The Teaching Dimension of Regularized Kernel Learners

Hong Qian [*1 2] **Xu-Hui Liu** [*3 4] **Chen-Xi Su** [1] **Aimin Zhou** [1 5] **Yang Yu** [3 4]

Abstract

Teaching dimension (TD) is a fundamental theoretical property for understanding machine teaching algorithms. It measures the sample complexity of teaching a target hypothesis to a learner. The TD of linear learners has been studied extensively, whereas the results of teaching non-linear learners are rare. A recent result investigates the TD of polynomial and Gaussian kernel learners. Unfortunately, the theoretical bounds therein show that the TD is high when teaching those non-linear learners. Inspired by the fact that regularization can reduce the learning complexity in machine learning, a natural question is whether the similar fact happens in machine teaching. To answer this essential question, this paper proposes a unified theoretical framework termed STARKE to analyze the TD of regularized kernel learners. On the basis of STARKE, we derive a generic result of any type of kernels. Furthermore, we disclose that the TD of regularized linear and regularized polynomial kernel learners can be strictly reduced. For regularized Gaussian kernel learners, we reveal that, although their TD is infinite, their ϵ-approximate TD can be exponentially reduced compared with that of the unregularized learners. The extensive experimental results of teaching the optimization-based learners verify the theoretical findings.

1. Introduction

Machine teaching (Zhu et al., 2018) is aimed at designing an optimal training set (aka teaching set) to steer a learner (aka student) towards a target hypothesis. It can be regarded as an inverse problem of machine learning. Machine teaching has various applications, such as reinforcement learning (Kamalaruban et al., 2019), trustworthy AI (Zhang et al., 2018), education (Patil et al., 2014) and cognitive psychology (Shafto et al., 2014). In those scenarios, although a teacher knows the target hypothesis, she cannot telepathize it into the learner's mind. For instance, a botanist intends to teach the students to categorize the flowers into peony, rose, and azalea. The botanist has the correct decision boundary in mind, but she could only teach via picking the representative and informative flower examples and showing them to the students. The choice of training set can be optimized if the teacher has a good understanding of how the students learn from the examples.

Related Work. Teaching dimension (TD) (Goldman & Kearns, 1991; Shinohara, 1991) is a fundamental theoretical property for understanding machine teaching algorithms. It measures the sample complexity of teaching, and is defined as the minimal number of training examples required in the worst case to teach a target hypothesis to a learner. Along the theoretical research direction of TD, one of the most considered settings is teaching a version space learner (Goldman & Kearns, 1991; Anthony et al., 1995; Chen et al., 2018; Kirkpatrick et al., 2019). A version space learner maintains a set of hypotheses, keeps removing those which are not consistent with the receiving training examples, and outputs a qualified hypothesis in the end. To reduce the teaching dimension, i.e., the teaching complexity, a series of teaching models are proposed, such as recursive teaching (Zilles et al., 2011), preference-based teaching (Gao et al., 2017; Mansouri et al., 2019) and non-clashing teaching (Kirkpatrick et al., 2019). However, the power of version space learners is limited (Goldman & Mathias, 1996), and it can hardly model the behavior of a wide range of modern learners.

To tackle the above issue, the notation of TD is extended to the optimization-based learners, i.e, the empirical risk minimization (ERM) learners. The scenario of teaching ERM learners is more realistic, and the version space learners are a special case of it provided that we optimize the 0-1 loss. Liu et al. (2016) extensively investigate the TD of ERM learners under the linear hypothesis space. Since the linear ERM learners may be restricted, recently, Kumar et al. (2021) generalize the hypothesis space to the ***non-linear*** ones by considering the kernel perceptrons in ERM. They disclose

[*]Equal contribution [1]School of Computer Science and Technology, East China Normal University, Shanghai, China. [2]Shanghai Key Laboratory of Multidimensional Information Processing. [3]School of Artificial Intelligence, Nanjing University, Nanjing, China. [4]National Key Laboratory for Novel Software Technology. [5]Shanghai Institute of AI for Education.. Correspondence to: Hong Qian <hqian@cs.ecnu.edu.cn>.

Proceedings of the 39^{th} International Conference on Machine Learning, Baltimore, Maryland, USA, PMLR 162, 2022. Copyright 2022 by the author(s).

(a) Hinge loss function *without* regularization

(b) Hinge loss function *with* regularization

Figure 1. Teaching a 1-D threshold classifier. The positive training examples are marked as red diamonds and the negative ones are marked as blue rectangles. The black point represents the target hypothesis θ^*. (a) A hinge loss learner *without* regularization and the output hypotheses marked by greed line under the given training set. (b) A hinge loss learner *with* L_2 regularization and the teaching set $\{(\theta^*, -1)\}$.

that the TD is $\Theta(d)$ for teaching linear kernel perceptrons in $\mathbb{R}^d$, and $\Theta(d^p)$ for polynomial kernel perceptrons with degree p. For Gaussian kernel perceptrons, they reveal that the exact teaching is impossible with a finite teaching set, and thus propose the ϵ-approximate TD (ϵ-TD). It is the minimal number of training examples required in the worst case to teach a target hypothesis to a learner that has no more than ϵ excess risk. Under the approximate teaching scenario, the ϵ-TD of Gaussian kernel perceptrons is $d^{O(\log^2(1/\epsilon))}$.

Problem & Motivation. The theoretical bounds derived in Kumar et al. (2021) indicate that the exact or ϵ-approximate TD is high when teaching linear, polynomial, and Gaussian kernel learners. The high sample complexity of teaching could lead to the low efficiency of machine teaching algorithms thus blocking their further applications. Inspired by the fact that ***regularization*** can reduce the learning complexity in machine learning (Bousquet & Elisseeff, 2002; Mohri et al., 2012), a natural and fundamental question is whether the similar fact happens in machine teaching. Furthermore, Kumar et al. (2021) only consider the linear, polynomial, and Gaussian kernel learners. Other types of widely-used kernels, e.g., exponential kernels (Feragen et al., 2015) and Laplacian kernels (Fadel et al., 2016; Drewnik & Pasternak-Winiarski, 2017), are omitted therein. Therefore, a generic theoretical analysis framework that is able to derive the TD of ***any*** type of (non-linear) kernel learners is quite appealing.

Our Contribution. In this paper, we focus on answering the above essential question: ***Can regularization help reduce the teaching complexity of kernel learners*?** Fortunately, our answer is ***YES***. Intuitively, consider a 1-dimensional threshold classifier $h_{\theta^*}(x) = 2(\mathbb{I}(x > \theta^*) - 0.5)$, i.e., $h_{\theta^*}(x)$ returns -1 if $x \leq \theta^*$ and $+1$ if $x > \theta^*$. As illustrated in Figure 1, the hinge loss is used. For a learner *without* regularization, the hypotheses in the interval between the closest and oppositely labeled two examples in the training set have the equal hinge loss. Thus, it is impossible for the learner to pick out the unique target hypothesis θ^* with a finite teaching set. In contrast, for a learner *with* L_2 regularization whose regularization coefficient is less than 1, one teaching example $(\theta^*, -1)$ is enough due to its preference to θ^* with less L_2 norm. The contribution of this paper is three folds:

Table 1. The teaching dimension (TD) of ERM linear, polynomial and Gaussian kernel learners *with* and *without* regularization (line 2-4). The ϵ-approximate teaching dimension (ϵ-TD) of ERM Gaussian kernel learners *with* and *without* regularization (line 5).

Kernel Type (TD Type)	*With* Regularization (This Paper)	*Without* Regularization (Kumar et al., 2021)
Linear (TD)	1	$\Theta(d)$
Polynomial (TD)	$\mathcal{G}^*(d,p)$	TD $\geq \binom{d+p-1}{p}$
Gaussian (TD)	∞	∞
Gaussian (ϵ-TD)	$O(1/\epsilon^2)$	$d^{O(\log^2(1/\epsilon))}$

- As a cornerstone of analyzing TD, we at first propose a unified theoretical framework termed STARKE. Based on the subset analysis and the extension technique, STARKE is able to analyze the exact or ϵ-TD of the regularized ERM linear and non-linear kernel learners. With the help of STARKE, a generic result of TD or ϵ-TD is derived for any type of kernels.

- Via specifying the kernel type in STARKE, we disclose the TD of the regularized ERM homogeneous linear and polynomial kernel learners. Their TD is strictly reduced compared with that of the unregularized learners. As shown in Table 1, the TD is reduced from $\Theta(d)$ to 1 for the linear kernel learners. For the polynomial kernel learners, we reveal that $\mathcal{G}^*(d,p) \leq \binom{d+p-1}{p}$, and their TD is strictly smaller than $\binom{d+p-1}{p}$ for some target hypotheses.

- For regularized ERM Gaussian kernel learners, we reveal that, although their TD is infinite, their ϵ-TD can be exponentially reduced compared with that of the unregularized learners, as shown in Table 1. Besides, the experiment results verify the theoretical findings.

To the best of our knowledge, the above results are the first known bounds on (approximately) teaching the regularized ERM non-linear kernel learners.

The consequent sections introduce the preliminaries, describe the proposed STARKE framework, present the case study on five types of kernels, show the experiment results, and finally give the discussion and conclusion.

2. Preliminaries

This section introduces the necessary notations, definitions, concepts, and the existing results, so as to pave the way for the consequent sections.

Basic Definitions. We denote $\mathcal{X} \subseteq \mathbb{R}^d$ as the input space and $\mathcal{Y}$ as the output space. A hypothesis is a mapping $h\colon \mathcal{X} \to \mathcal{Y}$. This paper assumes that the hypothesis $h_{\boldsymbol{\theta}}$ can be identified by its parameter $\boldsymbol{\theta}$. The hypothesis space $\mathcal{H}$ consists of a set of hypotheses. A training example is denoted as a pair $(\mathbf{x}, y) \in \mathcal{X} \times \mathcal{Y}$. A training set is a multiset $D = \{(\mathbf{x}_1, y_1), \cdots, (\mathbf{x}_n, y_n)\}$, where the repeated pairs are acceptable. Let $\mathbb{D}$ denote the set of all training sets with all sizes. An algorithm $\mathcal{A}\colon \mathbb{D} \to 2^{\mathcal{H}}$ learns from a training set $D \in \mathbb{D}$ and outputs a subset of the hypothesis space $\mathcal{H}$.

Let $\boldsymbol{\theta}^* \in \mathcal{H}$ be the target hypothesis. The exact teaching is successful if a helpful teacher identifies a training set $D \in \mathbb{D}$ such that $\mathcal{A}(D) = \{\boldsymbol{\theta}^*\}$. Such a D is called the *teaching set* (TS) of $\boldsymbol{\theta}^*$ with respect to $\mathcal{H}$. The *teaching dimension* of $\boldsymbol{\theta}^*$ is the minimum size of the teaching set, i.e.,

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

Furthermore, the teaching dimension of the whole hypothesis space $\mathcal{H}$ is defined as the teaching dimension of the hardest hypothesis, i.e., $\mathrm{TD}(\mathcal{H}) = \max_{\boldsymbol{\theta} \in \mathcal{H}} \mathrm{TD}(\boldsymbol{\theta})$.

Regularized Empirical Risk Minimization. Consider a training set $D = \{(\mathbf{x}_i, y_i)\}_{i=1}^n$, where $\mathbf{x}_i \in \mathbb{R}^d$. The optimization problem identified by the regularized ERM can be formulated as

$$\mathcal{A}(D) = \arg\min_{\boldsymbol{\theta} \in \mathcal{H}} \sum_{i=1}^{n} \ell(f_{\boldsymbol{\theta}}(\mathbf{x}_i), y_i) + \Omega(||\boldsymbol{\theta}||^2)\,, \qquad (1)$$

where ℓ is the loss function and $||\cdot||$ is the L_2 norm. This paper assumes that the regularization function $\Omega(x^2)$ is strictly increasing, differentiable and convex. This assumption is not strong in the field of teaching theory, since analyzing TD of the non-linear learners is still in its infancy. It can be easily satisfied by the widely-used $\Omega(x^2) = \frac{1}{2}\mu x^2$ regularization function. The regularized ERM learner outputs a set of hypotheses $\mathcal{A}(D)$.

Approximate Teaching. When a finite teaching set does not exist, exact teaching is impossible and the teaching dimension is meaningless. Therefore, Kumar et al. (2021) propose to consider the ϵ-approximate teaching set and the ϵ-approximate teaching dimension instead. Let $\{\hat{\boldsymbol{\theta}}\} = \mathcal{A}(D)$, and $\boldsymbol{\theta}^*$ is the target hypothesis. If $\hat{\boldsymbol{\theta}}$ satisfies

$$\left|F(\hat{\boldsymbol{\theta}}) - F(\boldsymbol{\theta}^*)\right| \leq \epsilon\,,$$

where $F(\boldsymbol{\theta}) = \mathbb{E}[\ell(f_{\boldsymbol{\theta}}(\mathbf{x}), y)] + \Omega(||\boldsymbol{\theta}||^2)$ and the expectation is over $(\mathbf{x}, y) \sim \mathcal{P}$, then we call D as the ϵ-approximate teaching set for the regularized ERM learners. In a similar way, the ϵ-approximate teaching dimension (ϵ-TD) of them can also be defined.

Reproducing kernel Hilbert space (RKHS). An RKHS H is uniquely determined by a reproducing kernel, which is a kernel operator $k\colon \mathcal{X} \times \mathcal{X} \to \mathbb{R}$ adheres to the Mercer's positive definite conditions (Vapnik, 1998). Let $H_{\text{pre}} = \{\sum_{i=1}^n \alpha_i k(\mathbf{x}_i, \cdot)\colon n \in \mathbb{N}, \alpha_i \in \mathbb{R}, \mathbf{x}_i \in \mathcal{X}\}$, then H is the closure of H_{pre}. An RKHS with k can be decomposed as $k(\mathbf{x}_1, \mathbf{x}_2) = \langle \Phi(\mathbf{x}_1), \Phi(\mathbf{x}_2) \rangle$ (Steinwart & Christmann, 2008) for any $\mathbf{x}_1, \mathbf{x}_2 \in \mathcal{X}$, where $\Phi(\cdot)$ is the feature map. Specifically, the equation holds for $\Phi(\mathbf{x}) = k(\mathbf{x}, \cdot)$, and this type of feature map is called the canonical feature map.

