The Teaching Dimension of Kernel Perceptron

Akash Kumar
MPI-SWS

Hanqi Zhang
University of Chicago

Adish Singla
MPI-SWS

Yuxin Chen
University of Chicago

Abstract

Algorithmic machine teaching has been studied under the linear setting where exact teaching is possible. However, little is known for teaching nonlinear learners. Here, we establish the sample complexity of teaching, aka teaching dimension, for kernelized perceptrons for different families of feature maps. As a warm-up, we show that the teaching complexity is $\Theta(d)$ for the exact teaching of linear perceptrons in $\mathbb{R}^d$, and $\Theta(d^k)$ for kernel perceptron with a polynomial kernel of order k. Furthermore, under certain smooth assumptions on the data distribution, we establish a rigorous bound on the complexity for approximately teaching a Gaussian kernel perceptron. We provide numerical examples of the optimal (approximate) teaching set under several canonical settings for linear, polynomial and Gaussian kernel perceptrons.

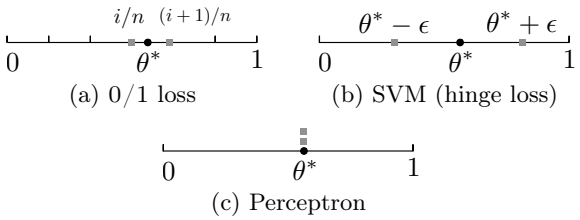
1 Introduction

Machine teaching studies the problem of finding an optimal training sequence to steer a learner towards a target concept (Zhu et al., 2018). An important learning-theoretic complexity measure of machine teaching is the *teaching dimension* (Goldman & Kearns, 1995), which specifies the minimal number of training examples required in the worst case to teach a target concept. Over the past few decades, the notion of teaching dimension has been investigated under a variety of learner's models and teaching protocols (e.g,. Cakmak & Lopes (2012); Singla et al. (2013; 2014); Liu et al. (2017); Haug et al. (2018); Tschiatschek et al. (2019); Liu et al. (2018); Kamalaruban et al. (2019); Hunziker et al. (2019); Devidze et al. (2020); Rakhsha et al. (2020)). One of the most studied scenarios is the case of teaching a version-space learner (Goldman & Kearns, 1995; Anthony et al., 1995; Zilles et al., 2008; Doliwa et al., 2014; Chen et al., 2018; Mansouri et al., 2019; Kirkpatrick et al., 2019). Upon receiving a sequence of training examples from the teacher, a version-space learner maintains a set of hypotheses that are consistent with the training examples, and outputs a *random* hypothesis from this set.

As a canonical example, consider teaching a 1-dimensional binary threshold function $f_{\theta^*}(x) = \mathbb{1}\{x - \theta^*\}$ for $x \in [0, 1]$. For a learner with a finite (or countable infinite) version space, e.g., $\theta \in \{\frac{i}{n}\}_{i=0,\ldots,n}$ where $n \in \mathbb{Z}^+$ (see Fig. 1a), a smallest training set is $\{(\frac{i}{n}, 0), (\frac{i+1}{n}, 1)\}$ where $\frac{i}{n} \le \theta^* < \frac{i+1}{n}$; thus the teaching dimension is 2. However, when the version space is continuous, the teaching dimension becomes ∞, because it is no longer possible for the learner to pick out a unique threshold θ^* with a finite training set. This is due to two key (limiting) modeling assumptions of the version-space learner: (1) all (consistent) hypotheses in the version space are treated equally, and (2) there exists a hypothesis in the version space that is consistent with all training examples. As one can see, these assumptions fail to capture the behavior of many modern learning algorithms, where the best hypotheses are often selected via *optimizing* certain loss functions, and the data is not perfectly separable (i.e. not realizable w.r.t. the hypothesis/model class).

To lift these modeling assumptions, a more realistic teaching scenario is to consider the learner as an *empirical risk minimizer* (ERM). In fact, under the realizable setting, the version-space learner could be viewed as an ERM that optimizes the 0-1 loss—one that finds all hypotheses with zero training error. Recently, Liu & Zhu (2016) studied the teaching dimension of linear ERM, and established values of teaching dimension for several classes of linear (regularized) ERM learners, including support vector machine (SVM), logistic regression and ridge regression. As illustrated in Fig. 1b, for the previous example it suffices to use $\{(\theta^* - \epsilon, 0), (\theta^* + \epsilon, 1)\}$ with any $\epsilon \le \min(1 - \theta^*, \theta^*)$ as training set to teach θ^* as an optimizer of the SVM objective (i.e., $l2$ regu-

Proceedings of the 24[th] International Conference on Artificial Intelligence and Statistics (AISTATS) 2021, San Diego, California, USA. PMLR: Volume 130.



(a) 0/1 loss (b) SVM (hinge loss) (c) Perceptron

Figure 1: Teaching a 1D threshold function to an ERM learner. Training instances are marked in grey. (a) Version-space learner with a finite hypothesis set. (b) SVM and training set $\{(\theta^* - \epsilon, 0), (\theta^* + \epsilon, 1)\}$. (c) ERM learner with (perceptron) loss and training set $\{(\theta^*, 0), (\theta^*, 1)\}$.

larized hinge loss); hence the teaching dimension is 2. In Fig. 1c, we consider teaching an ERM learner with perceptron loss, i.e., $\ell(f_\theta(x), y) = \max(-y \cdot (x - \theta), 0)$ (where $y \in \{-1, 1\}$). If the teacher is allowed to construct *any* training example with *any* labeling[1] , then it is easy to verify that the minimal training set is $\{(\theta^*, -1), (\theta^*, 1)\}$.

While these results show promise at understanding optimal teaching for ERM learners, existing work (Liu & Zhu, 2016) has focused exclusively on the linear setting with the goal to teach the exact hypothesis (e.g., teaching the exact model parameters or the exact decision boundary for classification tasks). Aligned with these results, we establish an upper bound as shown in §3.1. It remains a fundamental challenge to rigorously characterize the teaching complexity for nonlinear learners. Furthermore, in the cases where exact teaching is not possible with a finite training set, the classical teaching dimension no longer captures the fine-grained complexity of the teaching tasks, and hence one needs to relax the teaching goals and investigate new notions of teaching complexity.

In this paper, we aim to address the above challenges. We focus on kernel perceptron, a specific type of ERM learner that is less understood even under the linear setting. Following the convention in teaching ERM learners, we consider the *constructive* setting, where the teacher can construct arbitrary teaching examples in the support of the data distribution. Our contributions are highlighted below, with main theoretical results summarized in Table 1.

- We formally define approximate teaching of kernel perceptron, and propose a novel measure of teaching complexity, namely the *ϵ-approximate teaching dimension* (ϵ-TD), which captures the complexity of teaching a "relaxed" target that is close to the target hypothesis in terms of the expected risk. Our relaxed notion of teaching dimension strictly generalizes the teaching dimension of Liu & Zhu (2016), where it trades off the teaching complexity against the risk of the taught hypothesis, and hence is more practical in characterizing the complexity of a teaching task (§2).

- We show that exact teaching is feasible for kernel perceptrons with finite dimensional feature maps, such as linear kernel and polynomial kernel. Specifically, for data points in $\mathbb{R}^d$, we establish a $\Theta(d)$ bound on the teaching dimension of linear perceptron. Under a mild condition on data distribution, we provide a tight bound of $\Theta\left(\binom{d+k-1}{k}\right)$ for polynomial perceptron of order k. We also exhibit optimal training sets that match these teaching dimensions (§3.1 and §3.2).

