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.

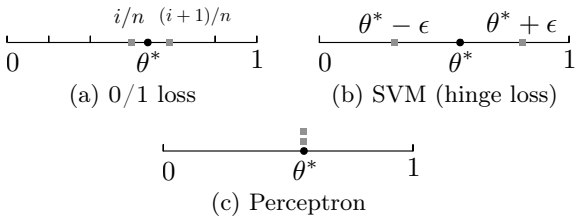
1 Introduction

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:

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

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

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

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 Appendix A.

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 $\mathbf{x}, \mathbf{x}' \in \mathbb{R}^d$

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

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 Appendix B.

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 $t > 0$*

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

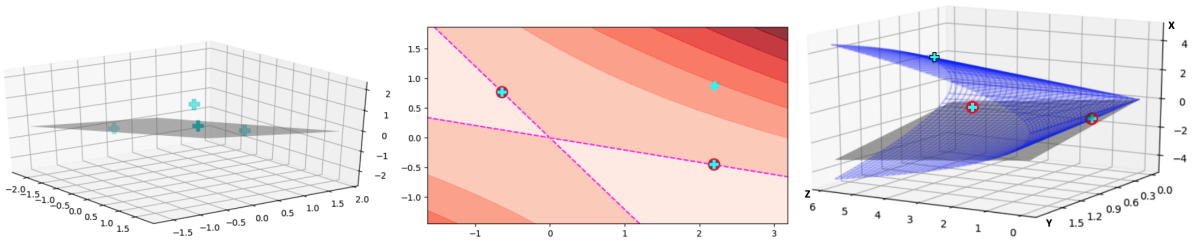
Lemma 1 shows that for *exact* teaching $\boldsymbol{\theta}^*$ should satisfy Assumption 3.2.1. Then, the natural question that arises is whether we can achieve arbitrarily ϵ-close *approximate* teaching for $\boldsymbol{\theta}^*$. In other words, we would like to find $\tilde{\boldsymbol{\theta}}^*$ 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 Appendix C. 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 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.



(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.

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

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

where $\mathcal{C}_{\boldsymbol{\lambda}}^k = \frac{k!}{\prod_{i=1}^d \boldsymbol{\lambda}_i!}$. Thus, we use Eq. (7) to obtain explicit feature representation to the Gaussian kernel in Eq. (6) as $\Phi_{k,\boldsymbol{\lambda}}(\mathbf{x}) = e^{-\frac{||\mathbf{x}||^2}{2\sigma^2}} \cdot \frac{\sqrt{\mathcal{C}_{\boldsymbol{\lambda}}^k}}{\sqrt{k!}\sigma^k} \cdot \mathbf{x}^{\boldsymbol{\lambda}}$. 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:

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

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 Appendix D, 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 Appendix C.

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 $i \in [2r-1]$

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

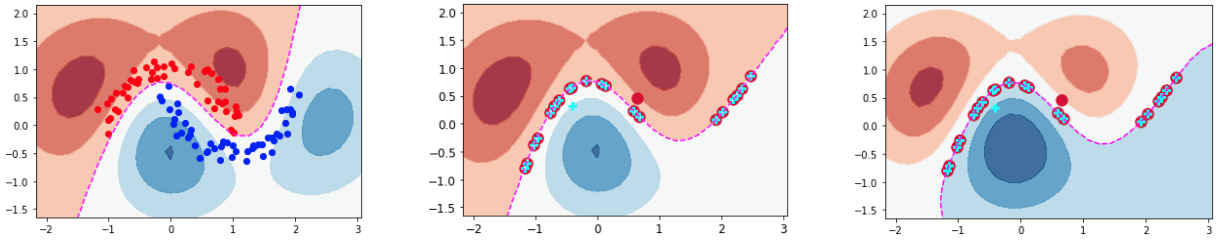
and with respect to the teaching set

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

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 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).



(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 $\boldsymbol{\theta}^*$ pointwise then they shouldn't be highly (or unboundedly) sensitive to datapoints involved in the classifiers. It is discussed in greater details in Appendix C. 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 Appendix D.1, 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 Appendix D:

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.

Dutta, J. and C. S., L. Bounded sets of kkt multipliers in vector optimization. *Journal of Global Optimization*, 36:425–437, 11 2006. doi: 10.1007/s10898-006-9019-y.

Gauvin, J. A necessary and sufficient regularity condition to have bounded multipliers in nonconvex programming. *Math. Program.*, 12(1):136–138, December 1977. ISSN 0025-5610. doi: 10.1007/BF01593777.

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.

Luksan, L., Matonoha, C., and Vlček, J. Interior point methods for large-scale nonlinear programming. *Optimization Methods and Software*, 20:569–582, 08 2005. doi: 10.1080/10556780500140508.

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.

Movahedian, N. Bounded lagrange multiplier rules for general nonsmooth problems and application to

mathematical programs with equilibrium constraints. *Journal of Global Optimization*, 67, 05 2016. doi: 10.1007/s10898-016-0442-4.

Nguyen, V., Strodiot, J.-J., and Mifflin, R. On conditions to have bounded multipliers in locally lipschitz programming. *Mathematical Programming*, 18:100–106, 12 1980. doi: 10.1007/BF01588302.

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.

A List of Appendices

First, we provide the proofs of our theoretical results in full detail in the subsequent sections. We follow these by the experimental evaluations section. The appendices are summarized as follows:

- Appendix B provides the proof of Theorem 2
- Appendix C provides the motivations and key insights into Assumptions [3.2.1, 3.4.1, 3.4.2]
- Appendix D provides relevant results and proofs of Theorem 3 and Theorem 4 (Approximate Teaching Set for Gaussian Learner)
- Appendix E provides the experimental evaluations for the theoretical results on various datasets

B Polynomial Kernel Perceptron

In this appendix, we would provide the proof for the main result of §3.2 i.e. Theorem 2. We would complete the proof by constructing a teaching set for exact teaching. Similar to the proof of Theorem 1, the key idea is to find linear independent polynomials on the orthogonal subspace defined by $\boldsymbol{\theta}^* \in \mathcal{H}_{\mathcal{K}}$. Our Assumption 3.2.1 would ensure that there are such $r-1$ linear independent polynomials. Rest of the work follows steps inspired as seen in the proof of Theorem 1. We assume that $\boldsymbol{\theta}^*$ is non-degenerate and has at least one point in $\mathbb{R}^d$ classified *strictly* positive, and provide the poof below.

Proof of Theorem 2. First, we would show the construction of a teaching set for a target model $\boldsymbol{\theta}^* \in \mathcal{H}_{\mathcal{K}}$. Denote by $\mathcal{V}_{\boldsymbol{\theta}^*}^{\perp} \subset \mathcal{H}_k$ ($\cong \mathcal{H}_{\mathcal{K}}$ using Proposition 1 i.e isomorphic as vector spaces) the orthogonal subspace of $\boldsymbol{\theta}^*$. Since $\boldsymbol{\theta}^*$ satisfies Assumption 3.2.1, thus there exists a set of $r-1$ linearly independent vectors (polynomials because of Proposition 1) of the form $\{\Phi(\mathbf{z}_i)\}_{i=1}^{r-1}$ in $\mathcal{V}_{\boldsymbol{\theta}^*}^{\perp}$ where $\{\mathbf{z}_i\}_{i=1}^{r-1} \in \mathbb{R}^d$. Note that $\boldsymbol{\theta}^* \cdot \Phi(\mathbf{z}_i) = 0$. Now, pick $\mathbf{a} \in \mathbb{R}^d$ such that $\boldsymbol{\theta}^* \cdot \Phi(\mathbf{a}) > 0$ (assuming non-degeneracy). We note that $\{(\mathbf{z}_i, 1)\}_{i=1}^{r-1} \cup \{(\mathbf{z}_i, -1)\}_{i=1}^{r-1} \cup \{(\mathbf{a}, 1)\}$ forms a teaching set for the decision boundary corresponding to $\boldsymbol{\theta}^*$. Using similar ideas from the proof of Theorem 1, we notice that any solution $\hat{\boldsymbol{\theta}}$ to Eq. (3) satisfies $\boldsymbol{\theta}^* \cdot \Phi(\mathbf{z}_i) = 0$ for the labelled datapoints corresponding to $\Phi(\mathbf{z}_i)$. Thus, $\hat{\boldsymbol{\theta}}$ doesn't have any component along $\Phi(\mathbf{z}_i)$. Eq. (3) is minimized if $\hat{\boldsymbol{\theta}} \cdot \Phi(\mathbf{a}) \geq 0$ implying $\hat{\boldsymbol{\theta}} = t\boldsymbol{\theta}^*$. Thus, under Assumption 3.2.1, we show an upper bound $\mathcal{O}\left(\binom{d+k-1}{k}\right)$ on the size of a teaching set for $\boldsymbol{\theta}^*$. □

C Motivation for Assumptions

In this appendix, we discuss the motivations and insights for the key Assumptions [3.2.1, 3.4.1, 3.4.2] made in §3.2 and §3.4. This appendix is organized in the following way: Appendix C.1 discusses Assumption 3.2.1 and provides the proofs of Lemma 1 and Lemma 2 in the context of polynomial kernel (see §3.2); Appendix C.2 discusses the Assumption 3.4.1 and Assumption 3.4.2 in the context of Gaussian kernel perceptron (see §3.4).

