Journal of Machine Learning Research 17 (2016) 1-25 Submitted 12/15; Revised 6/16; Published 9/16

The Teaching Dimension of Linear Learners

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

Xiaojin Zhu JERRYZHU@CS.WISC.EDU
Department of Computer Sciences
University of Wisconsin-Madison
Madison, WI 53706, USA

Editor: Sanjoy Dasgupta

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.

Keywords: Optimization based learner, Karush-Kuhn-Tucker conditions, VC-dimension

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 and Kearns, 1995; Shinohara and 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 $\theta^* = 2, 3, \ldots, 10$.

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:

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

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, $\lambda > 0$ is the regularization coefficient, and A is a positive semidefinite matrix. $\|\cdot\|_A$ is the Mahalanobis norm: $\|\boldsymbol{\theta}\|_A := \sqrt{\boldsymbol{\theta}^\top A \boldsymbol{\theta}}$. This covers both homogeneous (e.g. $A = I$) and inhomogeneous (e.g. $A = [I, 0; 0, I]$) learners. 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.

	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).

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.

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 section 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 $\mathbb{D} = \cup_{n=1}^{\infty} (\mathcal{X} \times \mathcal{Y})^n$ 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 $\mathcal{A}_{vs}$:

$$\mathcal{A}_{vs}(D) = \{h \in \mathcal{H} \mid \forall (\mathbf{x}, y) \in D, h(\mathbf{x}) = y\}. \tag{2}$$

That is, the learner $\mathcal{A}_{vs}$ keeps track of the version space consisting of all hypotheses h that are consistent with D. 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:

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

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

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

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 $\mathcal{A}(D)$. We define teaching dimension to be the size of the smallest training set D such that $\mathcal{A}(D) = \{h^*\}$, 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

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

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 $LB \le TD(\boldsymbol{\theta}^*, \mathcal{A}_{opt})$ 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)*:*

$$LB1 = \begin{cases} d - \mathit{Rank}(A) + 1, & \text{if } A\boldsymbol{\theta}^* \neq \mathbf{0} \\ d - \mathit{Rank}(A), & \text{otherwise.} \end{cases} \tag{3}$$

Proof Let n^* be the minimal number of training items to ensure a unique solution $\boldsymbol{\theta}^*$. First consider the case $n^* = 0$. It happens if and only if $\boldsymbol{\theta}^* = \mathbf{0}$ and $\mathrm{Rank}(A) = d$, which is a special case of $A\boldsymbol{\theta}^* = \mathbf{0}$. Clearly, this case is consistent with LB1. Next consider the case $n^* \geq 1$. Since $\boldsymbol{\theta}^*$ solves (1), the KKT condition holds:

$$-\lambda A\boldsymbol{\theta}^* \in \sum_{i=1}^{n^*} \partial_1 \ell(\mathbf{x}_i^\top \boldsymbol{\theta}^*, y_i)\mathbf{x}_i. \tag{4}$$

We seek all $\boldsymbol{\delta}$ such that $\boldsymbol{\theta}^* + \boldsymbol{\delta}$ satisfies

$$A(\boldsymbol{\theta}^* + \boldsymbol{\delta}) = A\boldsymbol{\theta}^* \quad \text{and} \quad \mathbf{x}_i^\top(\boldsymbol{\theta}^* + \boldsymbol{\delta}) = \mathbf{x}_i^\top \boldsymbol{\theta}^* \quad \forall i = 1, \cdots, n^*, \tag{5}$$

For any such $\boldsymbol{\delta}$, simple algebra verifies that $\boldsymbol{\theta}^* + t\boldsymbol{\delta}$ satisfies the KKT condition (4) for any $t \in [0, 1]$. Consequently, $\boldsymbol{\theta}^* + \boldsymbol{\delta}$ also solves the problem in (1). To see this, we consider two situations:

- If the loss function $\ell(\cdot, \cdot)$ is convex in the first argument, the KKT condition is a sufficient optimality condition, which means that $\boldsymbol{\theta}^* + \boldsymbol{\delta}$ solves (1).
- If the loss function $\ell(\cdot, \cdot)$ is smooth (not necessary convex) in the first argument, we have $f(\boldsymbol{\theta}^*) = f(\boldsymbol{\theta}^* + \boldsymbol{\delta})$ by using the Taylor expansion (recall f is defined in equation 1):

$$\begin{aligned}
f(\boldsymbol{\theta}^* + \boldsymbol{\delta}) =& f(\boldsymbol{\theta}^*) + \langle \nabla f(\boldsymbol{\theta}^* + t\boldsymbol{\delta}), \boldsymbol{\delta} \rangle \quad (\text{for some } t \in [0, 1]) \\
=& f(\boldsymbol{\theta}^*) + \left\langle \sum_{i=1}^{n^*} \nabla_1 \ell(\mathbf{x}_i^\top(\boldsymbol{\theta}^* + t\boldsymbol{\delta}),\ y_i)\mathbf{x}_i + \lambda A(\boldsymbol{\theta}^* + t\boldsymbol{\delta})),\ \boldsymbol{\delta} \right\rangle \\
=& f(\boldsymbol{\theta}^*) + \left\langle \underbrace{\sum_{i=1}^{n^*} \nabla_1 \ell(\mathbf{x}_i^\top \boldsymbol{\theta}^*,\ y_i)\mathbf{x}_i + \lambda A\boldsymbol{\theta}^*}_{=\mathbf{0} \text{ due to the KKT condition (4)}},\ \boldsymbol{\delta} \right\rangle \\
=& f(\boldsymbol{\theta}^*).
\end{aligned}$$

Therefore, $\boldsymbol{\theta}^* + \boldsymbol{\delta}$ also solves (1). However, the uniqueness of $\boldsymbol{\theta}^*$ requires $\boldsymbol{\delta} = \mathbf{0}$ to be the only value satisfying (5). This is equivalent to say

$$\mathrm{Null}(A) \cap \mathrm{Null}(\mathrm{Span}\{\mathbf{x}_1, \cdots, \mathbf{x}_{n^*}\}) = \{\mathbf{0}\}. \tag{6}$$

It indicates that

$$\mathrm{Rank}(A) + \mathrm{Dim}(\mathrm{Span}\{\mathbf{x}_1, \cdots, \mathbf{x}_{n^*}\}) \geq d.$$

From $n^* \geq \mathrm{Dim}(\mathrm{span}\{\mathbf{x}_1, \cdots, \mathbf{x}_{n^*}\})$, we have $n^* \geq d - \mathrm{Rank}(A)$. We proved the general case for LB1.

If we have $A\boldsymbol{\theta}^* \neq \mathbf{0}$, we can further improve LB1. Let $\mathbf{g}^* = (g_1^*, \ldots, g_{n^*}^*)^\top$ be the vector satisfying

$$-\lambda A\boldsymbol{\theta}^* = \sum_{i=1}^{n^*} g_i^* \mathbf{x}_i \quad \text{and} \quad g_i^* \in \partial_1 \ell(\mathbf{x}_i^\top \boldsymbol{\theta}^*, y_i) \quad \forall i = 1, 2, \cdots, n^*. \tag{7}$$

Since $\boldsymbol{\theta}^*$ satisfies the KKT condition, such vector $\mathbf{g}^*$ must exist. Applying $A\boldsymbol{\theta}^* \neq \mathbf{0}$ to (7), we have $\mathbf{g}^* \neq \mathbf{0}$ and

$$\text{Dim}\left(\text{Span}\{A_{.1}, A_{.2}, \cdots, A_{.d}\} \cap \text{Span}\{\mathbf{x}_1, \cdots, \mathbf{x}_{n^*}\}\right) \geq 1. \tag{8}$$

To satisfy (6), we must have

$$d = \text{Dim}\left(\text{Span}\{A_{.1}, A_{.2}, \cdots, A_{.d}, \mathbf{x}_1, \cdots, \mathbf{x}_{n^*}\}\right).$$