We can generalize the problem identified by Equation (1) to the non-linear setting by the kernel method, namely, rewriting $f_{\boldsymbol{\theta}}(\mathbf{x})$ as the inner product $\langle \boldsymbol{\theta}, \Phi(\mathbf{x}) \rangle$. If the canonical feature map is used, the optimization problem becomes

$$\mathcal{A}(D) = \arg\min_{\boldsymbol{\theta} \in H} \sum_{i=1}^{n} \ell(\langle \boldsymbol{\theta}, k(\mathbf{x}_i, \cdot) \rangle, y_i) + \Omega(||\boldsymbol{\theta}||_H^2), \quad (2)$$

where $\|\cdot\|_H$ is the RKHS norm. In this way, the hypothesis space becomes the RKHS, i.e., $\mathcal{H} = H$, and we are able to express the hypothesis as $\boldsymbol{\theta} = \sum_{i=1}^{\infty} \alpha_i k(\mathbf{x}_i, \cdot)$ (Steinwart & Christmann, 2008).

3. STARKE: A Unified Theoretical Framework

In this section, we introduce the proposed **s**ubse**t** **a**nalysis framework for deriving teaching dimension of **r**egularized **k**ernel **l**earners (STARKE). The STARKE theoretical framework consists of two essential parts. First, analyzing the teaching dimension of a considered subset of RKHS. Second, extending the teaching to the whole RKHS via the strong representation ability of the considered subset. The two parts of STARKE and the relationship among the main theoretical results therein are illustrated in Figure 2. We elaborate them respectively.

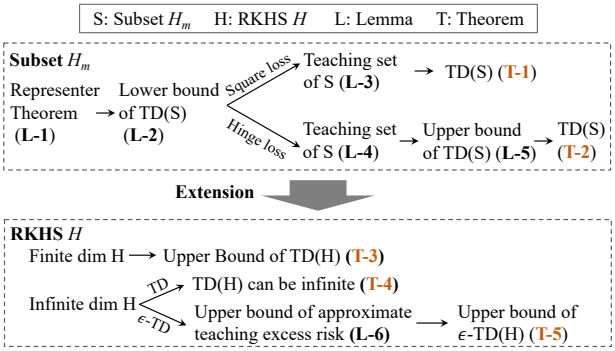


Figure 2. The STARKE framework and the relationship among the main theoretical results therein.

3.1. Subset Analysis

This paper considers the subset of RKHS defined as $H_m = \{\sum_{i=1}^{m} \alpha_i k(\mathbf{x}_i, \cdot) \colon \alpha_i \in \mathbb{R}, \mathbf{x}_i \in \mathcal{X}\}$. Notably, H_m is different from the aforementioned H_{pre} because m is fixed here. In the following analysis, we assume that $H_m - H_{m-1} \neq \emptyset$. The assumption means that we only consider the smallest m such that H_m stays unchanged. In fact, if the assumption does not hold, i.e., $H_m - H_{m-1} = \emptyset$, it implies that $H_m = H_{m-1}$ and thus H_m can be replaced by H_{m-1}. This process can be done recursively until we find an m_0 such that the assumption holds for H_{m_0}. Furthermore, the Representer Theorem is a canonical result in kernel methods. It is necessary in our proof process, and we state it in the following lemma for the purpose of self-contained.

Lemma 1 (Representer Theorem)**.** *Given a training set $\{\mathbf{x}_i, y_i\}_{i=1}^n$, if the regularization function Ω in Equation (2) is monotonically increasing, then the solution has the following form*

$$\boldsymbol{\theta}^* = \sum_{i=1}^{n} \alpha_i k(\mathbf{x}_i, \cdot)\,, \quad \alpha_i \in \mathbb{R}\,. \tag{3}$$

The proof of the above lemma can be found in (Schölkopf & Smola, 2002). Note that the regularization function Ω meets the condition of Lemma 1, and thus the solution of Equation (2) has the form of Equation (3). Next, we study the teaching dimension of H_m with square loss function and hinge loss function. It is easy to see that if the target hypothesis $\boldsymbol{\theta}^* = \mathbf{0}$, we do not need any training data to uniquely obtain the target hypothesis from Equation (2). Thus, we only consider the non-trivial case when $\boldsymbol{\theta}^* \neq \mathbf{0}$. At first, the lower bound of the teaching dimension of subset H_m is derived, and its proof can be found in Appendix A.1.

Lemma 2 (Lower Bound of TD(H_m))**.** *The lower bound of the teaching dimension of subset H_m is m.*

To determine the teaching dimension, it suffices to derive the upper bound if it matches the lower bound. The upper bound of TD(H_m) can be established by providing a teaching set (TS), the cardinality of which is its upper bound. We now construct the teaching sets for H_m, and the square loss function and the hinge loss function are analyzed respectively. The analysis is applied to both regression and classification tasks, since the classification tasks can be accomplished by simply set a threshold for the predicted value. The square loss function is $\ell(x, y) = (x - y)^2$. The teaching set is provided in the following lemma, and its proof can be found in Appendix A.2.

Lemma 3 (TS(H_m)-Square Loss)**.** *Given any $\boldsymbol{\theta}^* \in H_m - H_{m-1}$, where $\boldsymbol{\theta}^* = \sum_{i=1}^{m} \alpha_i^* k(\mathbf{x}_i^*, \cdot)$, then a teaching set of $\boldsymbol{\theta}^*$ with the square loss function is*

$$X = X^*, \quad Y = K^* \boldsymbol{\alpha}^* - \boldsymbol{\alpha}^* \Omega'((\boldsymbol{\alpha}^*)^T K^* \boldsymbol{\alpha}^*)\,,$$

where $X^* = (\mathbf{x}_1^*, \cdots, \mathbf{x}_m^*)^T$, $\boldsymbol{\alpha}^* = (\alpha_1^*, \cdots, \alpha_m^*)^T$, $\Omega'(x)$ is the derivative of the regularization function $\Omega(x)$ and K^* is the Gram matrix such that $K_{ij}^* = \alpha_i^* \alpha_j^* k(\mathbf{x}_i^*, \mathbf{x}_j^*)$.

Lemma 3 gives a teaching set for $\boldsymbol{\theta}^* \in H_m - H_{m-1}$. Noticing that $H_m = H_0 \cup (H_1 - H_0) \cup \cdots \cup (H_m - H_{m-1})$, we can derive the teaching set for all hypotheses in H_m by substituting different m. Since the teaching set has m elements, the upper bound of the teaching dimension of H_m is m. Combined it with Lemma 2 results in Theorem 1.

Theorem 1 (TD(H_m)-Square Loss)**.** *The teaching dimension of subset H_m with the square loss function is m.*

Theorem 1 indicates that whatever the kernel function we use, exact teaching can be performed in the subset H_m. We now turn to the hinge loss function defined as $\ell(x, y) = \max(1 - xy, 0)$. We denote $\langle k(\mathbf{x}, \cdot), \boldsymbol{\theta}^* \rangle$ as $g(\mathbf{x}, \boldsymbol{\theta}^*)$ and denote $\{1, 2, \ldots, n\}$ as $[n]$. The teaching set is constructed in Lemma 4, and its proof is in Appendix A.3.

Lemma 4 (TS(H_m)-Hinge Loss)**.** *If $\boldsymbol{\theta}^* \in H_m$, then a teaching set of $\boldsymbol{\theta}^*$ with the hinge loss function is*

$$\mathbf{x}_{ij} = \mathbf{x}_i^*,\ y_{ij} = \alpha_i \Omega_* / n_i, \quad i \in [m],\ j \in [n_i],$$

where $n_i = \lceil \max(1, \alpha_i g(\mathbf{x}_i^*, \boldsymbol{\theta}^*) \Omega_*) \rceil$ and Ω_* is defined as $2\Omega'(\|\boldsymbol{\theta}^*\|_H^2)$.

Lemma 5 (Upper Bound of TD(H_m)-Hinge Loss)**.** *If $\boldsymbol{\theta}^* \in H_m$, then the upper bound of TD(H_m) with the hinge loss function is $\sum_{i=1}^{m} \lceil \max(1, \alpha_i g(\mathbf{x}_i^*, \boldsymbol{\theta}^*) \Omega_*) \rceil$, where $\boldsymbol{\theta}^* = \sum_{i=1}^{m} \alpha_i^* k(\mathbf{x}_i^*, \cdot)$, $\Omega_* = 2\Omega'(\|\boldsymbol{\theta}^*\|_H^2)$.*

The corresponding upper bound of TD(H_m) is shown in Lemma 5, and it indicates that TD(H_m) $= m$ is non-trivial. It also implies the difference of TD between the square loss and the hinge loss function. That is to say, TD(H_m) $= m$ does not hold for all loss functions and regularization functions. Based on Lemma 5, the upper bound of TD(H_m) can be improved with properly selected regularization functions. The improved upper bound matches the lower bound so that we obtain TD(H_m) for the hinge loss function. Its proof is in Appendix A.4.

Theorem 2 (TD(H_m)-Hinge Loss)**.** *If the regularization function Ω satisfies $\Omega'(x) \leq \min_{i \in I^*} 1/\beta$, where $I^* = \{i \colon \alpha_i^* g(\mathbf{x}_i^*, \boldsymbol{\theta}^*) > 0\}$ and $\beta = 2\alpha_i^* \sqrt{k(\mathbf{x}_i^*, \mathbf{x}_i^*)}\, \|\boldsymbol{\theta}^*\|_H$, then TD($H_m$) with the hinge loss function is m.*

Remark. We would like to point out that, for the widely-used regularization function $\Omega(x) = \frac{1}{2}\mu x$, the condition of $\Omega'(x) \leq \min_{i \in I^*} 1/\beta$ implies $\mu \leq \min_{i \in I^*} 1/\beta$, thus the condition can be satisfied by choosing small enough μ.

In a nutshell, the teaching dimension of subset H_m for both square loss and hinge loss functions are m under the mild assumptions we have assumed.

3.2. Representation Ability of Subset H_m w.r.t. RKHS

In this section, we analyze the representation ability of subset H_m with respect to (w.r.t.) the RKHS H. Then the TD or ϵ-TD of any type of regularized ERM kernel learners can be disclosed consequently.

3.2.1. Teaching Dimension for Finite Dimensional RKHS

For any finite dimensional RKHS H, H can be represented by H_m with large enough m.

Theorem 3 (Upper Bound of TD(H)-Finite Dimension)**.** *If the dimension of the RKHS* $\dim(H) = d_H$*, then we have that* $H_m = H$ *with* $m \leq d_H$.

The proof can be found in Appendix B.1. From the theorems and the conclusion of the above section, the teaching dimension of the regularized kernel learners with finite dimensional RKHSs is $O(d_H)$ if $\dim(H) = d_H$.

3.2.2. Teaching Dimension for Infinite Dimensional RKHS

Kumar et al. (2021) prove that exact teaching requires the infinite teaching set for Gaussian kernel which induces an infinite dimensional RKHS. However, for the regularized learners, since the regularization function can be arbitrarily chosen, the result is not so obvious. In this section, we prove that even if using regularization, for some specific kernels, such as Gaussian, exponential and Laplacian kernels, we cannot find a finite teaching set for some target hypotheses. The pessimistic result is shown in Theorem 4, and its proof is in Appendix B.1.

Theorem 4 (TD(H)-Pessimistic Infinity)**.** *Assume the RKHS H of the considered kernel is infinite dimensional. For all regularization functions, the TD of the regularized learner is ∞ if there exists no $m_0 < \infty$ s.t. $H_{m_0} = H$.*

Remark 1. Theorem 4 does not require the assumption of Ω. In other words, this theorem can be applied to all types of regularization functions. Furthermore, this theorem can also be applied to all types of loss functions.

Remark 2. The RKHSs induced by Gaussian, exponential and Laplacian kernels are infinite dimensional and can not be represented by H_m if $m < \infty$. The proof can be found in the Appendix B.3.

3.2.3. Approximate Teaching from H_m to H

Although the whole RKHS cannot be exactly taught for some kernels, the result on H_m implies that we can realize exact teaching on the subset of H. In this way, the subset H_m can be seen as an approximation to H, and performing teaching in H_m is performing approximate teaching in H.

It has been clarified that there exists an RKHS that cannot be expressed as H_m with a finite m. In this case, we attempt to explore how well H_m can approximate H. Let $\boldsymbol{\theta}^*$ be the optimal solution to

$$\min_{\boldsymbol{\theta}\in H} F(\boldsymbol{\theta}) = \min_{\boldsymbol{\theta}\in H} \mathbb{E}[\ell(\langle k(\mathbf{x},\cdot), \boldsymbol{\theta}\rangle, y)] + \Omega(||\boldsymbol{\theta}||_H^2), \quad (4)$$

where the expectation is taken over the joint distribution $\mathcal{P}(\mathbf{x}, y)$. Following Koltchinskii (2011), we define the excess risk of any hypothesis $\boldsymbol{\theta}$ as

$$\Lambda(\boldsymbol{\theta}) = F(\boldsymbol{\theta}) - F(\boldsymbol{\theta}^*). \quad (5)$$

Let $\boldsymbol{\theta}_m^*$ be the optimal solution to $\min_{\boldsymbol{\theta}\in H_m} F(\boldsymbol{\theta})$, then the approximation error of H_m is $\Lambda(\boldsymbol{\theta}_m^*)$.

Similar to Yang et al. (2012), we assume $\max_{y\in\mathcal{Y}} \ell(0, y) \leq 1$ and $\ell(z, y)$ has a bounded partial derivative $|\nabla_z \ell(z, y)| \leq C_1$. Consider $\{(\mathbf{x}_i, y_i)_{i=1}^N\}$, which is i.i.d. and sampled from the joint distribution $\mathcal{P}(\mathbf{x}, y)$. Let K be the Gram matrix of $\{\mathbf{x}_i\}_{i=1}^N$, and $\sup_x k(\mathbf{x}, \mathbf{x}) \leq C(k)$ where $C(k)$ is a constant corresponding to kernel k, and $\{\lambda_i\}_{i=1}^N$ be the eigenvalues of K with $\lambda_1 \geq \cdots \geq \lambda_N$, Ω^{-1} be the inverse function of Ω, the existence of which is guaranteed by the monotonicity of Ω, we can upper bound the excess risk of approximate teaching as Lemma 6. Its proof is in Appendix B.4.

Lemma 6 (Upper Bound of Approximate Teaching Excess Risk)**.** *If $M_1 = \Omega^{-1}(\Omega(0) + 1) \leq e^{2N}/4\gamma^2$, where $\gamma \leq 1$, $\sup_{\|\boldsymbol{\theta}\|_H \leq M_1} 2\|\boldsymbol{\theta}\|_H \Omega'(\|\boldsymbol{\theta}\|_H^2)$ is bounded, and $\lambda_{m+1} \in O(N/\sqrt{m})$ for all N, then we have that*

$$\Lambda(\boldsymbol{\theta}_m^*) \in O\left(\frac{C(k)+1}{\sqrt{m}}\right).$$

Lemma 6 indicates that the excess risk raised by teaching in a subset rather than the whole RKHS is upper bounded by $O((C(k)+1)/\sqrt{m})$. Since the definition of ϵ-TD is that the excess risk between the taught hypothesis and the target hypothesis is no more than ϵ, we can derive the ϵ-TD based on Lemma 6 as shown in Theorem 5.

