The Teaching Dimension of Linear Learners

Ji Liu JI.LIU.UWISC@GMAIL.COM
University of Rochester, Rochester, NY 14627 USA

Xiaojin Zhu JERRYZHU@CS.WISC.EDU
H. Gorune Ohannessian OHANNESSIAN@WISC.EDU
University of Wisconsin-Madison, Madison, WI 53706 USA

Abstract

Teaching dimension is a learning theoretic quantity that specifies the minimum training set size to teach a target model to a learner. Previous studies on teaching dimension focused on version-space learners which maintain all hypotheses consistent with the training data, and cannot be applied to modern machine learners which select a specific hypothesis via optimization. This paper presents the first known teaching dimension for ridge regression, support vector machines, and logistic regression. We also exhibit optimal training sets that match these teaching dimensions. Our approach generalizes to other linear learners.

1 Introduction

Consider a teacher who knows both a target model and the learning algorithm used by a machine learner. The teacher wants to teach the target model to the learner by *constructing* a training set. The training set does not need to contain independent and identically distributed items drawn from some distribution. Furthermore, the teacher can construct any item in the input space. How many training items are needed? This is the question addressed by the *teaching dimension* (Goldman & Kearns, 1995; Shinohara & Miyano, 1991). We give the precise definition in section 2, but first illustrate the intuition with an example.

Consider integers $x \in \{1 \dots 10\}$ and threshold classifiers h_θ on them, so that $h_\theta(x)$ returns -1 if $x < \theta$ and 1 if $x \geq \theta$. Now let the hypothesis space $\mathcal{H}$ consist of eleven classifiers $\mathcal{H} = \{h_\theta \mid \theta \in \{1 \dots 11\}\}$. Let the learner be a version-space learner, namely it maintains a version space $\{h_\theta \in \mathcal{H} \mid h_\theta \text{ consistent with the training set}\}$. Equivalently, the learner is a 0-1 loss empirical risk minimizer (ERM) which finds all models with zero training error. If we want to teach a target model (in this paper we use hypothesis and model exchangeably), say h_9, to such a learner, we can construct a training set that results in a singleton version space $\{h_9\}$. It is easy to see that the training set $D = \{(x_1 = 8, y_1 = -1), (x_2 = 9, y_2 = 1)\}$ is the smallest set for this purpose. We say that the teaching dimension of h_9 with respect to $\mathcal{H}$ is $TD(h_9) = |D| = 2$. Similarly, $TD(h_{11}) = 1$ because $D = \{(x_1 = 10, y_1 = -1)\}$ suffices. In fact, $TD(h^*_\theta) = 1$ for target model $\theta^* = 1$ or 11, and 2 for $\theta^* = 2, 3, \dots, 10$.

The astute reader may notice that this example does not apply to continuous spaces. To see this, let us extend $x \in \mathbb{R}$ and $\mathcal{H} = \{h_\theta \mid \theta \in \mathbb{R}\}$. The learner's version space under any linearly separable training set would now be represented by the interval between the two closest oppositely labeled items. It is impossible for the version-space learner to pick out a unique target model h_{θ^*} with a finite training set. In other words, $TD(h_{\theta^*}) = \infty$ for all target models θ^*. This is counterintuitive because ostensibly we can teach any one of the "modern" machine learning algorithms such as a support vector machine (SVM) with only two training items: $D = \{(x_1 = \theta^* - \epsilon, y_1 = -1), (x_2 = \theta^* + \epsilon, y_2 = 1)\}$ with any $\epsilon > 0$. The issue here is that a version-space learner is not equipped with the ability to pick the max-margin (or any other specific) hypothesis from the version space. In contrast, an SVM is *not* a version-space learner in our terminology; we have stronger knowledge from optimization on how it picks a specific hypothesis from the hypothesis space. This paper will utilize such knowledge to derive teaching dimensions that are distinct from classic teaching dimension analysis (e.g. Doliwa et al. (2014)). Specifically, we extend teaching dimension to linear learners that learn by regularized surrogate-loss empirical risk minimization:

$$\mathcal{A}_{opt}(D) := \text{Argmin}_{\boldsymbol{\theta} \in \mathbb{R}^d} \underbrace{\sum_{i=1}^{n} \ell(\mathbf{x}_i^\top \boldsymbol{\theta}, y_i) + \frac{\lambda}{2} \|\boldsymbol{\theta}\|_A^2}_{=: f(\boldsymbol{\theta})} . \quad (1)$$

Here, we identify $\mathcal{H}$ with $\mathbb{R}^d$, h with $\boldsymbol{\theta}$, the surrogate loss function ℓ is either smooth or convex in the first argument,

Proceedings of the 33^{rd} International Conference on Machine Learning, New York, NY, USA, 2016. JMLR: W&CP volume 48.

	homogeneous			inhomogeneous		
	ridge	SVM	logistic	ridge	SVM	logistic
exact parameter	1	$\lceil \lambda \|\boldsymbol{\theta}^*\|^2 \rceil$	$\left\lceil \frac{\lambda \|\boldsymbol{\theta}^*\|^2}{\tau_{\max}} \right\rceil$	2	$2 \left\lceil \frac{\lambda \|\mathbf{w}^*\|^2}{2} \right\rceil^\dagger$	$2 \left\lceil \frac{\lambda \|\mathbf{w}^*\|^2}{2\tau_{\max}} \right\rceil^\dagger$
decision boundary	-	1	1	-	2	2

Table 1. The teaching dimension of ridge regression, SVM, and logistic regression. (†: up to rounding effect, see section 3.3).

$\lambda > 0$ is the regularization coefficient, and A is a positive semidefinite matrix. This covers both homogeneous (e.g. $A = I$) and inhomogeneous (e.g. $A = [I, 0; 0, I]$) learners. $\|\cdot\|_A$ is the Mahalanobis norm: $\|\boldsymbol{\theta}\|_A := \sqrt{\boldsymbol{\theta}^\top A \boldsymbol{\theta}}$. We follow the convention in optimization when we use the capitalized Argmin to emphasize that it returns a *set* that achieves the minimum. The teacher can construct a training set with any items in $\mathbb{R}^d$. The alternative pool-based teaching setting, where the teacher is given a finite pool of candidate training items and must select items from that pool, is not studied in this paper. By linear learners we mean the input $\mathbf{x}$ and the parameter $\boldsymbol{\theta}$ interact only via their inner product $\mathbf{x}^\top \boldsymbol{\theta}$. Linear learners include SVMs, logistic regression, and linear regression. Our analysis technique involves a novel application of the Karush-Kuhn-Tucker (KKT) conditions.