Using the fact in linear algebra

$$\begin{aligned}
&\text{Dim}\left(\text{Span}\{A_{.1}, A_{.2}, \cdots, A_{.d}, \mathbf{x}_1, \cdots, \mathbf{x}_{n^*}\}\right) \\
=&\underbrace{\text{Dim}\left(\text{Span}\{A_{.1}, A_{.2}, \cdots, A_{.d}\}\right)}_{=\text{Rank}(A)} + \\
&\underbrace{\text{Dim}\left(\text{Span}\{\mathbf{x}_1, \cdots, \mathbf{x}_{n^*}\}\right)}_{\leq n^*} - \\
&\underbrace{\text{Dim}\left(\text{Span}\{A_{.1}, A_{.2}, \cdots, A_{.d}\} \cap \text{Span}\{\mathbf{x}_1, \cdots, \mathbf{x}_{n^*}\}\right)}_{\geq 1 \text{ (from (8))}}
\end{aligned}$$

We conclude that $n^* \geq d - \text{Rank}(A) + 1$. We completed the proof for LB1. ■

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)*:*

$$LB2 = \begin{cases} \left\lceil \lambda \Big(\sup_{\alpha \in \mathbb{R}, y \in \mathcal{Y}, g \in -\partial_1 \ell(\alpha \|\boldsymbol{\theta}^*\|_A^2, y)} \alpha g \Big)^{-1} \right\rceil, & \text{if } A \text{ has full rank and } \boldsymbol{\theta}^* \neq 0 \\ 0, & \text{otherwise.} \end{cases} \tag{9}$$

Proof When A has full rank we have an equivalent expression for the KKT condition (4):

$$-\lambda A^{\frac{1}{2}} \boldsymbol{\theta}^* \in \sum_{i=1}^{n^*} A^{-\frac{1}{2}} \mathbf{x}_i \partial_1 \ell(\mathbf{x}_i^\top \boldsymbol{\theta}^*, y_i) \quad \forall i = 1, \cdots, n^*. \tag{10}$$

Let us decompose $A^{-\frac{1}{2}}\mathbf{x}_i$ for all $i = 1, \cdots, n^*$ into $A^{-\frac{1}{2}}\mathbf{x}_i = \alpha_i A^{\frac{1}{2}}\boldsymbol{\theta}^* + \mathbf{u}_i$, where $\mathbf{u}_i$ is orthogonal to $A^{\frac{1}{2}}\boldsymbol{\theta}^*$: $\mathbf{u}_i^\top A^{\frac{1}{2}}\boldsymbol{\theta}^* = 0$. Equivalently $\mathbf{x}_i = \alpha_i A\boldsymbol{\theta}^* + A^{\frac{1}{2}}\mathbf{u}_i$. Applying this decomposition, we have

$$\mathbf{x}_i^\top \boldsymbol{\theta}^* = \alpha_i \|\boldsymbol{\theta}^*\|_A^2 + \mathbf{u}_i^\top A^{\frac{1}{2}}\boldsymbol{\theta}^* = \alpha_i \|\boldsymbol{\theta}^*\|_A^2.$$

Putting it back in (10) we obtain

$$-\lambda A^{\frac{1}{2}} \boldsymbol{\theta}^* \in \sum_{i=1}^{n^*} \left(\alpha_i A^{\frac{1}{2}} \boldsymbol{\theta}^* + \mathbf{u}_i \right) \partial_1 \ell(\alpha_i \|\boldsymbol{\theta}^*\|_A^2, y_i) \quad \forall i = 1, \cdots, n^*. \tag{11}$$

Since $\mathbf{u}_i$ is orthogonal to $A^{\frac{1}{2}}\boldsymbol{\theta}^*$, (11) can be rewritten as

$$\begin{aligned} & \exists \alpha_i \in \mathbb{R},\ \exists y_i \in \mathcal{Y},\ \exists g_i \in \partial_1 \ell(\alpha_i \|\boldsymbol{\theta}^*\|_A^2, y_i) \quad \forall i = 1, \cdots, n^* \\ \text{satisfying} \quad & \sum_{i=1}^{n^*} g_i \mathbf{u}_i = 0 \\ & -\lambda A^{\frac{1}{2}}\boldsymbol{\theta}^* = A^{\frac{1}{2}}\boldsymbol{\theta}^* \sum_{i=1}^{n^*} \alpha_i g_i \end{aligned} \tag{12}$$

Since $A\boldsymbol{\theta}^* \neq 0$, we have $A^{\frac{1}{2}}\boldsymbol{\theta}^* \neq 0$ and (12) is equivalent to $-\lambda = \sum_{i=1}^{n^*} \alpha_i g_i$. It follows that

$$\lambda = -\sum_{i=1}^{n^*} \alpha_i g_i \leq n^* \sup_{\alpha \in \mathbb{R}, y \in \mathcal{Y}, g \in \partial_1 \ell(\alpha \|\boldsymbol{\theta}^*\|_A^2, y)} -\alpha g = n^* \sup_{\alpha \in \mathbb{R}, y \in \mathcal{Y}, g \in -\partial_1 \ell(\alpha \|\boldsymbol{\theta}^*\|_A^2, y)} \alpha g$$

It indicates the lower bound for n^*

$$n^* \geq \left\lceil \frac{\lambda}{\sup_{\alpha \in \mathbb{R}, y \in \mathcal{Y}, g \in -\partial_1 \ell(\alpha \|\boldsymbol{\theta}^*\|_A^2, y)} \alpha g} \right\rceil .$$

■

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):

$$\mathcal{A}_{opt}(D) = \text{Argmin}_{\mathbf{w}, b} \quad \sum_{i=1}^{n} \ell(y_i(\mathbf{x}_i^\top \mathbf{w} + b)) + \frac{\lambda}{2}\|\mathbf{w}\|_A^2. \tag{13}$$

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* (13) *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* (13)*:*

$$LB3 = \left\lceil \lambda \left(\sup_{\alpha \in \mathbb{R}, g \in -\partial \ell(\alpha \|\mathbf{w}^*\|_A^2)} \alpha g \right)^{-1} \right\rceil . \tag{14}$$

Proof Let $D = \{\mathbf{x}_i, y_i\}_{i=1,\cdots,n}$ be a teaching set for $[\mathbf{w}^*; b^*]$. The following KKT condition needs to be satisfied:

$$\mathbf{0} \in \sum_{i=1}^{n} \partial \ell(y_i(\mathbf{x}_i^\top \mathbf{w}^* + b^*)) y_i \begin{bmatrix} \mathbf{x}_i \\ 1 \end{bmatrix} + \begin{bmatrix} \lambda A \mathbf{w}^* \\ 0 \end{bmatrix}. \tag{15}$$

If we construct a new training set

$$\hat{D} = \left\{ \hat{\mathbf{x}}_i = \mathbf{x}_i + \frac{b^*}{\|\mathbf{w}^*\|_A^2} A\mathbf{w}^*,\ \hat{y}_i = y_i \right\}_{i=1,\cdots,n}$$

then $[\mathbf{w}^*; 0]$ satisfies the KKT condition defined on $\hat{D}$. This can be verified as follows:

$$\begin{aligned}
&\sum_{i=1}^{n} \partial \ell(\hat{y}_i(\hat{\mathbf{x}}_i^\top \mathbf{w}^*)) \hat{y}_i \begin{bmatrix} \hat{\mathbf{x}}_i \\ 1 \end{bmatrix} + \begin{bmatrix} \lambda A \mathbf{w}^* \\ 0 \end{bmatrix} \\
=&\sum_{i=1}^{n} \partial \ell(y_i(\mathbf{x}_i^\top \mathbf{w}^* + b^*)) y_i \begin{bmatrix} \mathbf{x}_i + \frac{b^*}{\|\mathbf{w}^*\|_A^2} A\mathbf{w}^* \\ 1 \end{bmatrix} + \begin{bmatrix} \lambda A \mathbf{w}^* \\ 0 \end{bmatrix} \\
=&\underbrace{\sum_{i=1}^{n} \partial \ell(y_i(\mathbf{x}_i^\top \mathbf{w}^* + b^*)) y_i \begin{bmatrix} \mathbf{x}_i \\ 1 \end{bmatrix} + \begin{bmatrix} \lambda A \mathbf{w}^* \\ 0 \end{bmatrix}}_{\ni \mathbf{0} \text{ from (15)}} + \begin{bmatrix} \frac{b^*}{\|\mathbf{w}^*\|_A^2} A\mathbf{w}^* \\ 0 \end{bmatrix} \underbrace{\sum_{i=1}^{n} \partial \ell(y_i(\mathbf{x}_i^\top \mathbf{w}^* + b^*)) y_i}_{\ni 0 \text{ from (15)}} \\
\ni&\mathbf{0}
\end{aligned}$$

