Nonparametric Iterative Machine Teaching

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

Machine Teaching

Machine teaching (MT) 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.

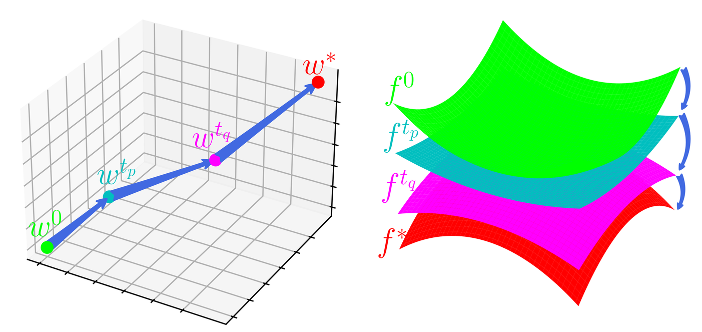
It can be considered an inverse problem of machine learning, where machine learning aims to learn model parameters from a dataset, while MT aims to find a minimal dataset from the target model parameters.

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

- batch fashion where the teacher is allowed to interact with the learner once, or
- iterative fashion where an iterative teacher would feed examples sequentially based on current status of the iterative learner.

Motivation

Previous iterative machine teaching algorithms 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

Figure: Comparison between parameterized and nonparametric IMT.

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, $\mathcal{O}(2\mathcal{L}(f^0)/(\tilde\eta\epsilon))$.

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:

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

where $\mathcal{M}$ denotes a discrepancy measure, 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 $t \le T$ and $\|f^t - f^*\|_{\mathcal{H}} \ge \epsilon$ **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 $x_j^{t*} \in \mathcal{X}$ randomly;

    (**GFT**) 1. Pick x_j^{t*} with the maximal difference between f^t and f^*:

$$x_j^{t*} = \arg\max_{x_i^t \in \mathcal{X} - \{x_i^{t*}\}_{i=1}^{j-1}} \left| f^t(x_i^t) - f^*(x_i^t) \right|;$$

    2. Add $(x_j^{t*}, y_j^{t*} = f^*(x_j^{t*}))$ 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

Analysis of Iterative Teaching Dimension

Assumption 1. The loss function $\mathcal{L}(f)$ is $L_{\mathcal{L}}$-Lipschitz smooth, *i.e.*, $\forall f, f' \in \mathcal{H}$ and $x \in \mathcal{X}$

$$|E_x[\nabla_f \mathcal{L}(f)] - E_x[\nabla_f \mathcal{L}(f')]| \le L_{\mathcal{L}} |E_x[f] - E_x[f']|,$$

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

Assumption 2. The kernel function $K(x, x') \in \mathcal{H}$ is bounded, *i.e.*, $\forall x, x' \in \mathcal{X}$, $K(x, x') \le M_K$, where $M_K \ge 0$ is a constant.

Lemma 3 (Sufficient Descent for **RFT**). Under Assumption 1 and 2, if $\eta^t \le 1/(2L_{\mathcal{L}}\cdot M_K)$, RFT teachers can reduce the loss $\mathcal{L}$ by $\mathcal{L}(f^{t+1}) - \mathcal{L}(f^t) \le -\eta^t/2 \cdot \mathbb{S}_{\mathcal{L}}(f^t; x^t)$.

Theorem 4 (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 $\mathbb{S}_{\mathcal{L}}(f^t; x^t)$ as $\min_t \mathbb{S}_{\mathcal{L}}(f^t; x^t) \le 2\mathcal{L}(f^0)/(\tilde\eta t)$, where $0 < \tilde\eta = \min_t \eta^t \le \frac{1}{2L_{\mathcal{L}}\cdot M_K}$.