To our knowledge, this paper gives the first known values of teaching dimension for ridge regression, SVM, and logistic regression. We summarize our main results in Table 1. The table separately lists homogeneous (without a bias term) and inhomogeneous (with a bias term) versions of the linear learners. The teaching goal refers to the intention of the teacher: is teaching considered successful only when the learner learns the exact target parameter, or when the learner learns the correct decision boundary (which can be achieved by any positive scaling of the target parameter). See section 3 for definition of the target parameters $\boldsymbol{\theta}^*, \mathbf{w}^*$ and the constant $\tau_{\max}$. The target parameters are assumed to be nonzero. We will also present the corresponding minimum teaching set construction in section 3. All proofs in this paper are provided in Supplemental Material.

2 Classic Teaching Dimension and its Limitations

Let $\mathcal{X}$ denote the input space and $\mathcal{Y} \subseteq \mathbb{R}$ the output space. A hypothesis is a function $h : \mathcal{X} \to \mathcal{Y}$. In this paper we identify a hypothesis $h_{\boldsymbol{\theta}}$ with its model parameter $\boldsymbol{\theta}$. The hypothesis space $\mathcal{H}$ is a set of hypotheses. By training item we mean a pair $(x, y) \in \mathcal{X} \times \mathcal{Y}$. A training set is a multiset $D = \{(\mathbf{x}_1, y_1) \dots (\mathbf{x}_n, y_n)\}$ where repeated items are allowed. Importantly, for the purpose of teaching we do *not* assume that D be drawn $i.i.d.$ from a distribution. Let $\mathbb{D}$ denote the set of all training sets of all sizes. A learning algorithm $\mathcal{A} : \mathbb{D} \to 2^{\mathcal{H}}$ takes in a training set $D \in \mathbb{D}$ and outputs a subset of the hypothesis space $\mathcal{H}$. That is, $\mathcal{A}$ does not necessarily return a unique hypothesis.

Classic teaching dimension analysis is restricted to the version-space learner $\mathcal{A}_{vs}$:

$$\mathcal{A}_{vs}(D) = \{h \in \mathcal{H} \mid \ h \text{ is consistent with } D\ \}. \quad (2)$$

That is, the learner $\mathcal{A}_{vs}$ keeps track of the version space. Let the target model be $h_{\boldsymbol{\theta}^*} \in \mathcal{H}$. Teaching is successful if the teacher identifies a training set $D \in \mathbb{D}$ such that $\mathcal{A}_{vs}(D) = \{h_{\boldsymbol{\theta}^*}\}$ the singleton set. Such a D is called a **teaching set** of $h_{\boldsymbol{\theta}^*}$ with respect to $\mathcal{H}$. The teaching dimension of the hypothesis $h_{\boldsymbol{\theta}^*}$ is the minimum size of the teaching set:

$$TD(h_{\boldsymbol{\theta}^*}) = \begin{cases} \min_{D \in \mathbb{D}} |D|, & \text{for } D \text{ a teaching set of } h_{\boldsymbol{\theta}^*} \\ \infty, & \text{if no teaching set exists} \end{cases} \quad (3)$$

Furthermore, the teaching dimension of the whole hypothesis space $\mathcal{H}$ is defined by the hardest hypothesis: $TD(\mathcal{H}) = \max_{h \in \mathcal{H}} TD(h)$. In this paper we will focus on the fine-grained teaching dimension of individual hypothesis $TD(h)$.

Classic teaching dimension analysis has several limitations: the learner is assumed to be a version-space learner $\mathcal{A}_{vs}$, and the hypothesis space is typically finite or countably infinite. As the example in section 1 showed, these fail to capture the teaching dimension of "modern" machine learners which has $\mathbb{R}^d$ as input space and picks a unique hypothesis via regularized empirical risk minimization (1). Furthermore, the target model can be ambiguous when the learner is a classifier: should the learner learn the exact target parameter $\boldsymbol{\theta}^*$, or the target decision boundary? In linear models any scaled parameter $c\boldsymbol{\theta}^*$ with $c > 0$ produces the same target decision boundary. These limitations motivate us to generalize the teaching dimension in the next section.

3 Main Results

To make our teaching dimension's dependency on the learning algorithm explicit, henceforth we write teaching dimension with two arguments as

$$TD(h^*, \mathcal{A}) \quad (4)$$

where $h^* \in \mathcal{H}$ is the target model, and $\mathcal{A} : \mathbb{D} \to 2^{\mathcal{H}}$ is the learning algorithm which given a training set $D \in \mathbb{D}$ returns a set of hypotheses $A(D)$. We define teaching dimension to be the size of the smallest training set D such that $A(D) = \{h^*\}$, the singleton set containing the target model. With this notation, the classic teaching dimension is $TD(h^*, \mathcal{A}_{vs})$ where $\mathcal{A}_{vs}$ is the version space learning algorithm (2). In this paper we focus on $\mathcal{A}_{opt}$ in (1) instead, namely linear learners in $\mathbb{R}^d$. Linear learners include many popular members such as both homogeneous and inhomogeneous versions of linear regression, SVM, and logistic

regression. In addition, the linear interaction between $\mathbf{x}$ and $\boldsymbol{\theta}$ makes the loss function subgradient easy to compute, though in principle our analysis technique is applicable to other optimization-based learners, too. In this section our goal is to teach the exact parameter $\boldsymbol{\theta}^*$, consequently our teaching dimension of interest is

$$TD(\boldsymbol{\theta}^*, \mathcal{A}_{opt}). \tag{5}$$

Later in section 4 for classification we will teach the decision boundary instead.

How to reason about our teaching dimension $TD(\boldsymbol{\theta}^*, \mathcal{A}_{opt})$? It is the size of the *smallest* training set D with which (1) has a unique solution $\boldsymbol{\theta}^*$. Our strategy is to first establish a number of lower bounds $LB \leq TD(\boldsymbol{\theta}^*, \mathcal{A}_{opt})$ by showing that any training set with which (1) has a unique solution $\boldsymbol{\theta}^*$ must have at least LB items. Section 3.1 is devoted to such lower bounds. The actual teaching dimension is learner dependent. In sections 3.2 and 3.3 we construct specific teaching sets for three popular learners: ridge regression, SVM, and logistic regression. These teaching sets uniquely returns $\boldsymbol{\theta}^*$ via (1). By definition, the size of these teaching sets is an upper bound on $TD(\boldsymbol{\theta}^*, \mathcal{A}_{opt})$, respectively. If the lower and upper bounds match, we would have identified the teaching dimension $TD(\boldsymbol{\theta}^*, \mathcal{A}_{opt})$.

3.1 Lower Bounds on Teaching Dimension $TD(\boldsymbol{\theta}^*, \mathcal{A}_{opt})$

In this section we provide three general lower bounds on the teaching dimension. These lower bounds capture different aspects of a teaching set, and should be used in conjunction (i.e. taking the maximum) when applicable. We will instantiate these lower bounds for specific learners in section 3.2. In the following let $\mathcal{X}$ and $\mathcal{Y}$ be the feasible region of all $\mathbf{x}_i$'s and y_i's respectively. We will use the notation $\partial_1 \ell(\cdot, \cdot)$ in the following way: if $\ell(\cdot, \cdot)$ is smooth, then it denotes a singleton set only containing the gradient w.r.t. the first argument; if $\ell(\cdot, \cdot)$ is convex, then it denotes the subdifferential w.r.t the first argument.

