dividing the class evenly into sections does not quite work). This highlights the importance of intelligent uneven partitions computed by machine teaching (center panel) to achieve optimality. Finally, the black curve shows the case where the students are randomly assigned to each section. This is a bad idea as *each* section now takes about the same training set size to teach as the whole original class! Under our setting it is important to keep the sections homogeneous.

## 5 Discussions

Readers familiar with Probably Approximately Correct (PAC) learning and VC theory may wonder how our work relates to those framework. The class TD in Definition 4 provides a good contrast. We *are* controlling the error to be uniformly below $\epsilon$. But a major difference is that we do not rely on concentration of measure over *i.i.d.* samples in the training set. Instead, we start with the target concept $\theta^*$ and then optimize the training set $D$. Therefore, we do not have the "P" in PAC learning. In terms of VC dimension, it is known that VC and TD are distinct measures [Goldman and Kearns, 1995], and the distinction carries over to our classroom teaching setting.

In order to make analysis tractable, we made several simplifying assumptions that should be relaxed in future work. A strong assumption is that the teacher knows everything about the students $\mathbb{A}$ (i.e., their learning algorithms). Future work may allow the teacher to probe the students if this is not the case, and jointly optimize the effort of teaching and probing.

A variant that is ethically questionable but unfortunately occurs in practice is for a teacher to give up on at most $\delta$ fraction of the students, to allow them to have arbitrarily large student failures. The remaining students must $\epsilon$-approximately learn. This type of $(\epsilon, \delta)$ class TD can be much smaller.

Interestingly, in the education literature it was suggested that some cultures prefer large classes, with the argument that students can observe and learn from each other's mistakes [Stigler and Hiebert, 1998]. It is possible to build this into machine teaching theory by formulating the class $\mathbb{A}$ as a multi-agent system. More research is needed to understand how academic diversity benefits teaching.

In summary, this paper advances machine teaching theory and provides a potential theoretical framework for optimizing real-world education resource allocation in the future.

## References


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

[Alfeld *et al.*, 2017] S. Alfeld, X. Zhu, and P. Barford. Explicit defense actions against test-set attacks. In *AAAI*, 2017.

[Angluin and Krikis, 1997] D. Angluin and M. Krikis. Teachers, learners and black boxes. *COLT*, pages 285–297, 1997.

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

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

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

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

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

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

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

[Clement *et al.*, 2016] B. Clement, P.-Y. Oudeyer, and M. Lopes. A comparison of automatic teaching strategies for heterogeneous student populations. In *Educational Data Mining (EDM)*, 2016.

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

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

[Goldman and Mathias, 1996] S. A. Goldman and H. D. Mathias. Teaching a smarter learner. *Journal of Computer and Systems Sciences*, 52(2):255267, 1996.

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

[Lee and Brunskill, 2012] J.I. Lee and E. Brunskill. The impact on individualizing student models on necessary practice opportunities. In *EDM*, 2012.

[Lindsey *et al.*, 2013] R. Lindsey, Michael Mozer, William J Huggins, and Harold Pashler. Optimizing instructional policies. In C.J.C. Burges, L. Bottou, M. Welling, Z. Ghahramani, and K.Q. Weinberger, editors, *Advances in Neural Information Processing Systems 26*, pages 2778–2786. 2013.

[Liu *et al.*, 2016] J. Liu, X. Zhu, and H. G. Ohannessian. The teaching dimension of linear learners. In *ICML*, 2016.

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

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

[Patil *et al.*, 2014] K. Patil, X. Zhu, L. Kopec, and B. C. Love. Optimal teaching for limited-capacity human learners. *NIPS*, pages 2465–2473, 2014.

[Rivest and Yin, 1995] R. L. Rivest and Y. L. Yin. Being taught can be faster than asking questions. *COLT*, pages 144–151, 1995.

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

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

[Stigler and Hiebert, 1998] J. W Stigler and J. Hiebert. Teaching is a cultural activity. *American Educator*, 22(4):4–11, 1998.

[Suh *et al.*, 2016] J. Suh, X. Zhu, and S. Amershi. The label complexity of mixed-initiative classifier training. *ICML*, 2016.

[Valiant, 1984] L. G. Valiant. A theory of the learnable. In *STOC '84: Proceedings of the sixteenth annual ACM symposium on Theory of computing*, pages 436–445, New York, NY, USA, 1984. ACM.

[Vapnik, 1995] Vladimir Vapnik. *The Nature of Statistical Learning Theory*. Springer, second edition, 1995.

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

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

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