Reformulation of a model $\boldsymbol{\theta}$ as a polynomial form As noted in §2, we consider the reproducing kernel Hilbert space (Scholkopf & Smola, 2001) $\mathcal{H}_{\mathcal{K}}$ which could be spanned by the linear combinations of kernel functions of the form $\mathcal{K}(\mathbf{x}, \cdot)$. More concretely,

$$\mathcal{H}_{\mathcal{K}} = \left\{ \sum_{i=1}^{m} \alpha_i \cdot \mathcal{K}(\mathbf{x}_i, \cdot) \ : \ m \in \mathbb{N},\, \mathbf{x}_i \in \mathcal{X},\, \alpha_i \in \mathbb{R}, i = 1, \cdots, m \right\}.$$

Thus, we could write any model $f_{\boldsymbol{\theta}} \in \mathcal{H}_{\mathcal{K}}$ (parametrized by $\boldsymbol{\theta}$) as $\sum_{i=1}^{n} \alpha_i \cdot \mathcal{K}(\mathbf{x}_i, \cdot)$ for some $n \in \mathbb{N}$, $\mathbf{x}_i \in \mathcal{X}$ for $i \in [n]$. This interesting because if $\mathcal{K}(\cdot, \cdot)$ is a polynomial kernel of degree k, then

$$f_{\boldsymbol{\theta}}(\mathbf{x}) = \sum_{i=1}^{n} \alpha_i \cdot \mathcal{K}(\mathbf{x}_i, \mathbf{x}) = \sum_{i=1}^{n} \alpha_i \cdot \langle \mathbf{x}_i,\, \mathbf{x} \rangle^k = \sum_{i=1}^{n} \alpha_i \cdot (\mathbf{x}_{i1}\mathbf{x}_1 + \cdots + \mathbf{x}_{id}\mathbf{x}_d)^k \tag{14}$$

where $\mathbf{x}_i = (\mathbf{x}_{i1}, \cdots, \mathbf{x}_{id})$. Thus, $f_{\boldsymbol{\theta}}(\cdot)$ could be reformulated as a homogeneous polynomial of degree k in d variables. Notice that for polynomial kernel in §3.2, for a target model $\boldsymbol{\theta}^*$ we study the orthogonal projections of the form $\Phi(\mathbf{x})$ for $\mathbf{x} \in \mathcal{X}$ such that $\boldsymbol{\theta}^* \cdot \Phi(\mathbf{x}) = 0$. Alternatively, using Eq. (14) we wish to solve the polynomial equation:

$$f_{\boldsymbol{\theta}^*}(\mathbf{x}) = 0 \implies \sum_{i=1}^{n} \alpha_i \cdot (\mathbf{x}_{i1}\mathbf{x}_1 + \cdots + \mathbf{x}_{id}\mathbf{x}_d)^k = 0$$

where we denote $\boldsymbol{\theta}^* := \sum_{i=1}^{n} \alpha_i \cdot \Phi(\mathbf{x}_i)$. For Assumption 3.2.1 we wish to find $\binom{d+k-1}{k}$ real solutions of the form $\mathbf{x}' \in \mathbb{R}^d$, of this equation which are linearly independent. It is well-studied in the literature of polynomial algebra that this equation might not satisfy the required assumption. We construct one such model for the proof of Lemma 2. This reformulation can be extended for sum of polynomial kernels of the form $\sum_{j=1}^{s} c_j \cdot \langle \mathbf{x},\, \mathbf{x}' \rangle^j$ where $c_j \geq 0$. In Assumption 3.4.1 the reformulation reduces to a variant of the above polynomial equation i.e.

$$\sum_{i=1}^{n} \alpha_i \left(\sum_{j=1}^{s} c_i \cdot (\mathbf{x}_{i1}\mathbf{x}_1 + \cdots + \mathbf{x}_{id}\mathbf{x}_d)^j \right) = 0$$

So far, we discussed a characterization of the notion of orthogonality for a target model $\boldsymbol{\theta}^*$ in the form of a polynomial equation. This characterization would help us understand Assumption 3.2.1 and Assumption 3.4.1. In §C.1, we discuss that Assumption 3.2.1 is the most natural step to make for exact teaching of a target model.

C.1 Limitation of Exact Teaching: Polynomial Kernel Perceptron

In this subsection, we provide the proofs of Lemma 1 and Lemma 2 as stated in §3.3. These results establish that in the realizable setting, Assumption 3.2.1 is required for exact teaching: Lemma 1. Furthermore, there are pathological cases where violation of the assumption leads to models which couldn't be approximately taught: Lemma 2.

Proof of Lemma 1. We would prove the result by contradiction. Assume that $\mathcal{TS}_{\boldsymbol{\theta}^*}$ be a teaching set which *exactly* teaches $\boldsymbol{\theta}^*$. WLOG we enumerate the teaching set as $\mathcal{TS}_{\boldsymbol{\theta}^*} = \{(\mathbf{x}_1, y_1), \cdots, (\mathbf{x}_n, y_n)\}$. For the sake of clarity, we would rewrite (3) again

$$\boldsymbol{\mathcal{A}}_{opt}(\mathcal{TS}_{\boldsymbol{\theta}^*}) := \operatorname*{arg\,min}_{\boldsymbol{\theta} \in \mathcal{H}_{\mathcal{K}}} \sum_{i=1}^{n} \max(-y_i \cdot \boldsymbol{\theta}^* \cdot \Phi(\mathbf{x}_i), 0) \tag{15}$$

Denote by $\mathcal{V}_{\boldsymbol{\theta}^*}^{\perp} \subset \mathcal{H}_k$ the orthogonal subspace of $\boldsymbol{\theta}^*$. We denote the objective value of Eq. (15) by $\mathbf{p}(\boldsymbol{\theta}) := \sum_{i=1}^{n} \max(-y_i \cdot \boldsymbol{\theta} \cdot \Phi(\mathbf{x}_i),\, 0)$. We further define *effective direction of a teaching point* $(\mathbf{x}_i, y_i) \in \mathcal{TS}_{\boldsymbol{\theta}^*}$ in the

RKHS $\mathcal{H}_\mathcal{K}$ as $\boldsymbol{d}_i := -y_i \cdot \Phi(\mathbf{x}_i)$. Because of the realizable setting i.e. all teaching points are correctly classified, it is clear that

$$-y_i \cdot \boldsymbol{\theta}^* \cdot \Phi(\mathbf{x}_i) \leq 0 \implies \boldsymbol{\theta}^* \cdot \boldsymbol{d}_i \leq 0.$$

Since $\boldsymbol{\theta}^*$ violates Assumption 3.2.1, thus $\exists$ a unit normalized direction $\hat{\boldsymbol{d}} \in \mathcal{V}_{\boldsymbol{\theta}^*}^\perp$ which can't be spanned by $\mathcal{S}_0 \triangleq \left\{\Phi(\mathbf{x}) \,:\, \Phi(\mathbf{x}) \in \mathcal{V}_{\boldsymbol{\theta}^*}^\perp \text{ for some } \mathbf{x} \in \mathcal{X}\right\}$. such that $\hat{\boldsymbol{d}} \perp \textbf{span}\left\langle \mathcal{S}_0 \right\rangle$. Now, we would show that $\exists \lambda > 0$ (real) such that

$$\left(\boldsymbol{\theta}^* + \lambda \hat{\boldsymbol{d}}\right) \in \boldsymbol{\mathcal{A}}_{opt}(\mathcal{TS}_{\boldsymbol{\theta}^*}) \tag{16}$$

Notice that for some i if $\boldsymbol{d}_i \in \mathcal{V}_{\boldsymbol{\theta}^*}^\perp$ then $(\boldsymbol{\theta}^* + \lambda \hat{\boldsymbol{d}}) \cdot \boldsymbol{d}_i = \boldsymbol{\theta}^* \cdot \boldsymbol{d}_i \leq 0$. Now, we consider the case when $\boldsymbol{d}_i \notin \mathcal{V}_{\boldsymbol{\theta}^*}^\perp$. We could expand $\boldsymbol{d}_i$ as follows:

$$\boldsymbol{d}_i = a_i \hat{\boldsymbol{d}}^\perp + b_i \hat{\boldsymbol{d}} \tag{17}$$

where a_i and b_i are real scalars and $\hat{\boldsymbol{d}}^\perp$ is normalized orthogonal projection of $\boldsymbol{d}_i$ to orthogonal complement (orthogonal subspace) of $\hat{\boldsymbol{d}}$. These constructions are illustrated in Fig. 4a. Now, we would compute the following dot product:

$$\begin{aligned}
&(\boldsymbol{\theta}^* + \lambda \hat{\boldsymbol{d}}) \cdot \boldsymbol{d}_i \\
\implies &\boldsymbol{\theta}^* \cdot \boldsymbol{d}_i + \lambda \hat{\boldsymbol{d}} \cdot (a_i \hat{\boldsymbol{d}}^\perp + b_i \hat{\boldsymbol{d}}) &(18)\\
\implies &\boldsymbol{\theta}^* \cdot \boldsymbol{d}_i + \lambda \hat{\boldsymbol{d}} \cdot b_i \hat{\boldsymbol{d}} &(19)
\end{aligned}$$

Eq. (18) follows using Eq. (17). In Eq. (19) we note that $\hat{\boldsymbol{d}} \perp \hat{\boldsymbol{d}}^\perp$. If $b_i \leq 0$ then $(\boldsymbol{\theta}^* + \lambda \hat{\boldsymbol{d}}) \cdot \boldsymbol{d}_i \leq 0$ as $\boldsymbol{\theta}^* \cdot \boldsymbol{d}_i < 0$. If $b_i > 0$, then to ensure $(\boldsymbol{\theta}^* + \lambda \hat{\boldsymbol{d}}) \cdot \boldsymbol{d}_i \leq 0$, using Eq. (19) we need

$$\lambda \leq \frac{-\boldsymbol{\theta}^* \cdot \boldsymbol{d}_i}{b_i}$$

Since, i is chosen arbitrarily thus for all the effective directions $\boldsymbol{d}_i \notin \mathcal{V}_{\boldsymbol{\theta}^*}^\perp$ where $b_i > 0$, we pick postive scalar λ such that:

$$\lambda := \min_{i:\, b_i > 0} \frac{-\boldsymbol{\theta}^* \cdot \boldsymbol{d}_i}{b_i}$$

For this choice of λ we show that $\boldsymbol{\theta}^* + \lambda \hat{\boldsymbol{d}} \in \boldsymbol{\mathcal{A}}_{opt}(\mathcal{TS}_{\boldsymbol{\theta}^*})$. Thus, by definition, $\mathcal{TS}_{\boldsymbol{\theta}^*}$ as stated above can't teach $\boldsymbol{\theta}^*$ exactly. Hence, if $\boldsymbol{\theta}^*$ violates Assumption 3.2.1 then we can't teach it exactly in the realizable setting. □

Now, we provide the proof of Lemma 2 for which we give a construction of a model $\boldsymbol{\theta}^*$ which violates Assumption 3.2.1 and also show that it can't be taught arbitrarily ϵ-close approximately. The proof is also illustrated in Fig. 4b.

