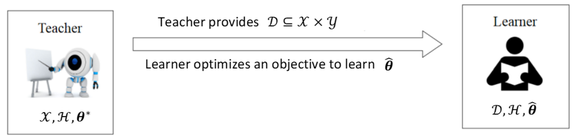
The teaching dimension of kernel perceptron

Akash Kumar[⋆] **Hanqi Zhang**[†] **Adish Singla**[⋆] **Yuxin Chen**[†]

[⋆]Max Planck Institute for Software Systems (MPI-SWS) [†]University of Chicago

Algorithmic Teaching: ERM Learner



Teaching 1D threshold function to an ERM learner

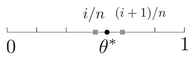


Figure 1: 0/1 loss

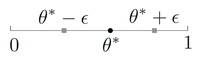


Figure 2: SVM (hinge loss)

Figure 3: Perceptron

	version-space	**SVM**	**perceptron**
TD	2 (discrete)	2 (continuous)	2 (continuous)
TS	$\{(\frac{i}{n}, 0), (\frac{i+1}{n}, 1)\}$	$\{(\theta^* - \epsilon, 0), (\theta^* + \epsilon, 1)\}$	$\{(\theta^*, -1), (\theta^*, 1)\}$

Constructive Setting

Teacher could provide *arbitrarily constructed* training set (teaching examples) $\mathcal{D}$ in the support of the data distribution $\mathcal{P}$. ‘

Kernel Perceptron Learner

Fix the training set $\mathcal{D} := \{(\mathbf{x}_i, y_i)\}_{i=1}^n$ where $\mathbf{x}_i \in \mathbb{R}^d$ and hypothesis $\boldsymbol{\theta} \in \mathbb{R}^d$.

Linear Perceptron

A homogeneous linear perceptron corresponding to hypothesis $\boldsymbol{\theta}$ is defined as:

$$f_{\boldsymbol{\theta}}(\mathbf{x}) := \text{sign}(\boldsymbol{\theta} \cdot \mathbf{x})$$

Perceptron Loss Function

For a labelled point $(\mathbf{x}, y)$, hypothesis $\boldsymbol{\theta}$ we consider the perceptron loss ℓ:

$$\ell(f_{\boldsymbol{\theta}}(\mathbf{x}), y) := \max(-y \cdot f_{\boldsymbol{\theta}}(\mathbf{x}), 0)$$

Optimal Perceptron Algorithm (Learner)

For a given training set $\mathcal{D}$, we consider the optimal perceptron algorithm $\boldsymbol{\mathcal{A}}_{opt}$ which minimizes the perceptron loss as follows:

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

Non-Linear Kernel Perceptron Learner

For a kernel operator $\mathcal{K} : \mathcal{X} \times \mathcal{X} \to \mathbb{R}$ inducing a *reproducing kernel Hilbert space* (RKHS) $\mathcal{H}_{\mathcal{K}}$, a non-linear kernel perceptron optimizes the following loss:

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

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. 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 $\mathcal{K}$. A reproducing kernel Hilbert space with $\mathcal{K}$ could be decomposed as $\mathcal{K}(\mathbf{x}, \mathbf{x}') = \langle \Phi(\mathbf{x}), \Phi(\mathbf{x}') \rangle$ [SS01] for any $\mathbf{x}, \mathbf{x}' \in \mathcal{X}$. Thus, we also identify $f_{\boldsymbol{\theta}}$ as $\sum_{i=1}^l \alpha_i \cdot \Phi(\mathbf{a}_i)$.

Power Non-Linear Kernels

(**polynomial**) $\mathcal{K}(\mathbf{x}, \mathbf{x}') = (\langle \mathbf{x}, \mathbf{x}' \rangle)^k$ (**Gaussian**) $\mathcal{K}(\mathbf{x}, \mathbf{x}') = e^{-\frac{\|\mathbf{x}-\mathbf{x}'\|^2}{2\sigma^2}}$

Our Contributions

We study the problem of teaching kernel perceptron (an ERM learner) under the *constructive* setting, where the teacher can construct arbitrary teaching examples in the support of the data distribution. This work extends the result of [LZ16] for teaching linear ERM learners.

- 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.
- We establish tight bounds on the teaching dimension of linear and polynomial perceptron. We exhibit optimal training sets that match these teaching dimensions.
- We show that for Gaussian kernelized perceptron, exact teaching is not possible with a finite set of examples, and then establish a $d^{\mathcal{O}(\log^2 \frac{1}{\epsilon})}$ bound on the ϵ-approximate teaching dimension.