- We further show that for Gaussian kernelized perceptron, exact teaching is not possible with a finite set of hypotheses, and then establish a $d^{\mathcal{O}\left(\log^2 \frac{1}{\epsilon}\right)}$ bound on the ϵ-approximate teaching dimension (§3.4). To the best of our knowledge, these results constitute the first known bounds on (approximately) teaching a non-linear ERM learner (§3).

	linear	**polynomial**	**Gaussian**
TD (exact)	$\Theta(d)$	$\Theta\left(\binom{d+k-1}{k}\right)$	∞
ϵ-approximate TD	-	-	$d^{\mathcal{O}\left(\log^2 \frac{1}{\epsilon}\right)}$
Assumption	-	3.2.1	3.4.1, 3.4.2

Table 1: Teaching dimension for kernel perceptron

2 Problem Statement

Basic definitions We denote by $\mathcal{X}$ the input space and $\mathcal{Y} := \{-1, 1\}$ the output space. A hypothesis is a function $h : \mathcal{X} \to \mathcal{Y}$. In this paper, we identify a hypothesis $h_{\boldsymbol{\theta}}$ with its model parameter $\boldsymbol{\theta}$. The hypothesis space $\mathcal{H}$ is a set of hypotheses. By training point we mean a pair $(\mathbf{x}, y) \in \mathcal{X} \times \mathcal{Y}$. We assume that the training points are drawn from an unknown distribution $\mathcal{P}$ over $\mathcal{X} \times \mathcal{Y}$. A training set is a multiset $\mathcal{D} = \{(\mathbf{x}_1, y_1), \cdots, (\mathbf{x}_n, y_n)\}$ where repeated pairs are allowed. Let $\mathbb{D}$ denote the set of all training sets of all sizes. A learning algorithm $\boldsymbol{\mathcal{A}} : \mathbb{D} \to 2^{\mathcal{H}}$ takes in a training set $D \in \mathbb{D}$ and outputs a subset of the hypothesis space $\mathcal{H}$. That is, $\boldsymbol{\mathcal{A}}$ doesn't necessarily return a unique hypothesis.

Kernel perceptron Consider a set of training points $\mathcal{D} := \{(\mathbf{x}_i, y_i)\}_{i=1}^n$ where $\mathbf{x}_i \in \mathbb{R}^d$ and hypothesis $\boldsymbol{\theta} \in$

[1] If the teacher is restricted to only provide consistent labels (i.e., the realizable setting), then the ERM with perceptron loss reduces to the version space learner, where the teaching dimension is ∞.

$\mathbb{R}^d$. A linear perceptron is defined as $f_{\boldsymbol{\theta}}(\mathbf{x}) := \text{sign}(\boldsymbol{\theta}\cdot\mathbf{x})$ in homogeneous setting. We consider the algorithm $\boldsymbol{\mathcal{A}}_{opt}$ to learn an optimal perceptron to classify $\mathcal{D}$ as defined below:

$$\boldsymbol{\mathcal{A}}_{opt}\left(\mathcal{D}\right) := \arg\min_{\boldsymbol{\theta}\in\mathbb{R}^d} \sum_{i=1}^{n} \ell(f_{\boldsymbol{\theta}}(\mathbf{x}_i), y_i). \tag{1}$$

where the loss function $\ell(f_{\boldsymbol{\theta}}(\mathbf{x}), y) := \max(-y\cdot f_{\boldsymbol{\theta}}(\mathbf{x}), 0)$. Similarly, we consider the non-linear setting via kernel-based hypotheses for perceptrons that are defined with respect to a kernel operator $\mathcal{K} : \mathcal{X}\times\mathcal{X} \to \mathbb{R}$ which adheres to Mercer's positive definite conditions (Vapnik, 1998). A kernel-based hypothesis has the form,

$$f(\mathbf{x}) = \sum_{i=1}^{k} \alpha_i \cdot \mathcal{K}(\mathbf{x}_i, \mathbf{x}) \tag{2}$$

where $\forall i\ \mathbf{x}_i \in \mathcal{X}$ and α_i are reals. In order to simplify the derivation of the algorithms and their analysis, we associate a *reproducing kernel Hilbert space* (RKHS) with $\mathcal{K}$ in the standard way common to all kernel methods. Formally, let $\mathcal{H}_{\mathcal{K}}$ be the closure of the set of all hypotheses of the form given in Eq. (2). A non-linear kernel perceptron corresponding to $\mathcal{K}$ optimizes Eq. (1) as follows:

$$\boldsymbol{\mathcal{A}}_{opt}(\mathcal{D}) := \arg\min_{\boldsymbol{\theta}\in\mathcal{H}_{\mathcal{K}}} \sum_{i=1}^{n} \ell(f_{\boldsymbol{\theta}}(\mathbf{x}_i), y_i) \tag{3}$$

where $f_{\boldsymbol{\theta}}(\cdot) = \sum_{i=1}^{l} \alpha_i \cdot \mathcal{K}(\mathbf{a}_i, \cdot)$ for some $\{\mathbf{a}_i\}_{i=1}^{l} \subset \mathcal{X}$ and α_i real. Alternatively, we also write $f_{\boldsymbol{\theta}}(\cdot) = \boldsymbol{\theta}\cdot\Phi(\cdot)$ where $\Phi : \mathcal{X} \to \mathcal{H}_{\mathcal{K}}$ is defined as feature map to the kernel function $\mathcal{K}$. A reproducing kernel Hilbert space with $\mathcal{K}$ could be decomposed as $\mathcal{K}(\mathbf{x}, \mathbf{x}') = \langle\Phi(\mathbf{x}), \Phi(\mathbf{x}')\rangle$ (Scholkopf & Smola, 2001) for any $\mathbf{x}, \mathbf{x}' \in \mathcal{X}$. Thus, we also identify $f_{\boldsymbol{\theta}}$ as $\sum_{i=1}^{l} \alpha_i \cdot \Phi(\mathbf{a}_i)$.

The teaching problem We are interested in the problem of teaching a target hypothesis $\boldsymbol{\theta}^*$ where a helpful *teacher* provides labelled data points $\mathcal{TS} \subseteq \mathcal{X}\times\mathcal{Y}$, also defined as a *teaching set*. Assuming the constructive setting (Liu & Zhu, 2016), to teach a kernel perceptron learner the teacher can construct a training set with any items in $\mathbb{R}^d$ i.e. for any $(\mathbf{x}', y') \in \mathcal{TS}$ we have $\mathbf{x}' \in \mathbb{R}^d$ and $y' \in \{-1, 1\}$. Importantly, for the purpose of teaching we do not *assume* that $\mathcal{TS}$ are drawn *i.i.d* from a distribution. We define the teaching dimension for *exact* parameter of $\boldsymbol{\theta}^*$ corresponding to a kernel perceptron as $TD(\boldsymbol{\theta}^*, \boldsymbol{\mathcal{A}}_{opt})$, which is the size of the smallest teaching set $\mathcal{TS}$ such that $\boldsymbol{\mathcal{A}}_{opt}\left(\mathcal{TS}\right) = \{\boldsymbol{\theta}^*\}$. We define teaching of exact parameters of a target hypothesis $\boldsymbol{\theta}^*$ as *exact teaching*. Since, a perceptron is agnostic to norms, we study the problem of teaching a target classifier *decision boundary* where $\boldsymbol{\mathcal{A}}_{opt}\left(\mathcal{TS}\right) = \{t\boldsymbol{\theta}^*\}$ for some real $t > 0$. Thus,

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

Since it can be stringent to construct a teaching set for decision boundary (see §3.4), exact teaching is not always feasible. We introduce and study *approximate teaching* which is formally defined as:

Definition 1 (ϵ-approximate teaching set)**.** *Consider a kernel perceptron learner, with a kernel $\mathcal{K} : \mathcal{X}\times\mathcal{X} \to \mathbb{R}$ and the corresponding RKHS feature map $\Phi(\cdot)$. For a target model $\boldsymbol{\theta}^* \in \mathcal{H}_{\mathcal{K}}$ and $\epsilon > 0$, we say $\mathcal{TS} \subseteq \mathcal{X}\times\mathcal{Y}$ is an ϵ-approximate teaching set wrt to $\mathcal{P}$ if the kernel perceptron $\hat{\boldsymbol{\theta}} \in \boldsymbol{\mathcal{A}}_{opt}(\mathcal{TS})$ satisfies*

$$\left|\mathbb{E}\left[\max(-y\cdot f^*(\mathbf{x}), 0)\right] - \mathbb{E}\left[\max(-y\cdot \hat{f}(\mathbf{x}), 0)\right]\right| \leq \epsilon \tag{4}$$

where the expectations are over $(\mathbf{x}, y) \sim \mathcal{P}$ and $f^*(\mathbf{x}) = \boldsymbol{\theta}^*\cdot\Phi(\mathbf{x})$ and $\hat{f}(\mathbf{x}) = \hat{\boldsymbol{\theta}}\cdot\Phi(\mathbf{x})$.