where $0 \in \sum_{i=1}^{n} \partial \ell(y_i(\mathbf{x}_i^\top \mathbf{w}^* + b^*)) y_i$ is from the bias dimension in (15). It follows that

$$\mathbf{0} \in \sum_{i=1}^{n} \partial \ell(\hat{y}_i \hat{\mathbf{x}}_i^\top \mathbf{w}^*) \hat{y}_i \hat{\mathbf{x}}_i + \lambda A \mathbf{w}^*$$

which is equivalent to

$$\begin{aligned}
\mathbf{0} &\in \sum_{i=1}^{n} \partial \ell(\hat{y}_i \hat{\mathbf{x}}_i^\top \mathbf{w}^*) A^{-\frac{1}{2}} \underbrace{\hat{y}_i \hat{\mathbf{x}}_i}_{=:\mathbf{z}_i} + \lambda A^{\frac{1}{2}} \mathbf{w}^* \\
&= \sum_{i=1}^{n} \partial \ell(\mathbf{z}_i^\top \mathbf{w}^*) A^{-\frac{1}{2}} \mathbf{z}_i + \lambda A^{\frac{1}{2}} \mathbf{w}^*.
\end{aligned} \tag{16}$$

We decompose $A^{-\frac{1}{2}} \mathbf{z}_i = \alpha_i A^{\frac{1}{2}} \mathbf{w}^* + \mathbf{u}_i$ where $\mathbf{u}_i$ satisfies $\mathbf{u}_i^\top A^{\frac{1}{2}} \mathbf{w}^* = 0$. Applying this decomposition to (16), we have

$$\lambda A^{\frac{1}{2}} \mathbf{w}^* \in \sum_{i=1}^{n} -\partial \ell(\alpha_i \|\mathbf{w}^*\|_A^2)(\alpha_i A^{\frac{1}{2}} \mathbf{w}^* + \mathbf{u}_i). \tag{17}$$

Since $\mathbf{u}_i$ is orthogonal to $A^{\frac{1}{2}} \mathbf{w}^*$, (17) implies that

$$\lambda A^{\frac{1}{2}} \mathbf{w}^* \in \sum_{i=1}^{n} -\partial \ell(\alpha_i \|\mathbf{w}^*\|_A^2) \alpha_i A^{\frac{1}{2}} \mathbf{w}^*.$$

Since $\mathbf{w}^* \neq \mathbf{0}$ we have

$$\lambda \in \sum_{i=1}^{n} -\partial \ell(\alpha_i \|\mathbf{w}^*\|_A^2)\alpha_i.$$

Together with

$$\sum_{i=1}^{n} -\partial \ell(\alpha_i \|\mathbf{w}^*\|_A^2)\alpha_i \leq n \sup_{\alpha \in \mathbb{R}, g \in -\partial \ell(\alpha \|\mathbf{w}^*\|_A^2)} \alpha g,$$

we obtain LB3. ■

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:

$$\min_{\boldsymbol{\theta} \in \mathbb{R}^d} \quad \sum_{i=1}^{n} \frac{1}{2}(\mathbf{x}_i^\top \boldsymbol{\theta} - y_i)^2 + \frac{\lambda}{2}\|\boldsymbol{\theta}\|^2. \tag{18}$$

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* (18)*:*

$$\mathbf{x}_1 = a\boldsymbol{\theta}^*, \quad y_1 = \frac{\lambda + \|\mathbf{x}_1\|^2}{a} \tag{19}$$

where a can be any nonzero real number.

Proof We simply verify the KKT condition to see that $\boldsymbol{\theta}^*$ is a solution to (18) by applying the construction in (19). The uniqueness of $\boldsymbol{\theta}^*$ is guaranteed by the strong convexity of (18). ■

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 (19), 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 (18).

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 (3), we obtain the lower bound $d - \text{Rank}(I) + 1 = 1$ which matches the teaching set size in (19). ■

Homogeneous SVM solves the problem:

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

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* (20)*. There are $n = \lceil \lambda \|\boldsymbol{\theta}^*\|^2 \rceil$ identical training items, each taking the form*

$$\mathbf{x}_i = \frac{\lambda \boldsymbol{\theta}^*}{\lceil \lambda \|\boldsymbol{\theta}^*\|^2 \rceil}, \quad y_i = 1. \tag{21}$$

Proof We only need to verify that the KKT condition holds for $\boldsymbol{\theta}^*$. Due to the strong convexity of (20) uniqueness is guaranteed automatically. We denote the subgradient

$\partial_a \max(1-a,0) = -\partial_1 \max(1-a,0) = -\mathbf{I}(a)$, where

$$\mathbf{I}(a) = \begin{cases} 1, & \text{if } a < 1 \\ [0,1], & \text{if } a = 1 \\ 0, & \text{otherwise} \end{cases} . \tag{22}$$

The KKT condition is

$$\begin{aligned} &\sum_{i=1}^{n} -y_i \mathbf{x}_i \partial_1 \max(1 - y_i \mathbf{x}_i^\top \boldsymbol{\theta}^*, 0) + \lambda \boldsymbol{\theta}^* \\ =&\sum_{i=1}^{n} -y_i \mathbf{x}_i \mathbf{I}(y_i \mathbf{x}_i^\top \boldsymbol{\theta}^*) + \lambda \boldsymbol{\theta}^* \\ =&-n \frac{\lambda \boldsymbol{\theta}^*}{\lceil \lambda \|\boldsymbol{\theta}^*\|^2 \rceil} \mathbf{I}\left(\frac{\lambda \|\boldsymbol{\theta}^*\|^2}{\lceil \lambda \|\boldsymbol{\theta}^*\|^2 \rceil} \right) + \lambda \boldsymbol{\theta}^* \\ =&-\lambda \boldsymbol{\theta}^* \mathbf{I}\left(\frac{\lambda \|\boldsymbol{\theta}^*\|^2}{\lceil \lambda \|\boldsymbol{\theta}^*\|^2 \rceil} \right) + \lambda \boldsymbol{\theta}^* \\ \ni&\mathbf{0} \end{aligned}$$

where the last line is due to $\mathbf{I}\left(\frac{\lambda \|\boldsymbol{\theta}^*\|^2}{\lceil \lambda \|\boldsymbol{\theta}^*\|^2 \rceil} \right)$ giving either the set $[0,1]$ or the value 1. ■

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

Proof We show this number matches LB2. Let $A = I$, $\ell(a,b) = \max(1-ab, 0)$, and consider the denominator of (9):

$$\begin{aligned} \sup_{\alpha \in \mathbb{R}, y \in \mathcal{Y}, g \in -\partial_1 \ell(\alpha \|\boldsymbol{\theta}^*\|^2, y)} \alpha g &= \sup_{\alpha, y \in \{-1,1\}, g \in y\mathbf{I}(y\alpha \|\boldsymbol{\theta}^*\|^2)} \alpha g \\ &= \sup_{\alpha, g \in \mathbf{I}(\alpha \|\boldsymbol{\theta}^*\|^2)} \alpha g \\ &= \frac{1}{\|\boldsymbol{\theta}^*\|^2} \end{aligned}$$

where the first equality is due to $\partial_1 \ell(a,b) = -b\mathbf{I}(ab)$. Therefore, $LB2 = \lceil \lambda \|\boldsymbol{\theta}^*\|^2 \rceil$ which matches the construction in (21). ■