Theorem 5 (Upper Bound of ϵ-TD(H))**.** *Under the conditions as Lemma 6, for the regularized ERM learner, the ϵ-TD of the RKHS H is upper bounded by*

$$\epsilon\text{-TD}(H) \in O\left(\left(\frac{C(k)+1}{\epsilon}\right)^2\right).$$

4. Theoretical Case Study

We take the linear, polynomial, Gaussian, exponential and Laplacian kernels as the case study to show how to apply the generic STARKE framework to determine the TD or ϵ-TD of the regularized kernel learners. The RKHSs of the linear

and polynomial kernels are finite dimensional. Based on Theorem 3, exact teaching is achievable under such kernels. However, the situation differs for Gaussian, exponential and Laplacian kernels, and Theorem 5 can be helpful.

4.1. Linear Kernel

The linear kernel is defined as $k(\mathbf{x}_1, \mathbf{x}_2) = \langle \mathbf{x}_1, \mathbf{x}_2 \rangle + c$. For simplicity, this paper considers the homogeneous scenario, i.e., $c = 0$. The result when $c \neq 0$ can be derived similarly. The canonical feature map of linear kernel is $\Phi(\mathbf{x}) = \langle \mathbf{x}, \cdot \rangle$. We denote the dimension of input space $\dim(\mathcal{X}) = d$. Based on the linearity of inner product, we have Lemma 7.

Lemma 7 (H Identification)**.** *Let H be the RKHS defined by the linear kernel, and $\dim(\mathcal{X}) = d$, then $H = H_1$.*

We leave the proof of Lemma 7 in Appendix C.1. According to Lemma 7 and the theoretical result of subset, we derive the following corollary.

Corollary 1 (TD under Linear Kernel)**.** *Under the conditions in Theorem 2, for the regularized ERM linear kernel learners with the considered two loss functions, the TD of the RKHS H is $\mathrm{TD}(H) = 1$.*

Remark. For the linear kernel learner without regularization, Kumar et al. (2021) derive the teaching dimension, which is $\Theta(d)$. When the regularization term is equipped, the TD is drastically reduced to 1. This implies that regularization is helpful to reduce the teaching complexity.

4.2. Polynomial Kernel

The polynomial kernel of degree $p \in \mathbb{N}$ is defined as $k(\mathbf{x}_1, \mathbf{x}_2) = (\langle \mathbf{x}_1, \mathbf{x}_2 \rangle + c)^p$, where $c \geq 0$ is a constant. We consider the homogeneous scenario for simplicity, i.e., $c = 0$. The result when $c > 0$ can be derived similarly.

For the input space $\mathcal{X} \subseteq \mathbb{R}^d$, the dimension of RKHS induced by polynomial kernel with degree p is $\phi = \binom{d+p-1}{p}$. By Theorem 3, we have that $\boldsymbol{\theta}^* = \sum_{i=1}^{\phi} \alpha_i k(\mathbf{x}_i, \cdot)$. Note that $\alpha_i k(\mathbf{x}_i, \cdot) = k(\sqrt[p]{\alpha_i}\mathbf{x}_i, \cdot)$. Without loss of generality, we assume $\alpha_i = 1$. Let $\mathbf{z} = (z_1, \cdots, z_d)$, we have

$$\begin{aligned}\langle \boldsymbol{\theta}^*, \mathbf{z} \rangle = C_1 z_1^p + \frac{p!}{(p-1)!1!} C_2 z_1^{p-1} z_2 \\ + \frac{p!}{(p-1)!1!} C_3 z_1^{p-1} z_3 + \cdots + C_p z_d^p \,.\end{aligned}$$

Consider the following system of polynomial equations

$$\begin{cases} y_{11}^p + y_{21}^p + \cdots + y_{m1}^p = C_1 \\ y_{11}^{p-1} \cdot y_{12} + \cdots + y_{m1}^{p-1} \cdot y_{m2} = C_2 \\ \cdots \\ y_{1d}^p + \cdots + y_{md}^p = C_\phi \,, \end{cases} \tag{6}$$

where y_{ij} are variables. If the solution for the polynomial system exists, let $\mathbf{y}_i = (y_{i1}, \cdots, y_{id})$, then $\boldsymbol{\theta}^* = \sum_{i=1}^m k(\mathbf{y}_i, \cdot)$. Namely, $\boldsymbol{\theta}^*$ can be expressed with no more than m components. For simplicity, let the polynomials be $f_1, f_2, \ldots, f_\phi$. Define the degree lexicographic order as $y_{11} \succ y_{12} \succ \cdots \succ y_{21} \succ \cdots \succ y_{md}$, and the derived Gröbner basis (Hartshorne, 1977) for f_i as $\mathcal{G}(\boldsymbol{\theta}^*, d, p, m)$. We have the following lemma.

Lemma 8 (H Identification)**.** *Let H be the RKHS defined by the homogeneous polynomial kernel with degree p, then $H = H_{\mathcal{G}^*(d,p)}$, where*

$$\begin{aligned}\mathcal{G}^*(d,p) &= \max_{\boldsymbol{\theta} \in H} \tilde{\mathcal{G}}(\boldsymbol{\theta}, d, p) \\ &= \max_{\boldsymbol{\theta} \in H} \left\{ \arg\min_m \{ m : \mathcal{G}(\boldsymbol{\theta}, d, p, m) \neq \{1\} \} \right\}.\end{aligned}$$

Its proof is in Appendix C.2. With Lemma 8 and the results of Theorem 1 and 2, the following corollary is derived.

Corollary 2 (TD under Polynomial Kernel)**.** *Under the conditions in Theorem 2, for the regularized ERM polynomial kernel learners with the considered two loss functions, the teaching dimension of $\boldsymbol{\theta}^*$ is $\mathrm{TD}(\boldsymbol{\theta}^*) = \tilde{\mathcal{G}}(\boldsymbol{\theta}^*, d, p)$ if $\boldsymbol{\theta}^* \neq \mathbf{0}$ ($\mathrm{TD}(\mathbf{0}) = 0$) , and the teaching dimension of the RKHS H is $\mathrm{TD}(H) = \mathcal{G}^*(d, p)$.*

Remark 1. $\mathcal{G}^*(d,p) \leq \binom{d+p-1}{p}$. This can be seen by letting $m = \binom{d+p-1}{p}$, then the polynomial system (6) has at least one solution $y_{ij} = x_{ij}$. By the Hilbert's Nullstellensatz (Hartshorne, 1977), $\mathcal{G}(\boldsymbol{\theta}^*, d, p, m) \neq \{1\}$, and this holds for all $\boldsymbol{\theta}^* \in H$.

Remark 2. $\tilde{\mathcal{G}}(\boldsymbol{\theta}^*, d, p)$ can be strictly smaller than $\binom{d+p-1}{p}$ for some $\boldsymbol{\theta}^*$. We can easily find out an example to satisfy it (cf. Appendix C.3 for more details of the example).

Remark 3. For the polynomial kernel learner without regularization, Kumar et al. (2021) derive the lower bound of the TD, which is $\binom{d+p-1}{p}$. As mentioned before, the TD for regularized learner meets $\mathcal{G}^*(d,p) \leq \binom{d+p-1}{p}$, and is strictly smaller than $\binom{d+p-1}{p}$ for some $\boldsymbol{\theta}^*$. Besides, the lower bound of ERM learner without regularization needs the assumptions on $\boldsymbol{\theta}^*$, which are stated in Appendix D. This indicates that regularization not only reduces the sample complexity of teaching, but also relaxes the conditions on exact teaching for polynomial kernel.

4.3. Gaussian Kernel

The Gaussian kernel with parameter $\sigma > 0$ is defined as $k(\mathbf{x}_1, \mathbf{x}_2) = \exp(-\frac{||\mathbf{x}_1 - \mathbf{x}_2||^2}{2\sigma^2})$. The TD of Gaussian kernel learners is infinity, and ϵ-TD is suitable for it. Corollary 3 shows the result of ϵ-TD for the regularized ERM Gaussian kernel learners, and we leave its proof in Appendix C.4.

Corollary 3 (ϵ-TD under Gaussian Kernel)**.** *Under the conditions in Theorem 2 and 5, for the regularized ERM Gaussian kernel learners with the considered two loss functions, the ϵ-approximate teaching dimension of the RKHS H is* ϵ-TD$(H) \in O(1/\epsilon^2)$.

Remark. For the ERM Gaussian kernel learners without regularization, Kumar et al. (2021) derive that the ϵ-TD is upper bounded by $d^{O(\log^2(1/\epsilon))}$. It can be seen that the ϵ-TD is exponentially reduced with regularization. Besides, the assumptions they made is not necessary in the regularization scenario (cf. Appendix D for more discussions).

4.4. Exponential and Laplacian Kernels

In addition to linear, polynomial and Gaussian kernels, our generic STARKE framework can also be applied to other types of kernels. We take exponential and Laplacian kernels as examples, and analyze TD and ϵ-TD of them. Notably, Kumar et al. (2021) do not involve those types of kernels.

As shown in Appendix B.3, the RKHSs of both exponential and Laplacian kernels cannot be expressed by H_m. Thus, the TD of exponential and Laplacian kernels is infinite. On the other hand, approximate teaching from H_m to H enables us to obtain the ϵ-TD of the two kernels as Corollary 4. Its proof can be found in Appendix C.4.

Corollary 4 (ϵ-TD under Exponential and Laplacian Kernels)**.** *Under the conditions in Theorem 2 and 5, for the regularized ERM exponential and Laplacian kernel learners with the considered two loss functions, the ϵ-approximate teaching dimension of the RKHS H is* ϵ-TD$(H) \in O(1/\epsilon^2)$.

5. Experiments

In this section, we perform the empirical study to verify the theoretical results. The code is available at https://github.com/liuxhym/STARKE.git.

For exact teaching, we provide numerical results of linear and polynomial kernel learners respectively. For the regularized linear learners, with the target hypothesis in $\mathbb{R}^3$, the teaching sets have only one element for both square loss and hinge loss. However, teaching the unregularized learners needs three elements for square loss and five elements for hinge loss. For the regularized polynomial learner, given a certain target hypothesis, the teaching sets have two elements for both square loss and hinge loss. While the unregularized learners need four elements for square loss and cannot be taught for hinge loss since the violation of assumption (cf. Appendix F for more details).

We next show the empirical results for approximate teaching with Gaussian kernel. Since STARKE can be applied to any type of kernels, we also conduct experiments on exponential and Laplacian kernels (cf. Appendix G for more details due to page limitation). For Gaussian kernel, we set $\sigma = 0.9$ and adopt square loss for regression while hinge loss for classification. For regression, we choose two synthetic datasets: the make-regression (MR) dataset from sklearn as well as the Sin dataset, and two real-world datasets: MPG from UCI (Blake et al., 1998) and Eunite. The regularization function $\Omega(x^2) = x^2$ is applied. For classification, we choose two synthetic datasets: the two-moon (Moon) dataset as well as the two-circles (Circle) dataset from sklearn, and two UCI binary classification datasets: Adult and Sonar. The classification threshold is set as zero in experiments. The regularization function $\Omega(x^2) = \frac{1}{200}x^2$ is applied.

According to the theoretical result, the performance of the teaching is measured by excess risk. However, this measurement varies dramatically among different datasets. In order to avoid the influence of datasets on the excess risk, we introduce the excess risk ratio $\bar{\Lambda}$ as the measurement, which is the value of the current excess risk Λ divided by a reference excess risk. For each dataset, the reference risk is the average of Λ calculated by $\boldsymbol{\theta}^*$ with 5% random samples for 100 times except for Adult. Considering the larger sample size of Adult compared with the others, if we set under 5% random samples as before, the teaching set will be very large even if the excess risk ratio is 100%, resulting in the relationship between teaching set size and excess risk ratio being expressed inappropriately. Thus, for Adult, the reference risk is calculated under 1% random samples for 100 times and then averaged. We use Nyström method (Williams & Seeger, 2000) to approximate $\boldsymbol{\theta}_m^* \in H_m$, which is an approximation of $\boldsymbol{\theta}^* \in H$. For approximate teaching, $\boldsymbol{\theta}_m^*$ is treated as the target hypothesis, and the approximate teaching set is constructed via Lemma 3 for square loss and Lemma 4 for hinge loss based on $\boldsymbol{\theta}_m^*$.

Visualization of the Teaching Set. For intuitive understanding, we visualize the teaching set for the regularized ERM learners with Gaussian kernel on the synthetic datasets. The cardinality of teaching sets is determined by whether it is enough to obtain a low risk learned hypothesis, and we choose the smallest possible one. Figure 3 shows the results for regression on Sin dataset, while Figure 4 shows the results for classification on Moon and Circle datasets. The left sub-figure of Figure 3 shows the data points in the dataset and the target hypothesis. The teaching set and the learned hypothesis is shown in the right sub-figure. The top sub-figures in Figure 4 are the results on the Circle dataset, and the bottom sub-figures are the results on the Moon dataset. The interface between the blue and red areas is the decision boundary of teacher in the sub-figure (a), and learner in the sub-figure (b). The positive and negative points in dataset are marked by red and blue dots respectively in the sub-figure (a). The constructed teaching set (TS) is shown by dark-red stars in the sub-figure (b). The results show that with much less data points than that of the dataset, the

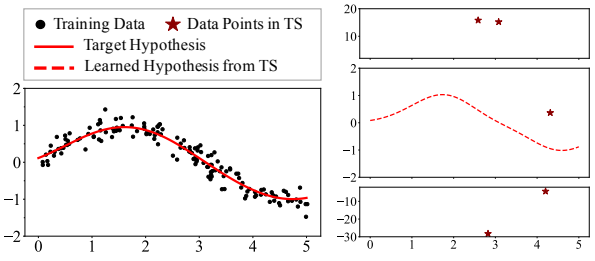
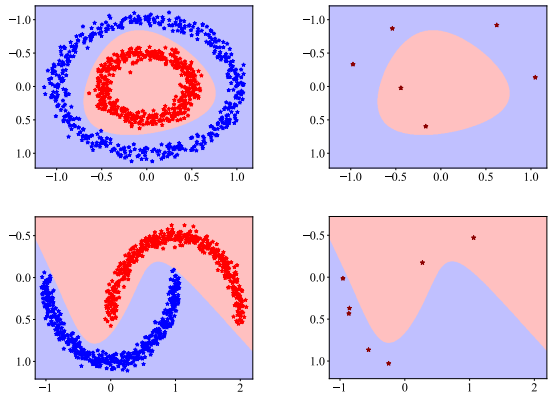


Figure 3. Approximate teaching on the Sin dataset. The left sub-figure: the target hypothesis $\boldsymbol{\theta}^*$ is marked as the red solid line and the data points is marked as the black dots. The right sub-figure: the learned hypothesis is marked as the dashed red line and the teaching set is marked as the dark-red stars.



