Nonparametric Iterative Machine Teaching

Chen Zhang[1], **Xiaofeng Cao**[1], **Weiyang Liu**[2,3], **Ivor W. Tsang**[4], **James T. Kwok**[5]

[1] Jilin University
[2] Max Planck Institute for Intelligent Systems
[3] University of Cambridge
[4] Agency for Science, Technology and Research
[5] Hong Kong University of Science and Technology

June 20, 2023

Source code is available at `https://github.com/chen2hang/NonparametricTeaching`.

Overview

What is Machine Teaching?

Machine teaching (MT) [10, 11] is the study of how to design the optimal teaching set, typically with minimal examples, so that learners can quickly learn target models based on these examples.

It can be considered as an inverse problem of machine learning, where machine learning aims to learn model parameters from a dataset, while MT aims to find a minimal dataset from the target model parameters.

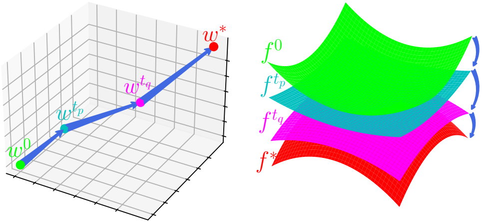
Considering the interaction manner between teachers and learners, MT can be conducted in either

- batch fashion [10, 5, 1, 6] where the teacher is allowed to interact with the learner once, or
- iterative fashion [2, 3, 4] where an iterative teacher would feed examples sequentially based on current status of the iterative learner.

Nonparametric Iterative Machine Teaching

Previous iterative machine teaching algorithms [2, 3, 9, 8] are solely based on parameterized families of target models. They mainly focus on convergence in the parameter space, resulting in difficulty when the target models are defined to be functions without dependency on parameters.

To address such a limitation, we study a more general task – **Nonparametric Iterative Machine Teaching**, which aims to teach nonparametric target models to learners in an iterative fashion.



(a) Parametric IMT (b) Nonparametric IMT

Cont.

Main Contribution:

- We comprehensively study **Nonparametric Iterative Machine Teaching**, which focuses on exploring iterative algorithms for teaching parameter-free target models from the optimization perspective.
- We propose two teaching algorithms, which are named Random Functional Teaching (RFT) and Greedy Functional Teaching (GFT), respectively. RFT is based on random sampling with ground truth labels, and the derivation of GFT is based on the maximization of an informative scalar.
- We theoretically analyze the asymptotic behavior of both RFT and GFT. We prove that per-iteration reduction of loss $\mathcal{L}$ for RFT and GFT has a negative upper bound expressed by the discrepancy of iterative teaching, and we derive that the iterative teaching dimension (ITD) of GFT is $\mathcal{O}(\psi(\frac{2\mathcal{L}(f^0)}{\tilde{\eta}\epsilon}))$, which is shown to be lower than the ITD of RFT, $\mathcal{O}(2\mathcal{L}(f^0)/\left(\tilde{\eta}\epsilon\right))$.

Teaching Settings

Functional Optimization: We define nonparametric iterative machine teaching as a functional minimization over the collection of potential teaching sequences $\mathbb{D}$ in the reproducing kernel Hilbert space:

$$\mathcal{D}^* = \arg\min_{\mathcal{D}\in\mathbb{D}} \quad \mathcal{M}(\hat{f}, f^*) + \lambda \cdot \mathrm{len}(\mathcal{D}) \qquad \text{s.t.} \quad \hat{f} = \mathcal{A}(\mathcal{D}), \tag{1}$$

where $\mathcal{M}$ denotes a discrepancy measure, $\mathrm{len}(\mathcal{D})$, which is regularized by a constant λ, is the length of the teaching sequence $\mathcal{D}$, and $\mathcal{A}$ represents the learning algorithm of learners.

Functional Teaching Algorithms

Algorithm 1 Random / Greedy Functional Teaching

Input: Target f^*, initial f^0, per-iteration pack size k, small constant $\epsilon > 0$ and maximal iteration number T.