Homogeneous logistic regression solves the problem:

$$\min_{\boldsymbol{\theta} \in \mathbb{R}^d} \quad \sum_{i=1}^{n} \log(1 + \exp\{-y_i \mathbf{x}_i^\top \boldsymbol{\theta}\}) + \frac{\lambda}{2} \|\boldsymbol{\theta}\|^2 \tag{23}$$

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

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

where the equality can be derived in following: The optimal t^* satisfies

$$1 + e^{t^*} = t^* e^{t^*} \quad \Leftrightarrow \quad (t^* - 1)e^{t^*-1} = 1/e$$

which suggests $t^* = W_{\text{lam}}(1/e) + 1$. We apply the optimality condition above and the optimal value of t^* to obtain

$$\max_t \frac{t}{1+e^t} = \frac{t^*}{1+e^{t^*}} = \frac{1}{e^{t^*}} = \frac{1}{e \cdot e^{W_{\text{lam}}(1/e)}} = W_{\text{lam}}(1/e).$$

For any value $a \le \tau_{\max}$, 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 $\tau^{-1}(a) \equiv a - W_{\text{lam}}(-ae^a)$, which can be derived from

$$\frac{t}{1+e^t} = \frac{a - W_{\text{lam}}(-ae^a)}{1 + e^{a - W_{\text{lam}}(-ae^a)}} = \frac{a + ae^a/e^{W_{\text{lam}}(-ae^a)}}{1 + e^{a - W_{\text{lam}}(-ae^a)}} = a.$$

Proposition 3 *Given any target model $\boldsymbol{\theta}^* \neq 0$, the following is a teaching set for homogeneous logistic regression* (23)*. There are $n = \left\lceil \frac{\lambda\|\boldsymbol{\theta}^*\|^2}{\tau_{\max}} \right\rceil$ identical training items, each taking the form*

$$\mathbf{x}_i = \tau^{-1}\left(\lambda\|\boldsymbol{\theta}^*\|^2 \left\lceil \frac{\lambda\|\boldsymbol{\theta}^*\|^2}{\tau_{\max}} \right\rceil^{-1}\right) \frac{\boldsymbol{\theta}^*}{\|\boldsymbol{\theta}^*\|^2}, \quad y_i = 1. \tag{24}$$

Proof We first verify that $\boldsymbol{\theta}^*$ is a solution to (23) based on the teaching set construction in (24). We only need to verify the gradient of (23) is zero. Computing the gradient of (23), we have

$$\begin{aligned}
&\sum_{i=1}^{n} \frac{-y_i \mathbf{x}_i}{1 + \exp\{y_i \mathbf{x}_i^\top \boldsymbol{\theta}^*\}} + \lambda\boldsymbol{\theta}^* \\
=& -n \frac{\mathbf{x}_i}{1 + \exp\left\{\tau^{-1}\left(\lambda\|\boldsymbol{\theta}^*\|^2 \left\lceil \frac{\lambda\|\boldsymbol{\theta}^*\|^2}{\tau_{\max}} \right\rceil^{-1}\right)\right\}} + \lambda\boldsymbol{\theta}^* \\
=& -n \frac{\tau^{-1}\left(\lambda\|\boldsymbol{\theta}^*\|^2 \left\lceil \frac{\lambda\|\boldsymbol{\theta}^*\|^2}{\tau_{\max}} \right\rceil^{-1}\right)}{1 + \exp\left\{\tau^{-1}\left(\lambda\|\boldsymbol{\theta}^*\|^2 \left\lceil \frac{\lambda\|\boldsymbol{\theta}^*\|^2}{\tau_{\max}} \right\rceil^{-1}\right)\right\}} \frac{\boldsymbol{\theta}^*}{\|\boldsymbol{\theta}^*\|^2} + \lambda\boldsymbol{\theta}^* \\
=& -n\lambda\|\boldsymbol{\theta}^*\|^2 \left\lceil \frac{\lambda\|\boldsymbol{\theta}^*\|^2}{\tau_{\max}} \right\rceil^{-1} \frac{\boldsymbol{\theta}^*}{\|\boldsymbol{\theta}^*\|^2} + \lambda\boldsymbol{\theta}^* \\
=& \mathbf{0},
\end{aligned}$$

where the third equality uses the fact $\lambda\|\boldsymbol{\theta}^*\|^2 \left\lceil \frac{\lambda\|\boldsymbol{\theta}^*\|^2}{\tau_{\max}} \right\rceil^{-1} \le \tau_{\max}$ and the property $a = \frac{\tau^{-1}(a)}{1+e^{\tau^{-1}(a)}}$. The strong convexity of (23) automatically implies uniqueness. ■

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* $\boldsymbol{\theta}^* \neq \mathbf{0}$.

Proof We show that the number matches LB2. In (9) let $A = I$ and $\ell(a, b) = \log(1 + \exp\{-ab\})$. The denominator of LB2 is:

$$\begin{aligned}
\sup_{\alpha\in\mathbb{R}, y\in\mathcal{Y}, g\in-\partial_1\ell(\alpha\|\boldsymbol{\theta}^*\|^2, y)} \alpha g =& \sup_{\alpha, y\in\{-1,1\}, g=y(1+\exp\{y\alpha\|\boldsymbol{\theta}^*\|^2\})^{-1}} \alpha g \\
=& \sup_{\alpha, g=(1+\exp\{\alpha\|\boldsymbol{\theta}^*\|^2\})^{-1}} \alpha g \\
=& \sup_{\alpha} \frac{\alpha}{1+\exp\{\alpha\|\boldsymbol{\theta}^*\|^2\}} \\
=& \|\boldsymbol{\theta}^*\|^{-2} \sup_{t} \frac{t}{1+\exp\{t\}} \\
=& \frac{\tau_{\max}}{\|\boldsymbol{\theta}^*\|^2},
\end{aligned}$$

which implies $LB2 = \left\lceil \frac{\lambda\|\boldsymbol{\theta}^*\|^2}{\tau_{\max}} \right\rceil$. ■

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:

$$\min_{\mathbf{w}\in\mathbb{R}^d, b\in\mathbb{R}} \quad \sum_{i=1}^{n} \frac{1}{2}(\mathbf{x}_i^\top \mathbf{w} + b - y_i)^2 + \frac{\lambda}{2}\|\mathbf{w}\|^2 \tag{25}$$

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* (25) *with* $n = 1$*:*

$$\mathbf{x}_1 = \mathbf{0}, \quad y_1 = b^*. \tag{26}$$

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

$$\mathbf{x}_1 - \mathbf{x}_2 = a\mathbf{w}^*, \quad y_1 = \mathbf{x}_1^\top \mathbf{w}^* + b^* + \frac{\lambda}{a}, \quad y_2 = y_1 - a\|\mathbf{w}^*\|^2 - 2\frac{\lambda}{a}. \tag{27}$$

Proof We first prove the case for $\mathbf{w}^* = \mathbf{0}$. We can verify that the KKT condition is satisfied by designing $\mathbf{x}_1$ and y_1 as in (26):

$$\begin{aligned}
(\mathbf{x}_1^\top \mathbf{w}^* + b^* - y_1)\mathbf{x}_1 + \lambda\mathbf{w}^* =& \mathbf{0} \\
\mathbf{x}_1^\top \mathbf{w}^* + b^* - y_1 =& 0.
\end{aligned}$$

The uniqueness of $[\mathbf{w}^*; b^*]$ is indicated by the strong convexity of (25) when $n = 1$.