(a) $\boldsymbol{\theta}^*$ (b) Learned hypothesis

Figure 4. Approximate teaching on the Moon and Circle datasets. The binary dataset is marked by red and blue dots. The interface between the blue and red regions is the decision boundary of the learned hypothesis. (a) The target hypothesis $\boldsymbol{\theta}^*$. (b) The learned hypothesis with the teaching set being marked as the dark-red stars.

learner can obtain nearly the same hypothesis as the target hypothesis generated by the dataset.

On Approximate Ability of H_m. In the approximate teaching setting, we treat H_m as an approximation to H. From theoretical perspective, the approximation error of H_m decreases as $O((C(k)+1)/\sqrt{m})$, implying H_m is a good approximation of H. We study the approximate ability of H_m empirically. To show the approximate ability of H_m, we perform Nyström method (Williams & Seeger, 2000) with m components to fit the kernel SVM and kernel ridge regression. The whole process is repeated for 100 times and the excess risk of H_m is represented by the best learned hypothesis to match the machine teaching setting.

The result is shown in Table 2. The value inside the table is the smallest m such that H_m achieves an excess risk ratio no more than that in the first line of the table. For the Sin dataset, the excess risk of H_8 can even achieve zero. All the

Table 2. The relation between excess risk ratio $\bar{\Lambda}$ and m of H_m.

Dataset	$\bar{\Lambda}=100\%$	80%	60%	40%	20%	0%
Sin	1	1	1	1	1	8
MR	1	1	2	5	7	>60
MPG	2	3	3	5	10	>60
Eunite	3	4	5	6	7	>60
Circle	1	2	3	3	4	>60
Moon	1	3	3	4	5	>60
Adult	9	139	396	>400	>400	>400
Sonar	4	28	62	103	149	> 200

regression datasets achieve less than 20% excess risk ratio within 10 samples. For classification, H_m achieves small excess risk in synthetic datasets with small m. Adult and Sonar are more difficult and need higher m to achieve small excess risk ratio. However, if we focus on the accuracy of classification rather than the excess risk, the hypothesis in H_m with small m also performs well, which is shown in Appendix H. Therefore, the approximate ability of H_m is good enough for approximate teaching. We can also observe that the number from left to right in the table increases quadratically as exposed by Corollary 3.

Regularization vs. Without Regularization. Before teaching, the target hypothesis $\boldsymbol{\theta}^* \in H$ should be obtained first. For regression, $\boldsymbol{\theta}^*$ is obtained by performing the Gaussian kernel ridge regression on the whole dataset, while the Gaussian kernel SVM is performed for classification. For approximate teaching, the estimated $\boldsymbol{\theta}^*$, which belongs to H_m, is obtained by performing Nyström method (Williams & Seeger, 2000) with m components on the dataset. In order to reduce the error caused by randomness, the Nyström method is repeated for 15 times and we choose the one with the lowest excess risk.

The comparison between the regularized learners and the unregularized ones is shown Table 3. The first line of Table 3 is as same as Table 2. The value/symbol inside the table shows the difference between the ϵ-TD of the regularized learners and that of the unregularized ones. "$\circ$" means the unregularized learner cannot reach such excess risk ratio within 60 samples, and "$\times$" means both regularized and unregularized learners cannot reach such excess risk ratio within 60 samples. The result shows that the regularized learner surpasses the unregularized one in approximate teaching, because the differences are all positive. The scalability of teaching for regularized learner is also better than unregularized one, as teaching fails for unregularized learner on hard datasets or under small excess risk ratio, as indicated by the "$\circ$". This is because as ϵ decreases, the ϵ-TD of unregularized learner increases exponentially while regularized learner increases quadratically. It matches our theoretical findings.

By comparing the difference of ϵ-TD on Sin, MR, MPG,

Table 3. The difference between ϵ-TD of the regularized learners and that of the unregularized learners under the excess risk ratio $\bar{\Lambda}$. "$\circ$" means that only unregularized learner cannot reach such ratio within 60 samples, and "$\times$" means that both learners cannot reach such ratio within 60 samples.

Dataset	$\bar{\Lambda} = 100\%$	80%	60%	40%	20%	0%
Sin	$\circ$	$\circ$	$\circ$	$\circ$	$\circ$	$\circ$
MR	4	5	4	9	8	$\times$
MPG	23	25	27	29	$\circ$	$\times$
Eunite	$\circ$	$\circ$	$\circ$	$\circ$	$\circ$	$\times$
Circle	$\circ$	$\circ$	$\circ$	$\circ$	$\circ$	$\circ$
Moon	25	26	$\circ$	$\circ$	$\circ$	$\circ$
Adult	$\circ$	$\circ$	$\circ$	$\times$	$\times$	$\times$
Sonar	$\circ$	$\circ$	$\circ$	$\circ$	$\times$	$\times$

Eunite, Circle, Moon, Adult and Sonar under a unified criterion, e.g., 40% excess risk ratio, it reveals that Sin, Eunite, Circle, Moon and Sonar are hard for teaching without regularization, while MR and MPG are easy (MR is easier than MPG). Adult is hard for both teaching with and without regularization. The excess risk ratio $\bar{\Lambda}$ enables us to explicitly observe the advantages of regularization on different datasets in a unified way.

6. Discussion and Conclusion

This paper gives an affirmative answer to the essential question of whether regularization can help reduce the teaching complexity in machine teaching. We propose a unified theoretical framework STARKE that is able to analyze any type of kernels. With the help of STARKE, we intensively analyze the popular regularized ERM (non-linear) kernel learners, e.g., kernel SVM and kernel ridge regression. Our theoretical findings reveal that, when equipped with regularization, the TD or ϵ-TD of them is substantially reduced compared with that of the unregularized ones. The results obtained may be beneficial for the researchers to have a deeper understanding of teaching the complex concepts.

We would like to point out that Kumar et al. (2021) inspire this work. They analyze the perceptron loss instead of the square loss and the hinge loss, whereas we do not include the perceptron loss in our framework. On the one hand, the square loss and the hinge loss function may be more popular. On the other hand, the power of perceptron loss may be weaker and the optimal solution of the optimization problem is dominated by the regularization function, so that it cannot learn some hypotheses. The possible future work could be generalizing the proposed framework to other loss functions (e.g., exponential loss and logistic loss), and imposing some realistic conditions on the teaching set under the real-world scenarios instead of arbitrary selection.

Acknowledgements

The authors would like to thank Wei Wang for the valuable suggestions, and the anonymous reviewers for their helpful comments. This work is supported by the National Natural Science Foundation of China (No. 62106076), the Natural Science Foundation of Shanghai (No. 21ZR1420300), the "Chenguang Program" sponsored by Shanghai Education Development Foundation and Shanghai Municipal Education Commission (No. 21CGA32), the Shanghai Key Laboratory of Multidimensional Information Processing at East China Normal University (No. MIP202101), the Fundamental Research Funds for the Central Universities, and the National Key Laboratory for Novel Software Technology at Nanjing University (No. KFKT2021B14).

References

Anthony, M., Brightwell, G. R., and Shawe-Taylor, J. On specifying boolean functions by labelled examples. *Discrete Applied Mathematics*, 61(1):1–25, 1995.

Blake, C. L., Keogh, E., and Merz, C. J. UCI Repository of machine learning databases. [http://www.ics.uci.edu/~mlearn/MLRepository.html], 1998.

Bousquet, O. and Elisseeff, A. Stability and generalization. *Journal of Machine Learning Research*, 2:499–526, 2002.

Chen, Y., Singla, A., Aodha, O. M., Perona, P., and Yue, Y. Understanding the role of adaptivity in machine teaching: The case of version space learners. In *Proceedings of the 31st International Conference on Neural Information Processing Systems (NeurIPS'18)*, pp. 1483–1493, Montreal, Canada, 2018.

Drewnik, M. and Pasternak-Winiarski, Z. SVM kernel configuration and optimization for the handwritten digit recognition. In *Proceedings of 16th Conference on Computer Information Systems and Industrial Management (CISIM'17)*, pp. 87–98, Bialystok, Poland, 2017.

Drineas, P. and Mahoney, M. W. On the nyström method for approximating a gram matrix for improved kernel-based learning. *Journal of Machine Learning Research*, 6(12): 2153–2175, 2005.

Fadel, S., Ghoniemy, S., Abdallah, M., Sorra, H. A., Ashour, A., and Ansary, A. Investigating the effect of different kernel functions on the performance of svm for recognizing arabic characters. *International Journal of Advanced Computer Science and Applications*, 7(1):446–450, 2016.

Feragen, A., Lauze, F., and Hauberg, S. Geodesic exponential kernels: When curvature and linearity conflict. In *IEEE Conference on Computer Vision and Pattern*

Recognition (CVPR'15), pp. 3032–3042, Boston, MA, 2015.

Gao, Z., Ries, C., Simon, H. U., and Zilles, S. Preference-based teaching. *Journal of Machine Learning Research*, 2017.

Goldman, S. A. and Kearns, M. J. On the complexity of teaching. In *Proceedings of the 4th Workshop on Computational Learning Theory (COLT'91)*, pp. 303–314, Santa Cruz, CA, 1991.

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

Hartshorne, R. *Algebraic Geometry*. Springer, 1977.

Kamalaruban, P., Devidze, R., Cevher, V., and Singla, A. Interactive teaching algorithms for inverse reinforcement learning. In *Proceedings of the 28th International Joint Conference on Artificial Intelligence (IJCAI'19)*, pp. 2692–2700, Macao, China, 2019.

Kirkpatrick, D. G., Simon, H. U., and Zilles, S. Optimal collusion-free teaching. In *Proceedings of 30th International Conference on Algorithmic Learning Theory (ALT'19)*, pp. 506–528, Chicago, IL, 2019.

Koltchinskii, V. *Oracle Inequalities in Empirical Risk Minimization and Sparse Recovery Problems*. Springer, 2011.

Koltchinskii, V. and Yuan, M. Sparsity in multiple kernel learning. *The Annals of Statistics*, 38(6):3660–3695, 2010.

Kumar, A., Zhang, H., Singla, A., and Chen, Y. The teaching dimension of kernel perceptron. In *Proceedings of the 24th International Conference on Artificial Intelligence and Statistics (AISTATS'21)*, pp. 2071–2079, Virtual Event, 2021.

Liu, J., Zhu, X., and Ohannessian, H. The teaching dimension of linear learners. In *Proceedings of the 33rd International Conference on Machine Learning (ICML'16)*, pp. 117–126, New York City, NY, 2016.

Mansouri, F., Chen, Y., Vartanian, A., Zhu, X. J., and Singla, A. Preference-based batch and sequential teaching: Towards a unified view of models. In *Proceedings of the 32nd International Conference on Neural Information Processing Systems (NeurIPS'19)*, pp. 9195–9205, Vancouver, Canada, 2019.

Mohri, M., Rostamizadeh, A., and Talwalkar, A. *Foundations of Machine Learning*. MIT Press, 2012.

Patil, K. R., Zhu, X., Kopec, L., and Love, B. C. Optimal teaching for limited-capacity human learners. In *Proceedings of the 27th International Conference on Neural Information Processing Systems (NeurIPS'14)*, pp. 2465–2473, Montreal, Canada, 2014.

Schölkopf, B. and Smola, A. J. *Learning with Kernels: Support Vector Machines, Regularization, Optimization, and beyond*. MIT Press, 2002.

Shafto, P., Goodman, N. D., and Griffiths, T. L. A rational account of pedagogical reasoning: Teaching by, and learning from, examples. *Cognitive Psychology*, 71:55–89, 2014.

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

Steinwart, I. and Christmann, A. *Support Vector Machines*. Springer, 2008.

Vapnik, V. *Statistical Learning Theory*. Wiley, 1998.

Williams, C. K. I. and Seeger, M. W. Using the nyström method to speed up kernel machines. In *Proceedings of the 13rd Conference on Neural Information Processing Systems (NeurIPS'20)*, pp. 682–688, Denver, CO, 2000.

Yang, T., Li, Y., Mahdavi, M., Jin, R., and Zhou, Z. Nyström method vs random fourier features: A theoretical and empirical comparison. In *Proceedings of the 26th Conference on Neural Information Processing Systems (ICML'12)*, pp. 485–493, Lake Tahoe, NV, 2012.

Zhang, X., Zhu, X., and Wright, S. J. Training set debugging using trusted items. In *Proceedings of the 32nd AAAI Conference on Artificial Intelligence (AAAI'18)*, pp. 4482–4489, New Orleans, LA, 2018.

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

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

A. Proofs in Section 3.1

A.1. Proof of Lemma 2

Proof of Lemma 2. Because $H_m - H_{m-1} \neq \emptyset$, there exists $\boldsymbol{\theta} \in H_m$ such that $\boldsymbol{\theta}$ cannot be expressed with less than m terms. However, by Lemma 1, the number of terms contained by the learned hypothesis is no more than the cardinality of the training set. Therefore, the cardinality of the teaching set is no less than m examples, and we conclude the proof. □

A.2. Proof of Lemma 3

We first introduce Lemma 9, which is necessary to prove Lemma 3 in the paper.