LB1 comes from a degree-of-freedom perspective. It is necessary to have this amount of training items for a unique solution to exist in (1).

Theorem 1. *Given any target model $\boldsymbol{\theta}^*$, there is a degree-of-freedom lower bound on the number of training items to obtain a unique solution $\boldsymbol{\theta}^*$ from solving (1):*

$$LB1 = \begin{cases} d - \mathit{Rank}(A) + 1, & \text{if } A\boldsymbol{\theta}^* \neq \mathbf{0} \\ d - \mathit{Rank}(A), & \text{otherwise.} \end{cases} \tag{6}$$

LB2 observes that the regularizer acts as a prior. If λ is large, more items are needed to sway the prior toward the target $\boldsymbol{\theta}^*$.

Theorem 2. *Given any target model $\boldsymbol{\theta}^*$, there is a strength-of-regularization lower bound on the required number of training items to obtain a unique solution $\boldsymbol{\theta}^*$ from solving (1):*

$$LB2 = \begin{cases} \left\lceil \lambda \left(\sup_{\alpha \in \mathbb{R}, y \in \mathcal{Y}, g \in -\partial_1 \ell(\alpha \|\boldsymbol{\theta}^*\|_A^2, y)} \alpha g \right)^{-1} \right\rceil, \\ \qquad \text{if } A \text{ has full rank and } \boldsymbol{\theta}^* \neq 0 \\ 0, \qquad \text{otherwise.} \end{cases} \tag{7}$$

LB1 and LB2 apply to all generalized linear learners. Due to the popularity of inhomogeneous margin-based linear learners (which include the standard form of SVM and logistic regression), we provide a tighter lower bound LB3 for such learners in Theorem 3. For inhomogeneous margin-based linear learners the learning algorithm $\mathcal{A}_{opt}$ solves a special form of (1):

$$\mathcal{A}_{opt}(D) = \operatorname*{Argmin}_{\mathbf{w}, b} \sum_{i=1}^{n} \ell(y_i(\mathbf{x}_i^\top \mathbf{w} + b)) + \frac{\lambda}{2} \|\mathbf{w}\|_A^2. \tag{8}$$

LB3 will prove to be instrumental in computing the teaching dimension for those learners. Following standard notation, we define $\boldsymbol{\theta} = [\mathbf{w}; b]$ where $\mathbf{w} \in \mathbb{R}^d$ is the weight vector and $b \in \mathbb{R}$ the bias (offset) term. Note $\boldsymbol{\theta} \in \mathbb{R}^{d+1}$ now. The $d \times d$ regularization matrix A applies only to $\mathbf{w}$ while b is not regularized. Furthermore, margin-based linear learners have loss functions defined on the margin $y(\mathbf{x}^\top \mathbf{w} + b)$. This loss function structure will play a key role in obtaining LB3.

Theorem 3. *Assume matrix A in (8) has full rank and $\mathbf{w}^* \neq \mathbf{0}$. Given any target model $[\mathbf{w}^*; b^*]$, there is an inhomogeneous-margin lower bound on the required number of training items to obtain a unique solution $[\mathbf{w}^*; b^*]$ from solving (8):*

$$LB3 = \left\lceil \lambda \left(\sup_{\alpha \in \mathbb{R}, g \in -\partial \ell(\alpha \|\mathbf{w}^*\|_A^2)} \alpha g \right)^{-1} \right\rceil. \tag{9}$$

3.2 The Teaching Dimension $TD(\boldsymbol{\theta}^*, \mathcal{A}_{opt})$ of Three Homogeneous Learners

We now turn to upper bounding teaching dimension by constructing teaching sets. To prove that we indeed have a teaching set for a target $\boldsymbol{\theta}^*$, we need to show that $\boldsymbol{\theta}^*$ is a solution of (1), and the solution is unique. The size of any such teaching set is an upper bound on the teaching dimension. The teaching dimension itself is determined if such an upper bound matches the corresponding lower bound. We show that this is indeed the case for our constructed teaching sets. For the sake of reference we preview in Table 2 the instantiated lower bounds that we will use in this section; their derivation will be shown below.

	homogeneous			inhomogeneous		
lower bound	ridge	SVM	logistic	ridge	SVM	logistic
LB1	1	1	1	2	2	2
LB2	0	$\lceil \lambda \|\boldsymbol{\theta}^*\|^2 \rceil$	$\left\lceil \frac{\lambda \|\boldsymbol{\theta}^*\|^2}{\tau_{\max}} \right\rceil$	0	0	0
LB3	-	-	-	-	$\lceil \lambda \|\mathbf{w}^*\|^2 \rceil$	$\left\lceil \frac{\lambda \|\mathbf{w}^*\|^2}{\tau_{\max}} \right\rceil$

Table 2. Lower bounds of teaching dimension $TD(\boldsymbol{\theta}^*, \mathcal{A}_{opt})$ for homogeneous and inhomogeneous versions of ridge regression, SVM, and logistic regression.

Teaching dimension is learner-dependent. We choose three learners to study their teaching dimension due to these learners' popularity in machine learning: ridge regression, SVM, and logistic regression. It turns out that homogeneous and inhomogeneous versions of these learners require different analysis. We devote this section to the homogeneous version where the regularizer matrix $A = I$ the identity matrix, and the next section to the inhomogeneous version. It is possible to extend our analysis to other linear learners of the form (1).

It is easy to see that if the target model $\boldsymbol{\theta}^* = \mathbf{0}$, we do not need any training data to uniquely obtain the target model from (1). In the following, we only consider the nontrivial case $\boldsymbol{\theta}^* \neq \mathbf{0}$.

Homogeneous ridge regression solves the following optimization problem:

$$\min_{\boldsymbol{\theta} \in \mathbb{R}^d} \quad \sum_{i=1}^{n} \frac{1}{2}(\mathbf{x}_i^\top \boldsymbol{\theta} - y_i)^2 + \frac{\lambda}{2}\|\boldsymbol{\theta}\|^2. \tag{10}$$

We only need one training item to uniquely obtain any nonzero target model $\boldsymbol{\theta}^*$, as the following construction shows.

Proposition 1. *Given any target model $\boldsymbol{\theta}^* \neq 0$, the following is a teaching set for homogeneous ridge regression (10):*

$$\mathbf{x}_1 = a\boldsymbol{\theta}^*, \quad y_1 = a^{-1}(\lambda + \|\mathbf{x}_1\|^2) \tag{11}$$

where a can be any nonzero real number.