We then prove the case for $\mathbf{w}^* \neq \mathbf{0}$. With simple algebra, we can verify the KKT condition holds via the construction in (27):

$$\begin{aligned}(\mathbf{x}_1^\top \mathbf{w}^* + b^* - y_1)\mathbf{x}_1 + (\mathbf{x}_2^\top \mathbf{w}^* + b^* - y_2)\mathbf{x}_2 + \lambda \mathbf{w}^* =& \mathbf{0} \\ (\mathbf{x}_1^\top \mathbf{w}^* + b^* - y_1) + (\mathbf{x}_2^\top \mathbf{w}^* + b^* - y_2) =& 0.\end{aligned}$$

Similarly, the uniqueness is implied by the strong convexity of (25) when $n = 2$. ■

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

Proof We match the lower bound LB1 in (3). Note $\boldsymbol{\theta}^* = [\mathbf{w}^*; b^*] \in \mathbb{R}^{d+1}$, and A in this case is a $(d+1) \times (d+1)$ matrix with the $d \times d$ identity matrix I_d padded with one additional row and column of zeros for the offset. Therefore $Rank(A) = Rank(I_d) = d$. When $\mathbf{w}^* = \mathbf{0}$, $A\boldsymbol{\theta}^* = \mathbf{0}$ and $LB1 = (d+1) - Rank(A) = 1$. When $\mathbf{w}^* \neq \mathbf{0}$, $A\boldsymbol{\theta}^* \neq \mathbf{0}$ and $LB1 = (d+1) - Rank(A) + 1 = 2$. These lower bounds match the teaching set sizes in (26) and (27), respectively. ■

Inhomogeneous SVM solves the problem:

$$\min_{\mathbf{w} \in \mathbb{R}^d, b \in \mathbb{R}} \quad \sum_{i=1}^{n} \max(1 - y_i(\mathbf{x}_i^\top \mathbf{w} + b),\ 0) + \frac{\lambda}{2}\|\mathbf{w}\|^2. \tag{28}$$

Proposition 5 *Given any target model $[\mathbf{w}^*; b^*]$ with $\mathbf{w}^* \neq 0$, the following is a teaching set for inhomogeneous SVM (28). 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*

$$\mathbf{x}_+^\top \mathbf{w}^* = 1 - b^*, \quad \mathbf{x}_- = \mathbf{x}_+ - \frac{2\mathbf{w}^*}{\|\mathbf{w}^*\|^2}. \tag{29}$$

Proof Unlike in previous learners (including homogeneous SVM), we no longer have strong convexity w.r.t. b. In order to prove that (29) is a teaching set, we need to verify the KKT condition and verify solution uniqueness.

We first verify the KKT condition to show that the solution under (29) includes the target model $[\mathbf{w}^*; b^*]$. From (29), we have

$$\mathbf{x}_+^\top \mathbf{w}^* + b^* = 1, \quad \mathbf{x}_-^\top \mathbf{w}^* + b^* = -1. \tag{30}$$

Applying them to the KKT condition and using the notation in (22) we obtain

$$\begin{aligned}
&-\frac{n}{2}\mathbf{I}(\mathbf{x}_+^\top \mathbf{w}^* + b^*)\begin{bmatrix}\mathbf{x}_+\\1\end{bmatrix} + \frac{n}{2}\mathbf{I}(-\mathbf{x}_-^\top \mathbf{w}^* - b^*)\begin{bmatrix}\mathbf{x}_-\\1\end{bmatrix} + \begin{bmatrix}\lambda\mathbf{w}^*\\0\end{bmatrix}\\
=&-\frac{n}{2}\mathbf{I}(1)\begin{bmatrix}\mathbf{x}_+\\1\end{bmatrix} + \frac{n}{2}\mathbf{I}(1)\begin{bmatrix}\mathbf{x}_-\\1\end{bmatrix} + \begin{bmatrix}\lambda\mathbf{w}^*\\0\end{bmatrix}\\
\supset&\frac{n}{2}\mathbf{I}(1)\begin{bmatrix}\mathbf{x}_- - \mathbf{x}_+\\0\end{bmatrix} + \begin{bmatrix}\lambda\mathbf{w}^*\\0\end{bmatrix} \quad \text{setting the last dimension to 0}\\
=&\mathbf{I}(1)\begin{bmatrix}-\frac{n}{\|\mathbf{w}^*\|^2}\mathbf{w}^*\\0\end{bmatrix} + \begin{bmatrix}\lambda\mathbf{w}^*\\0\end{bmatrix} \quad \text{applying (29)}\\
\supseteq&\mathbf{I}(1)\begin{bmatrix}-\lambda\mathbf{w}^*\\0\end{bmatrix} + \begin{bmatrix}\lambda\mathbf{w}^*\\0\end{bmatrix} \quad \text{observing } n \geq \lambda\|\mathbf{w}^*\|^2\\
\ni&\mathbf{0}.
\end{aligned}$$

It proves that $[\mathbf{w}^*; b^*]$ solves (28) by our teaching set construction.

Next we prove uniqueness by contradiction. We use $f(\mathbf{w}, b)$ to denote the objective function in (28) under the teaching set. It is easy to verify that $f(\mathbf{w}^*, b^*) = \frac{\lambda}{2}\|\mathbf{w}^*\|^2$. Assume that there exists another solution $[\bar{\mathbf{w}}; \bar{b}]$ different from $[\mathbf{w}^*; b^*]$. We can obtain $\|\bar{\mathbf{w}}\|^2 \leq \|\mathbf{w}^*\|^2$ due to

$$\frac{\lambda}{2}\|\mathbf{w}^*\|^2 = f(\mathbf{w}^*, b^*) = f(\bar{\mathbf{w}}, \bar{b}) \geq \frac{\lambda}{2}\|\bar{\mathbf{w}}\|^2.$$

The second equality is due to $[\bar{\mathbf{w}}; \bar{b}]$ being a solution; the inequality is due to whole-part relationship. Therefore, there are only two possibilities for the norm of $\bar{\mathbf{w}}$: $\|\bar{\mathbf{w}}\| = \|\mathbf{w}^*\|$ or $\|\bar{\mathbf{w}}\| = t\|\mathbf{w}^*\|$ for some $0 \leq t < 1$. Next we will show that both cases are impossible.

(Case 1) For the case $\|\bar{\mathbf{w}}\| = \|\mathbf{w}^*\|$, we have

$$\begin{aligned}
f(\bar{\mathbf{w}}, \bar{b}) =&\frac{n}{2}\max\left(1 - (\mathbf{x}_+^\top \bar{\mathbf{w}} + \bar{b}), 0\right) + \frac{n}{2}\max\left(1 + (\mathbf{x}_-^\top \bar{\mathbf{w}} + \bar{b}), 0\right) + \frac{\lambda}{2}\|\bar{\mathbf{w}}\|^2\\
=&\frac{n}{2}\max\left(\underbrace{\mathbf{x}_+^\top(\mathbf{w}^* - \bar{\mathbf{w}}) + (b^* - \bar{b})}_{=:\Delta_+}, 0\right) + \frac{n}{2}\max\left(\underbrace{-\mathbf{x}_-^\top(\mathbf{w}^* - \bar{\mathbf{w}}) - (b^* - \bar{b})}_{=:\Delta_-}, 0\right)\\
&+ \frac{\lambda}{2}\|\mathbf{w}^*\|^2\\
=&\frac{n}{2}\max(\Delta_+, 0) + \frac{n}{2}\max(\Delta_-, 0) + f(\mathbf{w}^*, b^*).
\end{aligned}$$

From $f(\bar{\mathbf{w}}, \bar{b}) = f(\mathbf{w}^*, b^*)$, it follows $\Delta_+ \leq 0$ and $\Delta_- \leq 0$. Since

$$0 \geq \Delta_+ + \Delta_- = (\mathbf{x}_+ - \mathbf{x}_-)^\top(\mathbf{w}^* - \bar{\mathbf{w}}) = \frac{2(\mathbf{w}^*)^\top(\mathbf{w}^* - \bar{\mathbf{w}})}{\|\mathbf{w}^*\|^2} = 2 - 2\frac{\bar{\mathbf{w}}^\top \mathbf{w}^*}{\|\mathbf{w}^*\|^2},$$

we have $\bar{\mathbf{w}}^\top \mathbf{w}^* \geq \|\mathbf{w}^*\|^2$. But because $\|\bar{\mathbf{w}}\| = \|\mathbf{w}^*\|$, we must have $\bar{\mathbf{w}} = \mathbf{w}^*$. Applying this new observation to $\Delta_+ \leq 0$ and $\Delta_- \leq 0$, we obtain $b^* = \bar{b}$. It means that $[\mathbf{w}^*; b^*] = [\bar{\mathbf{w}}; \bar{b}]$, contradicting our assumption $[\mathbf{w}^*; b^*] \neq [\bar{\mathbf{w}}; \bar{b}]$.