Proof of Lemma 2. Assume $\boldsymbol{\theta}^*$ is a target model which violates Assumption 3.2.1. If $\boldsymbol{\theta}^*$ can be taught *approximately* for arbitrarily small $\epsilon > 0$ then $\exists\ \tilde{\boldsymbol{\theta}}^*$ which can be taught *exactly* (i.e. satisfies Assumption 3.2.1) such that

$$\boldsymbol{\theta}^* \cdot \tilde{\boldsymbol{\theta}}^* \geq 1 - \cos a_\epsilon, \text{ where } \cos a_\epsilon = \epsilon$$

if $\boldsymbol{\theta}^*$ and $\tilde{\boldsymbol{\theta}}^*$ are unit normalized. This implies that if $\Phi(\mathbf{x}) \in \mathcal{V}_{\tilde{\boldsymbol{\theta}}^*}^\perp \subset \mathcal{H}_\mathcal{K}$ (orthogonal complement of $\tilde{\boldsymbol{\theta}}^*$) such that $||\Phi(\mathbf{x})|| \leq 1$ then the following holds:

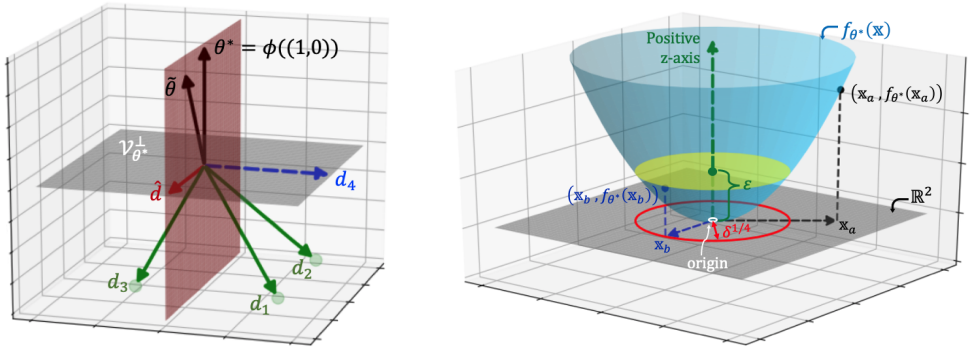
$$|\boldsymbol{\theta}^* \cdot \Phi(\mathbf{x})| \leq \epsilon \tag{20}$$

Alternatively, we could think of $\Phi(\mathbf{x})$ as being almost orthogonal to $\boldsymbol{\theta}^*$.

Now, we would show a construction of a target model when k has parity even, which not only violates Assumption 3.2.1 but can't be taught *approximately* such that Eq. (20) holds. The idea is to find $\boldsymbol{\theta}^*$ which doesn't have almost orthogonal projections in $\mathcal{H}_\mathcal{K}$ with norm lower-bounded by δ.

Consider the following construction for a target model $\boldsymbol{\theta}^* \in \mathcal{H}_\mathcal{K}$:

$$\boldsymbol{\theta}^* = \sum_{i=1}^{d} \frac{1}{\sqrt{d}} \cdot \Phi(\mathbf{e}_i), \quad ||\boldsymbol{\theta}^*|| = 1 \tag{21}$$



(a) (b)

Figure 4: Illustrations for Proofs of Lemma 1 and Lemma 2. (a) For Lemma 1, consider $\boldsymbol{\theta}^*$ as shown. $\mathbf{span}\,\langle \mathcal{S}_0 \rangle$ only covers the direction indicated by the dashed blue arrow. $\tilde{\boldsymbol{\theta}}$ correctly labels not only teaching examples on $\mathcal{V}_{\boldsymbol{\theta}^*}^{\perp}$ and within $\mathbf{span}\,\langle \mathcal{S}_0 \rangle$, but also those not on $\mathcal{V}_{\boldsymbol{\theta}^*}^{\perp}$, e.g. along effective directions d_1, d_2, d_3; (b) To visualize the proof idea of Lemma 2, we demonstrate an example in $\mathbb{R}^2$ with feature space of dimension 3 (where $k = 2$). We consider a model $\boldsymbol{\theta}^* = \frac{1}{\sqrt{2}} \cdot \Phi((1,0)) + \frac{1}{\sqrt{2}} \cdot \Phi((0,1))$. Since k is even, each point $\mathbf{x}$ in $\mathbb{R}^2$ corresponds to a non-negative value of $f_{\boldsymbol{\theta}^*}(\mathbf{x})$. This function is plotted as the blue surface in (b) along z-axis. Yellow surface represents a threshold of ϵ along z-axis; thus any point above it has value more than ϵ. A red δ-norm ring on the $\mathbb{R}^2$-plane denotes the constraint on the norm of a teaching point ($||\Phi(\mathbf{x})|| = \delta \implies ||\mathbf{x}|| = \delta^{1/4}$). If we are constrained to select only points outside of the red δ-norm ring, then the plot illustrates a situation where no points outside the ring satisfies $f_{\boldsymbol{\theta}^*}(\mathbf{x}) = \boldsymbol{\theta}^* \cdot \Phi(\mathbf{x}) < \epsilon$.

where $\{\mathbf{e}_i\}$'s form the standard basis in $\mathbb{R}^d$. Notice that for any $\mathbf{x} \in \mathcal{X}$,

$$\boldsymbol{\theta}^* \cdot \Phi(\mathbf{x}) = \sum_{i=1}^{d} \frac{1}{\sqrt{d}} \cdot \Phi(\mathbf{e}_i) \cdot \Phi(\mathbf{x}) = \sum_{i=1}^{d} \frac{1}{\sqrt{d}} \cdot \mathbf{x}_i^k. \tag{22}$$

RHS of the above equation is zero only when all the $\mathbf{x}_i^k = 0$ since k is even. Thus, the only projection orthogonal to $\boldsymbol{\theta}^*$ is the zero projection in $\mathcal{H}_{\mathcal{K}}$, thus violates Assumption 3.2.1. Now, we show that $\boldsymbol{\theta}^*$ as constructed in Eq. (21) can't be taught approximately for arbitrarily small $\epsilon > 0$.

If $\mathbf{x} \in \mathcal{X}$ is such that $||\Phi(\mathbf{x})|| \geq \delta$, then using Hölder's inequality:

$$\sum_{i=1}^{d} \mathbf{x}_i^2 \leq \sum_{i=1}^{d} 1 \cdot \mathbf{x}_i^2 \leq d^{\frac{k-2}{k}} \left(\sum_{i=1}^{d} \left(\mathbf{x}_i^2\right)^{\frac{k}{2}} \right)^{\frac{2}{k}}$$

But we have $\sum_{i=1}^{d} \mathbf{x}_i^2 \geq \delta^{\frac{2}{k}}$. Thus, using Eq. (21)-Eq. (22)

$$\left(\sum_{i=1}^{d} \mathbf{x}_i^k \right) \geq \frac{\delta}{d^{2(k-2)}} \implies \boldsymbol{\theta}^* \cdot \Phi(\mathbf{x}) \geq \frac{\delta}{d^{2k-\frac{7}{2}}}$$

Thus, if $\epsilon < \frac{\delta}{d^{2k-\frac{7}{2}}}$ then $\boldsymbol{\theta}^* \cdot \Phi(\mathbf{x}) > \epsilon$. This implies that $\Phi(\mathbf{x})$ can't be chosen almost orthogonal to $\boldsymbol{\theta}^*$ violating Eq. (20). Hence, $\nexists\, \tilde{\boldsymbol{\theta}}^*$ arbitrarily close to $\boldsymbol{\theta}^*$ which can be taught exactly.

Thus, the construction of $\boldsymbol{\theta}^*$ in Eq. (21) violates Assumption 3.2.1 and can't be taught approximately for arbitrarily small $\epsilon > 0$. □

Is the assumption of lower bound δ restrictive? Now, we would argue that the assumption of a lower bound on the norm of the teaching point for Lemma 2 is only for analysis of the proof presented above. Consider the target model $\boldsymbol{\theta}^*$ constructed in Eq. (21). Consider that $\exists\, \tilde{\boldsymbol{\theta}}^*$ which can be taught exactly using arbitrarily small normed teaching points (i.e. lower bound of δ is violated) such that

$$\boldsymbol{\theta}^* \cdot \tilde{\boldsymbol{\theta}}^* \geq 1 - \cos a_\epsilon, \text{ where } \cos a_\epsilon = \epsilon$$

for arbitrarily small $\epsilon > 0$. Define the teaching set as $\mathcal{TS}_{\tilde{\boldsymbol{\theta}}^*}$. But, even if we unit-normalize all the teaching points, call the normalized set $\mathcal{TS}^{unit}_{\tilde{\boldsymbol{\theta}}^*}$, Eq. (3) is still satisfied. Since in that case for any $(\mathbf{x}_i, y_i) \in \mathcal{TS}^{unit}_{\tilde{\boldsymbol{\theta}}^*}$, Eq. (20) is violated. Hence, violating the assumption of lower bound on the norm of the teaching points doesn't invalidate the claim of Lemma 2.

C.2 Approximate Teaching: Assumption 3.4.1 and Assumption 3.4.2

As noted in §3.4, the teaching dimension of a Gaussian kernel perceptron learner is ∞. This calls for studying these non-linear kernel in the setting of approximate teaching. Inspired by our discussion in the previous subsection, we argue that the underlying assumptions: Assumption 3.4.1 and Assumption 3.4.2 are fairly mild in order to establish strong results stated in Theorem 3 and Theorem 4 (cf. §3.4). This appendix subsection is divided into two paragraphs corresponding to the assumptions as follows:

Existence of orthogonal linear independent projections: Assumption 3.4.1. Notice that the projected polynomial space or the approximated kernel $\tilde{\mathcal{K}}$ is a sum of polynomial kernels. We rewrite Eq. (8) for ease of clarity:

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

If we replace $z = \frac{\langle \mathbf{x}, \mathbf{x}' \rangle}{\sigma^2}$, we could write

$$\tilde{\mathcal{K}}(\mathbf{x}, \mathbf{x}') = \boldsymbol{e}^{-\frac{||\mathbf{x}||^2}{2\sigma^2}} \boldsymbol{e}^{-\frac{||\mathbf{x}'||^2}{2\sigma^2}} \sum_{k=0}^{s} \frac{1}{k!} \cdot z^k$$

Since all the coefficients of the polynomial $\sum_{k=0}^{s} \frac{1}{k!} \cdot z^k$ are positive thus if s is even then $\tilde{\mathcal{K}}(\mathbf{x}, \mathbf{x}') > 0$. Thus, if $\boldsymbol{\theta}^* = \sum_{i=1}^{l} \alpha_i \cdot \mathcal{K}(\mathbf{a}_i, \cdot)$ for some $\{\mathbf{a}_i\}_{i=1}^{l} \subset \mathcal{X}$ such that α_i's are positive then Assumption 3.4.1 would be violated. Hence, there is a class of r values for which the assumption would be violated.

It is straight-forward to note that Lemma 1 could be extended to sum of polynomial kernels. Similar extension for Lemma 2 when the highest degree is of parity even could be established. These results follow by noting the polynomial space of homogeneous polynomials of degree k in d variables is isomorphic (Cucker & Smale, 2001) to the polynomial space of degree k in $(d-1)$ variables. Since Hilbert space $\tilde{\mathcal{K}}$ is a sum of polynomial kernels thus the extended results hold. This implies that there could be pathological cases where $\mathbb{P}\boldsymbol{\theta}^*$ could not be learnt approximately in $\tilde{\mathcal{K}}$. But this poses a problem because most of the information of a model in terms of the eigenvalues of the orthogonal basis of the Gaussian kernel is contained in the starting indices i.e. $\forall k \leq s, \quad \Phi_{k,\boldsymbol{\lambda}}(\mathbf{x}) = \boldsymbol{e}^{-\frac{||\mathbf{x}||^2}{2\sigma^2}} \cdot \frac{\sqrt{\mathcal{C}^k_{\boldsymbol{\lambda}}}}{\sqrt{k!}\sigma^k} \cdot \mathbf{x}^{\boldsymbol{\lambda}}$ where $\sum_{i=1}^{d} \boldsymbol{\lambda}_i = k$. It has been discussed in Appendix D. Since the fixed Hilbert space induced by $\tilde{\mathcal{K}}$ is spanned by these truncated projections, thus Assumption 3.4.1 gives a characterization for approximately teachable models. It is left to be understood if there is a more unified characterization which could incorporate approximately teachable models beyond Assumption 3.4.1.