It is worth to note that the teaching set is inconsistent with the target model, that is, $\mathbf{x}_1^\top \boldsymbol{\theta}^* = a\|\boldsymbol{\theta}^*\|^2 \neq y_1 = \frac{\lambda}{a} + a\|\boldsymbol{\theta}^*\|^2$, unless the regularization is absent $\lambda = 0$. The teacher intentionally overshoots the target in order to precisely counter the learner's regularizer. This has been observed before for Bayesian learners, too (Zhu, 2013).

We encourage the reader to distinguish two senses of uniqueness. The teaching set itself is not necessarily unique. In the construction (11), any $a \neq 0$ leads to a valid teaching set. Nonetheless, any one of the teaching sets will lead to the unique solution $\boldsymbol{\theta}^*$ in (10).

Corollary 1. *The teaching dimension $TD(\boldsymbol{\theta}^*, \mathcal{A}^{hom}_{ridge}) = 1$ for homogeneous ridge regression and target $\boldsymbol{\theta}^* \neq \mathbf{0}$.*

Proof. Substituting A by I in LB1 (6), we obtain the lower bound $d - \text{Rank}(I) + 1 = 1$ which matches the teaching set size in (11). □

Homogeneous SVM solves the problem:

$$\min_{\boldsymbol{\theta} \in \mathbb{R}^d} \quad \sum_{i=1}^{n} \max(1 - y_i \mathbf{x}_i^\top \boldsymbol{\theta},\ 0) + \frac{\lambda}{2}\|\boldsymbol{\theta}\|^2. \tag{12}$$

To teach this learner one training item is in general not enough: we will show that we need $\lceil \lambda \|\boldsymbol{\theta}^*\|^2 \rceil$ training items. In fact, we will construct such a teaching set consisting of *identical* training items. It is well-known in the teaching literature that a teaching set does not need to consist of $i.i.d.$ samples from a distribution, and can look unusual. It is possible to incorporate additional constraints into a teaching problem if one wants the training items to be diverse, but we do not consider that in the present paper.

Proposition 2. *Given any target model $\boldsymbol{\theta}^* \neq 0$, the following is a teaching set for homogeneous SVM (12). There are $n = \lceil \lambda \|\boldsymbol{\theta}^*\|^2 \rceil$ identical training items, each taking the form*

$$\mathbf{x}_i = \frac{\lambda \boldsymbol{\theta}^*}{\lceil \lambda \|\boldsymbol{\theta}^*\|^2 \rceil}, \quad y_i = 1. \tag{13}$$

Corollary 2. *The teaching dimension $TD(\boldsymbol{\theta}^*, \mathcal{A}^{hom}_{svm}) = \lceil \lambda \|\boldsymbol{\theta}^*\|^2 \rceil$ for homogeneous SVM and target $\boldsymbol{\theta}^* \neq \mathbf{0}$.*

Homogeneous logistic regression solves the problem:

$$\min_{\boldsymbol{\theta} \in \mathbb{R}^d} \quad \sum_{i=1}^{n} \log(1 + \exp\{-y_i \mathbf{x}_i^\top \boldsymbol{\theta}\}) + \frac{\lambda}{2}\|\boldsymbol{\theta}\|^2 \tag{14}$$

where log has base e. The situation is similar to homogeneous SVM. However, due to the negative log likelihood term we have a coefficient defined by the Lambert W function (Corless et al., 1996), which we denote by W_{lam}. Recall the defining equation for Lambert W function is $W_{\text{lam}}(x)e^{W_{\text{lam}}(x)} = x$. We further define

$$\tau_{\max} := \max_t \frac{t}{1 + e^t} = W_{\text{lam}}(1/e) \approx 0.2785. \tag{15}$$

For any value $a \leq \tau_{\max}$, we define $\tau^{-1}(a)$ as the solution to $a = \frac{t}{1+e^t}$. By using the Lambert W function $\tau^{-1}(a)$ can be expressed as $\tau^{-1}(a) \equiv a - W_{\text{lam}}(-ae^a)$.

Proposition 3. *Given any target model $\boldsymbol{\theta}^* \neq 0$, the following is a teaching set for homogeneous logistic regression (14). There are $n = \left\lceil \frac{\lambda\|\boldsymbol{\theta}^*\|^2}{\tau_{\max}} \right\rceil$ identical training items, each taking the form*

$$\mathbf{x}_i = \tau^{-1}\left(\lambda\|\boldsymbol{\theta}^*\|^2 \left\lceil \frac{\lambda\|\boldsymbol{\theta}^*\|^2}{\tau_{\max}} \right\rceil^{-1}\right) \frac{\boldsymbol{\theta}^*}{\|\boldsymbol{\theta}^*\|^2}, \quad y_i = 1. \tag{16}$$

Corollary 3. *The teaching dimension $TD(\boldsymbol{\theta}^*, \mathcal{A}_{log}^{hom}) = \left\lceil \frac{\lambda\|\boldsymbol{\theta}^*\|^2}{\tau_{\max}} \right\rceil$ for homogeneous logistic regression and target $\boldsymbol{\theta}^* \neq \mathbf{0}$.*

3.3 The Teaching Dimension $TD(\boldsymbol{\theta}^*, \mathcal{A}_{opt})$ of Three Inhomogeneous Learners

Inhomogeneous learners are defined by $\boldsymbol{\theta} = [\mathbf{w}; b]$ where the weight vector $\mathbf{w} \in \mathbb{R}^d$ and the scalar offset $b \in \mathbb{R}$. The offset b is not regularized. Similar to the previous section, we need to instantiate the teaching dimension lower bounds and design the teaching sets. We show that the size of our teaching set exactly matches the lower bound for inhomogeneous ridge regression, and differs from the lower bound of inhomogeneous SVM and logistic regression by at most one due to rounding. Therefore, up to rounding we also establish the teaching dimension for these inhomogeneous learners.

Inhomogeneous ridge regression solves the problem:

$$\min_{\mathbf{w}\in\mathbb{R}^d, b\in\mathbb{R}} \sum_{i=1}^{n} \frac{1}{2}(\mathbf{x}_i^\top \mathbf{w} + b - y_i)^2 + \frac{\lambda}{2}\|\mathbf{w}\|^2 \tag{17}$$

Proposition 4. *Given any target model $[\mathbf{w}^*; b^*]$, if $\mathbf{w}^* = \mathbf{0}$ (b^* can be an arbitrary value), the following is a teaching set for inhomogeneous ridge regression (17) with $n = 1$:*

$$\mathbf{x}_1 = \mathbf{0}, \quad y_1 = b^*. \tag{18}$$

If $\mathbf{w}^* \neq \mathbf{0}$, any $n = 2$ items satisfying the following are a teaching set for $a \neq 0$:

$$\begin{aligned} &\mathbf{x}_1 - \mathbf{x}_2 = a\mathbf{w}^*, \quad y_1 = \mathbf{x}_1^\top \mathbf{w}^* + b^* + a^{-1}\lambda, \\ &y_2 = y_1 - a\|\mathbf{w}^*\|^2 - 2a^{-1}\lambda. \end{aligned} \tag{19}$$