(Case 2) Next we turn to the case $\|\bar{\mathbf{w}}\| = t\|\mathbf{w}^*\|$ for some $t \in [0, 1)$. Recall our assumption that $[\bar{\mathbf{w}}; \bar{b}]$ solves (28). Then it follows that the following specific construction $[\hat{\mathbf{w}}, \hat{b}]$ solves (28) as well:

$$\hat{\mathbf{w}} = t\mathbf{w}^*, \quad \hat{b} = tb^*. \tag{31}$$

To see this, we consider the following optimization problem:

$$\begin{aligned} \min_{\mathbf{w},b} \quad & L(\mathbf{w}, b) := \frac{n}{2} \max(1 - (\mathbf{x}_+^\top \mathbf{w} + b), 0) + \frac{n}{2} \max(1 + (\mathbf{x}_-^\top \mathbf{w} + b), 0) \\ \text{s.t.} \quad & \|\mathbf{w}\| \le t\|\mathbf{w}^*\|. \end{aligned} \tag{32}$$

Since $[\bar{\mathbf{w}}; \bar{b}]$ solves (28), it is easy to see that $[\bar{\mathbf{w}}; \bar{b}]$ solves (32) too, otherwise there exists a solution for (32) which gives a lower function value on (28). Then we can verify that $[\hat{\mathbf{w}}; \hat{b}]$ solves (32) as well by showing the following geometric optimality condition holds:

$$-\left[\begin{array}{c} \frac{\partial L(\mathbf{w},b)}{\partial \mathbf{w}} \\ \frac{\partial L(\mathbf{w},b)}{\partial b} \end{array}\right]\Bigg|_{[\hat{\mathbf{w}};\hat{b}]} \cap \underbrace{\mathcal{N}_{\|\mathbf{w}\| \le t\|\mathbf{w}^*\|}(\hat{\mathbf{w}}, \hat{b})}_{\text{Normal cone to the set } \{[\mathbf{w}; b] : \|\mathbf{w}\| \le t\|\mathbf{w}^*\|\} \text{ at } [\hat{\mathbf{w}}; \hat{b}]} \neq \emptyset.$$

Given a convex closed set Ω and a point $\boldsymbol{\theta} \in \Omega$, the normal cone at point $\boldsymbol{\theta}$ is defined to be a set

$$\mathcal{N}_\Omega(\boldsymbol{\theta}) = \{\boldsymbol{\phi} : \ \langle \boldsymbol{\phi}, \boldsymbol{\psi} - \boldsymbol{\theta} \rangle \le 0 \ \forall \boldsymbol{\psi} \in \Omega\}.$$

The optimality condition basically suggests that at the optimal point, the negative (sub)gradient direction overlaps with the normal cone. In other words, there does not exist any valid direction to decrease the objective at the optimal point. Readers can refer to Nocedal and Wright (2006) or Bertsekas and Nedic (2003) for more explanations about the geometric optimality condition.

Because of (30) and (31), we have $\mathbf{x}_+^\top \hat{\mathbf{w}} + \hat{b} = t < 1$. Thus at $[\hat{\mathbf{w}}; \hat{b}]$ the subgradient is

$$-\left[\begin{array}{c} \frac{\partial L(\mathbf{w},b)}{\partial \mathbf{w}} \\ \frac{\partial L(\mathbf{w},b)}{\partial b} \end{array}\right]\Bigg|_{[\hat{\mathbf{w}};\hat{b}]} = \frac{n}{2} \begin{bmatrix} \mathbf{x}_+ - \mathbf{x}_- \\ 0 \end{bmatrix} = \begin{bmatrix} \frac{n\mathbf{w}^*}{\|\mathbf{w}^*\|^2} \\ 0 \end{bmatrix}$$

And the normal cone is

$$\mathcal{N}_{\|\mathbf{w}\| \le t\|\mathbf{w}^*\|}(\hat{\mathbf{w}}, \hat{b}) = \left\{ s \begin{bmatrix} \mathbf{w}^* \\ 0 \end{bmatrix} \ \middle| \ s \ge 0 \right\}.$$

The intersection is non-empty by choosing $s = \frac{n}{\|\mathbf{w}^*\|^2}$. Since both $[\hat{\mathbf{w}}; \hat{b}]$ and $[\bar{\mathbf{w}}; \bar{b}]$ solve (32), we have $L(\hat{\mathbf{w}}, \hat{b}) = L(\bar{\mathbf{w}}, \bar{b})$. Together with $\|\hat{\mathbf{w}}\| = \|\bar{\mathbf{w}}\|$, we have

$$f(\hat{\mathbf{w}}, \hat{b}) = L(\hat{\mathbf{w}}, \hat{b}) + \frac{\lambda}{2}\|\hat{\mathbf{w}}\|^2 = f(\bar{\mathbf{w}}, \bar{b}) = f(\mathbf{w}^*, b^*).$$

Therefore, we proved that $[\hat{\mathbf{w}}; \hat{b}]$ solves (28) as well. To see the contradiction, let us check the function value of $f(\hat{\mathbf{w}}, \hat{b})$ via a different route:

$$\begin{aligned}
f(\hat{\mathbf{w}}, \hat{b}) =& f(t\mathbf{w}^*, tb^*) \\
=& \sum_{i=1}^{\frac{n}{2}} \max\left(1 - t(\mathbf{x}_+^\top \mathbf{w}^* + b^*), 0\right) + \sum_{i=1}^{\frac{n}{2}} \max\left(1 + t(\mathbf{x}_-^\top \mathbf{w}^* + b^*), 0\right) + \frac{\lambda}{2}\|\mathbf{w}^*\|^2 t^2 \\
=& \sum_{i=1}^{\frac{n}{2}} \max\left(1 - t, 0\right) + \sum_{i=1}^{\frac{n}{2}} \max\left(1 - t, 0\right) + \frac{\lambda}{2}\|\mathbf{w}^*\|^2 t^2 \\
=& n(1-t) - \frac{\lambda}{2}\|\mathbf{w}^*\|^2(1-t^2) + \frac{\lambda}{2}\|\mathbf{w}^*\|^2 \\
\geq& n(1-t) - \frac{n}{2}(1-t^2) + \frac{\lambda}{2}\|\mathbf{w}^*\|^2 \\
=& \frac{n}{2}(1-t)^2 + f(\mathbf{w}^*, b^*) \\
>& f(\mathbf{w}^*, b^*),
\end{aligned}$$

where the first inequality uses the fact that $n \geq \lambda\|\mathbf{w}^*\|^2$. It contradicts our early assertion $f(\hat{\mathbf{w}}, \hat{b}) = f(\mathbf{w}^*, b^*)$. Putting cases 1 and 2 together we prove uniqueness. ■

Our construction of the teaching set in (29) 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 $LB3 = \left\lceil \lambda\|\mathbf{w}^*\|^2 \right\rceil$. 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* $\left\lceil \lambda\|\mathbf{w}^*\|^2 \right\rceil \leq TD(\boldsymbol{\theta}^*, \mathcal{A}_{svm}^{inh}) \leq 2\left\lceil \frac{\lambda\|\mathbf{w}^*\|^2}{2} \right\rceil$.