Table 1: Main Results

	linear	**polynomial**	**gaussian**
TD (exact)	$\Theta(d)$	$\Theta\left(\binom{d+k-1}{k}\right)$	∞
ϵ-approximate TD	-	-	$d^{\mathcal{O}(\log^2 \frac{1}{\epsilon})}$
Assumption	-	1	2, 3

Teaching Complexity

Fix a kernel perceptron learner, a kernel $\mathcal{K} : \mathcal{X} \times \mathcal{X} \to \mathbb{R}$ with the corresponding RKHS feature map $\Phi(\cdot)$ and a target model $\boldsymbol{\theta}^* \in \mathcal{H}_{\mathcal{K}}$ $\Phi(\cdot)$.

Teaching Set

A set of labelled points $\mathcal{TS} \subseteq \mathcal{X} \times \mathcal{Y}$ provided by a helpful teacher for teaching a target hypothesis $\boldsymbol{\theta}^*$ to a kernel perceptron learner.

Teaching Dimension for Exact Parameters

We define the teaching dimension for *exact* parameter (upto decision boundary) of $\boldsymbol{\theta}^*$ corresponding to a kernel perceptron as $TD(t\boldsymbol{\theta}^*, \boldsymbol{\mathcal{A}}_{opt})$, which is the size of the smallest teaching set $\mathcal{TS}$ such that $\boldsymbol{\mathcal{A}}_{opt}(\mathcal{TS}) = \{t\boldsymbol{\theta}^*\}$ for some real $t > 0$, where

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

Approximate Teaching

Definition 1 (ϵ-**approximate TS**). *For a given $\epsilon > 0$, 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*

$$\left| \mathop{\mathbb{E}}_{(\mathbf{x},y)\sim\mathcal{P}}[\max(-y \cdot f^*(\mathbf{x}), 0)] - \mathop{\mathbb{E}}_{(\mathbf{x},y)\sim\mathcal{P}}\left[\max(-y \cdot \hat{f}(\mathbf{x}), 0)\right] \right| \leq \epsilon \text{ where } f^*(\cdot) = \boldsymbol{\theta}^* \cdot \Phi(\cdot),\ \hat{f}(\cdot) = \hat{\boldsymbol{\theta}} \cdot \Phi(\cdot)$$

Definition 2 (ϵ-**approximate TD**). *For a given $\epsilon > 0$, 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 $\mathcal{P}$.*

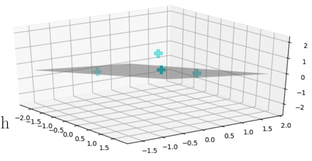
Teaching Homogeneous Linear Perceptron

We show $TD(\{t\boldsymbol{\theta}^*\}, \boldsymbol{\mathcal{A}}_{opt}) = \Theta(d)$ via constructing 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 $\mathbf{v}_d = \boldsymbol{\theta}^*$.



Teaching Non-Linear Kernel Perceptron

Main Ideas and Results for Polynomial Kernel Perceptron

- $\mathcal{H}_{\mathcal{K}}$ is isomorphic to space of homogeneous polynomials.
- **Assumption 1**: *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 $\forall i\ \mathbf{z}_i \in \mathcal{X}$.*
- Without **Assumption 1**, $\boldsymbol{\theta}^*$ *can't* be taught *exactly* and infeasibilty of *approximate* teaching for some pathological cases.
- Under **Assumption 1**, we achieve a tight bound on teaching dimension (Table 1)

Gaussian Kernel Perceptron

- Key Ideas
  - Truncate the taylor features of the Gaussian Kernel to obtain a finite dimensional kernel.
  - Inspired from polynomial setting, we make a necessary assumption on projection of $\boldsymbol{\theta}^*$ in truncated space.
  - Analyze the solutions of perceptron algorithm and pick bounded solutions. (**Assumption 3**)
- Truncated Taylor Features

$$\underbrace{\Phi_{k,\boldsymbol{\lambda}}(\mathbf{x}) = e^{-\frac{\|\mathbf{x}\|^2}{2\sigma^2}} \cdot \frac{\sqrt{c_{\boldsymbol{\lambda}}^k}}{\sqrt{k!}\sigma^k} \cdot \mathbf{x}^{\boldsymbol{\lambda}}}_{\text{coordinated Gaussian feature map: } \Phi(\cdot)} \longrightarrow \underbrace{\tilde{\Phi}_{k,\boldsymbol{\lambda}}(\mathbf{x}) = \Phi_{k,\boldsymbol{\lambda}}(\mathbf{x})}_{\text{taylor features } \forall k \leq s} \longrightarrow \underbrace{\tilde{\mathcal{K}}(\mathbf{x}, \mathbf{x}') = e^{-\frac{\|\mathbf{x}\|^2 + \|\mathbf{x}'\|^2}{2\sigma^2}} \sum_{k=0}^{s} \frac{1}{k!} \left(\frac{\langle \mathbf{x}, \mathbf{x}' \rangle}{\sigma^2}\right)^k}_{\text{truncated kernel with projection map}} \longrightarrow \mathbb{P}\text{ (projection map)}$$