Lemma 9. *For $\{\mathbf{x}_i\}_{i=1}^m$, denote $\boldsymbol{\theta} = \sum_{i=1}^m \alpha_i k(\mathbf{x}_i, \cdot)$, if $\boldsymbol{\theta} \in H_m - H_{m-1}$, then the Gram matrix K with $K_{ij} = \langle k(\mathbf{x}_i, \cdot), k(\mathbf{x}_j, \cdot)\rangle$ is invertible.*

Proof. Let $F = \text{span}\{k(\mathbf{x}_1, \cdot), \cdots, k(\mathbf{x}_m, \cdot)\}$, then F has finite dimension and we can find a standard orthogonal basis of F via Schmidt orthogonalization. We denote the basis as $e_1, e_2, e_3, \cdots, e_m$, then $(k(\mathbf{x}_1, \cdot), k(\mathbf{x}_2, \cdot), \cdots, k(\mathbf{x}_m, \cdot)) = (e_1, e_2, \cdots, e_m)P$, where P is invertible. We have that

$$\begin{aligned}\langle \alpha_i k(\mathbf{x}_i, \cdot), \alpha_j k(\mathbf{x}_j, \cdot)\rangle &= \left\langle \sum_{k=1}^m p_{ki} e_k, \sum_{r=1}^m p_{rj} e_r \right\rangle \\ &= \sum_{k=1}^m \sum_{r=1}^m p_{ki} p_{rj} \langle e_k, e_r\rangle \\ &= \sum_{k=1}^m p_{ki} p_{kj} \langle e_k, e_k\rangle \\ &= \sum_{k=1}^m p_{ki} p_{kj}.\end{aligned}$$

Thus, $K = P^T P$ and P is invertible $\Rightarrow$ K is invertible. □

Proof of Lemma 3. Provided $X = X^*$, the estimated $\hat{\boldsymbol{\theta}}$ can be expressed as $\sum_{i=1}^m \alpha_i k(\mathbf{x}_i^*, \cdot)$ according to Lemma 1 in the main paper. Therefore, we can rewrite the optimization problem as

$$\mathcal{A}(D) = \arg\min_{\alpha} \sum_{i=1}^m \ell(\beta_i, y_i) + \Omega(\boldsymbol{\alpha}^T K \boldsymbol{\alpha}), \quad (7)$$

where β_i is the i-th element of $K\boldsymbol{\alpha}$. Furthermore, the KKT (Karush–Kuhn–Tucker) condition states that

$$2K\boldsymbol{\alpha}\Omega'(\boldsymbol{\alpha}^T K \boldsymbol{\alpha}) \in \sum_{i=1}^m (K)_i \nabla_{\beta_i} \ell(\beta_i, y_i), \quad (8)$$

where $\nabla_{\beta_i} \ell(\beta_i, y_i) = \frac{\partial \ell(\beta_i, y_i)}{\partial \beta_i}$ and K_i is the i-th column of K. By substituting $\ell(x, y) = (x - y)^2$ into Formula (10), we have that

$$2K\boldsymbol{\alpha}\Omega'(\boldsymbol{\alpha}^T K \boldsymbol{\alpha}) \in 2K(K\boldsymbol{\alpha} - Y). \quad (9)$$

By Lemma 9, we have that

$$2\boldsymbol{\alpha}\Omega'(\boldsymbol{\alpha}^T K \boldsymbol{\alpha}) \in 2K(K\boldsymbol{\alpha} - Y). \quad (10)$$

By substituting $\ell(x, y) = (x - y)^2$ into Formula (10), we have that

$$2\boldsymbol{\alpha}\Omega'(\boldsymbol{\alpha}^T K \boldsymbol{\alpha}) \in 2(K\boldsymbol{\alpha} - Y). \quad (11)$$

Obviously, $\boldsymbol{\alpha} = \boldsymbol{\alpha}^*$ satisfies Equation (11). By the assumption of Ω made in the paper and the strong convexity of square loss function as well as the Hilbert norm, we have that the optimization problem (2) in the paper is strongly convex. Therefore, $\boldsymbol{\alpha}^*$ is the only solution to this problem, which proves the lemma. □

A.3. Proof of Lemma 4

We denote the sub-gradient $\nabla_u \max(1 - u, 0) = -\mathbf{I}(u)$, where

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

Proof of Lemma 4. The KKT (Karush–Kuhn–Tucker) condition of Equation (2) is

$$2\boldsymbol{\theta}^* \Omega'(\|\boldsymbol{\theta}^*\|^2) \in \sum_{i=1}^{n^*} k(\mathbf{x}_i, \cdot) y_i \mathbf{I}(y_i f(\mathbf{x}_i, \boldsymbol{\theta}^*)). \quad (12)$$

Remember that $\Omega_* = 2\Omega'(\|\theta^*\|^2)$, the teaching dimension is the minimum of n^* such that

$$\boldsymbol{\theta}^* = \sum_{i=1}^{n^*} \frac{y_i}{\Omega_*} \mathbf{I}(y_i g(\mathbf{x}_i, \boldsymbol{\theta}^*)) k(\mathbf{x}_i, \cdot). \quad (13)$$

Remember $\boldsymbol{\theta}^*$ can be expressed as

$$\boldsymbol{\theta^*} = \sum_{i=1}^m \alpha_i k(\mathbf{x}_i^*, \cdot),$$

we can construct the teaching set using $\{\mathbf{x}_i\}_{i=1}^m$.

If $\alpha_i \Omega_* g(\mathbf{x}_i, \boldsymbol{\theta}^*) \leq 1$, let $y_i = \alpha_i \Omega_*$ and $\mathbf{x}_i = \mathbf{x}_i^*$, we have

$$\frac{y_i}{\Omega_*} \mathbf{I}(y_i g(\mathbf{x}_i, \boldsymbol{\theta}^*)) k(\mathbf{x}_i, \cdot) = \alpha_i k(\mathbf{x}_i^*, \cdot).$$

If $\alpha_i \Omega_* g(\mathbf{x}_i, \boldsymbol{\theta}^*) > 1$, then the construction of teaching set can be expressed as the following optimization problem

$$\begin{aligned} &\min n_i \\ \text{s.t.} &\sum_{i=1}^{n} y_i = \alpha_i \Omega_* \\ &y_i g(\mathbf{x}_i, \boldsymbol{\theta}^*) \leq 1 \quad \forall i \in [n_i] \,. \end{aligned}$$

The solution to the above optimization problem is $n_i = \lceil \max(1, \alpha_i g(\mathbf{x}_i^*, \boldsymbol{\theta}^*) \Omega_*) \rceil$ and the corresponding teaching items can be $\mathbf{x}_{ij} = \mathbf{x}_i^*,\ y_{ij} = \alpha_i \Omega_* / n$, for $j \in [n_i]$.

According to the assumption of Ω made in the main paper and the convexity of hinge loss and strong convexity of Hilbert norm, the uniqueness of $\boldsymbol{\theta}^*$ is guaranteed.

Then we have that

$$\mathbf{x}_{ij} = \mathbf{x}_i^*,\ y_{ij} = \alpha_i \Omega_* / n_i,\ \ i \in [m],\ j \in [n_i] \,,$$

is the teaching set. □

A.4. Proof of Theorem 2

Proof of Theorem 2. Note that

$$\begin{aligned} \alpha_i^* g(\mathbf{x}_i^*, \boldsymbol{\theta}^*) &= \alpha_i^* \sqrt{k(\mathbf{x}_i^*, \mathbf{x}_i^*)} \left\langle \frac{k(\mathbf{x}_i^*, \cdot)}{\sqrt{k(\mathbf{x}_i^*, \mathbf{x}_i^*)}}, \boldsymbol{\theta}^* \right\rangle \\ &\leq \alpha_i^* \sqrt{k(\mathbf{x}_i, \mathbf{x}_i)} \left\langle \frac{\boldsymbol{\theta}^*}{\|\boldsymbol{\theta}^*\|_H}, \boldsymbol{\theta}^* \right\rangle \\ &= \alpha_i^* \sqrt{k(\mathbf{x}_i^*, \mathbf{x}_i^*)} \, \|\boldsymbol{\theta}^*\|_H \ , \end{aligned}$$

then $\Omega'(x) \leq \min_{i \in I^*} 1/(2\alpha_i^* \sqrt{k(\mathbf{x}_i^*, \mathbf{x}_i^*)} \, \|\boldsymbol{\theta}^*\|_H)$ implies $\Omega'(x) \leq 1/(2\alpha_i^* g(\mathbf{x}_i^*, \boldsymbol{\theta}^*))$. Combine the conclusion of Lemma 5, we have that the teaching dimension is m. □

B. Proofs in Section 3.2

B.1. Proof of Theorem 3

Proof of Theorem 3. Suppose $\exists \boldsymbol{\theta} \in H_{d_H + r} - H_{d_H}$, where $r \in \mathbb{Z}_+$, then we have that

$$\boldsymbol{\theta} = \sum_{i=1}^{d_H + r} \alpha_i k(\mathbf{x}_i, \cdot),$$

where at least $d_H + 1$ elements in $\{k(\mathbf{x}_i, \cdot)\}_{i=1}^{d_H + r}$ are linear independent, or $\boldsymbol{\theta}$ can be expressed with no more than d_H components, which implies $\boldsymbol{\theta} \in H_{d_H}$, a contradiction.

However, the linear independence of $d_H + 1$ elements is contradict to $\dim(H) = d_H$. Then $H_{d_H + r} - H_{d_H} = \emptyset$. Because the conclusion holds for all $r \in \mathbb{Z}_+$, and H is the closure of H_{pre}, we have that $H_{d_H} = H$. Thus, $H_m = H$ with $m \leq d_H$. □

B.2. Proof of Theorem 4

Proof of Theorem 4. We prove the result by contradiction. Assume that $D^* = \{(\mathbf{x}_1, y_1), \cdots, (\mathbf{x}_n, y_n)\}$ is a finite teaching set for a target hypothesis $\boldsymbol{\theta}^*$. Let

$$H_n = \left\{ \sum_{i=1}^{n} \alpha_i k(\mathbf{x}_i, \cdot) \colon \alpha_i \in \mathbb{R} \right\}.$$

Denote $H_n^\perp$ the orthogonal subspace of H_n, we have $H_n \oplus H_n^\perp = H$, where $\oplus$ denotes direct sum. By the definition of H_n, we obtain $\dim(H_n) \leq n$. Since $\dim(H) = \infty > \dim(H_n)$, we have $H_n^\perp \neq \emptyset$. Then, there exists $\boldsymbol{d} \in H_n^\perp$, such that $\boldsymbol{d} \neq \mathbf{0}$ and $\boldsymbol{d} \perp H_n$. For all $\lambda \in \mathbb{R}$, since $k(\mathbf{x}_i, \cdot) \in H_n$, we have that

$$\begin{aligned} &\sum_{i=1}^{n} \ell(\langle \boldsymbol{\theta}^*, k(\mathbf{x}_i, \cdot) \rangle, y_i) \\ &= \sum_{i=1}^{n} \ell(\langle \boldsymbol{\theta}^* + \lambda \boldsymbol{d}, k(\mathbf{x}_i, \cdot) \rangle, y_i). \end{aligned} \tag{14}$$

By the definition of norm and inner product, we have that

$$||\boldsymbol{\theta}^* + \lambda \boldsymbol{d}||_H^2 = ||\boldsymbol{\theta}^*||_H^2 + \lambda^2 ||\boldsymbol{d}||_H^2 + 2\lambda \langle \boldsymbol{\theta}^*, \boldsymbol{d} \rangle \,.$$

The problem can be divided into two cases.

Case 1: $\langle \boldsymbol{\theta}^*, \boldsymbol{d} \rangle \neq 0$. In this case, let $\lambda = -\frac{2\langle \boldsymbol{\theta}^*, \boldsymbol{d} \rangle}{||\boldsymbol{d}||_H^2} \neq 0$, then $||\boldsymbol{\theta}^*||_H^2 = ||\boldsymbol{\theta}^* + \lambda \boldsymbol{d}||_H^2$, but $\boldsymbol{\theta}^* \neq \boldsymbol{\theta}^* + \lambda \boldsymbol{d}$. Combined this result with Equation (14), we have that $\boldsymbol{\theta}^* + \lambda \boldsymbol{d}$ is also a solution to Equation (2) in the paper, thus a contradiction.

Case 2: $\langle \boldsymbol{\theta}^*, \boldsymbol{d} \rangle = 0$. Because of the uniqueness of $\boldsymbol{\theta}^*$, we have that

$$\Omega(||\boldsymbol{\theta}^*||_H^2) < \Omega(||\boldsymbol{\theta}^* + \lambda \boldsymbol{d}||_H^2)$$

holds for all $\lambda \in \mathbb{R}$. This is equivalent to

$$\Omega(x) < \Omega(x'),$$

for $x = ||\boldsymbol{\theta}^*||_H^2$ and all $x' > x$. Since $\boldsymbol{\theta}^*$ can be arbitrary chosen in H, x ranges from 0 to infinity. Therefore, Ω is a monotonically increasing for $x \geq 0$. By Lemma 1 in the paper, the hypothesis has the following form

$$\boldsymbol{\theta}^* = \sum_{i=1}^{n} \alpha_i k(\mathbf{x}_i, \cdot).$$

Let $r(\boldsymbol{\theta})$ be the infimum number of data points such that $\boldsymbol{\theta} \in H$ can be expressed as the linear combination of the functions mapped from the data points by the canonical feature map. The assumption that the teaching dimension is finite implies that $\sup_{\boldsymbol{\theta} \in H} r(\boldsymbol{\theta}) = N < \infty$. Let

$$H_N = \left\{ \sum_{i=1}^{N} \alpha_i k(\mathbf{x}_i, \cdot) \colon \alpha_i \in \mathbb{R}, \mathbf{x}_i \in \mathcal{X} \right\},$$

then $H_N = H$, hence a contradiction to the non-existence of finite m_0. □

B.3. Proof of Remark of Theorem 4

We at first recall the expression of Gaussian, exponential and Laplacian kernels,

$$\text{Gaussian kernel: } e^{-\frac{||\mathbf{x}_1-\mathbf{x}_2||^2}{2\sigma^2}},$$
$$\text{exponential kernel: } e^{-\frac{||\mathbf{x}_1-\mathbf{x}_2||}{2\sigma^2}},$$
$$\text{Laplacian kernel: } e^{-\frac{||\mathbf{x}_1-\mathbf{x}_2||}{\sigma}}.$$

Proposition 1. *The RKHS induced by Gaussian kernel, exponential kernel and Laplacian kernel cannot be represented by H_m with $m \leq \infty$.*

Proof of Proposition 1. Let H, H' and H'' be the RKHS induced by Gaussian kernel, exponential kernel and Laplacian kernel, respectively. It suffices for us to show for the case $d = 1$, i.e., $\dim(\mathcal{X}) = 1$. Without loss of generality, we assume the parameters for Gaussian kernel, exponential kernel and Laplacian kernel are $\frac{\sqrt{2}}{2}$, $\frac{\sqrt{2}}{2}$ and 1.

The following function belongs to H by its definition,

$$f(x) = \sum_{n=-\infty}^{+\infty} e^{-\|x-n\|^2}, \quad n \in \mathbb{Z},$$

then $f(p) = f(q) > 0$, if p and $q \in \mathbb{Z}$. It implies that $\lim_{x\to\infty} f(x) \neq 0$.

If $H = H_m$ for some finite m, we have

$$f(x) = \sum_{i=1}^{m} \alpha_i e^{-\|x-x_i\|^2}.$$

However,

$$\lim_{x\to\infty} \sum_{i=1}^{m} \alpha_i e^{-\|x-x_i\|^2} = 0,$$

a contradiction.