Naturally, we define approximate teaching dimension as:

Definition 2 (ϵ-approximate teaching dimension)**.** *Consider a kernel perceptron learner, with a kernel $\mathcal{K} : \mathcal{X}\times\mathcal{X} \to \mathbb{R}$ and the corresponding RKHS feature map $\Phi(\cdot)$. For a target model $\boldsymbol{\theta}^* \in \mathcal{H}_{\mathcal{K}}$ and $\epsilon > 0$, we define ϵ-$TD(\boldsymbol{\theta}^*, \boldsymbol{\mathcal{A}}_{opt})$ as the teaching dimension which is the size of the smallest teaching set for ϵ-approximate teaching of $\boldsymbol{\theta}^*$ wrt $\mathcal{P}$.*

According to Definition 2, exact teaching corresponds to constructing a 0-approximate teaching set for a target classifier (e.g., the decision boundary of a kernel perceptron). We study linear and polynomial kernelized perceptrons in the exact teaching setting. Under some mild assumptions on the smoothness of the data distribution, we establish approximate teaching bound on approximate teaching dimension for Gaussian kernelized perceptron.

3 Teaching Dimension for Kernel Perceptron

In this section, we study the generic problem of teaching kernel perceptrons in three different settings: 1) linear (in §3.1); 2) polynomial (in §3.2); and Gaussian (in §3.4). Before establishing our main result for Gaussian kernelized perceptrons, we first introduce two important results for linear and polynomial perceptrons inherently connected to the Gaussian perceptron. Our proofs are inspired by ideas from linear algebra and projective geometry as detailed in the supplemental materials.

3.1 Homogeneous Linear Perceptron

In this subsection, we study the problem of teaching a linear perceptron. First, we consider an optimization problem similar to Eq. (1) as shown in Liu & Zhu (2016):

$$\boldsymbol{\mathcal{A}}_{opt} := \arg\min_{\boldsymbol{\theta}\in\mathbb{R}^d} \sum_{i=1}^{n} \ell(\boldsymbol{\theta}\cdot\mathbf{x}_i, y_i) + \frac{\lambda}{2}||\boldsymbol{\theta}||_A^2 \qquad (5)$$

where $\ell(\cdot,\cdot)$ is a convex loss function, A is a positive semi-definite matrix, $||\boldsymbol{\theta}||_A$ is defined as $\sqrt{\boldsymbol{\theta}^\top A\boldsymbol{\theta}}$, and $\lambda > 0$. For convex loss function $\ell(\cdot,\cdot)$, Theorem 1 (Liu & Zhu, 2016) established a degree-of-freedom lower bound on the number of training items to obtain a unique solution $\boldsymbol{\theta}^*$. Since, the loss function for linear perceptron is convex thus we immediately obtain a lower bound on the teaching dimension as follows:

Corollary 1. *If $A = 0$ and $\lambda = 1$, then Eq. (1) can be solved as Eq. (5). Moreover, teaching dimension for decision boundary corresponding to a target model $\boldsymbol{\theta}^*$ is lower-bounded by $\Omega(d)$.*

Now, we would establish an upper bound on $TD(\boldsymbol{\mathcal{A}}_{opt}, \boldsymbol{\theta}^*)$ for exact teaching of the decision boundary of a target model $\boldsymbol{\theta}^*$. The key idea is to find a set of points which span the orthogonal subspace of $\boldsymbol{\theta}^*$, which we use to force a solution $\hat{\boldsymbol{\theta}} \in \boldsymbol{\mathcal{A}}_{opt}$ such that it has a component only along $\boldsymbol{\theta}^*$. Formally, we state the claim of the result with proof as follows:

Theorem 1. *Given any target model $\boldsymbol{\theta}^*$, for solving Eq. (1) the teaching dimension for the decision boundary corresponding to $\boldsymbol{\theta}^*$ is $\Theta(d)$. The following is a teaching set:*

$$\mathbf{x}_i = \mathbf{v}_i, \quad y_i = 1 \quad \forall\, i \in [d-1];$$
$$\mathbf{x}_d = -\sum_{i=1}^{d-1} \mathbf{v}_i, \quad y_d = 1; \quad \mathbf{x}_{d+1} = \boldsymbol{\theta}^*, \quad y_{d+1} = 1$$

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

Proof. Using Corollary 1, the lower bound for solving Eq. (1) is immediate. Thus, if we show that the mentioned labeled set of training points form a teaching set, then we can show an upper bound which would imply a tight bound of $\Theta(d)$ on the teaching dimension for finding the decision boundary. Denote the set of labeled data points as $\mathcal{D}$. Denote by $\mathbf{p}(\boldsymbol{\theta}) := \sum_{i=1}^{d+1} \max(-y_i\cdot\boldsymbol{\theta}\cdot\mathbf{x}_i,\, 0)$. Since $\{\mathbf{v}_i\}_{i=1}^d$ is an orthogonal basis, thus $\forall\, i \in [d-1] \quad \mathbf{v}_i\cdot\boldsymbol{\theta}^* = 0$, thus it is not very difficult to show that $\mathbf{p}(t\boldsymbol{\theta}^*) = 0$ for some positive scalar t. Note, if $\hat{\boldsymbol{\theta}}$ is a solution to Eq. (1) then:

$$\hat{\boldsymbol{\theta}} \in \arg\min_{\boldsymbol{\theta}\in\mathbb{R}^d} \sum_{i=1}^{d+1} \max(-y_i\cdot\boldsymbol{\theta}\cdot\mathbf{x}_i,\, 0)$$

Also, $\mathbf{p}(\hat{\boldsymbol{\theta}}) = 0 \implies \mathbf{x}_i\cdot\hat{\boldsymbol{\theta}} \geq 0\ \forall i \in [d]$ but then $\mathbf{x}_d = -\sum_{i=1}^{d-1}\mathbf{x}_i \implies \forall i \in [d] \quad \mathbf{x}_i\cdot\hat{\boldsymbol{\theta}} = 0$. Note that, $\hat{\boldsymbol{\theta}}\cdot\boldsymbol{\theta}^* \geq 0$ forces $\hat{\boldsymbol{\theta}} = t\boldsymbol{\theta}^*$ for some positive constant t. Thus, $\mathcal{D}$ is a teaching set for the decision boundary of $\boldsymbol{\theta}^*$. This establishes the upper bound, and hence the theorem follows. □

Numerical example To illustrate Theorem 1, we provide a numerical example for teaching a linear perceptron in $\mathbb{R}^3$, with $\boldsymbol{\theta}^* = (-3, 3, 5)^\top$ (illustrated in Fig. 2a). To construct the teaching set, we first obtain an orthogonal basis $\{(0.46, 0.86, -0.24)^\top, (0.76, -0.24, 0.6)^\top\}$ for the subspace orthogonal to $\boldsymbol{\theta}^*$, and add a vector $(-1.22, -0.62, -0.36)^\top$ which is in the exact opposite direction of the first two combined. Finally we add to $\mathcal{TS}$ an arbitrary vector which has a positive dot product with the normal vector, e.g. $(-0.46, 0.46, 0.76)^\top$. Labeling all examples positive, we obtain $\mathcal{TS}$ of size 4.