Corollary 4. *The teaching dimension for inhomogeneous ridge regression with target $\boldsymbol{\theta}^* = [\mathbf{w}^*; b^*]$ is $TD(\boldsymbol{\theta}^*, \mathcal{A}_{ridge}^{inh}) = 1$ if target $\mathbf{w}^* = \mathbf{0}$, or $TD(\boldsymbol{\theta}^*, \mathcal{A}_{ridge}^{inh}) = 2$ if $\mathbf{w}^* \neq \mathbf{0}$, regardless of the target offset b^*.*

Inhomogeneous SVM solves the problem:

$$\min_{\substack{\mathbf{w}\in\mathbb{R}^d \\ b\in\mathbb{R}}} \sum_{i=1}^{n} \max(1 - y_i(\mathbf{x}_i^\top \mathbf{w} + b),\, 0) + \frac{\lambda}{2}\|\mathbf{w}\|^2. \tag{20}$$

Proposition 5. *Given any target model $[\mathbf{w}^*; b^*]$ with $\mathbf{w}^* \neq 0$, the following is a teaching set for inhomogeneous SVM (20). We need $n = 2\left\lceil \frac{\lambda\|\mathbf{w}^*\|^2}{2} \right\rceil$ training items, half of which are identical positive items $\mathbf{x}_i = \mathbf{x}_+, \quad y_i = 1, \quad \forall i \in \left\{1, \cdots, \frac{n}{2}\right\}$ and half identical negative items $\mathbf{x}_i = \mathbf{x}_-, \quad y_i = -1, \quad \forall i \in \left\{\frac{n}{2}+1, \cdots, n\right\}$. $\mathbf{x}_+$ and $\mathbf{x}_-$ can be designed as any vectors satisfying*

$$\mathbf{x}_+^\top \mathbf{w}^* = 1 - b^*, \quad \mathbf{x}_- = \mathbf{x}_+ - 2\mathbf{w}^*\|\mathbf{w}^*\|^{-2}. \tag{21}$$

Our construction of the teaching set in (21) requires $n = 2\left\lceil \frac{\lambda\|\mathbf{w}^*\|^2}{2} \right\rceil$ training items. This is an upper bound on the teaching dimension. Meanwhile, we show below that the inhomogeneous SVM lower bound is $LB3 = \lceil \lambda\|\mathbf{w}^*\|^2 \rceil$. There can be a difference of at most one between the lower and upper bounds, which we call the "rounding effect." We suspect that this small gap is a technicality and not intrinsic. However, at present we do not have a teaching set construction that bridges this gap. Therefore, we state the teaching dimension as an interval in the following corollary and leave the precise value as an open question for future research.

Corollary 5. *The teaching dimension for inhomogeneous SVM and target $\boldsymbol{\theta}^* = [\mathbf{w}^*; b^*]$ where $\mathbf{w}^* \neq \mathbf{0}$ is in the interval $\lceil \lambda\|\mathbf{w}^*\|^2 \rceil \leq TD(\boldsymbol{\theta}^*, \mathcal{A}_{svm}^{inh}) \leq 2\left\lceil \frac{\lambda\|\mathbf{w}^*\|^2}{2} \right\rceil$.*

Inhomogeneous logistic regression solves the problem

$$\min_{\substack{\mathbf{w}\in\mathbb{R}^d \\ b\in\mathbb{R}}} \sum_{i=1}^{n} \log(1 + \exp\{-y_i(\mathbf{x}_i^\top \mathbf{w} + b)\}) + \frac{\lambda}{2}\|\mathbf{w}\|^2. \tag{22}$$

Proposition 6. *To create a teaching set for target model $[\mathbf{w}^*; b^*]$ with nonzero $\mathbf{w}^*$ for inhomogeneous logistic regression (22), we can use $n = 2\left\lceil \frac{\lambda\|\mathbf{w}^*\|^2}{2\tau_{\max}} \right\rceil$ training items where $\mathbf{x}_i = \mathbf{x}_+, \quad y_i = 1, \quad \forall i \in \left\{1, \cdots, \frac{n}{2}\right\}$ and $\mathbf{x}_i = \mathbf{x}_-, \quad y_i = -1, \quad \forall i \in \left\{\frac{n}{2}+1, \cdots, n\right\}$. $\mathbf{x}_+$ and $\mathbf{x}_-$ can be designed as any vectors satisfying*

$$\mathbf{x}_+^\top \mathbf{w}^* = t - b^*, \quad \mathbf{x}_- = \mathbf{x}_+ - 2t\mathbf{w}^*\|\mathbf{w}^*\|^{-2}, \tag{23}$$

where the constant t is defined by $t := \tau^{-1}\left(\frac{\lambda\|\mathbf{w}^*\|^2}{n}\right)$.

Corollary 6. *The teaching dimension for inhomogeneous logistic regression and target $\boldsymbol{\theta}^* = [\mathbf{w}^*; b^*]$ where $\mathbf{w}^* \neq \mathbf{0}$ is in the interval $\left\lceil \frac{\lambda\|\mathbf{w}^*\|^2}{\tau_{\max}} \right\rceil \leq TD(\boldsymbol{\theta}^*, \mathcal{A}_{log}^{inh}) \leq 2\left\lceil \frac{\lambda\|\mathbf{w}^*\|^2}{2\tau_{\max}} \right\rceil$.*

4 Teaching a Decision Boundary Instead of a Parameter

In section 3 we considered the teaching goal where the learner is required to learn the exact *target parameter* $\boldsymbol{\theta}^*$. But when the learner is a classifier often a weaker teaching goal is sufficient, namely teaching the learner a *target decision boundary*. In this section we consider this teaching

goal. Equivalently, such a goal is defined by the set of parameters that produce the target decision boundary. Teaching is successful if the learner arrives at any one parameter within that set.