Set $f^t \leftarrow f^0, t = 0$.

while $t \leq T$ and $\|f^t - f^*\|_{\mathcal{H}} \geq \epsilon$ **do**

  The teacher selects k teaching examples:
  Initialize the pack of teaching examples $\mathcal{K} = \emptyset$;
  for $j = 1$ **to** k **do**

    (**RFT**) 1. Pick $\boldsymbol{x}_j^{t*} \in \mathcal{X}$ randomly;

    (**GFT**) 1. Pick $\boldsymbol{x}_j^{t*}$ with the maximal difference between f^t and f^*:

$$\boldsymbol{x}_j^{t*} = \underset{\boldsymbol{x}_i^t \in \mathcal{X} - \{\boldsymbol{x}_i^{t*}\}_{i=1}^{j-1}}{\arg\max} \left| f^t(\boldsymbol{x}_i^t) - f^*(\boldsymbol{x}_i^t) \right|;$$

    2. Add $\left(\boldsymbol{x}_j^{t*}, y_j^{t*} = f^*\left(\boldsymbol{x}_j^{t*}\right)\right)$ into $\mathcal{K}$.

  end
  Provide $\mathcal{K}$ to learners.

  The learner updates f^t based on received $\mathcal{K}$:
  $f^t \leftarrow f^t - \eta^t \mathcal{G}(\mathcal{L}; f^t; \mathcal{K})$.

  Set $t \leftarrow t + 1$.

end

- It is straightforward for teachers to pick examples randomly and feed them to learners, which derives a simple teaching baseline called **Random Functional Teaching**.
- **Greedy Functional Teaching** is to search examples with steeper gradients, since the gradient norm at the optimal example should be maximal at every iteration.

Analysis of Iterative Teaching Dimension

To conduct *theoretical analysis* on the iterative teaching dimension, we have listed the assumptions [7] on $\mathcal{L}$ and the kernel function $K(\boldsymbol{x}, \boldsymbol{x}') \in \mathcal{H}$ below.

Assumption 1

The loss function $\mathcal{L}(f)$ is $L_{\mathcal{L}}$-Lipschitz smooth, *i.e.*, $\forall f, f' \in \mathcal{H}$ and $\boldsymbol{x} \in \mathcal{X}$

$$\left| E_{\boldsymbol{x}}\left[\nabla_f \mathcal{L}(f)\right] - E_{\boldsymbol{x}}\left[\nabla_f \mathcal{L}(f')\right] \right| \leq L_{\mathcal{L}} \left| E_{\boldsymbol{x}}\left[f\right] - E_{\boldsymbol{x}}\left[f'\right] \right|, \tag{2}$$

where $L_{\mathcal{L}} \geq 0$ is a constant.

Assumption 2

The kernel function $K(\boldsymbol{x}, \boldsymbol{x}') \in \mathcal{H}$ is bounded, *i.e.*, $\forall \boldsymbol{x}, \boldsymbol{x}' \in \mathcal{X}$, $K(\boldsymbol{x}, \boldsymbol{x}') \leq M_K$, where $M_K \geq 0$ is a constant.

Cont. (RFT)

Lemma (Sufficient Descent for RFT)

Under Assumption 1 and 2, if $\eta^t \leq 1/(2L_{\mathcal{L}} \cdot M_K)$, RFT teachers can reduce the loss $\mathcal{L}$ by $\mathcal{L}(f^{t+1}) - \mathcal{L}(f^t) \leq -\eta^t/2 \cdot \mathbb{S}_{\mathcal{L}}(f^t; \boldsymbol{x}^t)$.

Theorem (Convergence for RFT)

Suppose the model of learners is initialized with $f^0 \in \mathcal{H}$ and returns $f^t \in \mathcal{H}$ after t iterations, we have the upper bound of minimal $\mathbb{S}_{\mathcal{L}}(f^t; \boldsymbol{x}^t)$ as $\min_t \mathbb{S}_{\mathcal{L}}(f^t; \boldsymbol{x}^t) \leq 2\mathcal{L}(f^0)/\left(\tilde{\eta}t\right)$, *where* $0 < \tilde{\eta} = \min_t \eta^t \leq \frac{1}{2L_{\mathcal{L}} \cdot M_K}$.