3.2 Homogeneous Polynomial Kernelized Perceptron

In this subsection, we study the problem of teaching a polynomial kernelized perceptron in realizable setting. Similar to §3.1, we establish an exact teaching bound on the teaching dimension under a mild condition on the data distribution. We consider homogeneous polynomial kernel $\mathcal{K}$ of degree k in which for any $\mathbf{x}, \mathbf{x}' \in \mathbb{R}^d$

$$\mathcal{K}(\mathbf{x}, \mathbf{x}') = (\langle \mathbf{x}, \mathbf{x}'\rangle)^k$$

If $\Phi(\cdot)$ denotes the *feature map* for the corresponding RKHS, then we know that the dimension of the map is $\binom{d+k-1}{k}$ where each component of the map can be represented by $\Phi_{\boldsymbol{\lambda}}(\mathbf{x}) = \sqrt{\frac{k!}{\prod_{i=1}^d \boldsymbol{\lambda}_i!}}\mathbf{x}^{\boldsymbol{\lambda}}$ where $\boldsymbol{\lambda} \in (\mathbb{N}\cup\{0\})^d$ and $\sum_i \boldsymbol{\lambda}_i = k$. Denote by $\mathcal{H}_{\mathcal{K}}$ the RKHS corresponding to the polynomial kernel $\mathcal{K}$. We use $\mathcal{H}_k := \mathcal{H}_k(\mathbb{R}^d)$ to represent the linear space of homogeneous polynomials of degree k over $\mathbb{R}^d$. We mention an important result which shows the RKHS for polynomial kernels is isomorphic to the space of homogeneous polynomials of degree k in d variables.

Proposition 1 (Chapter III.2, Proposition 6 (Cucker & Smale, 2001))**.** *$\mathcal{H}_k = \mathcal{H}_{\mathcal{K}}$ as function spaces and inner product spaces.*

The dimension $\dim\left(\mathcal{H}_k(\mathbb{R}^d)\right)$ of the linear space of homogeneous polynomials of degree k over $\mathbb{R}^d$ is $\binom{d+k-1}{k}$. Denote by $r := \binom{d+k-1}{k}$. Since $\mathcal{H}_{\mathcal{K}}$ is a vector space for polynomial kernel $\mathcal{K}$, thus for exact teaching there is an obvious lower bound of $\Omega\left(\binom{d+k-1}{k}\right)$ on the teaching dimension.

Before we establish the main result of this subsection we state a mild assumption on the target model we consider for exact teaching which is as follows:

Assumption 3.2.1 (Existence of orthogonal polynomials)**.** *For the target model $\boldsymbol{\theta}^* \in \mathcal{H}_\mathcal{K}$, we assume that there exist $(r-1)$ linearly independent polynomials on the orthogonal subspace of $\boldsymbol{\theta}^*$ in $\mathcal{H}_\mathcal{K}$ of the form $\{\Phi(\mathbf{z}_i)\}_{i=1}^{r-1}$ where $\forall i\ \mathbf{z}_i \in \mathcal{X}$.*

Similar to Theorem 1, the key insight in having Assumption 3.2.1 is to find independent polynomial on the orthogonal subspace defined by $\boldsymbol{\theta}^*$. We state the claim here with proof established in the supplemental materials.

Theorem 2. *For all target models $\boldsymbol{\theta}^* \in \mathcal{H}_\mathcal{K}$ for which the Assumption 3.2.1 holds, for solving Eq. (3), the exact teaching dimension for the decision boundary corresponding to $\boldsymbol{\theta}^*$ is $\mathcal{O}\left(\binom{d+k-1}{k}\right)$.*

Numerical example For constructing $\mathcal{TS}$ in the polynomial case, we follow a similar strategy in the higher dimensional space that the original data is projected into. The only difference is that we need to ensure the teaching examples have pre-images in the original space. For that, we adopt a randomized algorithm that solves for $r-1$ boundary points in the original space (i.e. solve for $\theta^* \cdot \Phi(\mathbf{x}) = 0$) , while checking the images of these points are linearly independent. Also, instead of adding a vector in the opposite direction of these points combined, we simply repeat the $r-1$ points in the teaching set, while assigning one copy of them positive labels and the other copy negative labels. Finally, we need one last vector (label it positive) whose image has a positive component in θ^*, and we obtain $\mathcal{TS}$ of size $2r-1$.

Fig. 2b and Fig. 2c demonstrate the above constructive procedure on a numerical example with $d = 2$, homogeneous polynomial kernel of degree 2, and $\boldsymbol{\theta}^* = (1, 4, 4)^\top$. In Fig. 2b we show the decision boundary (red lines) and the level sets (polynomial contours) of this quadratic perceptron, as well as the teaching set identified via the above algorithmic procedure. In Fig. 2b, we visualize the decision boundary (grey plane) in the feature space (after applying the feature map). The blue surface corresponds to all the data points that have pre-images in the original space $\mathbb{R}^2$.

3.3 Limitations in Exact Teaching of Polynomial Kernel Perceptron

In the previous section §3.2, we imposed the Assumption 3.2.1 on the target models $\boldsymbol{\theta}^*$. It turns out that we couldn't do better than this. More concretely, we need to impose this assumption for exact teaching of polynomial kernel perceptron learner. Further, there are pathological cases where violation of the assumption leads to models which couldn't be approximately taught.

Intuitively, solving Eq. (3) in the paradigm of exact teaching reduces to nullifying the orthogonal subspace of $\boldsymbol{\theta}^*$ i.e. any component of $\boldsymbol{\theta}^*$ along the subspace is nullified. Since the information of the span of the subspace has to be encoded into the datapoints chosen for teaching, Assumption 3.2.1 is a natural step to make. Interestingly, we show that the step is not so stringent. In the realizable setting in which all the teaching points are correctly classified, if we lift the assumption then exact teaching is not possible.We state the claim in the following lemma:

Lemma 1. *Consider a target model $\boldsymbol{\theta}^*$ that doesn't satisfy Assumption 3.2.1. Then, there doesn't exist a teaching set $\mathcal{TS}_{\boldsymbol{\theta}^*}$ which exactly teaches $\boldsymbol{\theta}^*$ i.e. for any $\mathcal{TS}_{\boldsymbol{\theta}^*}$ and any real $t > 0$*

$$\boldsymbol{\mathcal{A}}_{opt}\left(\mathcal{TS}_{\boldsymbol{\theta}^*}\right) \neq \{t\boldsymbol{\theta}^*\}.$$

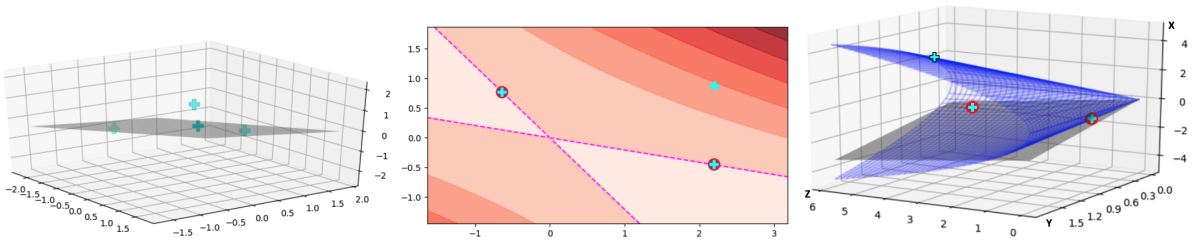
Lemma 1 shows that for *exact* teaching $\boldsymbol{\theta}^*$ should satisfy Assumption 3.2.1. Then, the natural question that arises is whether we can achieve arbitrarily ϵ-close *approximate* teaching for $\boldsymbol{\theta}^*$. In other words, we would like to find $\tilde{\boldsymbol{\theta}}^*$ that satisfies Assumption 3.2.1 and is in ϵ-neighbourhood of $\boldsymbol{\theta}^*$. We show a negative result for this when k is even. For this we assume that, the datapoints in the teaching set $\mathcal{TS}_{\tilde{\boldsymbol{\theta}}^*}$ have lower-bounded norm, call it, $\delta > 0$ i.e. if $(\mathbf{x}_i, y_i) \in \mathcal{TS}_{\tilde{\boldsymbol{\theta}}^*}$ then $||\Phi(\mathbf{x}_i)|| \geq \delta$. We require this additional assumption only for the purpose of analysis. We would show that it wouldn't lead to any pathological cases where the constructed target model $\boldsymbol{\theta}^*$ incorporates approximate teaching.