Proof The upper bound directly follows Proposition 5. We only need to show the lower bound $LB3 = \left\lceil \lambda\|\mathbf{w}^*\|^2 \right\rceil$ in Theorem 3. Let $A = I$, $\ell(a) = \max(1-a, 0)$, and consider the denominator of (14):

$$\sup_{\alpha\in\mathbb{R}, g\in -\partial\ell(\alpha\|\mathbf{w}^*\|^2)} \alpha g = \sup_{\alpha, g\in\mathbf{I}(\alpha\|\mathbf{w}^*\|^2)} \alpha g = \frac{1}{\|\mathbf{w}^*\|^2}$$

where the first equality is due to $\partial\ell(a) = -\mathbf{I}(a)$. Therefore, $LB3 = \left\lceil \lambda\|\mathbf{w}^*\|^2 \right\rceil$ which proves the lower bound. ■

Inhomogeneous logistic regression solves the problem

$$\min_{\mathbf{w}\in\mathbb{R}^d, b\in\mathbb{R}} \quad \sum_{i=1}^{n} \log(1 + \exp\{-y_i(\mathbf{x}_i^\top \mathbf{w} + b)\}) + \frac{\lambda}{2}\|\mathbf{w}\|^2. \tag{33}$$

Proposition 6 *To create a teaching set for target model* $[\mathbf{w}^*; b^*]$ *with nonzero* $\mathbf{w}^*$ *for inhomogeneous logistic regression* (33)*, 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*

$$\mathbf{x}_+^\top \mathbf{w}^* = t - b^*, \quad \mathbf{x}_- = \mathbf{x}_+ - \frac{2t}{\|\mathbf{w}^*\|^2}\mathbf{w}^*, \tag{34}$$

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

Proof We first point out that for t to be well-defined the argument to $\tau^{-1}()$ has to be bounded $\frac{\lambda\|\mathbf{w}^*\|^2}{n} \le \tau_{\max}$. This implies $n \ge \frac{\lambda\|\mathbf{w}^*\|^2}{\tau_{\max}}$. The size of our proposed teaching set is the smallest among all such symmetric construction that satisfy this constraint.

We verify that the KKT condition to show the construction in (34) includes the solution $[\mathbf{w}^*; b^*]$. From (34), we have

$$\mathbf{x}_+^\top \mathbf{w}^* + b^* = t \quad \mathbf{x}_-^\top \mathbf{w}^* + b^* = -t.$$

We apply them and the teaching set construction to compute the gradient of (33):

$$\begin{aligned}
&-\frac{n}{2}\frac{1}{1+\exp\{\mathbf{x}_+^\top \mathbf{w}^* + b^*\}}\begin{bmatrix}\mathbf{x}_+\\ 1\end{bmatrix} + \frac{n}{2}\frac{1}{1+\exp\{-\mathbf{x}_-^\top \mathbf{w}^* - b^*\}}\begin{bmatrix}\mathbf{x}_-\\ 1\end{bmatrix} + \begin{bmatrix}\lambda\mathbf{w}^*\\ 0\end{bmatrix}\\
=&-\frac{n}{2}\frac{1}{1+\exp\{t\}}\begin{bmatrix}\mathbf{x}_+\\ 1\end{bmatrix} + \frac{n}{2}\frac{1}{1+\exp\{t\}}\begin{bmatrix}\mathbf{x}_-\\ 1\end{bmatrix} + \begin{bmatrix}\lambda\mathbf{w}^*\\ 0\end{bmatrix}\\
=&-\frac{n}{\|\mathbf{w}^*\|^2}\frac{t}{1+\exp\{t\}}\begin{bmatrix}\mathbf{w}^*\\ 0\end{bmatrix} + \begin{bmatrix}\lambda\mathbf{w}^*\\ 0\end{bmatrix}\\
=&-\frac{n}{\|\mathbf{w}^*\|^2}\frac{\lambda\|\mathbf{w}^*\|^2}{n}\begin{bmatrix}\mathbf{w}^*\\ 0\end{bmatrix} + \begin{bmatrix}\lambda\mathbf{w}^*\\ 0\end{bmatrix}\\
=&\mathbf{0}.
\end{aligned}$$

This verifies the KKT condition.

Finally we show uniqueness. The Hessian matrix of the objective function (33) under our training set (34) is:

$$\underbrace{\frac{n}{2}\frac{\exp\{t\}}{(1+\exp\{t\})^2}}_{:=a}\underbrace{\begin{bmatrix}\mathbf{x}_+\mathbf{x}_+^\top + \mathbf{x}_-\mathbf{x}_-^\top & \mathbf{x}_+ + \mathbf{x}_-\\ \mathbf{x}_+^\top + \mathbf{x}_-^\top & 2\end{bmatrix}}_{:=A} + \lambda\underbrace{\begin{bmatrix}I & \mathbf{0}\\ \mathbf{0}^\top & 0\end{bmatrix}}_{:=B}.$$

Note $a > 0$ and $A = \begin{bmatrix}\mathbf{x}_+\\ 1\end{bmatrix}\begin{bmatrix}\mathbf{x}_+ & 1\end{bmatrix} + \begin{bmatrix}\mathbf{x}_-\\ 1\end{bmatrix}\begin{bmatrix}\mathbf{x}_- & 1\end{bmatrix}$ is positive semi-definite. We show that $aA + \lambda B$ is positive definite. Suppose not. Then there exists $[\mathbf{u}; v] \ne \mathbf{0}$ such that $[\mathbf{u}; v]^\top (aA + \lambda B)[\mathbf{u}; v] = 0$. This implies $[\mathbf{u}; v]^\top (aA)[\mathbf{u}; v] + \lambda\mathbf{u}^\top\mathbf{u} = 0$. Since the first term is non-negative due to A being positive semi-definite, $\mathbf{u} = \mathbf{0}$. But then we have $2av^2 = 0$ which implies $[\mathbf{u}; v] = \mathbf{0}$, a contradiction. Therefore uniqueness is guaranteed.

■

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* $\left\lceil \frac{\lambda \|\mathbf{w}^*\|^2}{\tau_{\max}} \right\rceil \leq TD(\boldsymbol{\theta}^*, \mathcal{A}^{inh}_{log}) \leq 2 \left\lceil \frac{\lambda \|\mathbf{w}^*\|^2}{2\tau_{\max}} \right\rceil$.

Proof The upper bound directly follows Proposition 6. We only need to show the lower bound $\left\lceil \frac{\lambda \|\mathbf{w}^*\|^2}{\tau_{\max}} \right\rceil$ by applying $LB3$ in Theorem 3. Let $A = I$ and $\ell(a) = \log(1 + \exp\{-a\})$ and consider the denominator of (14):

$$\begin{aligned}
\sup_{\alpha \in \mathbb{R}, g \in \partial \ell(-\alpha \|\mathbf{w}^*\|^2)} \alpha g &= \sup_{\alpha, g = (1+\exp\{\alpha \|\mathbf{w}^*\|^2\})^{-1}} \alpha g \\
&= \sup_{\alpha} \frac{\alpha}{1 + \exp\{\alpha \|\mathbf{w}^*\|^2\}} \\
&= \|\mathbf{w}^*\|^{-2} \sup_{t} \frac{t}{1 + \exp\{t\}} \\
&= \frac{\tau_{\max}}{\|\mathbf{w}^*\|^2},
\end{aligned}$$

which implies $LB3 = \left\lceil \frac{\lambda \|\mathbf{w}^*\|^2}{\tau_{\max}} \right\rceil$. ■

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

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

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 $TD(\{t\boldsymbol{\theta}^*\}, \mathcal{A}_{log}^{hom}) = 1$;
- (inhomogeneous SVM): we choose $t \leq \frac{\sqrt{2}}{\sqrt{\lambda}\|\mathbf{w}^*\|}$ so that $TD(\{t\boldsymbol{\theta}^*\}, \mathcal{A}_{svm}^{inh}) = 2$ (note LB1=2 in Table 2);
- (inhomogeneous logistic regression): we choose $t \leq \frac{\sqrt{2\tau_{\max}}}{\sqrt{\lambda}\|\mathbf{w}^*\|}$ so that $TD(\{t\boldsymbol{\theta}^*\}, \mathcal{A}_{log}^{inh}) = 2$ (note LB1=2 in Table 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 and Kearns (1995); Shinohara and Miyano (1991). Subsequent theoretical developments can be found in e.g. 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); 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. Section 6 discusses a new view to unify our work and some existing optimal teaching work.