Boundedness of weights: Assumption 3.4.2. It is fairly natural in the sense that in Theorem 3 we are bounding (approximating) the error values of the function point-wise i.e. f^* (using $\hat{f}$) for a fixed ϵ. If for some $\hat{\boldsymbol{\theta}} \in \mathcal{A}_{opt}$, $\hat{f}$ $(= \hat{\boldsymbol{\theta}} \cdot \Phi(\cdot))$ is unboundedly sensitive to some teaching (training) point, then bounding error becomes stringent. Further, we show that there exists a unique solution up to a positive constant scaling to Eq. (12) which satisfy the assumption in Appendix D.1.

The weights $\{\alpha_i\}_{i=1}^{l}$ and $[\beta_0, \gamma]$ could be thought of as Lagrange multipliers for Gaussian kernel perceptron. Boundedness of the multipliers is a well-studied problem in the mathematical programming and optimization literature (Gauvin, 1977; Movahedian, 2016; Dutta & C. S., 2006; Nguyen et al., 1980). Interestingly, Luksan et al. (2005) demonstrated the importance of the boundedness of the Lagrange multipliers for the study of interior point methods for non-linear programming. On the other hand, Assumption 3.4.2 as a regularity condition provides new insights into solving problems where task is to universally approximate the underlying functions as discussed in the proof of Theorem 3 in Appendix D.2.

D Gaussian Kernel Perceptron

In this appendix, we would provide the proofs to the key results: Theorem 3 and Theorem 4, as shown in §3.4.2. The key to establishing the results is to provide a constructive procedure for an approximate teaching set. Under the Assumption 3.4.1 and Assumption 3.4.2, when the Gaussian learner optimizes Eq. (12) w.r.t the teaching set, any solution $\hat{\boldsymbol{\theta}} \in \boldsymbol{\mathcal{A}}_{opt}$ would be ϵ-close to the optimal classifier point-wise, thereby bounding the error on the data distribution $\mathcal{P}$ on the input space $\mathcal{X}$. We organize this appendix as follows: in Appendix D.1 we show that there exists a solution to Eq. (12); in Appendix D.2 we provide the proofs for our key results Theorem 3 and Theorem 4.

Truncating the Taylor features of Gaussian kernel. In §3.4.1, we showed the construction of the projection $\mathbb{P}$ such that $\mathbb{P}\Phi$ forms a feature map for the kernel $\tilde{\mathcal{K}}$. We denote the orthogonal projection to $\mathbb{P}$ by $\mathbb{P}^{\perp}$. Thus, we can write $\Phi(\mathbf{x}) = \mathbb{P}\Phi(\mathbf{x}) + \mathbb{P}^{\perp}\Phi(\mathbf{x})$ for any $\mathbf{x} \in \mathbb{R}^d$. We discussed the choice of R and s. The primary motivation to pick them in the certain way is to retain maximum information in the first $\binom{d+s}{s}$ coordinates of $\Phi(\cdot)$. This is in line with the observation that the eigenvalues of the canonical orthogonal basis (Cucker & Smale, 2001) (also eigenvectors) for the Gaussian reproducing kernel Hilbert space $\mathcal{H}_{\mathcal{K}}$ decays with higher-indexed coordinates, thus the more sensitive eigenvalues are in the first $\binom{d+s}{s}$ coordinates. Thus, if we could show that $\mathbb{P}\hat{\boldsymbol{\theta}}$ is ϵ-approximately close to $\mathbb{P}\boldsymbol{\theta}^*$ where $\hat{\boldsymbol{\theta}} \in \boldsymbol{\mathcal{A}}_{opt}$ is a solution to Eq. (12), then $\hat{\boldsymbol{\theta}}$ also would be ϵ-approximately close to $\boldsymbol{\theta}^*$.

What should be an optimal R vs. choice of the index s? In §3.4.1, we solved for s such that

$$\frac{1}{(s+1)!} \cdot (R)^{s+1} \leq \epsilon$$

If $\mathbf{x} \in \mathcal{B}(\sqrt{2\sqrt{R}\sigma^2}, 0)$ then using Lemma 3 we have

$$\left|\left|\mathbb{P}^{\perp}\Phi(\mathbf{x})\right|\right|^2 \leq \frac{1}{(s+1)!} \cdot \left(\sqrt{R}\right)^{s+1} \leq \frac{\epsilon}{\left(\sqrt{R}\right)^{s+1}} \leq \frac{\epsilon}{\left(\sqrt{d}\right)^{s}} \tag{23}$$

where the last inequality follows as $R := \max\left\{\frac{\log^2 \frac{1}{\epsilon}}{e^2}, d\right\}$. We define $\epsilon_s := \frac{\epsilon}{\left(\sqrt{d}\right)^s}$. Note that Eq. (23) holds for all $\mathbf{x} \in \mathcal{X}$ since $\frac{\langle \mathbf{x}, \mathbf{x} \rangle_{\mathcal{H}_{\mathcal{K}}}}{\sigma^2} \leq 2\sqrt{R}$. This factor $\left(\sqrt{d}\right)^s$ in the denominator of ϵ_s would be useful in nullifying any $\sqrt{r}$ term since $r = \binom{d+s}{s} = \mathcal{O}\left(d^s\right)$.

D.1 Construction of a Solution to Eq. (12)

In this subsection, we would show that Eq. (12) has a minimizer $\hat{\boldsymbol{\theta}} \in \boldsymbol{\mathcal{A}}_{opt}$ such that $\mathbf{p}(\hat{\boldsymbol{\theta}}) = 0$ where $\mathbf{p}(\cdot)$ is the objective value. Notice that for any i the teaching points $\{(\mathbf{z}_i, 1), (\mathbf{z}_i, -1)\}$ are correctly classified only if $\hat{\boldsymbol{\theta}} \cdot \Phi(\mathbf{z}_i) = 0$ and $\hat{\boldsymbol{\theta}} \cdot \Phi(\mathbf{a}) > 0$. We define the set $\mathbf{B} = \{\mathbf{b}_1, \mathbf{b}_2, \cdots, \mathbf{b}_r\}$ to represent $\{\mathbf{z}_i\}_{i=1}^{r-1} \cup \{\mathbf{a}\}$ in that order. We define the Gaussian kernel Gram matrix $\boldsymbol{\Lambda}$ corresponding to $\mathbf{B}$ as follows:

$$\boldsymbol{\Lambda}[i, j] = \mathcal{K}(\mathbf{b}_i, \mathbf{b}_j) \quad \forall i, j \in [r] \tag{24}$$

Since $\{\mathbf{z}_i\}_{i=1}^{r-1}$ and $\mathbf{a}$ could be chosen from $\mathcal{B}(\sqrt{2\sqrt{R}\sigma^2}, 0)$ as for any two points $\mathbf{x}, \mathbf{x}'$ in the teaching set $\frac{\left|\left|\mathbf{x}-\mathbf{x}'\right|\right|^2}{2\sigma^2} = \Theta\left(\log \frac{1}{\epsilon}\right)$ thus all the non-diagonal entries of $\boldsymbol{\Lambda}$ could be bounded as $\Theta(\epsilon)$. Thus, the non-diagonal entries of $\boldsymbol{\Lambda}$ are upper bounded w.r.t to the choice of ϵ. We denote the concatenated vector of γ and β_0 by $\boldsymbol{\eta}$ as $\boldsymbol{\eta} := (\gamma^\top, \beta_0)^\top$. Consider the following matrix equation:

$$\boldsymbol{\Lambda} \cdot \boldsymbol{\eta} = \big(\underbrace{0, \cdots, 0}_{\text{For } \mathbf{z}_i's}, \underbrace{1}_{\text{For } \mathbf{a}}\big)^\top \tag{25}$$

Notice that any solution $\boldsymbol{\eta}$ to Eq. (25) has zero objective value for Eq. (12). Since $\sum_{i=1}^{r} (\boldsymbol{\Lambda}[i, r] \cdot \boldsymbol{\eta}_i) = \hat{\boldsymbol{\theta}} \cdot \Phi(\mathbf{a}) > 0$ thus we scale the last component of Eq. (25) to 1. First, we observe that Eq. (25) has a solution because Gaussian

kernel Gram matrix $\mathbf{\Lambda}$ to the finite set of points is *strictly positive definite* implying $\mathbf{\Lambda}$ is invertible. Thus, there is a unique solution $\boldsymbol{\eta}_0 \in \mathbb{R}^r$ such that:

$$\mathbf{\Lambda} \cdot \boldsymbol{\eta}_0 = \nu^\top$$

where $\nu := (0, 0, \cdots, 1)$ as shown in Eq. (25). Also, $\boldsymbol{\eta}_0^\top \cdot \mathbf{\Lambda} \cdot \boldsymbol{\eta}_0 = \beta_0 > 0$. Now, we need to ensure that $\boldsymbol{\eta}_0$ satisfies Assumption 3.4.2. To analyse the boundedness, we rewrite the above equation as:

$$\boldsymbol{\eta}_0 = \mathbf{\Lambda}^{-1} \cdot \nu^\top$$

To evaluate the entries of $\boldsymbol{\eta}_0$, we only need to understand the last column of $\mathbf{\Lambda}^{-1}$ (since $\mathbf{\Lambda}^{-1} \cdot \nu^\top$ contains entries from the last column of $\mathbf{\Lambda}^{-1}$). Using the construction of the inverse using the minors of $\mathbf{\Lambda}$, we note that $\mathbf{\Lambda}^{-1}[i, r] = \frac{1}{\det(\mathbf{\Lambda})} \cdot M_{(i,r)}$, where $M_{(i,r)}$ is the minor of $\mathbf{\Lambda}$ corresponding to entry indexed as (i, r) (determinant of the submatrix of $\mathbf{\Lambda}$ formed by removing the ith row and rth column). Note, determinant is an alternating form, which for a square matrix T of dimension n has the explicit sum $\sum_{\sigma \in S^n} sign(\sigma) \cdot \prod_i^n u_{i,\sigma(i)}$, where $T[i, j] = u_{i,j}$. Since the non-diagonal entries of $\mathbf{\Lambda}$ are bounded by ϵ, thus we can bound the minors. Note, $|M_{r,r}| \geq 1 - \mathcal{O}\left(\epsilon^2\right) + \cdots + (-1)^{r-1}\mathcal{O}\left(\epsilon^{r-1}\right)$ and for $i \neq r$ $|M_{i,r}| \leq \mathcal{O}\left(\epsilon\right) + \mathcal{O}\left(\epsilon^2\right) + \cdots + \mathcal{O}\left(\epsilon^{r-1}\right)$. Since, ϵ is sufficiently small, thus $|M_{r,r}|$ majorizes over $|M_{i,r}|$ for $i \neq r$. But then $\boldsymbol{\eta}_0 = \frac{1}{\det(\mathbf{\Lambda})}\left((-1)^{1+r}M_{1,r}, (-1)^{2+r}M_{2,r}, \cdots, (-1)^{r+r}M_{r,r}\right)^\top$. When we normalize $\boldsymbol{\theta}$, we get $\bar{\boldsymbol{\eta}}_0 = \boldsymbol{\eta}_0 / ||\boldsymbol{\theta}||$. We note that, $||\boldsymbol{\theta}||^2 = \boldsymbol{\eta}_0^\top \cdot \mathbf{\Lambda} \cdot \boldsymbol{\eta}_0 = \beta_0$, implying $\bar{\boldsymbol{\eta}}_0 = \boldsymbol{\eta}_0 / \sqrt{\beta_0}$. Since, $\beta_0 = \frac{1}{\det(\mathbf{\Lambda})} \cdot M_{r,r}$, thus entries of $\bar{\boldsymbol{\eta}}_0$ satisfy Assumption 3.4.2. Thus, we have a solution to Eq. (12) which satisfies Assumption 3.4.2.