Lemma 2. *Let $\mathcal{X} \subseteq \mathbb{R}^d$ and $\mathcal{H}_\mathcal{K}$ be the reproducing kernel Hilbert space such that kernel function $\mathcal{K}$ is of degree k. If k has parity even then there exists a target model $\boldsymbol{\theta}^*$ which violates Assumption 3.2.1 and can't be taught approximately.*

The results are discussed in details with proofs in the supplemental materials. Assumption 3.2.1 and the stated lemmas provide insights into understanding the problem of teaching for non-linear perceptron kernels. In the next section, we study Gaussian kernel and the ideas generated here would be useful in devising a teaching set in the paradigm of approximate teaching.

3.4 Gaussian Kernelized Perceptron

In this subsection, we consider the Gaussian kernel. Under mild assumptions inspired by the analysis of teaching dimension for exact teaching of linear and polynomial kernel perceptrons, we would establish as



(a) Linear ($\mathcal{TS}$) (b) Polynomial ($\mathcal{TS}$) (c) Polynomial (feature space)

Figure 2: Numerical examples of exact teaching for linear and polynomial perceptrons. Cyan plus marks and red dots correspond to positive and negative teaching examples respectively.

our main result an upper bound on the ϵ-approximate teaching dimension of Gaussian kernel perceptrons using a construction of an ϵ-approximate teaching set.

Preliminaries of Gaussian kernel A Gaussian kernel $\mathcal{K}$ is a function of the form

$$\mathcal{K}(\mathbf{x}, \mathbf{x}') = e^{-\frac{||\mathbf{x}-\mathbf{x}'||^2}{2\sigma^2}} \tag{6}$$

for any $\mathbf{x}, \mathbf{x}' \in \mathbb{R}^d$ and parameter σ. First, we would try to understand the feature map before we find an approximation to it. Notice:

$$e^{-\frac{||\mathbf{x}-\mathbf{x}'||^2}{2\sigma^2}} = e^{-\frac{||\mathbf{x}||^2}{2\sigma^2}} e^{-\frac{||\mathbf{x}'||^2}{2\sigma^2}} e^{\frac{\langle \mathbf{x},\, \mathbf{x}'\rangle}{\sigma^2}}$$

Consider the scalar term $z = \langle \mathbf{x},\, \mathbf{x}' \rangle / \sigma^2$. We can expand the term of the product using the Taylor expansion of e^z near $z = 0$ as shown in Cotter et al. (2011), which amounts to $e^{\frac{\langle \mathbf{x},\, \mathbf{x}'\rangle}{\sigma^2}} = \sum_{k=0}^{\infty} \frac{1}{k!} \left(\frac{\langle \mathbf{x},\, \mathbf{x}'\rangle}{\sigma^2} \right)^k$. We can further expand the previous sum as

$$\begin{aligned} e^{\frac{\langle \mathbf{x},\, \mathbf{x}'\rangle}{\sigma^2}} &= \sum_{k=0}^{\infty} \frac{1}{k!} \left(\frac{\langle \mathbf{x},\, \mathbf{x}'\rangle}{\sigma^2} \right)^k \\ &= \sum_{k=0}^{\infty} \frac{1}{k! \sigma^{2k}} \left(\sum_{l=1}^{d} \mathbf{x}_l \cdot \mathbf{x}'_l \right)^k \\ &= \sum_{k=0}^{\infty} \frac{1}{k! \sigma^{2k}} \sum_{|\boldsymbol{\lambda}|=k} \mathcal{C}^k_{\boldsymbol{\lambda}} \cdot \mathbf{x}^{\boldsymbol{\lambda}} \cdot (\mathbf{x}')^{\boldsymbol{\lambda}} \end{aligned} \tag{7}$$

where $\mathcal{C}^k_{\boldsymbol{\lambda}} = \frac{k!}{\prod_{i=1}^d \boldsymbol{\lambda}_i!}$. Thus, we use Eq. (7) to obtain explicit feature representation to the Gaussian kernel in Eq. (6) as $\Phi_{k,\boldsymbol{\lambda}}(\mathbf{x}) = e^{-\frac{||\mathbf{x}||^2}{2\sigma^2}} \cdot \frac{\sqrt{\mathcal{C}^k_{\boldsymbol{\lambda}}}}{\sqrt{k!}\sigma^k} \cdot \mathbf{x}^{\boldsymbol{\lambda}}$. We get the explicit feature map $\Phi(\cdot)$ for the Gaussian kernel with coordinates as specified. Theorem 1 of Ha Quang (2010) characterizes the RKHS of Gaussian kernel. It establishes that $\dim(\mathcal{H}_{\mathcal{K}}) = \infty$. Thus, we note that the exact teaching for an arbitrary target classifier f^* in this setting has an infinite lower bound. This calls for analysing the teaching problem of a Gaussian kernel in the *approximate* teaching setting.

Definitions and notations for approximate teaching For any classifier $f \in \mathcal{H}_{\mathcal{K}}$, we define $\mathbf{err}(f) = \mathbb{E}_{(\mathbf{x},y)\sim\mathcal{P}(\mathbf{x},y)}[\max(-y \cdot f(\mathbf{x}), 0)]$. Our goal is to find a classifier f with the property that its expected true loss $\mathbf{err}(f)$ is as small as possible. In the realizable setting, we assume that there exists an optimal separator f^* such that for any data instances sampled from the data distribution the labels are consistent i.e. $\mathcal{P}(y \cdot f^*(\mathbf{x}) \leq 0) = 0$. In addition, we also experiment for the non-realizable setting. In the rest of the subsection, we would study the relationship between the teaching complexity for an optimal Gaussian kernel perceptron for Eq. (3) and $|\mathbf{err}(f^*) - \mathbf{err}(\hat{f})|$ where f^* is the optimal separator and $\hat{f}$ is the solution to $\boldsymbol{\mathcal{A}}_{opt}(\mathcal{TS}_{\boldsymbol{\theta}^*})$ for the constructed teaching set $\mathcal{TS}_{\boldsymbol{\theta}^*}$.

3.4.1 Gaussian Kernel Approximation

Now, we would talk about finite-dimensional polynomial approximation $\tilde{\Phi}$ to the Gaussian feature map Φ via projection as shown in Cotter et al. (2011). Consider

$$\tilde{\Phi} : \mathbb{R}^d \longrightarrow \mathbb{R}^q$$
$$\tilde{\mathcal{K}}(\mathbf{x}, \mathbf{x}') = \tilde{\Phi}(\mathbf{x}) \cdot \tilde{\Phi}(\mathbf{x}')$$

With these approximations, we consider classifiers of the form $\tilde{f}(\mathbf{x}) = \tilde{\boldsymbol{\theta}} \cdot \tilde{\Phi}(\mathbf{x})$ such that $\tilde{\boldsymbol{\theta}} \in \mathbb{R}^q$. Now, assume that there is a projection map $\mathbb{P}$ such that $\tilde{\Phi} = \mathbb{P}\Phi$. In Cotter et al. (2011), authors used the following approximation to the Gaussian kernel:

$$\tilde{\mathcal{K}}(\mathbf{x}, \mathbf{x}') = e^{-\frac{||\mathbf{x}||^2}{2\sigma^2}} e^{-\frac{||\mathbf{x}'||^2}{2\sigma^2}} \sum_{k=0}^{s} \frac{1}{k!} \left(\frac{\langle \mathbf{x},\, \mathbf{x}'\rangle}{\sigma^2} \right)^k \tag{8}$$