For exponential kernel and Laplacian kernel, consider the following function,

$$g(x) = \sum_{n=-\infty}^{+\infty} e^{-\|x-n\|}, \quad n \in \mathbb{Z}.$$

It is easy to see that $g(x) \in H'$ and $g(x) \in H''$. And then the proof is same as the case for Gaussian kernel. □

B.4. Proof of Lemma 6

In order to calculate $\Lambda(\boldsymbol{\theta}_m^*)$, we introduce $\boldsymbol{\theta}_N \in H$, such that $\boldsymbol{\theta}_N$ is the optimal solution to

$$\mathcal{L}(\boldsymbol{\theta}) = \min_{\boldsymbol{\theta}\in H} \frac{1}{N} \sum_{i=1}^{N} \ell(\langle k(\mathbf{x}_i, \cdot), \boldsymbol{\theta}\rangle, y_i) + \Omega(\|\boldsymbol{\theta}\|_H^2), \quad (15)$$

where $\{(\mathbf{x}_i, y_i)_{i=1}^N\}$ is i.i.d. and sampled from the joint distribution $\mathcal{P}(\mathbf{x}, y)$. We next bound the generalization performance of $\boldsymbol{\theta}_m$ via the generalization performance of $\boldsymbol{\theta}_N$.

Following the work of (Yang et al., 2012), we now exploit the local Rademacher complexity. We define $\psi(\delta) = (\frac{2}{N}\sum_{i=1}^{N} \min(\delta^2, \lambda_i))^{1/2}$. Let $\widetilde{\varepsilon}$ denote the solution to $\delta^2 = \psi(\delta)$, where the existence and uniqueness of $\widetilde{\varepsilon}$ are determined by the sub-root property of $\psi(\delta)$. Denote $\gamma = \max(\widetilde{\varepsilon}, \sqrt{6\ln N/N})$. According to (Koltchinskii, 2011), we have that $\gamma^2 \in O(N^{-1/2})$, and when the eigenvalues of kernel function follow the p-power law, it can be improved to $\gamma^2 \in O(N^{-p/(p+1)})$. The following Lemma 10 bounds $\Lambda(\boldsymbol{\theta}_m^*)$ by $\Lambda(\boldsymbol{\theta}_N^*)$.

Lemma 10. *For $M_1 = \Omega^{-1}(\Omega(0) + 1) \leq e^{2N}/4\gamma^2$ with $\gamma \leq 1$, $\sup_{\|\boldsymbol{\theta}\|_H \leq M_1} 2\|\boldsymbol{\theta}\|_H \Omega'(\|\boldsymbol{\theta}\|_H^2) \leq M_2$, and $\lambda_{m+1} \in O(N/\sqrt{m})$, if $\Omega(x^2)$ is μ-strongly convex, then with a probability $1 - 2N^{-3}$,*

$$\Lambda(\boldsymbol{\theta}_m^*) \leq \Lambda(\boldsymbol{\theta}_N^*) + O\left(\gamma + \frac{C(k)+1}{\sqrt{m}} + e^{-N}\right).$$

Note that in our setting N can be arbitrarily chosen, we can let $N \to \infty$. In this way, we have $\Lambda(\boldsymbol{\theta}_N^*) \to 0$, $\gamma \to 0$ and $e^{-N} \to 0$, and complete the proof of Lemma 6.

We denote $\frac{1}{N}\sum_{i=1}^{N} \ell(\langle k(\mathbf{x}_i, \cdot), \boldsymbol{\theta}\rangle, y_i) + \Omega(\|\boldsymbol{\theta}\|_H^2)$ as $\widetilde{\mathcal{L}}(\boldsymbol{\theta})$. Let K be the Gram matrix of $\{\mathbf{x}_i\}_{i=1}^N$ under kernel k, i.e., $K = [k(\mathbf{x}_i, \mathbf{x}_j)]_{N\times N}$. Then we sample m construct a low rank matrix with the teaching set $\{\mathbf{x}_i^*, y_i^*\}_{i=1}^m$ as $\hat{K}_m = K_b K_m^\dagger K_b^T$, where $K_b = [k(\mathbf{x}_i^*, \mathbf{x}_j)]_{N\times m}$, $K_m = [k(\mathbf{x}_i^*, \mathbf{x}_j^*)]_{m\times m}$, and $K_m^\dagger$ is the pseudo inverse of $K_m^\dagger$. We first introduce Lemma 11, 12, 13, and 14, which are necessary to prove Lemma 10.

We now give a recap of the concept of Fenchel conjugate. Let $f(x)$ be a function of x, its Fenchel conjugate $f^*(\alpha)$ is defined as

$$f^*(\alpha) = \sup_z (\alpha z - f(z)).$$

Suppose f is convex and differentiable over $\mathbb{R}^n$. Any maximizer z^* of $\alpha z - f(z)$ satisfies $\alpha = \nabla_z f(z^*)$. It implies α falls in the range of the mapping $\nabla_z f(z)\colon \mathbb{R} \to \mathbb{R}$.

Lemma 11. *If f is closed and strong convex with parameter μ, then f^* has a Lipschitz continuous gradient with parameter $\frac{1}{\mu}$.*

Proof. By implication of strong convexity, we have

$$\|s_x - s_y\| \geq \mu \|x - y\| \quad \forall s_x \in \partial f(x), s_y \in \partial f(y),$$

which implies

$$\|s_x - s_y\| \geq \mu \|\nabla f^*(s_x) - \nabla f^*(s_y)\|.$$

Hence, f^* has a Lipschitz continuous gradient with constant value $\frac{1}{\mu}$. □

Lemma 12. *Under the assumption of Ω made in the paper, we have that*

$$0 \leq \widetilde{\mathcal{L}}(\boldsymbol{\theta}_m^*) - \widetilde{\mathcal{L}}(\boldsymbol{\theta}_N^*) \leq \frac{C_1 C_2 \mu}{N} \left\| K - \hat{K}_m \right\|_2,$$

where C_1 is the upper bound of $\nabla_x \ell(x, y)$, and C_2 is the upper bound of $|y|$ with $y \in \mathcal{Y}$. $\|\cdot\|_2$ stands for the spectral norm of a matrix. $\boldsymbol{\theta}_N^*$ is the solution that minimizes $\widetilde{\mathcal{L}}(\boldsymbol{\theta})$.

Proof. Let $\ell^*(\alpha)$ and Ω^* denote the Fenchel conjugate of $\ell(z, y)$ and $\Omega(z^2)$ in terms of z, respectively, i.e.,

$$\ell^*(\alpha) = \sup_z(\alpha z - \ell(z, y)),$$

$$\Omega^*(\alpha) = \sup_z(\alpha z - \Omega(z^2)).$$

By the property of conjugate function, the derivative of Ω^* at 0 is $\arg\min_x \Omega(x^2) = 0$.

According to Lemma 11, $(\Omega^*)'$ has a Lipschitz constant $\frac{1}{\mu}$, thus

$$(\Omega^*)'(x) \leq (\Omega^*)'(0) + \frac{1}{\mu} x = \frac{x}{\mu}. \tag{16}$$

Using the conjugates, we are able to rewrite $\widetilde{\mathcal{L}}(\boldsymbol{\theta}_N^*)$ as an equivalent form

$$\widetilde{\mathcal{L}}(\boldsymbol{\theta}_N^*) = \max_{\boldsymbol{\alpha}} -\frac{1}{N} \sum_{i=1}^N \ell^*(\alpha_i) - \Omega^* \left(\sqrt{\frac{1}{N^2} (\boldsymbol{\alpha} \circ \mathbf{y})^T K (\boldsymbol{\alpha} \circ \mathbf{y})} \right),$$

where $\boldsymbol{\alpha} = (\alpha_1, \cdots, \alpha_N)^T$ and $\circ$ denotes the element-wise dot product. Similarly, we can rewrite $\widetilde{\mathcal{L}}(\boldsymbol{\theta}_m^*)$ as

$$\widetilde{\mathcal{L}}(\boldsymbol{\theta}_m^*) = \max_{\boldsymbol{\alpha}} -\frac{1}{N} \sum_{i=1}^N \ell^*(\alpha_i) - \Omega^* \left(\sqrt{\frac{1}{N^2} (\boldsymbol{\alpha} \circ \mathbf{y})^T \hat{K}_m (\boldsymbol{\alpha} \circ \mathbf{y})} \right).$$

Then, we have that

$$\begin{aligned}
\widetilde{\mathcal{L}}(\boldsymbol{\theta}_m^*) = \max_{\boldsymbol{\alpha}} & -\frac{1}{N} \sum_{i=1}^N \ell^*(\alpha_i) \\
& - \Omega^* \left(\sqrt{\frac{1}{N^2} (\boldsymbol{\alpha} \circ \mathbf{y})^T K (\boldsymbol{\alpha} \circ \mathbf{y})} \right) \\
& + \Omega^* \left(\sqrt{\frac{1}{N^2} (\boldsymbol{\alpha} \circ \mathbf{y})^T K (\boldsymbol{\alpha} \circ \mathbf{y})} \right) \\
& - \Omega^* \left(\sqrt{\frac{1}{N^2} (\boldsymbol{\alpha} \circ \mathbf{y})^T \hat{K}_m (\boldsymbol{\alpha} \circ \mathbf{y})} \right),
\end{aligned}$$

and then

$$\begin{aligned}
\widetilde{\mathcal{L}}(\boldsymbol{\theta}_m^*) \leq \max_{\boldsymbol{\alpha}} & \left(-\frac{1}{N} \sum_{i=1}^N \ell^*(\alpha_i) \right. \\
& - \Omega^* \left(\sqrt{\frac{1}{N^2} (\boldsymbol{\alpha} \circ \mathbf{y})^T K (\boldsymbol{\alpha} \circ \mathbf{y})} \right) \\
& + \max_{\boldsymbol{\alpha}} \left(\Omega^* \left(\sqrt{\frac{1}{N^2} (\boldsymbol{\alpha} \circ \mathbf{y})^T K (\boldsymbol{\alpha} \circ \mathbf{y})} \right) \right) \\
& \left. - \Omega^* \left(\sqrt{\frac{1}{N^2} (\boldsymbol{\alpha} \circ \mathbf{y})^T \hat{K}_m (\boldsymbol{\alpha} \circ \mathbf{y})} \right) \right) \\
\overset{(a)}{=} & \widetilde{\mathcal{L}}(\boldsymbol{\theta}_N^*) \\
& + \max_{\boldsymbol{\alpha}} \frac{(\Omega^*)'(\sqrt{x_0})}{2N^2 \sqrt{x_0}} (\boldsymbol{\alpha} \circ \mathbf{y})^T (K - \hat{K}_m)(\boldsymbol{\alpha} \circ \mathbf{y}) \\
\overset{(b)}{\leq} & \widetilde{\mathcal{L}}(\boldsymbol{\theta}_N^*) \\
& + \max_{\boldsymbol{\alpha}} \frac{1}{2\mu N^2} (\boldsymbol{\alpha} \circ \mathbf{y})^T (K - \hat{K}_m)(\boldsymbol{\alpha} \circ \mathbf{y}) \\
\leq & \widetilde{\mathcal{L}}(\boldsymbol{\theta}_N^*) + \max_{\boldsymbol{\alpha}} \frac{C_2}{2\mu N^2} \|\boldsymbol{\alpha}\|^2 \left\| K - \hat{K}_m \right\|_2 \\
\overset{(c)}{\leq} & \widetilde{\mathcal{L}}(\boldsymbol{\theta}_N^*) + \frac{C_1 C_2}{2\mu N} \left\| K - \hat{K}_m \right\|_2,
\end{aligned}$$

where (a) uses Lagrange's mean value theorem, (b) is derived from Equation (16), and (c) follows $|\alpha_i| \leq C_1$ duo to $\nabla_x \ell(x, y) \leq C_1$. □

Lemma 13. *For $M_1 = \Omega^{-1}(\Omega(0) + 1) \leq e^{2N}/(4\gamma^2)$ with $\gamma \leq 1$ and $\sup_{\|\boldsymbol{\theta}\|_H^2 \leq M_1} 2 \|\boldsymbol{\theta}\|_H \Omega'(\|\boldsymbol{\theta}\|_H^2) \leq M_2$, with a probability $1 - 2N^{-3}$, we have that*

$$\Lambda(\boldsymbol{\theta}_m^*) \leq \Lambda(\boldsymbol{\theta}_N^*) + C_3 \left(\gamma + \frac{\left\| K - \widehat{K}_m \right\|_2}{N} + e^{-N} \right),$$

where C_3 is a constant.