Teaching dimension is distinct from VC dimension. For a finite hypothesis space $\mathcal{H}$, Goldman and 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 in Floyd and 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, and (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 specified by 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). In one experiment, Patil *et al.* showed that real 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 and 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 supports research on defenses.

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

6. A New View on Teaching

The optimal teaching literature has been cautious about the so-called collusion or coding tricks between the teacher and the learner. Nonetheless, what constitutes collusion does not have a fully satisfactory definition. Goldman and Mathias (1996) defined the teacher and the learner as collusion-free if (i) the teaching set is consistent with the target concept; (ii) any superset of the teaching set will make the learner learn the target concept, too. While this definition of collusion-free is useful, it does not capture all interesting learning behaviors. For example, Zilles et al. (2011, section 4) had to introduce a different notion of collusion in order to allow benign cooperation between the teacher and the learner. As another example, standard machine learning algorithms such as ridge regression does not satisfy either of the two properties: the teaching set (19) is inconsistent in that $y_1 \neq \mathbf{x}_1^\top \boldsymbol{\theta}^*$, and adding more consistent training items will in general produce a different model due to regularization.

We advance an alternative view on the relation between the teacher and the learner. Under this view, the learner publishes his learning algorithm $\mathcal{A} : \mathbb{D} \to 2^{\mathcal{H}}$. Recall $\mathcal{A}$ takes in a training set $D \in \mathbb{D}$ and outputs a subset of the hypothesis space $\mathcal{H}$. The teacher then uses a fixed strategy: she simply solves the training set cardinality minimization problem under the constraint that $\mathcal{A}$ returns the target hypothesis set Θ^*. For example, to teach a specific parameter vector $\boldsymbol{\theta}^*$ the target is the singleton set $\Theta^* = \{\boldsymbol{\theta}^*\}$; to teach a decision boundary the target is the set $\Theta^* = \{t\boldsymbol{\theta}^* \mid t > 0\}$. More precisely, the teacher's strategy is to solve the following optimization problem, whose objective value is the (learner-dependent) teaching dimension $TD(\Theta^*, \mathcal{A})$:

$$\begin{aligned} \min_{D \in \mathbb{D}} \quad & |D| \\ \text{s.t.} \quad & \Theta^* = \mathcal{A}(D). \end{aligned} \tag{35}$$

Our teaching dimension for linear learners clearly fits this view, with $\mathcal{A}_{opt}$ being a regularized empirical risk minimizer (1). Let us look at a few other interesting learners $\mathcal{A}$ under this view. We will use the following hypothesis space as it is historically used to contrast those learners (Goldman and Kearns, 1995; Zilles et al., 2011). Let $\mathcal{X} = \{x_1, \ldots, x_n\}$. Let $h_i(x) = 1$ if $x = x_i$ and 0 otherwise, for $i = 1 \ldots n$. In other words, h_i is the indicator concept on x_i. Let the all-negative concept be $h_0(x) = 0$ for all x. Let $\mathcal{H} = \{h_0, h_1, \ldots, h_n\}$.

- **The version-space learner $\mathcal{A}_{vs}$ as defined by** (2)**.** This is the learner behind the teaching dimension defined by Goldman and Kearns (1995). We have $\mathcal{A}_{vs}(\{(x_i, 1)\}) = \{h_i\}$ for $i = 1 \ldots n$, such that these target concepts have classic teaching dimension $TD(h_i, \mathcal{A}_{vs}) = 1$. But note that $\mathcal{A}_{vs}(\{(x_i, 0)\}) = \{h_0, \ldots, h_{i-1}, h_{i+1}, \ldots\}$ which does not reduce the version space to a single element. To specify the all-negative concept we need $\mathcal{A}_{vs}(\{(x_1, 0), \ldots, (x_n, 0)\}) = \{h_0\}$. That is, h_0's classic teaching dimension is $TD(h_0, \mathcal{A}_{vs}) = n$. These teaching dimensions are the objective values in our view (35) when we plug in $\mathcal{A}_{vs}$.

- **The Balbach learner $\mathcal{A}_B$ (Balbach, 2008).** Balbach noticed that $h_1, \ldots, h_n$ can each be taught with one item. The reasoning goes that as soon as the teaching set contains more than one item, it must be a helpful teacher's hint that the target concept is h_0. That is, the size of the training set carries useful information about the target concept. In the view of (35), we may define $\mathcal{A}_B(\{(x_i, 1)\}) = \{h_i\}$ for $i = 1 \ldots n$, and $\mathcal{A}_B(\{(x_i, 0), (x_j, 0)\}) = \{h_0\}$ for any $i \neq j$. For the sake of completeness, here and below for all other $D \in \mathbb{D}$ not explicitly mentioned we simply define $\mathcal{A}_.(D) = \{h \text{ consistent with D}\}$. When we plug $\mathcal{A}_B$ into (35) we obtain Balbach's teaching dimension $TD(h_i, \mathcal{A}_B)$ of 1 for $h_1, \ldots, h_n$, and 2 for h_0.

- **The subset learner $\mathcal{A}_s$ (Zilles et al., 2011).** Since the teaching sets for $h_1, \ldots, h_n$ each contain a positive item, it stands to reason that h_0 is the target concept as soon as a single negative training item is observed. We can define $\mathcal{A}_s(\{(x_i, 1)\}) = \{h_i\}$ and $\mathcal{A}_s(\{(x_i, 0)\}) = \{h_0\}$ for $i = 1 \ldots n$. When we plug $\mathcal{A}_s$ into (35) we obtain the subset teaching dimension of $TD(h, \mathcal{A}_s) = 1$ for all $h \in \mathcal{H}$, which is an improvement over the Balbach teaching dimension by a certain benign cooperation.

- **A coding-trick learner $\mathcal{A}_{c1}$.** This "learner" uses x to encode hypothesis: $\mathcal{A}_{c1}(\{(x_i, y)\}) = \{h_i\}$ for $i = 1 \dots n$ regardless of y, and all non-singleton training set maps to h_0: $\mathcal{A}_{c1}(D) = \{h_0\}$ if $|D| \neq 1$. $\mathcal{A}_{c1}$ is mathematically well-defined for teaching in (35), but one can argue that it does not seem like a reasonable learner: it ignores y completely and thus is inconsistent (although recall modern regularized empirical risk minimizers (1) can be inconsistent, too).

- **Another coding-trick learner $\mathcal{A}_{c2}$.** This "learner" uses training set size to encode the hypothesis, while ignoring the content of the training set: $\mathcal{A}_{c2}(D) = \{h_{|D|}\}$ if $|D| \leq n$, and $\emptyset$ if $|D| > n$. Again, $\mathcal{A}_{c2}$ is mathematically well-defined but does not seem like a reasonable learner.

As the examples above show, our alternative view of teaching in (35) does not resolve the issue of what constitutes coding-tricks. All the learners $\mathcal{A}$ are well-defined functions mapping a training set to a subset of hypotheses, so that the optimization problem (35) is also well-defined even for "unreasonable" learners like $\mathcal{A}_{c1}$ and $\mathcal{A}_{c2}$. However, our alternative view does provide two benefits:

- Because the teacher employs a fixed strategy (35), this view removes the notion of "collusion" altogether. Instead, the question becomes what learning algorithm $\mathcal{A}$ one would consider as admissible. This view point can be more natural when we extend teaching to richer, more complex learners.

- There can be a misconception that the classic teaching dimension defined by Goldman and Kearns (1995) is learner-independent and a property of $\mathcal{H}$ only, in part perhaps fueled by the original notation $TD(\mathcal{H})$. Our view highlights classic teaching dimension's dependency on the version space learner $\mathcal{A}_{vs}$. It is true that $\mathcal{A}_{vs}$ is a particularly simple and elegant learner with very nice properties. But, as others have observed (e.g. Balbach (2008); Zilles et al. (2011)), it does not capture all natural teaching and learning behaviors.

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.

Acknowledgments

The authors thank the editor and referees for their valuable comments. Special thanks to the production editor Dr. Charles Sutton for his help to prepare the final version of this paper. 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

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

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

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

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

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

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

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.

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

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

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

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

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

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.

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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