D.2 Proof of Theorem 3 and Theorem 4

In this section, we would establish our key results for the approximate teaching of a Gaussian kernel perceptron. Under the Assumption 3.4.1 and Assumption 3.4.2, we would show to teach a target model $\boldsymbol{\theta}^*$ ϵ-approximately we only require at most $d^{\mathcal{O}\left(\log^2 \frac{1}{\epsilon}\right)}$ labelled teaching points from $\mathcal{X}$. In order to achieve the ϵ-approximate teaching set, we would show that the teaching set $\mathcal{TS}_{\boldsymbol{\theta}^*}$ as constructed in Eq. (13) achieves an ϵ-closeness between $f^* = \boldsymbol{\theta}^* \cdot \Phi(\cdot)$ and $\hat{f} = \hat{\boldsymbol{\theta}} \cdot \Phi(\cdot)$ i.e $\left|f^*(\mathbf{x}) - \hat{f}(\mathbf{x})\right| \leq \epsilon$ point-wise.

Before we move to the proofs of the key results, we state the following relevant lemma which bounds the length (norm) of a vector spanned by a basis with the smoothness condition on the basis as mentioned in §3.4.2.

Lemma 4. *Consider the Euclidian space $\mathbb{R}^n$. Assume $\{\mathbf{v}_i\}_{i=1}^n$ forms a basis with unit norms. Additionally, for any i, j $|\mathbf{v}_i \cdot \mathbf{v}_j| \leq \cos\theta_0$ where $\cos\theta_0 \leq \frac{1}{2n}$. Fix a small real scalar $\epsilon > 0$. Now, consider any random vector $\mathbf{p} \in \mathbb{R}^n$ such that $\forall i \in [n]$ $|\mathbf{p} \cdot \mathbf{v}_i| \leq \epsilon$. Then the following bound on $\mathbf{p}$ holds:*

$$||\mathbf{p}||_2 \leq \sqrt{2n} \cdot \epsilon.$$

Proof. We define $M = \mathbb{R}^n$ as the space in which $\mathbf{p}$ and $\mathbf{v}_i$'s are embedded. Consider another copy of the space $N = \mathbb{R}^n$ with standard orthogonal basis $\{e_1, \cdots, e_n\}$. We define the map $\mathbf{W} : M \longrightarrow N$ as follows:

$$\begin{aligned} \mathbf{W} : M &\longrightarrow N \\ q &\mapsto (\mathbf{v}_1 \cdot q, \mathbf{v}_2 \cdot q, \cdots, \mathbf{v}_n \cdot q) \end{aligned}$$

Since $\{\mathbf{v}_i\}_{i=1}^n$ forms a basis, thus $\mathbf{W}$ is invertible. To ease the analysis, we could assume $\epsilon = 1$ (follows by scaling symmetry). Thus, it is clear that $w := \mathbf{W}\mathbf{p}$ has all its entries bounded in absolute value by 1.

We could write $\mathbf{p} = \mathbf{W}^{-1}w$, thus $||\mathbf{p}||^2 = \left(\mathbf{W}^{-1}w\right)^\top \left(\mathbf{W}^{-1}w\right) = w^\top \left(\mathbf{W}^\top \mathbf{W}\right)^{-1} w$. Thus, showing the bound for $w^\top \left(\mathbf{W}^\top \mathbf{W}\right)^{-1} w$ where $w \in [-1, 1]^n$ suffices. We note that $\left(\mathbf{W}^\top \mathbf{W}\right)$ is a symmetric $(n \times n)$ matrix with diagonal entries 1 and non-diagonal entries bounded in absolute value by $\cos\theta_0$.

Using convergence of the Neumann series $\sum_{k=0}^\infty \left(\mathbb{I}_{(n\times n)} - \left(\mathbf{W}^\top \mathbf{W}\right)\right)^k$ as $\left(\mathbb{I}_{(n\times n)} - \left(\mathbf{W}^\top \mathbf{W}\right)\right)$ is a bounded operator, we have:

$$\left|\left(\mathbf{W}^\top \mathbf{W}\right)^{-1} - \mathbb{I}_{(n\times n)}\right|_{\ell_\infty} \leq \left|\sum_{k=1}^\infty \left(\mathbb{I}_{(n\times n)} - \left(\mathbf{W}^\top \mathbf{W}\right)\right)^k\right|_{\ell_\infty} \tag{26}$$

$$\leq \sum_{k=1}^{\infty} \left| \left(\mathbb{I}_{(n\times n)} - \left(\mathbf{W}^\top \mathbf{W} \right) \right)^k \right|_{\ell_\infty} \tag{27}$$

$$\leq \sum_{k=1}^{\infty} n^{k-1} \cos^k \theta_0 \tag{28}$$

$$= \frac{\cos\theta_0}{1 - n\cos\theta_0} \tag{29}$$

where $|B|_{\ell_\infty}$ refers to the maximum absolute value of any entry of B. Eq. (26) follows using the Neumann series $\left(\mathbf{W}^\top \mathbf{W}\right)^{-1} = \sum_{k=0}^{\infty} \left(\mathbb{I}_{(n\times n)} - \left(\mathbf{W}^\top \mathbf{W} \right) \right)^k$. Eq. (27) is a direct consequence of triangle inequality. Since entries of $\left(\mathbb{I}_{(n\times n)} - \left(\mathbf{W}^\top \mathbf{W} \right) \right)^k$ are bounded in absolute value by $\cos\theta_0$ thus Eq. (28) follows. Using a straight-forward geometric sum we get an upper bound on the maximum absolute value of any entry in $\left(\mathbf{W}^\top \mathbf{W}\right)^{-1} - \mathbb{I}_{(n\times n)}$ in Eq. (29).

Now, we note that

$$\begin{aligned}
w^\top \left(\mathbf{W}^\top \mathbf{W}\right)^{-1} w &= w^\top \left(\left(\mathbf{W}^\top \mathbf{W}\right)^{-1} - \mathbb{I}_{(n\times n)} \right) w + w^\top \left(\mathbb{I}_{(n\times n)} \right) w \\
&\leq \sum_{i,j} w_{ij} \left(\left(\mathbf{W}^\top \mathbf{W}\right)^{-1} - \mathbb{I}_{(n\times n)} \right)_{ij} w_{ij} + ||w||^2 \\
&\leq \sum_{i,j} \left| w_{ij} \left(\left(\mathbf{W}^\top \mathbf{W}\right)^{-1} - \mathbb{I}_{(n\times n)} \right)_{ij} w_{ij} \right| + ||w||^2 \\
&\leq \sum_{i,j} \left| \left(\left(\mathbf{W}^\top \mathbf{W}\right)^{-1} - \mathbb{I}_{(n\times n)} \right)_{ij} \right| + n \qquad (30) \\
&\leq \frac{n^2 \cos\theta_0}{1 - n\cos\theta_0} + n \qquad (31) \\
&= \frac{n}{1 - n\cos\theta_0} \qquad (32)
\end{aligned}$$

In Eq. (30) we use $w \in [-1, 1]^n$. Eq. (31) follows using Eq. (29).

Since we have $||\mathbf{p}||_2^2 = w^\top \left(\mathbf{W}^\top \mathbf{W}\right)^{-1} w$ and $\cos\theta_0 \leq \frac{1}{2n}$, thus using Eq. (32)

$$||\mathbf{p}||_2 \leq \sqrt{\frac{n}{1 - n\cos\theta_0}} \leq \sqrt{2n}.$$

Scaling the map $\mathbf{W}$ to ϵ yields the stated claim.

□

Under the Assumption 3.4.1 and Assumption 3.4.2, and bounded norm of $\boldsymbol{\theta}^*$ and $\hat{\boldsymbol{\theta}}$, we would establish that $\mathcal{TS}_{\boldsymbol{\theta}^*}$ is a $d^{\mathcal{O}\left(\log^2 \frac{1}{\epsilon}\right)}$ size ϵ-approximate teaching set for $\boldsymbol{\theta}^*$. Before establishing the main result, we show the proof of Theorem 3 below. Using Eq. (23), we note that:

$$\begin{aligned}
\forall\, \mathbf{x} \in \mathcal{X} \quad & \left|\left| \mathbb{P}^\perp \Phi(\mathbf{x}) \right|\right| \leq \sqrt{\epsilon_s} \implies ||\mathbb{P}\Phi(\mathbf{x})|| \geq \sqrt{1 - \epsilon_s} \\
\forall\, (\mathbf{z}, y) \in \mathcal{TS}_{\boldsymbol{\theta}^*} \quad & \left|\left| \mathbb{P}^\perp \Phi(\mathbf{z}) \right|\right| \leq \sqrt{\epsilon_s} \implies ||\mathbb{P}\Phi(\mathbf{z})|| \geq \sqrt{1 - \epsilon_s}
\end{aligned}$$