This gives the following explicit feature representation for the approximated kernel:

$$\forall k \leq s, \quad \tilde{\Phi}_{k,\boldsymbol{\lambda}}(\mathbf{x}) = \Phi_{k,\boldsymbol{\lambda}}(\mathbf{x}) = e^{-\frac{||\mathbf{x}||^2}{2\sigma^2}} \cdot \frac{\sqrt{\mathcal{C}^k_{\boldsymbol{\lambda}}}}{\sqrt{k!}\sigma^k} \cdot \mathbf{x}^{\boldsymbol{\lambda}} \tag{9}$$

where $\Phi_{k,\boldsymbol{\lambda}}(\mathbf{x})$ is the coordinate for Gaussian feature map. Note that the feature map $\tilde{\Phi}$ defined by the

explicit features in Eq. (9) has dimension $\binom{d+s}{d}$. Thus, $\mathbb{P}\Phi = \tilde{\Phi}$ where the first $\binom{d+s}{d}$ coordinates are retained. We denote the RKHS corresponding to $\tilde{\mathcal{K}}$ as $\mathcal{H}_{\tilde{\mathcal{K}}}$. A simple property of the approximated kernel map is stated in the following lemma which was proven in Cotter et al. (2011).

Lemma 3 (Cotter et al. (2011))**.** *For the approximated map $\tilde{\mathcal{K}}$, we obtain the following upper bound:*

$$\left|\mathcal{K}(\mathbf{x},\mathbf{x}) - \tilde{\mathcal{K}}(\mathbf{x},\mathbf{x})\right| \leq \frac{1}{(s+1)!}\left(\frac{||\mathbf{x}|| \cdot ||\mathbf{x}'||}{\sigma^2}\right)^{s+1} \tag{10}$$

Note that if s is chosen large enough and the points $\mathbf{x}, \mathbf{x}'$ are bounded wrt σ^2, then RHS of Eq. (10) can be bounded by any $\epsilon > 0$. Since $\left|\mathcal{K}(\mathbf{x},\mathbf{x}) - \tilde{\mathcal{K}}(\mathbf{x},\mathbf{x})\right| = \left|\left|\mathbb{P}^{\perp}\Phi(\mathbf{x})\right|\right|^2$, thus for a Gaussian kernel, information theoretically, the first $\binom{d+s}{s}$ coordinates are highly sensitive. We would try to analyze this observation under some mild assumptions on the data distribution to construct an ϵ-approximate teaching set. As discussed in the supplemental materials, we would find the value of s as if the datapoints are coming from a ball of radius $R := \max\left\{\frac{\log^2\frac{1}{\epsilon}}{e^2}, d\right\}$ in $\mathbb{R}^d$ i.e. $\frac{||\mathbf{x}||^2}{\sigma^2} \leq R$. Thus, we wish to solve for the value of s such that $\frac{1}{(s+1)!} \cdot (R)^{s+1} \leq \epsilon$.

To approximate s we use Sterling's approximation, which states that for all positive integers n, we have

$$\sqrt{2\pi}n^{n+1/2}e^{-n} \leq n! \leq en^{n+1/2}e^{-n}.$$

Using the bound stated in Lemma 3, we fix the value for s as $e^2 \cdot R$. We would assume that $R = \frac{\log^2\frac{1}{\epsilon}}{e^2}$ since we wish to achieve arbitrarily small ϵ-approximate[2] teaching set. We define $r := r(\boldsymbol{\theta}^*, \epsilon) = \binom{d+s}{s}$.

3.4.2 Bounding the Error

In this subsection, we discuss our key results on approximate teaching of a Gaussian kernel perceptron learner under some mild assumptions on the target model $\boldsymbol{\theta}^*$. In order to show $\left|\mathbf{err}(f^*) - \mathbf{err}(\hat{f})\right| \leq \epsilon$ via optimizing to a solution $\hat{\boldsymbol{\theta}}$ for Eq. (3), we would achieve a pointwise ϵ-closeness between f^* and $\hat{f}$. Specifically, we show that $\left|f^*(\mathbf{x}) - \hat{f}(\mathbf{x})\right| \leq \epsilon$ universally which is similar in spirit to universal approximation theorems (Liang & Srikant, 2017; Lu & Lu, 2020; Yarotsky, 2017) for neural networks. We prove that this universal approximation could be achieved with $d^{\mathcal{O}\left(\log^2\frac{1}{\epsilon}\right)}$ size teaching set.

We assume that the input space $\mathcal{X}$ is bounded such that $\forall \mathbf{x} \in \mathcal{X}$ $\frac{\langle \mathbf{x}, \mathbf{x}\rangle_{\mathcal{H}_{\mathcal{K}}}}{\sigma^2} \leq 2\sqrt{R}$. Since the motivation is to find classifiers which are close to the optimal one pointwise, thus we assume that target model $\boldsymbol{\theta}^*$ has unit norm. As mentioned in Eq. (2), we can write the target model $\boldsymbol{\theta}^* \in \mathcal{H}_{\mathcal{K}}$ as $\boldsymbol{\theta}^* = \sum_{i=1}^{l} \alpha_i \cdot \mathcal{K}(\mathbf{a}_i, \cdot)$ for some $\{\mathbf{a}_i\}_{i=1}^{l} \subset \mathcal{X}$ and $\alpha_i \in \mathbb{R}$. The classifier corresponding to $\boldsymbol{\theta}^*$ is represented by f^*. Eq. (3) can be rewritten corresponding to a teaching set $\mathcal{D} := \{(\mathbf{x}_i,\ y_i)\}_{i=1}^{n}$ as:

$$\boldsymbol{\mathcal{A}}_{opt} := \arg\min_{\beta \in \mathbb{R}^l} \sum_{i=1}^{n} \max\left(-y_i \cdot \sum_{j=1}^{l} \beta_j \cdot \mathcal{K}(\mathbf{a}_j, \mathbf{x}_i),\ 0\right) \tag{11}$$

Similar to Assumption 3.2.1 (cf §3.2), to construct an approximate teaching set we assume a target model $\boldsymbol{\theta}^*$ has the property that for some truncated polynomial space $\mathcal{H}_{\tilde{\mathcal{K}}}$ defined by feature map $\tilde{\Phi}$ there are linearly independent projections in the orthogonal complement of $\mathbb{P}\boldsymbol{\theta}^*$ in $\mathcal{H}_{\tilde{\mathcal{K}}}$. More formally, we state the property as an assumption which is discussed in details in the supplemental materials.

Assumption 3.4.1 (Existence of orthogonal classifiers)**.** *For the target model $\boldsymbol{\theta}^*$ and some $\epsilon > 0$, we assume that there exists $r = r(\boldsymbol{\theta}^*, \epsilon)$ such that $\mathbb{P}\boldsymbol{\theta}^*$ has $r-1$ linear independent projections on the orthogonal subspace of $\mathbb{P}\boldsymbol{\theta}^*$ in $\mathcal{H}_{\tilde{\mathcal{K}}}$ of the form $\{\tilde{\Phi}(\mathbf{z}_i)\}_{i=1}^{r-1}$ such that $\forall i$ $\mathbf{z}_i \in \mathcal{X}$.*

For the analysis of the key results, we impose a smoothness condition on the linear independent projections $\{\tilde{\Phi}(\mathbf{z}_i)\}_{i=1}^{r-1}$ that they are oriented away by a factor of $\frac{1}{r-1}$. Concretely, for any i, j $\left|\tilde{\Phi}(\mathbf{z}_i) \cdot \tilde{\Phi}(\mathbf{z}_j)\right| \leq \frac{1}{2(r-1)}$. This smoothness condition is discussed in the supplemental. Now, we consider the following reformulation of the optimization problem in Eq. (11) as follows:

$$\boldsymbol{\mathcal{A}}_{opt} := \arg\min_{\beta_0 \in \mathbb{R},\, \gamma \in \mathbb{R}^{r-1}} \sum_{i=1}^{2r-1} \max\left(\ell(\beta_0, \gamma, \mathbf{x}_i, y_i),\ 0\right) \tag{12}$$

where for any $i \in [2r-1]$

$$\ell(\beta_0, \gamma, \mathbf{x}_i, y_i) = -y_i \cdot \left(\beta_0 \cdot \mathcal{K}(\mathbf{a}, \mathbf{x}_i) + \sum_{j=1}^{r-1} \gamma_j \cdot \mathcal{K}(\mathbf{z}_j, \mathbf{x}_i)\right)$$