In the case of inhomogeneous linear learners, the linear decision boundary $\{\mathbf{x} \mid \mathbf{x}^\top \mathbf{w}^* + b^* = 0\}$ is identified with the parameter set $\{t[\mathbf{w}^*; b^*] : t > 0\}$. Here we assume $\mathbf{w}^*$ is nonzero. The parameter $\boldsymbol{\theta}^* = [\mathbf{w}^*; b^*]$ is just a representative member of the set. Homogeneous linear learners are similar without b^*. We denote the corresponding "decision boundary" teaching dimension by $TD(\{t\boldsymbol{\theta}^*\}, \mathcal{A}_{opt})$. This notation extends our earlier definition of TD by allowing the first argument to be a set, with the understanding that the teaching goal is for the learned model to be an element in the set. It immediately follows that

$$TD(\{t\boldsymbol{\theta}^*\}, \mathcal{A}_{opt}) = \min_{t>0} TD(t\boldsymbol{\theta}^*, \mathcal{A}_{opt}). \quad (24)$$

Since it is sufficient to teach the parameter $t\boldsymbol{\theta}^*$ for some $t > 0$ in order to teach the decision boundary, we can choose the best t that minimizes $TD(t\boldsymbol{\theta}^*, \mathcal{A}_{opt})$. For SVM and logistic regression – either homogeneous or inhomogeneous – the teaching dimension $TD(t\boldsymbol{\theta}^*, \mathcal{A}_{opt})$ depends on $\|t\boldsymbol{\theta}^*\|$ (see Table 1). We can choose t sufficiently small to drive down the teaching set size toward its possible minimum indicated by the LB1 value in Table 2 (which is nonzero because of the ceiling function). Specifically, for any fixed parameter $\boldsymbol{\theta}^*$ representing the target decision boundary:

- (homogeneous SVM): we choose $t \leq \frac{1}{\sqrt{\lambda}\|\boldsymbol{\theta}^*\|}$ so that $TD(\{t\boldsymbol{\theta}^*\}, \mathcal{A}^{hom}_{svm}) = 1$;
- (homogeneous logistic regression): we choose $t \leq \frac{\sqrt{\tau_{\max}}}{\sqrt{\lambda}\|\boldsymbol{\theta}^*\|}$ so that $TD(\{t\boldsymbol{\theta}^*\}, \mathcal{A}^{hom}_{log}) = 1$;
- (inhomogeneous SVM): we choose $t \leq \frac{\sqrt{2}}{\sqrt{\lambda}\|\mathbf{w}^*\|}$ so that $TD(\{t\boldsymbol{\theta}^*\}, \mathcal{A}^{inh}_{svm}) = 2$ (note LB1=2 in Table 2) ;
- (inhomogeneous logistic regression): we choose $t \leq \frac{\sqrt{2\tau_{\max}}}{\sqrt{\lambda}\|\mathbf{w}^*\|}$ so that $TD(\{t\boldsymbol{\theta}^*\}, \mathcal{A}^{inh}_{log}) = 2$.

The resulting teaching dimension $TD(\{t\boldsymbol{\theta}^*\}, \mathcal{A}_{opt})$ is listed in Table 1 on the row marked by "decision boundary." The teaching set construction is the same as in sections 3.2 and 3.3, respectively, but with $t\boldsymbol{\theta}^*$.

5 Related Work

Teaching dimension as a learning-theoretic quantity has attracted a long history of research. It was proposed independently in Goldman & Kearns (1995); Shinohara & Miyano (1991). Subsequent theoretical developments can be found in e.g. Zilles et al. (2011); Balbach & Zeugmann (2009); Angluin (2004); Angluin & Krikis (1997); Goldman & Mathias (1996); Mathias (1997); Balbach & Zeugmann (2006); Balbach (2008); Kobayashi & Shinohara (2009); Angluin & Krikis (2003); Rivest & Yin (1995); Ben-David & Eiron (1998); Doliwa et al. (2014). Many of them assume little extra knowledge on the learner other than that it is consistent with the training data; though Zilles et al. (2011); Balbach (2008) allow the teacher and the learner to cooperate. These theoretically elegant teaching definitions diverge from the practice of modern machine learning where the learner solves an optimization problem to find a single model that is not necessarily the 0-1 loss ERM. Teaching such modern learners is our goal.

Teaching dimension is distinct from VC dimension. For a finite hypothesis space $\mathcal{H}$, Goldman & Kearns (1995) proved the relation

$$VC(\mathcal{H})/\log(|\mathcal{H}|) \leq TD(\mathcal{H}) \leq VC(\mathcal{H}) + |\mathcal{H}| - 2^{VC(\mathcal{H})}.$$

These inequalities are somewhat weak, as Goldman and Kearns had shown both cases where one quantity is much larger than the other. The distinction between TD and VC dimension is also present in our setting. For example, by inspecting the inhomogeneous SVM column in Table 1 we note that TD does not depend on the dimensionality d of the feature space $\mathbb{R}^d$. To see why this makes intuitive sense, note two d-dimensional points are sufficient to specify any bisecting hyperplane in $\mathbb{R}^d$. On the other hand, recall that the VC dimension for inhomogeneous hyperplanes in $\mathbb{R}^d$ is $d + 1$. Furthermore, there is an interesting connection to sample compression (Floyd & Warmuth, 1995). Our teaching set can be viewed as the compressed sample, but with two generalizations: (i) the original "sample" is the whole input space, (ii) the labels is allowed to diverge from the target model. Further quantification of these connections remains an open research question.

The teaching setting we considered is also distinct from active learning. In teaching the teacher knows the target model *a priori* and her goal is to *encode* the target model as a training set, knowing that the decoder is special (namely a specific machine learning algorithm). This communication perspective highlights the difference to active learning, which must explore the hypothesis space to find the target model. Consequently, the teaching dimension can be dramatically smaller than the active learning query complexity for the same learner and hypothesis space. For example, Zhu (2013) demonstrated that to learn a 1D threshold classifier within ϵ error, the teaching dimension is a constant TD=2 regardless of ϵ, while active learning would require $O(\log \frac{1}{\epsilon})$ queries which can be arbitrarily larger than TD.

While the present paper focused on the theory of optimal teaching, there are practical applications, too. One such application is computer-aided personalized education. The human student is modeled by a computational cognitive model, or equivalently the learning algorithm. The educational goal is encoded in the target model. The optimal teaching set is then well-defined, and represents the best personalized lesson for the student (Zhu, 2015; 2013;

Khan et al., 2011). Patil *et al.* showed that human students learn statistically significantly better under such optimal teaching set compared to an $i.i.d.$ training set (Patil et al., 2014). Because contemporary cognitive models often employ optimization-based machine learners, our teaching dimension study helps to characterize these optimal lessons.

Another application of optimal teaching is in computer security. In particular, optimal teaching is the mathematical formalism to study the so-called data poisoning attacks (Barreno et al., 2010; Mei & Zhu, 2015a;b; Alfeld et al., 2016). Here the "teacher" is an attacker who has a nefarious target model in mind. The "student" is a learning agent (such as a spam filter) which accepts data and adapts itself. The attacker wants to minimally manipulate the input data in order to manipulate the learning agent toward the attacker's target model. Teaching dimension quantifies the difficulty of data-poisoning attacks, and enables research on defenses.

Teaching dimension also has applications in interactive machine learning to quantify the minimum human interaction necessary (Suh et al., 2016; Cakmak & Thomaz, 2011), and in formal synthesis to generate computer programs satisfying a specification (Jha & Seshia, 2015).

6 Experiments

We illustrate some of the teaching dimensions by examples. These numerical experiments complement the theory to help build intuition and understanding.