Now, we could further bound the norms of $\mathbb{P}^\perp \hat{\boldsymbol{\theta}}$ and $\mathbb{P}^\perp \boldsymbol{\theta}^*$ using triangle inequality and boundedness of $\{\alpha_i\}_{i=1}^l$ and $[\beta_0, \gamma]$ (as shown in Assumption 3.4.2):

$$\left|\left| \mathbb{P}^\perp \boldsymbol{\theta}^* \right|\right| = \left|\left| \sum_{i=1}^{l} \alpha_i \cdot \mathbb{P}^\perp \Phi(\mathbf{a}_i) \right|\right| \leq \sum_{i=1}^{l} \left|\left| \alpha_i \cdot \mathbb{P}^\perp \Phi(\mathbf{a}_i) \right|\right| \leq \left(\sum_{i=1}^{l} |\alpha_i| \right) \cdot \sqrt{\epsilon_s} = \boldsymbol{C}_\epsilon \cdot \sqrt{\epsilon_s}$$

$$\left|\left|\mathbb{P}^{\perp}\hat{\boldsymbol{\theta}}\right|\right| = \left|\left|\beta_0 \cdot \mathbb{P}^{\perp}\Phi(\mathbf{a}) + \sum_{j=1}^{r-1}\gamma_j \cdot \mathbb{P}^{\perp}\Phi(\mathbf{z}_j)\right|\right| \leq \left|\left|\beta_0 \cdot \mathbb{P}^{\perp}\Phi(\mathbf{a})\right|\right| + \sum_{i=1}^{r-1}\left|\left|\gamma_i \cdot \mathbb{P}^{\perp}\Phi(\mathbf{z}_i)\right|\right| \leq \left(|\beta_0| + \sum_{i=1}^{r-1}|\gamma_i|\right) \cdot \sqrt{\epsilon_s} = \boldsymbol{D}_{\epsilon} \cdot \sqrt{\epsilon_s}$$

Proof of Theorem 3. In the following, we would bound $\left|f^*(\mathbf{x}) - \hat{f}(\mathbf{x})\right|$ by $\sqrt{\epsilon}$. In order to bound the modulus, we would split the difference using $\mathbb{P}$ and $\mathbb{P}^{\perp}$ and then analyze the terms correspondingly. We can write any classifier f as $f(\mathbf{x}) = \boldsymbol{\theta} \cdot \Phi(\mathbf{x}) = \mathbb{P}\boldsymbol{\theta} \cdot \mathbb{P}\Phi(\mathbf{x}) + \mathbb{P}^{\perp}\boldsymbol{\theta} \cdot \mathbb{P}^{\perp}\Phi(\mathbf{x})$. Thus, we have:

$$\begin{aligned}
\left|f^*(\mathbf{x}) - \hat{f}(\mathbf{x})\right| &= \left|\mathbb{P}\boldsymbol{\theta}^* \cdot \mathbb{P}\Phi(\mathbf{x}) + \mathbb{P}^{\perp}\boldsymbol{\theta}^* \cdot \mathbb{P}^{\perp}\Phi(\mathbf{x}) - \mathbb{P}\hat{\boldsymbol{\theta}} \cdot \mathbb{P}\Phi(\mathbf{x}) - \mathbb{P}^{\perp}\hat{\boldsymbol{\theta}} \cdot \mathbb{P}^{\perp}\Phi(\mathbf{x})\right| \\
&\leq \left|\mathbb{P}\boldsymbol{\theta}^* \cdot \mathbb{P}\Phi(\mathbf{x}) - \mathbb{P}\hat{\boldsymbol{\theta}} \cdot \mathbb{P}\Phi(\mathbf{x})\right| + \left|\mathbb{P}^{\perp}\boldsymbol{\theta}^* \cdot \mathbb{P}^{\perp}\Phi(\mathbf{x}) - \mathbb{P}^{\perp}\hat{\boldsymbol{\theta}} \cdot \mathbb{P}^{\perp}\Phi(\mathbf{x})\right| && (33) \\
&\leq \left|\mathbb{P}\boldsymbol{\theta}^* \cdot \mathbb{P}\Phi(\mathbf{x}) - \mathbb{P}\hat{\boldsymbol{\theta}} \cdot \mathbb{P}\Phi(\mathbf{x})\right| + \left|\left|\mathbb{P}^{\perp}\boldsymbol{\theta}^* - \mathbb{P}^{\perp}\hat{\boldsymbol{\theta}}\right|\right| \cdot \left|\left|\mathbb{P}^{\perp}\Phi(\mathbf{x})\right|\right| && (34) \\
&\leq \underbrace{\left|\mathbb{P}\boldsymbol{\theta}^* \cdot \mathbb{P}\Phi(\mathbf{x}) - \mathbb{P}\hat{\boldsymbol{\theta}} \cdot \mathbb{P}\Phi(\mathbf{x})\right|}_{\bigstar} + (\boldsymbol{C}_{\epsilon} + \boldsymbol{D}_{\epsilon}) \cdot \epsilon_s && (35)
\end{aligned}$$

Eq. (33) follows using triangle inequality. We can further bound $\left|\mathbb{P}^{\perp}\boldsymbol{\theta}^* \cdot \mathbb{P}^{\perp}\Phi(\mathbf{x}) - \mathbb{P}^{\perp}\hat{\boldsymbol{\theta}} \cdot \mathbb{P}^{\perp}\Phi(\mathbf{x})\right|$ using Cauchy-Schwarz inequality and thus Eq. (34) follows. Using the observations: $||\mathbb{P}^{\perp}\hat{\boldsymbol{\theta}}|| \leq \boldsymbol{D}_{\epsilon} \cdot \sqrt{\epsilon_s}$ and $\left|\left|\mathbb{P}^{\perp}\boldsymbol{\theta}^*\right|\right| \leq \boldsymbol{C}_{\epsilon} \cdot \sqrt{\epsilon_s}$, we could upper bound $\left|\left|\mathbb{P}^{\perp}\boldsymbol{\theta}^* - \mathbb{P}^{\perp}\hat{\boldsymbol{\theta}}\right|\right|$ by $(\boldsymbol{C}_{\epsilon} + \boldsymbol{D}_{\epsilon}) \cdot \sqrt{\epsilon_s}$. Since $\mathbf{x} \in \mathcal{X}$ thus $\left|\left|\mathbb{P}^{\perp}\Phi(\mathbf{x})\right|\right| \leq \sqrt{\epsilon_s}$ (as shown in Eq. (23)), which gives Eq. (35).

Now, the key is to bound the ($\bigstar$) appropriately and then the result would be proven. We would rewrite $\mathbb{P}\hat{\boldsymbol{\theta}} \cdot \mathbb{P}\Phi(\mathbf{x})$ in terms of the basis formed by $\{\mathbb{P}\Phi(\mathbf{z}_i)\}_{i=1}^{r-1} \cup \{\mathbb{P}\theta^*\}$ (by Assumption 3.4.1 $\{\mathbb{P}\Phi(\mathbf{z}_i)\}_{i=1}^{l}$ are linearly independent and orthogonal to $\mathbb{P}\theta^*$). Using the basis, we can write $\mathbb{P}\hat{\boldsymbol{\theta}} = \sum_{i=1}^{r-1} c_i \cdot \mathbb{P}\Phi(\mathbf{z}_i) + \lambda_r \cdot \mathbb{P}\boldsymbol{\theta}^*$ for some scalars $c_1, c_2, \cdots, \lambda_r$. Alternatively, we could rewrite $\mathbb{P}\hat{\boldsymbol{\theta}} = \beta_0 \cdot \mathbb{P}\Phi(\mathbf{a}) + \sum_{j=1}^{r-1}\gamma_j \cdot \mathbb{P}\Phi(\mathbf{z}_j)$ where $\beta_0 > 0$ (as shown in Appendix D.1). This could be used to note that $\lambda_r > 0$ because $\mathbb{P}\boldsymbol{\theta}^* \cdot \mathbb{P}\Phi(\mathbf{a}) > 0$ (cf §3.4).

We study the decomposition of $\mathbb{P}\hat{\boldsymbol{\theta}}$ in terms of the basis in order to understand the component of $\mathbb{P}\hat{\boldsymbol{\theta}}$ along $\mathbb{P}\boldsymbol{\theta}^*$. We observe that:

$$\left|\left|\mathbb{P}\hat{\boldsymbol{\theta}}\right|\right|^2 = \left|\left|\sum_{i=1}^{r-1} c_i \cdot \mathbb{P}\Phi(\mathbf{z}_i)\right|\right|^2 + ||\lambda_r \cdot \mathbb{P}\boldsymbol{\theta}^*||^2 \tag{36}$$

Since $\hat{\boldsymbol{\theta}}$ is a solution to Eq. (12), $\hat{\boldsymbol{\theta}} \cdot \Phi(\mathbf{z}_i) = 0$ for any $i \in [r-1]$. Now, we can write the equation in terms of projections as:

$$\forall i \quad \mathbb{P}\hat{\boldsymbol{\theta}} \cdot \mathbb{P}\Phi(\mathbf{z}_i) + \mathbb{P}^{\perp}\hat{\boldsymbol{\theta}} \cdot \mathbb{P}^{\perp}\Phi(\mathbf{z}_i) = 0 \tag{37}$$

Using Cauchy-Schwarz inequality on the product $|\mathbb{P}^{\perp}\hat{\boldsymbol{\theta}} \cdot \mathbb{P}^{\perp}\Phi(\mathbf{z}_i)|$ we obtain:

$$|\mathbb{P}^{\perp}\hat{\boldsymbol{\theta}} \cdot \mathbb{P}^{\perp}\Phi(\mathbf{z}_i)| \leq ||\mathbb{P}^{\perp}\hat{\boldsymbol{\theta}}|| \cdot ||\mathbb{P}^{\perp}\Phi(\mathbf{z}_i)|| \leq \boldsymbol{D}_{\epsilon} \cdot \epsilon_s$$

Plugging this into Eq. (37), we get the following bound on $|\mathbb{P}\hat{\boldsymbol{\theta}} \cdot \mathbb{P}\Phi(\mathbf{z}_i)|$:

$$|\mathbb{P}\hat{\boldsymbol{\theta}} \cdot \mathbb{P}\Phi(\mathbf{z}_i)| \leq \boldsymbol{D}_{\epsilon} \cdot \epsilon_s \tag{38}$$

We denote $V_O := \sum_{i=1}^{r-1} c_i \cdot \mathbb{P}\Phi(\mathbf{z}_i)$. Notice that V_O is the orthogonal projection of $\mathbb{P}\hat{\boldsymbol{\theta}}$ along the subspace $\mathbf{span}\langle\mathbb{P}\Phi(\mathbf{z}_1), \cdots, \mathbb{P}\Phi(\mathbf{z}_{r-1})\rangle$. Thus, we could rewrite Eq. (38) further as:

$$|\mathbb{P}\hat{\boldsymbol{\theta}} \cdot \mathbb{P}\Phi(\mathbf{z}_i)| = |\left(V_O + \lambda_r \cdot \mathbb{P}\boldsymbol{\theta}^*\right) \cdot \mathbb{P}\Phi(\mathbf{z}_i)| = |V_O \cdot \mathbb{P}\Phi(\mathbf{z}_i)| \leq \boldsymbol{D}_{\epsilon} \cdot \epsilon_s$$