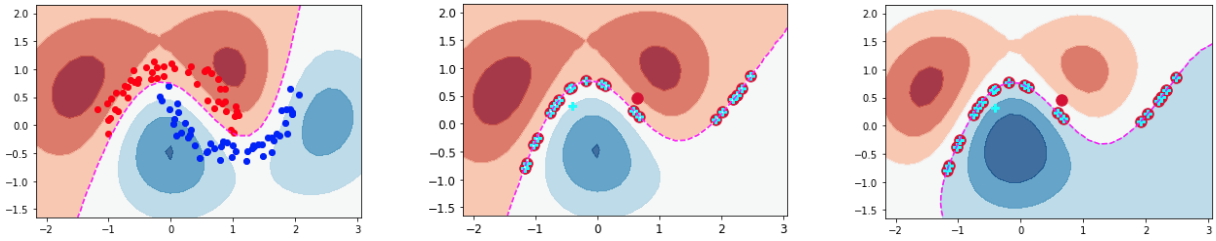
and with respect to the teaching set

$$\mathcal{TS}_{\boldsymbol{\theta}^*} := \left\{(\mathbf{z}_i, 1),\ (\mathbf{z}_i, -1)\right\}_{i=1}^{r-1} \cup \left\{(\mathbf{a}, 1)\right\} \tag{13}$$

where $\mathbf{a}$ is chosen such that $\mathbb{P}\boldsymbol{\theta}^* \cdot \mathbb{P}\Phi(\mathbf{a}) > 0$[3] and $\Phi(\mathbf{a}) \cdot \Phi(\mathbf{z}_i) \leq Q \cdot \epsilon$ (where Q is a constant). $\mathbf{a}$ could be chosen from a $\mathcal{B}(\sqrt{2\sqrt{R}\sigma^2}, 0)$ spherical ball in $\mathbb{R}^d$. We index the set $\mathcal{TS}_{\boldsymbol{\theta}^*}$ as $\{(\mathbf{x}_i, y_i)\}_{i=1}^{2r-1}$. Eq. (12) is optimized over $\hat{\boldsymbol{\theta}} = \beta_0 \cdot \mathcal{K}(\mathbf{a}, \cdot) + \sum_{j=1}^{r-1} \gamma_j \cdot \mathcal{K}(\mathbf{z}_j, \cdot)$ such that $\hat{\boldsymbol{\theta}} \cdot \Phi(\mathbf{a}) > 0$ and $\{\Phi(\mathbf{z}_i)\}_{i=1}^{r-1}$ satisfy Assumption 3.4.1 where $\left|\left|\hat{\boldsymbol{\theta}}\right|\right| = \mathcal{O}(1)$.

[2] When $R = d$ all the key results follow the same analysis.

[3] We assume $\boldsymbol{\theta}^*$ is non-degenerate in $\tilde{\mathcal{K}}$ (as for polynomial



(a) Optimal Gaussian boundary (b) Polynomial approximation (c) Taught Gaussian boundary

Figure 3: Approximate teaching for Gaussian kernel perceptron. (a) Teacher "receives" $\boldsymbol{\theta}^*$ by training from the complete data set; (b) Teacher identifies a polynomial approximation of the Gaussian decision boundary and generates the teaching set $\mathcal{TS}_{\boldsymbol{\theta}^*}$ (marked by red dots and cyan crosses); (c) Learner learns a Gaussian kernel perceptron from $\mathcal{TS}_{\boldsymbol{\theta}^*}$.

Note that any solution to Eq. (12) can have unbounded norm and can extend in arbitrary directions, thus we make an assumption on the learner which would be essential to bound the error of optimal separator of Eq. (12).

Assumption 3.4.2 (Bounded Cone)**.** *For the target model* $\boldsymbol{\theta}^* = \sum_{i=1}^{l} \alpha_i \cdot \mathcal{K}(\mathbf{a}_i, \cdot)$*, the learner optimizes to a solution* $\hat{\boldsymbol{\theta}}$ *for Eq. (12) with bounded coefficients. Alternatively, the sums* $\sum_{i=1}^{l} |\alpha_i|$ *and* $|\beta_0| + \sum_{j=1}^{r-1} |\gamma_j|$ *are bounded where* $\hat{\boldsymbol{\theta}} \in \mathcal{H}_{\mathcal{K}}$ *has the form* $\hat{\boldsymbol{\theta}} = \beta_0 \cdot \mathcal{K}(\mathbf{a}_j, \cdot) + \sum_{j=1}^{r-1} \gamma_j \cdot \mathcal{K}(\mathbf{z}_j, \cdot)$.

This assumption is fairly mild or natural in the sense that for $\hat{\boldsymbol{\theta}} \in \boldsymbol{\mathcal{A}}_{opt}$ as a classifier approximates $\boldsymbol{\theta}^*$ point-wise then they shouldn't be highly (or unboundedly) sensitive to datapoints involved in the classifiers. It is discussed in greater details in the supplemental materials. We denote by $\boldsymbol{C}_\epsilon := \sum_{i=1}^{l} |\alpha_i|$ and $\boldsymbol{D}_\epsilon := |\beta_0| + \sum_{j=1}^{r-1} |\gamma_j|$. In the supplemental materials, we show that there exists a unique solution (upto a positive scaling) to Eq. (12) which satisfies Assumption 3.4.2. We would show that $\mathcal{TS}_{\boldsymbol{\theta}^*}$ is an ϵ-approximate teaching set with $r = d^{\mathcal{O}\left(\log^2 \frac{1}{\epsilon}\right)}$ on the ϵ-approximate teaching dimension. To achieve this, we first establish the *ϵ-closeness* of $\hat{f}$ (classifier $\hat{f}(\mathbf{x}) := \hat{\boldsymbol{\theta}} \cdot \Phi(\mathbf{x})$ where $\hat{\boldsymbol{\theta}} \in \boldsymbol{\mathcal{A}}_{opt}$) to f^*. Formally, we state the result as follows:

Theorem 3. *For any target* $\boldsymbol{\theta}^* \in \mathcal{H}_{\mathcal{K}}$ *that satisfies Assumption 3.4.1-3.4.2 and* $\epsilon > 0$*, the teaching set* $\mathcal{TS}_{\boldsymbol{\theta}^*}$ *constructed for Eq. (12) satisfies* $\left| f^*(\mathbf{x}) - \hat{f}(\mathbf{x}) \right| \leq \epsilon$ *for any* $\mathbf{x} \in \mathcal{X}$ *and any* $\hat{f} \in \boldsymbol{\mathcal{A}}_{opt}(\mathcal{TS}_{\boldsymbol{\theta}^*})$.

Using Theorem 3, we can obtain the main result of the subsection which gives an $d^{\mathcal{O}\left(\log^2 \frac{1}{\epsilon}\right)}$ bound on ϵ-approximate teaching dimension. We detail the proofs in the supplemental materials:

kernels in §3.2) i.e. has points $\mathbf{a} \in \mathcal{X}$ such that $\mathbb{P}\boldsymbol{\theta}^* \cdot \mathbb{P}\Phi(\mathbf{a}) > 0$ (classified with label 1).