6.1 Sometimes SVM can be Taught by One Item

We first demonstrate the interesting fact that homogeneous SVMs (12) can sometimes be trained with a single training item. Training here is in the stricter sense of learning the exact parameter (as opposed to merely the decision boundary). Specifically, consider a homogeneous SVM in $\mathbb{R}^d$ with regularization weight $\frac{\lambda}{2}$. Consider a target parameter $\boldsymbol{\theta}^* \in \mathbb{R}^d$. Table 1 gives the teaching dimension $TD = \lceil \lambda \|\boldsymbol{\theta}^*\|^2 \rceil$. Therefore, when $\lambda\|\boldsymbol{\theta}^*\|^2 \leq 1$ then $TD = 1$ and there exists a teaching set of size one. In this case, our proposed teaching set construction in (13) consists of one positive training item: $\left(\mathbf{x}_1 = \frac{\lambda \boldsymbol{\theta}^*}{\lceil \lambda \|\boldsymbol{\theta}^*\|^2 \rceil}, y_1 = 1\right)$. We now numerically illustrate one such task.

We consider a high dimensional feature space $d = 10000$, the all-one target parameter $\boldsymbol{\theta}^* = [1, \ldots, 1]$ in $\mathbb{R}^d$, and a homogeneous SVM with $\lambda = 5 \times 10^{-5}$. Note $\lambda\|\boldsymbol{\theta}^*\|^2 = 0.5 < 1$ so $TD = 1$. Our proposed teaching set is $D_0 = \{\left(\mathbf{x}_1 = [5 \times 10^{-5} \ldots 5 \times 10^{-5}], y_1 = 1\right)\}$. To verify this, we train our SVM with training set D_0 by solving the standard SVM optimization problem: $\min_{\boldsymbol{\theta}, \xi} \frac{\lambda}{2}\|\boldsymbol{\theta}\|^2 + \xi$ subject to $y_1 \mathbf{x}_1^\top \boldsymbol{\theta} - 1 + \xi \geq 0$ and $\xi \geq 0$. This was implemented with CVX (Grant & Boyd, 2014; 2008). The SVM learned $\hat{\boldsymbol{\theta}} = [1.0001, \ldots, 1.0001]$, which is very close to the target parameter (relative error $\|\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}^*\| / \|\boldsymbol{\theta}^*\| = 10^{-4}$). Therefore, we numerically accept D_0 as a singleton training set for our SVM on target $\boldsymbol{\theta}^*$. In the following we will call the procedure of training the learner with some D and comparing the learned model $\hat{\boldsymbol{\theta}}$ with $\boldsymbol{\theta}^*$ "teaching-set-verification."

6.2 The Teaching Set may not be Unique

As mentioned after Proposition 1 any teaching set uniquely specifies the target model, but the teaching set itself may not be unique. We demonstrate this nonuniqueness with another parameter teaching task. Specifically, the learner is an inhomogeneous SVM (20) with $\lambda = 1$. We want to teach the target parameter $\mathbf{w}^* = [2, -1]$, $b^* = 2$. Corollary 5 indicates that $\lceil \lambda \|\mathbf{w}^*\|^2 \rceil = 2 \lceil \lambda \|\mathbf{w}^*\|^2 / 2 \rceil = TD = 6$. We now exhibit three such teaching sets.

The first teaching set D_1 is constructed using (21) in Proposition 5. We place three positive points at $\mathbf{x}_+ = [1, 3]$ and three negative points at $\mathbf{x}_- = [1/5, 17/5]$.

The second teaching set D_2 is obtained automatically by solving a machine teaching optimization problem. Conceptually, we fix $n = 6$, $y_1 = y_2 = y_3 = 1$, and $y_4 = y_5 = y_6 = -1$ but find $\mathbf{x}_1 \ldots \mathbf{x}_6 \in \mathbb{R}^2$ in a bilevel feasibility problem:

$$\min_{\mathbf{x}_1 \ldots \mathbf{x}_n} 0 \quad \text{subject to } \{[\mathbf{w}^*, b^*]\} = \tag{25}$$
$$\operatorname{argmin}_{\mathbf{w}, b} \quad \sum_{i=1}^{n} \max(1 - y_i(\mathbf{x}_i^\top \mathbf{w} + b),\ 0) + \frac{\lambda}{2}\|\mathbf{w}\|^2.$$

The constraint says that the SVM solution must be unique and equal $[\mathbf{w}^*, b^*]$ (recall argmin returns a set). In practice, we replace the lower level SVM optimization problem by its KKT conditions so that we have a single level feasibility problem:

$$\min_{\mathbf{x}_1 \ldots \mathbf{x}_n, \mu, \alpha, \xi} 0 \quad \text{subject to} \tag{26}$$
$$\begin{aligned}
\lambda w_j^* - \sum \alpha_i y_i x_{i,j} &= 0, j = 1 \ldots d \\
-\sum \alpha_i y_i &= 0 \\
1 - \alpha_i - \mu_i &= 0, \forall i \\
y_i(x_i^\top \mathbf{w}^* + b^*) - 1 + \xi_i &\geq 0, \forall i \\
\alpha_i(y_i(x_i^\top \mathbf{w}^* + b^*) - 1 + \xi_i) &= 0, \forall i \\
\mu_i \xi_i &\geq 0, \forall i \\
\mu_i, \alpha_i, \xi_i &\geq 0, \forall i
\end{aligned}$$

We implement this "teaching-set-finding" problem in GAMS (using the nonlinear solvers SNOPT and MINOS) which, unlike CVX, has the necessary optimization tools for solving this nonconvex program (GAMS, 2013).

The third teaching set D_3 is obtained like D_2, but we demonstrate additional controls one may pose over the teaching points. For simplicity, we assume the teacher prefers points close to the origin. This is implemented by replacing the feasibility objective of 0 in (26) with the squared Frobenius norm on the design matrix:

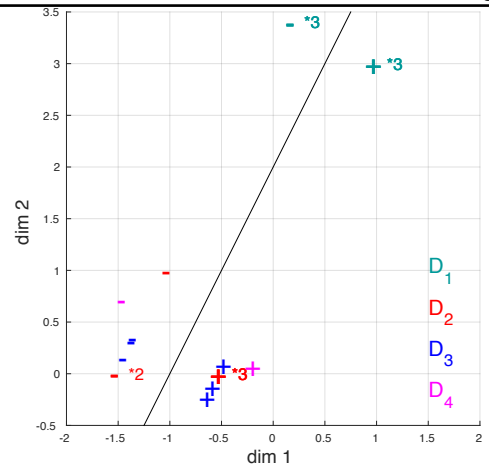

Figure 1. Different teaching sets D_1, D_2, D_3 teach the same target SVM parameters $\mathbf{w}^* = [2, -1]$, $b^* = 2$; A smaller D_4 teaches the corresponding decision boundary (the black line). "$*k$" stands for k overlapping points. Best seen in color.

$$\min_{\mathbf{x}_1 \dots \mathbf{x}_n, \mu, \alpha, \xi} \quad \sum_{i=1}^{n} \|\mathbf{x}_i\|^2. \tag{27}$$

Indeed, the squared Frobenius norm on D_3 is 6.8, smaller than that on D_2 (7.3) or D_1 (64.8).