Cont. (GFT)

Lemma (Sufficient Descent for GFT)

Under Assumption 1 and 2, if $\eta^t \leq 1/(2L_{\mathcal{L}} \cdot M_K)$, GFT teachers can reduce the loss $\mathcal{L}$ at a faster speed, $\mathcal{L}(f^{t+1}) - \mathcal{L}(f^t) \leq -\eta^t/2 \cdot \mathbb{S}_{\mathcal{L}}(f^t; \boldsymbol{x}^{t*}) \leq -\eta^t/2 \cdot \mathbb{S}_{\mathcal{L}}(f^t; \boldsymbol{x}^t)$.

Theorem (Convergence for GFT)

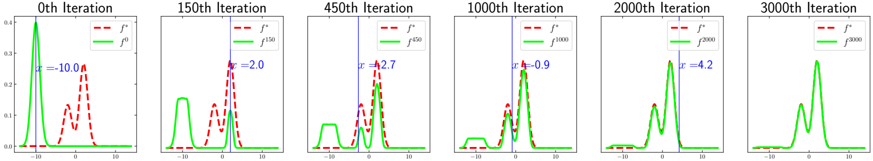
Suppose the model of learners is initialized with $f^0 \in \mathcal{H}$ and returns $f^t \in \mathcal{H}$ after t iterations, we have the upper bound of minimal $\mathbb{S}_{\mathcal{L}}(f^t; \boldsymbol{x}^{j*})$ as $\min_j \mathbb{S}_{\mathcal{L}}(f^j; \boldsymbol{x}^{j*}) \leq \frac{2}{\tilde{\eta}\psi(t)} \mathcal{L}(f^0)$, where $0 < \tilde{\eta} = \min\limits_t \eta^t \leq \frac{1}{2L_{\mathcal{L}} \cdot M_K}$, $\psi(t) = \sum_{j=0}^{t-1} \gamma^j$ and $\gamma^j = \frac{\mathbb{S}_{\mathcal{L}}(f^j; \boldsymbol{x}^j)}{\mathbb{S}_{\mathcal{L}}(f^j; \boldsymbol{x}^{j*})} \in (0, 1]$ named greedy ratio.

Experiments and Results

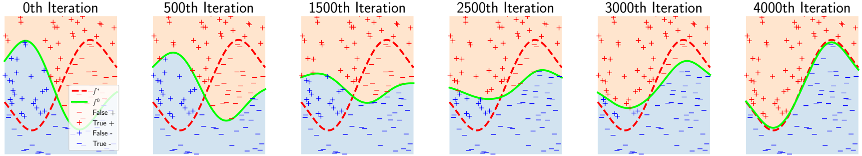
We test our RFT and GFT on both synthetic and real-world data, on which we find these two algorithms present satisfactory capability to tackle nonparametric teaching tasks.

- **Synthetic data.**

1D Gaussian Mixture.



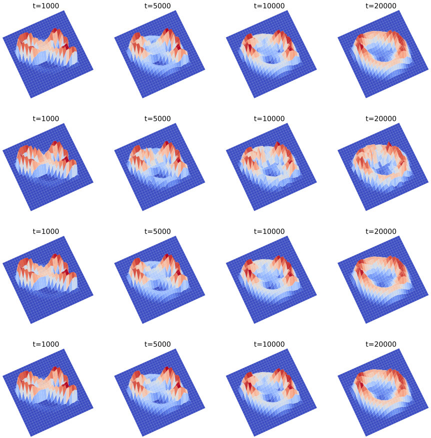
2D Classification.