Notice that $||\mathbb{P}\Phi(\mathbf{z}_i)|| \geq \sqrt{1-\epsilon_s}$. Hence, component of V_O along $\mathbb{P}\Phi(\mathbf{z}_i)$ is upper bounded by $\frac{\boldsymbol{D}_{\epsilon} \cdot \epsilon_s}{\sqrt{1-\epsilon_s}}$. Since $\{\mathbb{P}\Phi(\mathbf{z}_1), \cdots, \mathbb{P}\Phi(\mathbf{z}_{r-1})\}$ satisfy the conditions of Lemma 4 (the smoothness condition mentioned in §3.4) thus we could bound the norm of V_O as follows:

$$||V_O|| \leq \sqrt{2(r-1)} \cdot \frac{\boldsymbol{D}_{\epsilon} \cdot \epsilon_s}{\sqrt{1-\epsilon_s}} \tag{39}$$

Using Eq. (36) and Eq. (39) we can lower bound the norm of $\lambda_r \cdot \mathbb{P}\boldsymbol{\theta}^*$ as follows:

$$\left|\left|\mathbb{P}\hat{\boldsymbol{\theta}}\right|\right|^2 = \left|\left|\sum_{i=1}^{r-1} c_i \cdot \mathbb{P}\Phi(\mathbf{z}_i)\right|\right|^2 + ||\lambda_r \cdot \mathbb{P}\boldsymbol{\theta}^*||^2 = ||V_O||^2 + ||\lambda_r \cdot \mathbb{P}\boldsymbol{\theta}^*||^2$$

$$\implies ||\lambda_r \cdot \mathbb{P}\boldsymbol{\theta}^*||^2 \geq \left(1 - \boldsymbol{D}_\epsilon^2 \cdot \epsilon_s\right) - 2\,(r-1) \cdot \frac{\boldsymbol{D}_\epsilon^2 \cdot \epsilon_s^2}{(1-\epsilon_s)} \geq 1 - 2\boldsymbol{D}_\epsilon^2 \cdot \epsilon_s \tag{40}$$

This follows because $\left|\left|\mathbb{P}\hat{\boldsymbol{\theta}}\right|\right|^2 \geq \left(1 - \boldsymbol{D}_\epsilon^2 \cdot \epsilon_s\right)$ as $\left|\left|\hat{\boldsymbol{\theta}}\right|\right| = \mathcal{O}(1)$ and $\sqrt{2(r-1)} \cdot \epsilon_s \leq \sqrt{2(r-1)} \cdot \frac{\epsilon}{(\sqrt{d})^s} \leq \epsilon$. With these observations we can rewrite ($\bigstar$) as follows:

$$\left|\mathbb{P}\boldsymbol{\theta}^* \cdot \mathbb{P}\Phi(\mathbf{x}) - \mathbb{P}\hat{\boldsymbol{\theta}} \cdot \mathbb{P}\Phi(\mathbf{x})\right| = \left|\mathbb{P}\boldsymbol{\theta}^* \cdot \mathbb{P}\Phi(\mathbf{x}) - \sum_{i=1}^{r-1} c_i \cdot \mathbb{P}\Phi(\mathbf{z}_i) \cdot \mathbb{P}\Phi(\mathbf{x}) - \lambda_r \cdot \mathbb{P}\boldsymbol{\theta}^* \cdot \mathbb{P}\Phi(\mathbf{x})\right|$$

$$\leq |\mathbb{P}\boldsymbol{\theta}^* \cdot \mathbb{P}\Phi(\mathbf{x}) - \lambda_r \cdot \mathbb{P}\boldsymbol{\theta}^* \cdot \mathbb{P}\Phi(\mathbf{x})| + \left|\sum_{i=1}^{r-1} c_i \cdot \mathbb{P}\Phi(\mathbf{z}_i) \cdot \mathbb{P}\Phi(\mathbf{x})\right| \tag{41}$$

$$\leq \sqrt{|\boldsymbol{C}_\epsilon^2 - 2\boldsymbol{D}_\epsilon^2|} \cdot \sqrt{\epsilon_s} + \left|\left|\sum_{i=1}^{r-1} c_i \cdot \mathbb{P}\Phi(\mathbf{z}_i)\right|\right| \cdot ||\mathbb{P}\Phi(\mathbf{x})|| \tag{42}$$

$$\leq \sqrt{|\boldsymbol{C}_\epsilon^2 - 2\boldsymbol{D}_\epsilon^2|} \cdot \sqrt{\epsilon_s} + \sqrt{2(r-1)} \cdot \frac{\boldsymbol{D}_\epsilon \cdot \epsilon_s}{\sqrt{1-\epsilon_s}} \tag{43}$$

$$\leq \frac{\sqrt{|\boldsymbol{C}_\epsilon^2 - 2\boldsymbol{D}_\epsilon^2|} \cdot \sqrt{\epsilon}}{(\sqrt{d})^{s/2}} + 2\boldsymbol{D}_\epsilon \cdot \epsilon \tag{44}$$

$$\leq 2\max\left\{\frac{\sqrt{|\boldsymbol{C}_\epsilon^2 - 2\boldsymbol{D}_\epsilon^2|}}{(\sqrt{d})^{s/2}},\, 2\boldsymbol{D}_\epsilon \cdot \sqrt{\epsilon}\right\} \cdot \sqrt{\epsilon} \tag{45}$$

Eq. (41) is a direct implication of triangle inequality. In Eq. (42), in the first term we note that $\lambda_r > 0$ () $||\mathbb{P}\boldsymbol{\theta}^*|| \geq \sqrt{1 - \boldsymbol{C}_\epsilon^2 \cdot \epsilon_s}$ and use Eq. (40), and in the second use Cauchy-Schwarz inequality. Eq. (43) follows using Eq. (39) and that $||\mathbb{P}\Phi(\mathbf{x})||$ is bounded by 1. We could unfold the value of $\epsilon_s \leq \frac{\epsilon}{(\sqrt{d})^s}$. This gives us Eq. (44). We could rewrite Eq. (44) to get a bound in terms of $\sqrt{\epsilon}$ to obtain Eq. (45).

Now, using Eq. (35) and Eq. (45), can bound $\left|f^*(\mathbf{x}) - \hat{f}(\mathbf{x})\right|$ as follows:

$$\left|f^*(\mathbf{x}) - \hat{f}(\mathbf{x})\right| \leq 2\max\left\{\frac{\sqrt{|\boldsymbol{C}_\epsilon^2 - 2\boldsymbol{D}_\epsilon^2|}}{(\sqrt{d})^{s/2}},\, 2\boldsymbol{D}_\epsilon \cdot \sqrt{\epsilon}\right\} \cdot \sqrt{\epsilon} + (\boldsymbol{C}_\epsilon + \boldsymbol{D}_\epsilon) \cdot \frac{\epsilon}{(\sqrt{d})^s}$$

$$\leq 3\max\left\{\frac{\sqrt{|\boldsymbol{C}_\epsilon^2 - 2\boldsymbol{D}_\epsilon^2|}}{(\sqrt{d})^{s/2}},\, 2\boldsymbol{D}_\epsilon \cdot \sqrt{\epsilon},\, (\boldsymbol{C}_\epsilon + \boldsymbol{D}_\epsilon) \cdot \frac{\sqrt{\epsilon}}{(\sqrt{d})^s}\right\} \cdot \sqrt{\epsilon}$$

$$\leq 3\boldsymbol{C}' \cdot \sqrt{\epsilon}$$

where $\boldsymbol{C}' := \max\left\{\frac{\sqrt{|\boldsymbol{C}_\epsilon^2 - 2\boldsymbol{D}_\epsilon^2|}}{(\sqrt{d})^{s/2}},\, 2\boldsymbol{D}_\epsilon \cdot \sqrt{\epsilon},\, (\boldsymbol{C}_\epsilon + \boldsymbol{D}_\epsilon) \cdot \frac{\sqrt{\epsilon}}{(\sqrt{d})^s}\right\}$.

Notice that all the terms in $\max\left\{\frac{\sqrt{|\boldsymbol{C}_\epsilon^2 - 2\boldsymbol{D}_\epsilon^2|}}{(\sqrt{d})^{s/2}},\, 2\boldsymbol{D}_\epsilon \cdot \sqrt{\epsilon},\, (\boldsymbol{C}_\epsilon + \boldsymbol{D}_\epsilon) \cdot \frac{\sqrt{\epsilon}}{(\sqrt{d})^s}\right\}$ are smaller than 1 because of boundedness of $\boldsymbol{C}_\epsilon$ and $\boldsymbol{D}_\epsilon$. Thus, we have shown a $3C' \cdot \sqrt{\epsilon}$ (where C' is a constant smaller than 1) bound on the point-wise difference of $\hat{f}$ and f^*. Now, if we scale the ϵ and solve for $\epsilon^2/3$, we get the desired bound. Hence, the main claim of Theorem 3 is proven i.e. $\left|f^*(\mathbf{x}) - \hat{f}(\mathbf{x})\right| \leq \epsilon$. □

Now, we would complete the proof of the main result of §3.4 which bounds the error incurred by the solution $\hat{\boldsymbol{\theta}} \in \boldsymbol{\mathcal{A}}_{opt}(\mathcal{TS}_{\boldsymbol{\theta}^*})$ i.e. Theorem 4. The point-wise closeness of f^* and $\hat{f}$ established in Theorem 3 would be key in bounding the error. We complete the proof as follows:

Proof of Theorem 4. We show the error analysis when data-points are sampled from the data distribution $\mathcal{P}$.

$$\left|\mathbf{err}(f^*) - \mathbf{err}(\hat{f})\right| = \left|\mathop{\mathbb{E}}_{(\mathbf{x},y)\sim\mathcal{P}}\Big[\max(-y\cdot f^*(\mathbf{x}),0)\Big] - \mathop{\mathbb{E}}_{(\mathbf{x},y)\sim\mathcal{P}}\Big[\max(-y\cdot \hat{f}(\mathbf{x}),0)\Big]\right| \tag{46}$$

$$= \left|\mathop{\mathbb{E}}_{(\mathbf{x},y)\sim\mathcal{P}}\Big[\max(-y\cdot f^*(\mathbf{x}),0) - \max(-y\cdot \hat{f}(\mathbf{x}),0)\Big]\right| \tag{47}$$

$$\leq \mathop{\mathbb{E}}_{(\mathbf{x},y)\sim\mathcal{P}}\Big[\left|f^*(\mathbf{x}) - \hat{f}(\mathbf{x})\right|\Big] \tag{48}$$

$$\leq \epsilon \tag{49}$$

Eq. (46) follows using the definition of $\mathbf{err}(\cdot)$ function. Because of linearity of expectation, we get Eq. (47). In Eq. (48), we use the observation that modulus of an expectation is bounded by the expectation of the modulus of the random variable $f^*(\mathbf{x}) - \hat{f}(\mathbf{x})$. In Theorem 3, we showed that for any $\mathbf{x}\in\mathcal{X}$, $\left|f^*(\mathbf{x}) - \hat{f}(\mathbf{x})\right| \leq \epsilon$. Thus, the main claim follows. □

E Experimental Evaluation

In this section, we provide an algorithmic procedure for constructing the ϵ-approximate teaching set, and quantitatively evaluate our theoretical results as presented in Theorem 3 and Theorem 4 (cf. §3.4.2).