Assumptions and Reformulation of Perceptron Algorithm

Assumption 2 (**Existence of orthogonal classifiers**): *For some $\epsilon > 0$, $\exists\ r = r(\boldsymbol{\theta}^*, \epsilon)$ 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 $\forall i\ \mathbf{z}_i \in \mathcal{X}$.*

Fix $\mathcal{TS}_{\boldsymbol{\theta}^*} := \{(\mathbf{z}_i, 1), (\mathbf{z}_i, -1)\}_{i=1}^{r-1} \cup \{(\mathbf{a}, 1)\}$ where $\mathbb{P}\boldsymbol{\theta}^* \cdot \mathbb{P}\Phi(\mathbf{a}) > 0$ and $\Phi(\mathbf{a}) \cdot \Phi(\mathbf{z}_i) \leq Q\epsilon$ (constant Q).

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

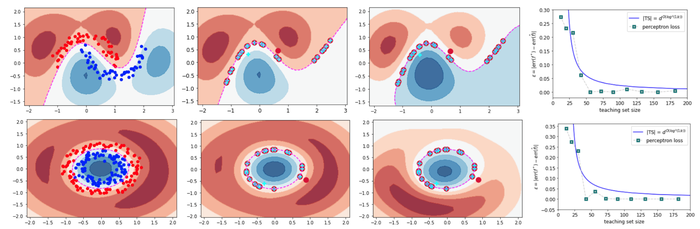
where for any $i \in [2r-1]$ $\ell(\beta_0, \gamma, \mathbf{x}_i, y_i) = -y_i \cdot \left(\beta_0 \cdot \mathcal{K}(\mathbf{a}, \mathbf{x}_i) + \sum_{j=1}^{r-1} \gamma_j \cdot \mathcal{K}(\mathbf{z}_j, \mathbf{x}_i)\right)$.

Note, $\hat{\boldsymbol{\theta}} \in \boldsymbol{\mathcal{A}}_{opt}(\mathcal{TS}_{\boldsymbol{\theta}^*})$ has the form $\beta_0 \cdot \mathcal{K}(\mathbf{a}, \cdot) + \sum_{j=1}^{r-1} \gamma_j \cdot \mathcal{K}(\mathbf{z}_j, \cdot)$ and target model has the form $\boldsymbol{\theta}^* = \sum_{i=1}^{l} \alpha_i \cdot \mathcal{K}(\mathbf{a}_i, \cdot)$.

Assumption 3 (**Bounded Cone**): *The learner optimizes to a solution $\hat{\boldsymbol{\theta}}$ for Eq. (1) with bounded coefficients, i.e., $\sum_{i=1}^{l} |\alpha_i|$ and $|\beta_0| + \sum_{j=1}^{r-1} |\gamma_j|$ are bounded where $\hat{\boldsymbol{\theta}} \in \mathcal{H}_{\mathcal{K}}$ has the form $\hat{\boldsymbol{\theta}} = \beta_0 \cdot \mathcal{K}(\mathbf{a}_j, \cdot) + \sum_{j=1}^{r-1} \gamma_j \cdot \mathcal{K}(\mathbf{z}_j, \cdot)$.*

Main Results

- Under **Assumptions 2-3** and $\epsilon > 0$, for any $\hat{f} \in \boldsymbol{\mathcal{A}}_{opt}(\mathcal{TS}_{\boldsymbol{\theta}^*})$ $\left|f^*(\mathbf{x}) - \hat{f}(\mathbf{x})\right| \leq \epsilon$.
- Under **Assumptions 2-3** and $\epsilon > 0$, the teaching set $\mathcal{TS}_{\boldsymbol{\theta}^*}$ constructed for Eq. (1) is an ϵ-approximate teaching set. (Table 1)



References

[SS01] Bernhard Scholkopf and Alexander J. Smola. *Learning with Kernels: Support Vector Machines, Regularization, Optimization, and Beyond*. Cambridge, MA, USA: MIT Press, 2001. ISBN: 0262194759.

[LZ16] Ji Liu and Xiaojin Zhu. “The Teaching Dimension of Linear Learners”. In: *Journal of Machine Learning Research* 17.162 (2016), pp. 1–25.