Theorem 4. *For any target* $\boldsymbol{\theta}^* \in \mathcal{H}_{\mathcal{K}}$ *that satisfies Assumption 3.4.1-3.4.2 and* $\epsilon > 0$*, the teaching set* $\mathcal{TS}_{\boldsymbol{\theta}^*}$ *constructed for Eq. (12) is an ϵ-approximate teaching set with* ϵ*-*$TD(\boldsymbol{\theta}^*, \boldsymbol{\mathcal{A}}_{opt}) = d^{\mathcal{O}\left(\log^2 \frac{1}{\epsilon}\right)}$ *i.e. for any* $\hat{f} \in \boldsymbol{\mathcal{A}}_{opt}(\mathcal{TS}_{\boldsymbol{\theta}^*})$,

$$\left| \boldsymbol{err}(f^*) - \boldsymbol{err}(\hat{f}) \right| \leq \epsilon.$$

Numerical example Fig. 3 demonstrates the approximate teaching process for a Gaussian learner. We aim to teach the optimal model $\boldsymbol{\theta}^*$ (infinite-dimensional) trained on a pre-collected dataset with Gaussian parameter $\sigma = 0.9$, whose corresponding boundary is shown in Fig. 3a. Now, for approximate teaching, the teacher calculates $\tilde{\boldsymbol{\theta}}$ using the polynomial approximated kernel (i.e. $\tilde{\mathcal{K}}$, and in this case, k=5) in Eq. (8) and the corresponding feature map in Eq. (9). To ensure Assumption 3.4.1 is met while generating teaching examples for $\tilde{\boldsymbol{\theta}}$, we employ the randomized algorithm (as was used in §3.2) with the key idea of ensuring that the teaching examples on the boundary are linearly independent in the approximated polynomial feature space, i.e. $\tilde{\mathcal{K}}(\mathbf{z}_i, \mathbf{z}_j) = 0$. Finally, the Gaussian learner receives $\mathcal{TS}_{\boldsymbol{\theta}^*}$ and learns the boundary shown in Fig. 3c. Note the slight difference between the boundaries in Fig. 3b and in Fig. 3c as the learner learns with a Gaussian kernel.

4 Conclusion

We have studied and extended the notion of teaching dimension for optimization-based perceptron learner. We also studied a more general notion of approximate teaching which encompasses the notion of exact teaching. To the best of our knowledge, our exact teaching dimension for linear and polynomial perceptron learner is new; so is the upper bound on the approximate teaching dimension of Gaussian kernel perceptron learner and our analysis technique in general. There are many

possible extensions to the present work. For example, one may extend our analysis to relaxing the assumptions imposed on the data distribution for polynomial and Gaussian kernel perceptrons. This can potentially be achieved by analysing the linear perceptron and finding ways to nullify subspaces other than orthogonal vectors. This could enhance the results for both the exact teaching of polynomial perceptron learner to more general case and a tighter bound on the approximate teaching dimension of Gaussian kernel perceptron learner. On the other hand, a natural extension of our work is to understand the approximate teaching complexity for other types of ERM learners, e.g. kernel SVM, kernel ridge, and kernel logistic regression. We believe the current work and its extensions would enrich our understanding of optimal and approximate teaching and enable novel applications.

5 Acknowledgements

Yuxin Chen is supported by NSF 2040989 and a C3.ai DTI Research Award 049755.

References

Anthony, M., Brightwell, G., and Shawe-Taylor, J. On specifying boolean functions by labelled examples. *Discrete Applied Mathematics*, 61:1–25, 07 1995. doi: 10.1016/0166-218X(94)00007-Z.

Cakmak, M. and Lopes, M. Algorithmic and human teaching of sequential decision tasks. In *AAAI*, 2012.

Chen, Y., Singla, A., Mac Aodha, O., Perona, P., and Yue, Y. Understanding the role of adaptivity in machine teaching: The case of version space learners. In *Advances in Neural Information Processing Systems*, pp. 1476–1486, 2018.

Cotter, A., Keshet, J., and Srebro, N. Explicit approximations of the gaussian kernel. *CoRR*, abs/1109.4603, 2011.

Cucker, F. and Smale, S. On the mathematical foundations of learning. *BULLETIN*, 39, 11 2001. doi: 10.1090/S0273-0979-01-00923-5.

Devidze, R., Mansouri, F., Haug, L., Chen, Y., and Singla, A. Understanding the power and limitations of teaching with imperfect knowledge. In *IJCAI*, 2020.

Doliwa, T., Fan, G., Simon, H. U., and Zilles, S. Recursive teaching dimension, vc-dimension and sample compression. *JMLR*, 15(1):3107–3131, 2014.

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

Ha Quang, M. Some properties of gaussian reproducing kernel hilbert spaces and their implications for function approximation and learning theory. *Constructive Approximation - CONSTR APPROX*, 32: 307–338, 10 2010. doi: 10.1007/s00365-009-9080-0.

Haug, L., Tschiatschek, S., and Singla, A. Teaching inverse reinforcement learners via features and demonstrations. In *Advances in Neural Information Processing Systems*, pp. 8464–8473, 2018.

Hunziker, A., Chen, Y., Aodha, O. M., Rodriguez, M. G., Krause, A., Perona, P., Yue, Y., and Singla, A. Teaching multiple concepts to a forgetful learner. In *Advances in Neural Information Processing Systems*, pp. 4050–4060, 2019.

Kamalaruban, P., Devidze, R., Cevher, V., and Singla, A. Interactive teaching algorithms for inverse reinforcement learning. In *IJCAI*, pp. 2692–2700, 2019.

Kirkpatrick, D., Simon, H. U., and Zilles, S. Optimal collusion-free teaching. In *Proceedings of the 30th International Conference on Algorithmic Learning Theory*, volume 98, pp. 506–528, 2019.

Liang, S. and Srikant, R. Why deep neural networks for function approximation? In *ICLR*, 2017.

Liu, J. and Zhu, X. The teaching dimension of linear learners. *Journal of Machine Learning Research*, 17 (162):1–25, 2016.

Liu, W., Dai, B., Humayun, A., Tay, C., Yu, C., Smith, L. B., Rehg, J. M., and Song, L. Iterative machine teaching. In *ICML*, pp. 2149–2158, 2017.

Liu, W., Dai, B., Li, X., Liu, Z., Rehg, J. M., and Song, L. Towards black-box iterative machine teaching. In *ICML*, pp. 3147–3155, 2018.

Lu, Y. and Lu, J. A universal approximation theorem of deep neural networks for expressing distributions, 2020.

Mansouri, F., Chen, Y., Vartanian, A., Zhu, J., and Singla, A. Preference-based batch and sequential teaching: Towards a unified view of models. In *Advances in Neural Information Processing Systems*, pp. 9195–9205, 2019.

Rakhsha, A., Radanovic, G., Devidze, R., Zhu, X., and Singla, A. Policy teaching via environment poisoning: Training-time adversarial attacks against reinforcement learning. In *ICML*, volume 119, pp. 7974–7984, 2020.

Scholkopf, B. and Smola, A. J. *Learning with Kernels: Support Vector Machines, Regularization, Optimization, and Beyond*. MIT Press, Cambridge, MA, USA, 2001. ISBN 0262194759.

Singla, A., Bogunovic, I., Bartók, G., Karbasi, A., and Krause, A. On actively teaching the crowd to classify. In *NIPS Workshop on Data Driven Education*, 2013.

Singla, A., Bogunovic, I., Bartók, G., Karbasi, A., and Krause, A. Near-optimally teaching the crowd to classify. In *ICML*, pp. 154–162, 2014.

Tschiatschek, S., Ghosh, A., Haug, L., Devidze, R., and Singla, A. Learner-aware teaching: Inverse reinforcement learning with preferences and constraints. In *Advances in Neural Information Processing Systems*, 2019.

Vapnik, V. *Statistical Learning Theory*. Adaptive and learning systems for signal processing, communications, and control. Wiley, 1998. ISBN 9788126528929.

Yarotsky, D. Error bounds for approximations with deep relu networks. *Neural networks : the official journal of the International Neural Network Society*, 94:103–114, 2017.

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. Teaching dimensions based on cooperative learning. In *COLT*, pp. 135–146, 2008.