Proof. Define the loss function

$$\bar{\ell}(\langle \boldsymbol{\theta}, k(\mathbf{x}, \cdot\rangle, y) = \ell(\langle \boldsymbol{\theta}, k(\mathbf{x}, \cdot\rangle, y) + \Omega(\|\boldsymbol{\theta}\|_H^2).$$

To simplify our notations, we define P_N and P as

$$P_N(\bar{\ell} \circ \boldsymbol{\theta}) = \frac{1}{N} \sum_{i=1}^N \bar{\ell}(\langle \boldsymbol{\theta}, k(\mathbf{x}, \cdot\rangle, y_i) = \widetilde{\mathcal{L}}(\boldsymbol{\theta}),$$

$$P(\bar{\ell} \circ \boldsymbol{\theta}) = \mathbb{E}[\bar{\ell}(\langle \boldsymbol{\theta}, k(\mathbf{x}, \cdot\rangle, y)] = F(\boldsymbol{\theta}).$$

Using those notations, we have that

$$\begin{aligned}
&\Lambda\left(\boldsymbol{\theta}_m^*\right)-\Lambda\left(\boldsymbol{\theta}_N^*\right)\\
&=P\left(\bar{\ell}\circ\boldsymbol{\theta}_m^*-\bar{\ell}\circ\boldsymbol{\theta}_N^*\right)\\
&=P_N\left(\bar{\ell}\circ\boldsymbol{\theta}_m^*-\bar{\ell}\circ\boldsymbol{\theta}_N^*\right)+\left(P-P_N\right)\left(\bar{\ell}\circ\boldsymbol{\theta}_m^*-\bar{\ell}\circ\boldsymbol{\theta}_N^*\right)\\
&\leq P_N\left(\bar{\ell}\circ\boldsymbol{\theta}_m^*-\bar{\ell}\circ\boldsymbol{\theta}_N^*\right)\\
&\qquad+\max_{(\boldsymbol{\theta},\boldsymbol{\theta}')\in\mathcal{G}}\left(P-P_N\right)\left(\bar{\ell}\circ\boldsymbol{\theta}-\bar{\ell}\circ\boldsymbol{\theta}'\right).
\end{aligned}$$

In the above formula, $\mathcal{G}$ is defined as

$$\mathcal{G}=\left\{(\boldsymbol{\theta},\boldsymbol{\theta}'):\ \|\boldsymbol{\theta}-\boldsymbol{\theta}'\|_{\ell_2}\leq r,\|\boldsymbol{\theta}-\boldsymbol{\theta}'\|_H\leq R\right\},$$

where $\|\boldsymbol{\theta}\|_{\ell_2}=\sqrt{\mathbb{E}[\boldsymbol{\theta}^2]}\leq\|\boldsymbol{\theta}\|_H$, and r as well as R are given by $r=R=\|\boldsymbol{\theta}_m^*-\boldsymbol{\theta}_N^*\|_H\leq 2M_1^{1/2}$. Using Lemma 9 from (Koltchinskii & Yuan, 2010), we have that, with probability $1-2N^{-3}$, for any $\gamma r\leq e^N$, $\gamma^2R\leq e^N$,

$$\begin{aligned}
&\sup_{(\boldsymbol{\theta},\boldsymbol{\theta}')\in\mathcal{G}}\left(P-P_N\right)\left(\bar{\ell}\circ\boldsymbol{\theta}'-\bar{\ell}\circ\boldsymbol{\theta}'\right)\\
&\leq C_4L\left(r\gamma+R\gamma^2+e^{-N}\right)\\
&\leq C_5L\left(r\gamma+e^{-N}\right),
\end{aligned}$$

where C_4, C_5 are constants, and L is the upper bound of the gradient of $\bar{\ell}$ for functions in $\mathcal{G}$ and is given by $L\leq\sup_{\|\boldsymbol{\theta}\|_H^2\leq M_1^{1/2}}2\|\boldsymbol{\theta}\|_H\Omega'(\|\boldsymbol{\theta}\|_H^2)+C_1\leq M_2+C_1$. Since $\max(\|\boldsymbol{\theta}_N^*\|_H,\|\boldsymbol{\theta}_m^*\|_H)\leq M_1$ and $M_1\leq e^{2N}/(4\gamma^2)$, we have the condition $\gamma r\leq e^N$ satisfied. Therefore, with a probability $1-2N^{-3}$, we have that

$$\begin{aligned}
&P\left(\bar{\ell}\circ\boldsymbol{\theta}_m^*-\bar{\ell}\circ\boldsymbol{\theta}_N^*\right)\\
&\leq P_N\left(\bar{\ell}\circ\boldsymbol{\theta}_m^*-\bar{\ell}\circ\boldsymbol{\theta}_N^*\right)+C_5(M_2+C_1)\left(r\gamma+e^{-N}\right)\\
&\leq\frac{C\left\|K-\widehat{K}_m\right\|_2}{2\mu N}+C_5(M_2+C_1)\left(r\gamma+e^{-N}\right),
\end{aligned}$$

where we use the result in Lemma 12. By the fact of $\Lambda(\boldsymbol{\theta})=P(\bar{\ell}\circ\boldsymbol{\theta}-\bar{\ell}\circ\boldsymbol{\theta}^*)$, we have that

$$\begin{aligned}
\Lambda\left(\boldsymbol{\theta}_m^*\right)\leq\Lambda\left(\boldsymbol{\theta}_N^*\right)+\frac{C\left\|K-\widehat{K}_m\right\|_2}{2\mu N}\\
+C_5(M_2+C_1)\left(2M_1^{1/2}\gamma+e^{-N}\right).
\end{aligned}$$

We complete the proof by absorbing the constant terms into a constant C_3. □

Now we only need to bound $\|K-\widehat{K}_m\|_2$. To this end, we apply the conclusion from (Drineas & Mahoney, 2005), and we state it in the following lemma.

Lemma 14. *Suppose G is an $N\times N$ symmetric positive semi-definite (SPSD) matrix, let $\widehat{G}_m$ be constructed by sampling m columns of G with a given probability. In addition, let G_m be the best rank-m approximation to G. With $\eta=1+\sqrt{8\log(1/\delta)}$, we have that, with probability at least $1-\delta$,*

$$\left\|G-\widehat{G}_m\right\|_2\leq\|G-G_m\|_2+\frac{2\eta}{\sqrt{m}}\sum_{i=1}^N G_{ii}^2,$$

where $\|\cdot\|_2$ is the spectral norm of matrix.

With the property that Gram matrix K is symmetric and positive semi-definite, we can apply Lemma 14 to K and $\widehat{K}_m$. Notably, in machine teaching setting, $\widehat{K}_m$ is constructed optimally, while in the setting of Lemma 14, $\widehat{G}_m$ is constructed by sampling. Therefore, we can achieve this inequality if $\delta>0$. Then, we have that

$$\left\|K-\widehat{K}_m\right\|_2\leq\left\|K-\bar{K}_m\right\|_2+\frac{2}{\sqrt{m}}\sum_{i=1}^N K_{ii}^2 \qquad (17)$$

for our setting, where $\bar{K}_m$ is the best rank-m approximation to K. Using the property of spectral norm of a matrix, we have $\|K-K_m\|_2=\lambda_{m+1}$. By substituting Inequality (17) into the result of Lemma 13, and letting $2\sup_{\mathbf{x}}k^2(\mathbf{x},\mathbf{x})=C(k)$, we complete the proof of Lemma 10.

C. Proofs in Section 4

C.1. Proof of Lemma 7

Proof of Lemma 7. For $\dim(\mathcal{X})=d$, the dimension of the induced RKHS H is $\dim(H)=d$. By Theorem 3 in the paper, $\boldsymbol{\theta}^*=\sum_{i=1}^d\alpha_ik(\mathbf{x}_i,\cdot)$. Then we have that

$$\boldsymbol{\theta}^*=\sum_{i=1}^d\alpha_i\left\langle\mathbf{x}_i,\cdot\right\rangle\overset{(a)}{=}\left\langle\sum_{i=1}^d\alpha_i\mathbf{x}_i,\cdot\right\rangle\overset{(b)}{=}\left\langle\mathbf{x},\cdot\right\rangle,$$

where (a) holds because of the linearity of inner product, and (b) holds because $\mathcal{X}$ is a linear space and there exists $\mathbf{x}\in\mathcal{X}$ such that $\mathbf{x}=\sum_{i=1}^d\alpha_i\mathbf{x}_i$. Therefore, all hypotheses in H can be expressed by one term, which ends the proof. □

C.2. Proof of Lemma 8

Proof of Lemma 8. According to the definition of $\mathcal{G}^*(d,p)$, we have that $\mathcal{G}(\boldsymbol{\theta},d,p,\mathcal{G}^*(d,p))\neq\{1\}$. By the Hilbert's Nullstellensatz (Hartshorne, 1977), the solution to the polynomial system exists when Gröbner basis is not $\{1\}$. □

C.3. An example for Remark 2 of Lemma 8

Let $\boldsymbol{\theta}^*=\sum_{i=1}^4k(\mathbf{x}_i,\cdot)$, where $\mathbf{x}_1=(1,2)$, $\mathbf{x}_2=(3,4)$, $\mathbf{x}_3=(5,6)$, $\mathbf{x}_4=(7,8)$, i.e., $d=2$. Let $p=3$, then $\binom{d+p-1}{p}=4$. Instantiate (6) with this example and let

$m = 2$, (6) becomes

$$\begin{cases} y_{11}^3 + y_{21}^3 = 496 \\ y_{11}^2 \cdot y_{12} + y_{21}^2 \cdot y_{22} = 580 \\ y_{11} \cdot y_{12}^2 + y_{21} \cdot y_{22}^2 = 680 \\ y_{12}^3 + y_{22}^3 = 800. \end{cases} \tag{18}$$

Then the Gröbner basis is $\mathcal{G}(\boldsymbol{\theta}^*, 2, 3, 2) = \{y_{12} \cdot y_{21}^3 - y_{11} \cdot y_{22} \cdot y_{21}^2, y_{12} \cdot y_{21}^2 \cdot y_{22} - y_{11} \cdot y_{21} \cdot y_{22}^2, y_{12} \cdot y_{21} \cdot y_{22}^2 - y_{11} \cdot y_{22}^3, y_{11}^3 + y_{21}^3, y_{12} \cdot y_{11}^2 + y_{22} \cdot y_{21}^2, y_{11} \cdot y_{12}^2 + y_{21} \cdot y_{22}^2, y_{12}^3 + y_{22}^3\} \neq \{1\}$. Similarly, we have $\mathcal{G}(\boldsymbol{\theta}^*, 2, 3, 1) = \{1\}$. According to the definition of $\mathcal{G}(\boldsymbol{\theta}^*, 2, 3)$, we have $\mathcal{G}(\boldsymbol{\theta}^*, 2, 3) = 2 < 4$.

C.4. Proof of Corollary 3 and 4

Proof of Corollary 3 and 4. According to Theorem 5 in the paper, the ϵ-approximate teaching dimension is $O((C(k)+1)^2/\epsilon^2)$. For the Gaussian kernel, exponential kernel and Laplacian kernel, $C(k) = 2\sup_{\mathbf{x}} k^2(\mathbf{x}, \mathbf{x}) = 2$. Then, the ϵ-TD can be simplified as $O(1/\epsilon^2)$. □

D. For the Unregularized Learners

We derive the teaching set for unregularized learners with square loss and hinge loss in this section. We also mention the assumptions needed by perceptron loss.

D.1. Square Loss

We start by providing teaching set for linear kernel.

Proposition 2. *For the target hypothesis $\boldsymbol{\theta}^*$, $\{\mathbf{x}_i, (\boldsymbol{\theta}^*)^T\mathbf{x}_i\}_{i=1}^d$ is a teaching set, where $\{\mathbf{x}_i\}_{i=1}^d$ are linearly independent.*

Proof. For

$$\mathcal{L}(\boldsymbol{\theta}) = \sum_{i=1}^{d} (\boldsymbol{\theta}^T\mathbf{x}_i - y_i)^2,$$

it is easy to see that $\mathcal{L}(\boldsymbol{\theta}^*) = 0$. Therefore, $\boldsymbol{\theta}^*$ is in the solution set.

If $\mathcal{L}(\boldsymbol{\theta}^* + \boldsymbol{\delta}) = 0$, then

$$\mathcal{L}(\boldsymbol{\theta}) = \sum_{i=1}^{d} \left((\boldsymbol{\theta}^* + \boldsymbol{\delta})^T\mathbf{x}_i - y_i\right)^2 = \sum_{i=1}^{d} (\boldsymbol{\delta}^T\mathbf{x}_i)^2 = 0\,.$$

Since $\{\mathbf{x}_i\}_{i=1}^d$ are linearly independent and $\dim(\boldsymbol{\delta}) = d$, we have $\boldsymbol{\delta} = \mathbf{0}$. This guarantees $\boldsymbol{\theta}^*$ is the unique solution. □

According to (Kumar et al., 2021), the RKHS of polynomial kernel is isomorphic to that of a higher dimensional linear kernel. Therefore, if the following assumption is satisfied, the teaching set for polynomial kernel can be derived the same as linear kernel.

Assumption 1. *For the target hypothesis $\boldsymbol{\theta}^* \in H$, we assume that there exist $r = dim(H)$ linearly independent polynomials of the form $\{\Phi(\mathbf{z}_i)\}_{i=1}^r$, where $\forall i,\ \mathbf{z}_i \in \mathcal{X}$ and Φ is the feature map of polynomial kernel.*

For Gaussian kernel, (Kumar et al., 2021) uses

$$\tilde{k}(\mathbf{x}_1, \mathbf{x}_2) = e^{-\frac{\|\mathbf{x}_1\|_H^2}{2\sigma^2}} e^{-\frac{\|\mathbf{x}_2\|_H^2}{2\sigma^2}} \sum_{t=0}^{s} \frac{1}{t!} \left(\frac{\langle \mathbf{x}_1, \mathbf{x}_2 \rangle}{\sigma^2}\right)^t$$

as an approximation of Gaussian kernel. The approximation error is small when the following assumption holds.

Assumption 2 (Assumption 3.4.2 in Kumar et al. (2021))**.** *For the target hypothesis $\boldsymbol{\theta}^* = \sum_{i=1}^l \alpha_i k(\mathbf{x}_i, \cdot)$, the learner optimizes to a solution $\hat{\boldsymbol{\theta}}$ with bounded coefficients. Alternatively, the sums $\sum_{i=1}^r |\alpha_i|$ and $|\beta| + \sum_{j=2}^r |\gamma_j|$ are bounded where $\hat{\boldsymbol{\theta}} \in H$ has the form $\hat{\boldsymbol{\theta}} = \beta k(\mathbf{x}_1, \cdot) + \sum_{j=2}^r \gamma_j k(\mathbf{x}_j, \cdot)$ and $r = dim(H)$.*

We denote $\tilde{H}$ as the RKHS induced by $\tilde{k}$, $\mathbb{P}\boldsymbol{\theta}^*$ as the projection of $\boldsymbol{\theta}^*$ in $\tilde{H}$. Similar to Assumption 1, we need an assumption for the approximate kernel.

Assumption 3. *For the target hypothesis $\boldsymbol{\theta}^* \in H$, we assume that there exist $r = dim(\tilde{H})$ linearly independent elements in $\tilde{H}$ of the form $\{\Phi(\mathbf{z}_i)\}_{i=1}^r$, where $\forall i,\ \mathbf{z}_i \in \mathcal{X}$ and Φ is the feature map of the approximate Gaussian kernel.*

With the assumption, we can provide teaching set for the approximate kernel as the linear kernel, and the teaching set is also the approximate teaching set for Gaussian kernel.

D.2. Hinge Loss

Similar to the square loss, we start by providing teaching set for linear kernel.

Proposition 3. *For the target hypothesis $\boldsymbol{\theta}^*$, the following is a teaching set*

$$\mathbf{x}_i = \mathbf{v}_i + \frac{\boldsymbol{\theta}^*}{\|\boldsymbol{\theta}^*\|_2},\ \ y_i = 1 \qquad \forall i \in \{1, \cdots, d-1\};$$

$$\mathbf{x}_i = -\mathbf{v}_i + \frac{\boldsymbol{\theta}^*}{\|\boldsymbol{\theta}^*\|_2},\ \ y_i = 1 \qquad \forall i \in \{d, \cdots, 2d-2\};$$

$$\mathbf{x}_{2d-1} = -\frac{\boldsymbol{\theta}^*}{\|\boldsymbol{\theta}^*\|_2},\ \ y_{2d-1} = 1\,,$$

where $\{\mathbf{v}_i\}_{i=1}^d$ is an orthogonal basis for $\mathbb{R}^d$ which extends with $\mathbf{v}_d = \boldsymbol{\theta}^*$.