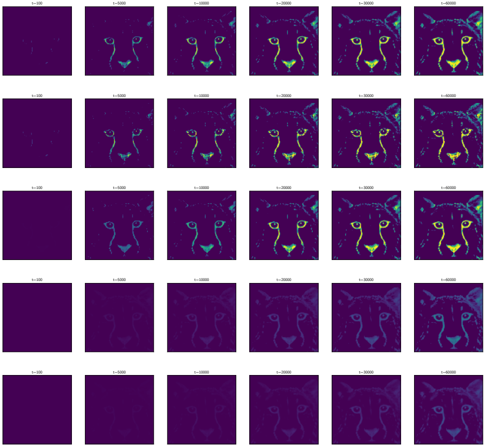
Cont.

- **Real-world data.**

Digit Correction.

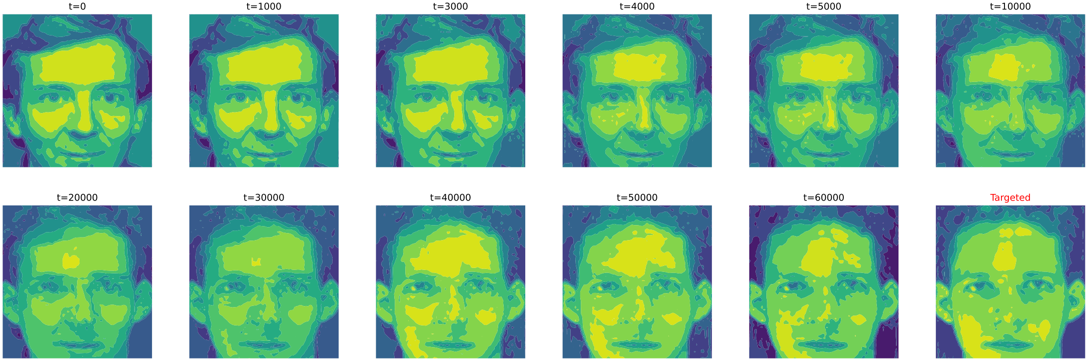


Cheetah Impartation.



Cont.

Sketch for Missing Person Report.



Thank you for listening!

References I

[1] Akash Kumar, Hanqi Zhang, Adish Singla, and Yuxin Chen. The teaching dimension of kernel perceptron. In AISTATS, 2021.

[2] Weiyang Liu, Bo Dai, Ahmad Humayun, Charlene Tay, Chen Yu, Linda B Smith, James M Rehg, and Le Song. Iterative machine teaching. In ICML, 2017.

[3] Weiyang Liu, Bo Dai, Xingguo Li, Zhen Liu, James Rehg, and Le Song. Towards black-box iterative machine teaching. In ICML, 2018.

[4] Weiyang Liu, Zhen Liu, Hanchen Wang, Liam Paull, Bernhard Schölkopf, and Adrian Weller. Iterative teaching by label synthesis. In NeurIPS, 2021.

[5] Farnam Mansouri, Yuxin Chen, Ara Vartanian, Jerry Zhu, and Adish Singla. Preference-based batch and sequential teaching: Towards a unified view of models. In NeurIPS, 2019.

[6] Hong Qian, Xu-Hui Liu, Chen-Xi Su, Aimin Zhou, and Yang Yu. The teaching dimension of regularized kernel learners. In ICML, 2022.

[7] Zebang Shen, Zhenfu Wang, Alejandro Ribeiro, and Hamed Hassani. Sinkhorn barycenter via functional gradient descent. In NeurIPS, 2020.

References II

[8] Pei Wang, Kabir Nagrecha, and Nuno Vasconcelos. Gradient-based algorithms for machine teaching. In CVPR, 2021.

[9] Zhaozhuo Xu, Beidi Chen, Chaojian Li, Weiyang Liu, Le Song, Yingyan Lin, and Anshumali Shrivastava. Locality sensitive teaching. In NeurIPS, 2021.

[10] Xiaojin Zhu. Machine teaching: An inverse problem to machine learning and an approach toward optimal education. In AAAI, 2015.

[11] Xiaojin Zhu, Adish Singla, Sandra Zilles, and Anna N Rafferty. An overview of machine teaching. arXiv preprint arXiv:1801.05927, 2018.