We visualize D_1, D_2, D_3 in Figure 1. Recall our goal is to teach the exact *parameters* and not just the decision boundary, that is why we need $TD = 6$ instead of 2 teaching points. The learned SVM parameters $[\hat{\mathbf{w}}, \hat{b}]$ have relative error 2×10^{-5}, 5×10^{-8}, 7×10^{-8}, respectively w.r.t. the target parameters $[\mathbf{w}^*, b^*]$. Therefore, D_1, D_2, D_3 all pass teaching-set-verification, and are different teaching sets for the same teaching task.

6.3 Decision Boundary Easier than Parameters

We now show that it is easier to teach the decision boundary than the model parameters. We teach the same inhomogeneous SVM learner with $\lambda = 1$ from section 6.2. Our target decision boundary is represented by the same parameters $\mathbf{w}^* = [2, -1]$, $b^* = 2$, though recall any positive scaling of the parameters is equivalent. Indeed, Section 4 suggests that we only need $TD = 2$ items to teaching the decision boundary, by reverting back to teaching the parameters $[t\mathbf{w}^*, tb^*]$ for any $t \le \frac{\sqrt{2}}{\sqrt{\lambda}\|\mathbf{w}^*\|} \approx 0.63246$.

We demonstrate this by letting the GAMS program find this largest t value while still be able to teach with two items. Specifically, we fix $n = 2, y_1 = 1, y_2 = -1$. We replace the feasibility objective of (26) by

$$\max_{t>0, \mathbf{x}_1, \mathbf{x}_2, \mu, \alpha, \xi} \quad t \tag{28}$$

and replace all occurrence of $\mathbf{w}^*$ and b^* by $t\mathbf{w}^*$ and tb^*, respectively, in the constraints of (26). This teaching-set-finding problem finds $t \approx 0.63246$ as the theory predicts. The two-item teaching set D_4 consists of ($\mathbf{x}_1 = [-0.17, 0.08], y_1 = 1$), ($\mathbf{x}_2 = [-1.43, 0.72], y_2 = -1$) and is visualized in Figure 1, too. Teaching-set-verification by training the SVM on D_4 returns $\hat{\mathbf{w}} = [1.2648, -0.6324], \hat{b} = 1.2648$, which is numerically the same as $[t\mathbf{w}^*, tb^*]$ (relative error 6×10^{-5}). Therefore, we successfully taught the decision boundary with two items.

6.4 TD is Tight up to Rounding Effect

Let $L = \lceil \lambda \|\mathbf{w}^*\|^2 \rceil$ and $U = 2 \lceil \lambda \|\mathbf{w}^*\|^2 / 2 \rceil$. L and U can differ by at most one, and Corollary 5 says $L \le TD \le U$. We show TD can be either L or U on different teaching tasks, but not $L - 1$.

Consider teaching the target parameters $\mathbf{w}^* = [-0.5; 0.8]$, $b^* = 2$. Let the learner be an inhomogeneous SVM (20) with $\lambda = 2.247192$. We choose this value because it leads to $L \neq U$ and it will turn out $TD = L$. Specifically, for this task $L = 3, U = 4$. Teaching-set-finding with GAMS succeeded on $n = L$ with the teaching set $D_5 = \{([0, -1.25], 1), ([-30, -22.5], -1), ([6, 0], -1)\}$. Teaching-set-verification relative error is 5×10^{-5}. It also succeeded on $n = U$. But it failed on $n = L - 1 = 2$. To show teaching-set-finding fails on a training set size n, we run $n + 1$ separate GAMS programs to enumerate all label configurations: all negative, first item positive, ..., all positive. We also employ multiple solvers in GAMS: SNOPT, MINOS, CONOPT. Failure means all these fail to find a teaching set. Therefore, we empirically showed that this task has $TD = L$.

Now consider a different SVM learner with $\lambda = 3$. The target parameters are the same as before. This task also has $L = 3, U = 4$. Teaching-set-finding with GAMS succeeded on $n = U$ with the teaching set $D_6 = \{([0, -1.25], 1), ([2, 0], 1), ([-20.1, -16.3], -1), ([9.6, 2.3], -1)\}$. Teaching-set-verification relative error is 7×10^{-9}. However, teaching-set-finding failed on $n = L$ or $n = L - 1$. Therefore, we empirically showed this task has $TD = U$.

7 Conclusion

We have presented a generalization on teaching dimension to optimization-based learners. To the best of our knowledge, our teaching dimension for ridge regression, SVM, and logistic regression is new; so are the lower bounds and our analysis technique in general.

There are many possible extensions to the present work. For example, one may extend our analysis to nonlinear learners. This can potentially be achieved by using the kernel trick on the linear learners. As another example, one may allow "approximate teaching" by relaxing the teaching goal, such that teaching is considered successful if the learner arrives at a model close enough to the target model. Taken together, the present paper and its extensions are expected to enrich our understanding of optimal teaching and enable novel applications.

Acknowledgements

We thank the reviewers for constructive comments. This work is supported in part by NSF grants CNS-1548078, IIS-0953219, DGE-1545481, and by the University of Wisconsin-Madison Graduate School with funding from the Wisconsin Alumni Research Foundation.

References

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

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

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

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

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

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

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

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

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

Bertsekas, D. and Nedic, A. *Convex analysis and optimization (conservative)*. Athena Scientific, 2003.

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

Corless, R. M., Gonnet, G. H., Hare, D. E. G., Jeffrey, D. J., and Knuth, D. E. On the LambertW function. *Advances in Computational Mathematics*, 5(1):329–359, 1996.

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

Floyd, S. and Warmuth, M. Sample compression, learnability, and the Vapnik-Chervonenkis dimension. *Machine learning*, 21(3):269–304, 1995.

GAMS. General Algebraic Modeling System (GAMS) Release 24.2.1. Washington, DC, USA, 2013.

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

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

Grant, M. and Boyd, S. Graph implementations for nonsmooth convex programs. In *Recent Advances in Learning and Control*, Lecture Notes in Control and Information Sciences, pp. 95–110. Springer-Verlag Limited, 2008.

Grant, M. and Boyd, S. CVX: Matlab software for disciplined convex programming, version 2.1. http://cvxr.com/cvx, March 2014.

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

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

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

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

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

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

Nocedal, J. and Wright, S. J. *Numerical Optimization (2nd edition)*. Springer, 2006.

Patil, K., Zhu, X., Kopec, L., and Love, B. C. Optimal teaching for limited-capacity human learners. *NIPS*, 2014.

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

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

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

Zhu, X. Machine teaching for Bayesian learners in the exponential family. *NIPS*, 2013.

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

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