Proof. For

$$\mathcal{L}(\boldsymbol{\theta}) = \sum_{i=1}^{2d-1} \max(1 - y_i\boldsymbol{\theta}^T\mathbf{x}_i, 0),$$

we can calculate $\mathcal{L}(\boldsymbol{\theta}^*) = 2$.

Suppose $\boldsymbol{\theta}_0 = \boldsymbol{\theta}^* + \boldsymbol{\delta}$ is the optimal solution for $\min \mathcal{L}(\boldsymbol{\theta})$, where $\boldsymbol{\delta}$ can be represented as $\sum_{i=1}^{d-1} t_i \mathbf{v}_i + t_d \boldsymbol{\theta}^*$.

Note that $\max(1 - y_{2d-1}\boldsymbol{\theta}_0^T \mathbf{x}_{2d-1}) = 2 + t_d$, the optimality of $\boldsymbol{\theta}_0$ implies $t_d \leq 0$. However,

$$\sum_{i=1}^{2d-2} \max(1 - y_i \boldsymbol{\theta}_0^T \mathbf{x}_i, 0) \geq -(2d-2)t_d \,,$$

so $t_d = 0$.

Based on the fact $t_d = 0$, we have

$$\sum_{i=1}^{2d-2} \max(1 - y_i \boldsymbol{\theta}_0^T \mathbf{x}_i, 0) \geq \sum_{i=1}^{2d-2} \|t_i \mathbf{v}_i\|_2 \,,$$

this implies $t_i = 0$.

Then we get $\boldsymbol{\delta} = \mathbf{0}$, and $\boldsymbol{\theta}^*$ is the unique solution to $\min \mathcal{L}(\boldsymbol{\theta})$. □

The derivation of polynomial kernel and Gaussian kernel is similar to that of the square loss except for some modifications of Assumption 1 and 3. The modification comes from the appearance of $\boldsymbol{\theta}^*$ in the teaching set. We states the modified assumptions as follows.

Assumption 4. *For the target hypothesis $\boldsymbol{\theta}^* \in H$, let Φ be the feature map of polynomial kernel, we assume that*

1. *There exist $(r-1)$ linearly independent polynomials on the orthogonal subspace of $\boldsymbol{\theta}^*$ in H of the form $\{\Phi(\mathbf{z}_i)\}_{i=1}^{r-1}$, where $\forall i,\ \mathbf{z}_i \in \mathcal{X}$ and $r = dim(H)$.*
2. *There exists polynomial such that $\boldsymbol{\theta}^* = \Phi(\mathbf{z})$, where $\mathbf{z} \in \mathcal{X}$.*

Assumption 5. *For the target hypothesis $\boldsymbol{\theta}^* \in H$, let Φ be the feature map of approximate Gaussian kernel, we assume that*

1. *There exist $(r-1)$ linearly independent elements on the orthogonal subspace of $\boldsymbol{\theta}^*$ in H of the form $\{\Phi(\mathbf{z}_i)\}_{i=1}^{r-1}$, where $\forall i,\ \mathbf{z}_i \in \mathcal{X}$ and $r = dim(\tilde{H})$.*
2. *There exists polynomial such that $\boldsymbol{\theta}^* = \Phi(\mathbf{z})$, where $\mathbf{z} \in \mathcal{X}$.*

D.3. Perceptron Loss

Because we only focus on square loss and hinge loss in this paper, the teaching set for perceptron loss is omitted and can be found in (Kumar et al., 2021). In this section, we provide the assumptions needed by perceptron loss for the purpose of self-contained. For polynomial kernel, we need one assumption.

Assumption 6 (Assumption 3.2.1 in (Kumar et al., 2021))**.** *For the target hypothesis $\boldsymbol{\theta}^* \in H$, let Φ be the feature map of polynomial kernel, we assume that there exist $(r-1)$ linearly independent polynomials on the orthogonal subspace of $\boldsymbol{\theta}^*$ in H of the form $\{\Phi(\mathbf{z}_i)\}_{i=1}^{r-1}$, where $\forall i,\ \mathbf{z}_i \in \mathcal{X}$ and $r = dim(H)$.*

For Gaussian kernel, Assumption 2 and the following assumption are needed.

Assumption 7 (Assumption 3.4.1 in (Kumar et al., 2021))**.** *For the target hypothesis $\boldsymbol{\theta}^* \in H$, let Φ be the feature map of approximate Gaussian kernel, we assume that there exists $(r-1)$ linearly independent elements such that on the orthogonal subspace of $\mathbb{P}\boldsymbol{\theta}^*$ in $\tilde{H}$ of the form $\{\Phi(\mathbf{z}_i)\}_{i=1}^{r-1}$, where $\forall i,\ \mathbf{z}_i \in \mathcal{X}$ and $r = dim(\tilde{H})$.*

E. Experiment Details

This section introduces the selected datasets. Sin: It is generated by sin operator with $(\mu = 0, \sigma = 0.15)$ Gaussian noise in $[0, 5]$. It has 150 samples and the dimension of input space is 1. Make-regression: An optionally-sparse random linear combination of random features with noise. It has 250 samples and the dimension of input space is 2. MPG: A dataset taken from the StatLib library maintained at Carnegie Mellon University, concerns city-cycle fuel consumption, is to be predicted in terms of both multivalued discrete and continuous attributes. It has 392 samples and the dimension of input space is 7. Eunite[1]: The Eunite 2001 competition dataset. Given load and some other information in previous years, the task is to predict daily maximum load in the next January. It has 336 samples and the dimension of input space is 16. Two-moon: Two interleaving half circles. It has 250 samples and the dimension of input space is 2. Two-circle: A large circle containing a smaller circle in 2D. It has 250 samples and the dimension of input space is 2. Adult: High dimensional binary classification problem with both continuous and discrete features. It has 1605 samples and the dimension of input space is 123. Sonar: Discriminate between sonar signals bounced off a metal cylinder and those bounced off a roughly cylindrical rock, the data is derived from 111 patterns, each pattern consists of 60 numbers representing the energy within a particular frequency band. It has 208 samples and the dimension of input space is 60.

F. Numerical Results for Exact Teaching

To illustrate exact teaching, we provide numerical examples for teaching the linear kernel learner and the polynomial kernel learner.

[1] https://www.csie.ntu.edu.tw/~cjlin/libsvmtools/datasets/regression.html#eunite2001

Linear Kernel. For the target hypothesis $\boldsymbol{\theta}^* = (-1, 1, 0) \in \mathbb{R}^3$, we consider the regularization function $\Omega(x^2) = \frac{1}{2}x^2$. According to Lemma 3 and 4, the teaching set for square loss is $\{(-1, 1, 0), 3/2\}$, for hinge loss is $\{(-1, 1, 0), 1\}$.

However, for unregularized learner, the teaching set can be $\{\{(1, 0, 0), -1\}, \{(0, 1, 0), 1\}, \{(0, 0, 1), 0\}\}$ for square loss. For hinge loss, we first obtain an orthogonal basis $\{(1/2, 1/2, 0), (0, 0, 1)\}$ for the subspace orthogonal to $\boldsymbol{\theta}^*$, and the teaching set is

$$\{(\frac{1+\sqrt{2}}{2}, \frac{1-\sqrt{2}}{2}, 0), 1\}, \{(\frac{\sqrt{2}}{2}, \frac{\sqrt{2}}{2}, 1), 1\},$$
$$\{(\frac{-1+\sqrt{2}}{2}, \frac{-1-\sqrt{2}}{2}, 0), 1\}, \{(\frac{\sqrt{2}}{2}, \frac{\sqrt{2}}{2}, -1), 1\},$$
$$\{(-\frac{\sqrt{2}}{2}, -\frac{\sqrt{2}}{2}, 0), 1\}.$$

Polynomial Kernel. For the target hypothesis $\boldsymbol{\theta}^* = \sum_{i=1}^{4} k(\mathbf{x}_i, \cdot)$, where $\mathbf{x}_1 = (1, 2)$, $\mathbf{x}_2 = (3, 4)$, $\mathbf{x}_3 = (5, 6)$, $\mathbf{x}_4 = (7, 8)$. Let the degree of the polynomial be 3, i.e., $p = 3$. We consider the regularization function $\Omega(x^2) = \frac{1}{10^7}x^2$. By solving the polynomial system (6), we have $\boldsymbol{\theta}^* = k((2.22, 3.48), \cdot) + k((7.86, 9.12), \cdot)$. Then we obtain the teaching set $\{\{(2.22, 3.18), 123946.37\}, \{(7.86, 9.12), 3164724.12\}\}$ for square loss. Note that for classification problem, $t\boldsymbol{\theta}^*$ is equivalent to $\boldsymbol{\theta}^*$ if t is a positive constant. We can construct the teaching set for hinge loss as $\{\{(2.22, 3.18), 1\}, \{(7.86, 9.12), 1\}\}$.

For the unregularized learner, we first consider the teaching set for square loss. Note that $(1, 0)$, $(0, 1)$, $(1, 2)$, $(2, 1) \in \mathcal{X}$, and the functions induced by mapping the four elements to RKHS with canonical feature map are linearly independent, we can construct the teaching set as

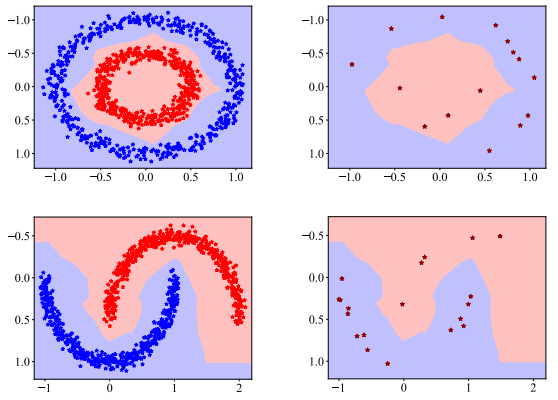
$$\{(1, 0), 496\}, \{(0, 1), 800\},$$
$$\{(1, 2), 18536\}, \{(2, 1), 15808\}.$$

For hinge loss, the teaching set can not be constructed because Assumption 4 is not satisfied.

G. Experiments on Laplacian and Exponential and Kernels

In this section, we perform experiments on Laplacian kernel and exponential kernel. The parameter σ of Laplacian kernel is set to be d, and the σ of exponential kernel is set to be 0.9, where d is the dimension of input space.

Figure 5 and Figure 6 visualize the teaching sets for Laplacian kernel learner and exponential kernel learner with hinge loss function respectively. Similar to Gaussian kernel, the top sub-figures are the results on the Circle dataset and the bottom sub-figures are the results on the Moon dataset. The interface between the blue and red areas is the decision boundary of teacher in the sub-figure (a), and learner in the sub-figure (b). The positive and negative points in dataset are marked by red and blue dots respectively in the sub-figure (a). The constructed teaching set (TS) is shown by dark-red stars in the sub-figure (b).



(a) $\boldsymbol{\theta}^*$ (b) Learned hypothesis

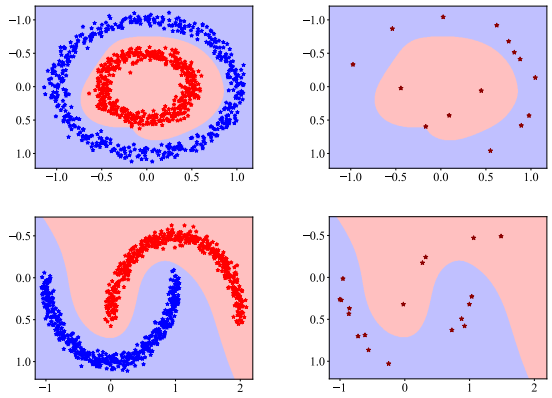
Figure 5. Approximate teaching with Laplacian kernel learner on Moon and Circle datasets. The binary dataset is marked by red and blue dots. The interface between the blue and red areas is the decision boundary of the learned hypothesis. (a) The target hypothesis $\boldsymbol{\theta}^*$. (b) The learned hypothesis with the teaching set being marked as the dark-red stars.

The results show that with much less data points than that of the dataset, both Laplacian kernel learners and exponential kernel learners can obtain nearly the same hypothesis as the target hypothesis generated by the dataset.

Table 4 and Table 5 indicate the relationship between the excess risk ratio $\bar{\Lambda}$ and ϵ-TD for Laplacian kernel learner and exponential kernel learner respectively. Each line of the table represents a dataset and each column indicates a ratio of reference risk. The number inside table shows the TD of our method.

Table 4. ϵ-TD of the regularized Laplacian kernel learner under the excess risk ratio $\bar{\Lambda}$.

Dataset	$\bar{\Lambda} = 100\%$	80%	60%	40%	20%	0%
Sin	2	2	2	2	4	>60
MR	2	2	3	5	10	>60
MPG	1	1	1	1	2	>60
Eunite	1	1	1	1	1	>60
Circle	2	3	4	5	6	17
Moon	2	2	3	5	6	18
Adult	1	3	8	15	37	>60
Sonar	1	3	5	11	36	>60



(a) $\boldsymbol{\theta}^*$ (b) Learned hypothesis

Figure 6. Approximate teaching with exponential kernel learner on Moon and Circle datasets. The binary dataset is marked by red and blue dots. The interface between the blue and red areas is the decision boundary of the learned hypothesis. (a) The target hypothesis $\boldsymbol{\theta}^*$. (b) The learned hypothesis with the teaching set being marked as the dark-red stars.

Table 5. ϵ-TD of the regularized exponential kernel learner under the excess risk ratio $\bar{\Lambda}$.

Dataset	$\bar{\Lambda} = 100\%$	80%	60%	40%	20%	0%
Sin	2	2	2	2	3	>60
MR	2	2	3	5	9	>60
MPG	1	1	1	2	4	>60
Eunite	1	1	2	3	5	>60
Circle	2	2	3	4	5	26
Moon	2	2	2	3	5	16
Adult	1	1	1	36	>60	>60
Sonar	1	4	8	16	42	>60

H. Extra Experiments on Gaussian Kernel with Hinge Loss

To better illustrate the ability of H_m when it comes to hinge loss, we also provide empirical results on the relationship between error rate (ER) and the m of H_m, where ER is defined as $1 -$ accuracy. The results are shown in Table 6.

Each line in the table stands for a binary classification dataset and each column indicates an error rate. The value inside the table is the smallest m such that H_m achieves ER no more than that in the first line of the table.

Table 6. The relationship between ER and m of H_m.

Dataset	ER = 0.5	0.4	0.3	0.2	0.1	0
Circle	1	1	2	3	3	4
Moon	1	1	1	1	3	7
Adult	1	1	1	8	>400	>400
Sonar	1	1	3	17	52	182