Results in this section are supplementary to Fig. 3. For a qualitative evaluation of the *ϵ-approximated teaching set*, please refer to Fig. 5c, which illustrates the learner's Gaussian kernel perceptron learned from the ϵ-approximated teaching sets on different classification tasks.

E.1 Experimental Setup

Our experiments are carried out on 4 different datasets: the two-moon dataset (2 interleaving half-circles with noise), the two-circles dataset (a large circle containing a small circle, with noise) from sklearn[4], the Banana dataset[5] where the two classes are not perfectly separable, and the Iris dataset[6] with one of the three classes removed. For each dataset, the following steps are performed:

1. For a given set of data, we, assuming the role of the teacher, find the optimal Gaussian (with $\sigma = 0.9$) separator $\boldsymbol{\theta}^*$ and plot the corresponding boundaries. We estimate the perceptron loss $\mathbf{err}(f^*)$ for this separator by summing up the total perceptron loss on the dataset and averaging over the size of the dataset.

2. For some s, we use the degree s polynomial approximation of the Gaussian separator to determine the approximate polynomial boundaries and select $r = \binom{2+s}{s} - 1$ points on the boundaries such that their images in the polynomial feature space are linearly independent. We make a copy of these points and assign positive labels to one copy and negative labels to the other. In addition, we pick 2 more points arbitrarily, one on each side of the boundaries (i.e. with opposite labels). Thus $\mathcal{TS}_{\theta^*}$ of size $2r$ is constructed.

3. Following Assumption 3.4.1-3.4.2, the Gaussian kernel perceptron learner (with the same σ parameter value 0.9) uses only $\mathcal{TS}_{\theta^*}$ to learn a separator $\hat{\boldsymbol{\theta}}$. The perceptron loss $\mathbf{err}(\hat{f})$ w.r.t. the original dataset is calculated by averaging the total perceptron loss over the number of points in the dataset.

4. Repeat Step 2 and Step 3 for $s = 2, 3, \cdots, 12$ and record the perceptron loss (i.e the max error function as shown in §2) for the corresponding teaching set sizes $2r$ (where $r = \binom{2+s}{s} - 1$). Then we plot the error $\left|\mathbf{err}(f^*) - \mathbf{err}(\hat{f})\right|$ as a function of the teaching set size.

The corresponding plots for Steps 1-4 are shown in columns (a)-(d) of Fig. 5, where for Step 2 (column (b)), the plots all correspond to when $s = 5$.

E.2 Implementation Details

In this subsection we provide more algorithmic and numeric details about the implementation of the experiments.

First we describe how the first $r - 1$ points are generated in Step 2. Given that the approximate polynomial separator has been found using the kernel and feature map approximation described in Eq. (8) and Eq. (9), we are able to plot the corresponding boundaries, and by the same reasoning as in the case of teaching set generation for the polynomial learner, we need to locate points on the boundaries such that their images in the r-dimensional feature space are linearly independent. We achieve this by sampling points on the zero-contour line and row-reducing the matrix formed by the image of all such points. This way, $r - 1$ qualified points can be efficiently located. In addition, as discussed in §3.4.2, the teaching points are selected within the radius of some small constant multiple of $\sqrt{R}$ consistently across the experiments. In this case, we have arbitrarily picked the constant to be 4.

In Step 3, when the learner learns the separator, we need to ensure Assumption 3.4.1-3.4.2 are satisfied. This is made possible by adding the corresponding constraints to the learner's optimization procedure. Specifically, we

[4] https://scikit-learn.org/stable/modules/classes.html#module-sklearn.datasets
[5] https://www.scilab.org/tutorials/machine-learning-—classification-svm
[6] https://archive.ics.uci.edu/ml/datasets/iris

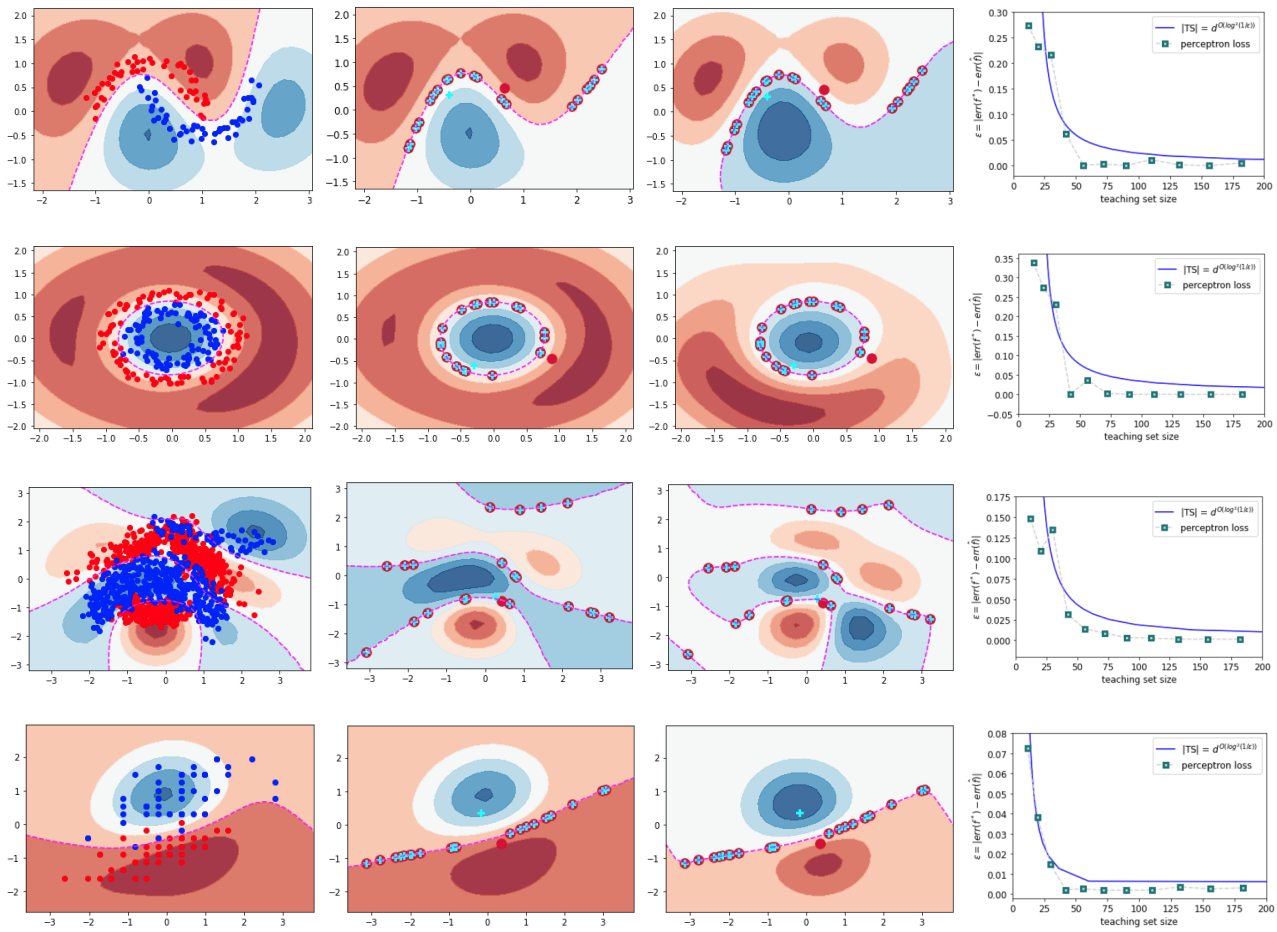
need to enforce that 1) the norm of $\hat{\boldsymbol{\theta}}$ is not far from 1, and 2) β[7] and γ are bounded absolutely as mentioned in Assumption 3.4.2. This is achieved by adjusting the specified bound higher or lower as the current-iteration $\hat{\boldsymbol{\theta}}$ norm varies during the optimization procedure. Eventually, we normalize $\hat{\boldsymbol{\theta}}$ and check that the final β and γ are indeed bounded (i.e. Assumption 3.4.2 is satisfied).

Finally, the perceptron loss calculated for each value of s is based on 5 separate runs of Step 2, while for each run, the learner's kernelized Gaussian perceptron learning algorithm is repeated 5 times. The learner's perceptron loss is then averaged over the 25 epochs of the algorithm to prevent numerical inaccuracies that may arise during the learner's constrained optimization process and possibly the teaching set generation process.

E.3 Results

We present the experiment results in Fig. 5. In the right-most plot of each row, the estimates of $\left|\mathbf{err}(f^*) - \mathbf{err}(\hat{f})\right|$ are plotted against the teaching set sizes $2r$ corresponding to $s = 2, \cdots, 12$ (as discussed in §3.4.1). As can be observed from the shape of the curves in plots of column (d), indeed, our experimental results confirm that the number of teaching examples needed for ϵ-approximate teaching is upper-bounded by $d^{\mathcal{O}\left(\log^2 \frac{1}{\epsilon}\right)}$ for Gaussian kernel perceptrons.

[7] We pick two points outside the orthogonal complement of $\mathbb{P}\boldsymbol{\theta}^*$, one with positive label and another with negative label. Thus, in place of β_0 (as used in §3.4) we use $\beta \in \mathbb{R}^2$ here.



(a) (b) (c) (d)

Figure 5: Constructing the ϵ-approximate teaching set $\mathcal{TS}_{\boldsymbol{\theta}^*}$ for a Gaussian kernel perceptron learner. Each row corresponds to the numerical results on a different dataset as described in the beginning of Appendix E. For each row from left to right: (a) optimal Gaussian boundary and the data set; (b) Teacher identifies a (degree-5) polynomial approximation of the Gaussian decision boundary and finds the ϵ-approximate teaching set $\mathcal{TS}_{\boldsymbol{\theta}^*}$ (marked by cyan plus markers and red dots); (c) Learner learns a Gaussian kernel perceptron from the optimal teaching set in the previous plot; (d) $|\mathcal{TS}|$-ϵ plot for degree-2 to degree-12 polynomial approximation teaching results. The blue curve corresponds to $d^{\mathcal{O}\left(\log^2 \frac{1}{\epsilon}\right)} = 2^{\mathcal{O}\left(\log^2 \frac{1}{\epsilon}\right)}$ where $d = 2$.