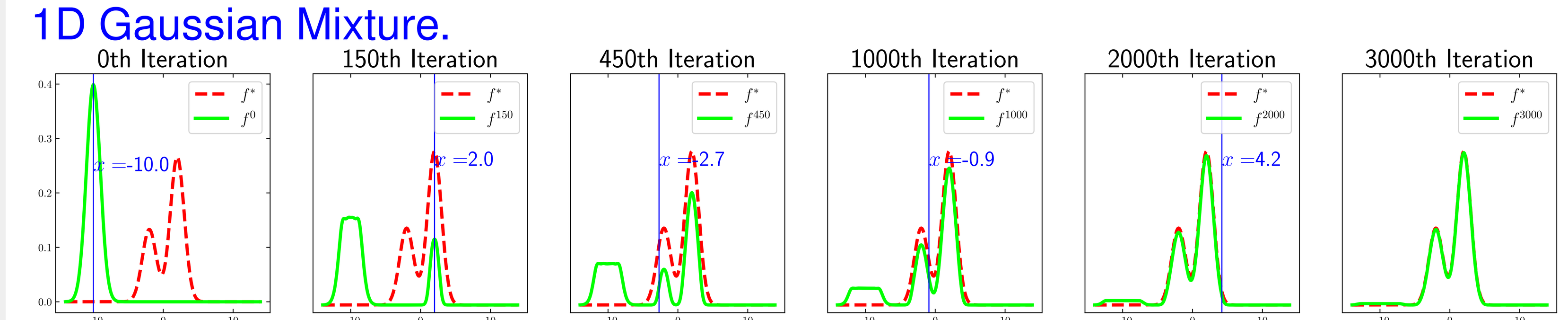
Lemma 5 (Sufficient Descent for **GFT**). Under Assumption 1 and 2, if $\eta^t \le 1/(2L_{\mathcal{L}}\cdot M_K)$, GFT teachers can reduce the loss $\mathcal{L}$ at a faster speed, $\mathcal{L}(f^{t+1}) - \mathcal{L}(f^t) \le -\eta^t/2 \cdot \mathbb{S}_{\mathcal{L}}(f^t; x^{t*}) \le -\eta^t/2 \cdot \mathbb{S}_{\mathcal{L}}(f^t; x^t)$.

Theorem 6 (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 $\mathbb{S}_{\mathcal{L}}(f^t; x^{j*})$ as $\min_j \mathbb{S}_{\mathcal{L}}(f^j; x^{j*}) \le \frac{2}{\tilde\eta\psi(t)}\mathcal{L}(f^0)$, where $0 < \tilde\eta = \min_t \eta^t \le \frac{1}{2L_{\mathcal{L}}\cdot M_K}$, $\psi(t) = \sum_{j=0}^{t-1}\gamma^j$ and $\gamma^j = \frac{\mathbb{S}_{\mathcal{L}}(f^j; x^j)}{\mathbb{S}_{\mathcal{L}}(f^j; x^{j*})} \in (0, 1]$ named greedy ratio.

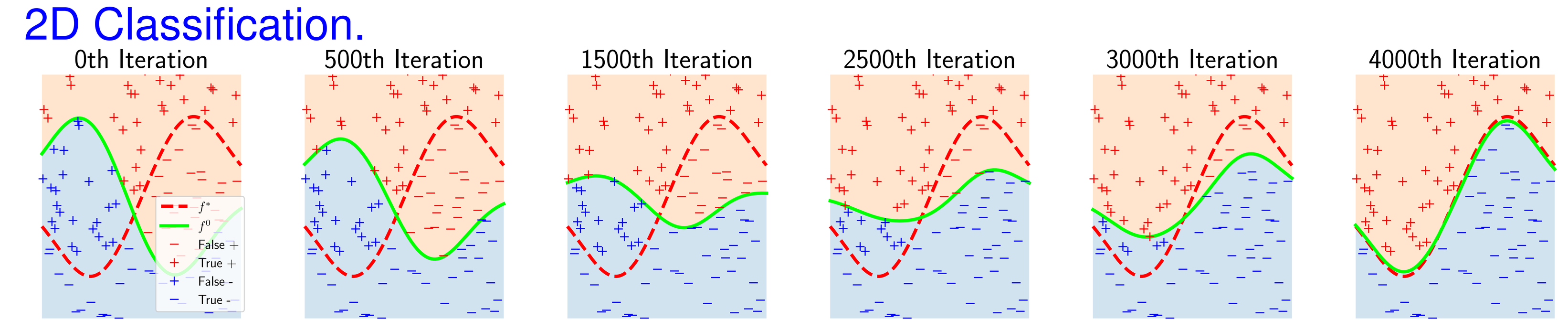
Experiments and Results

Synthetic data.

1D Gaussian Mixture.

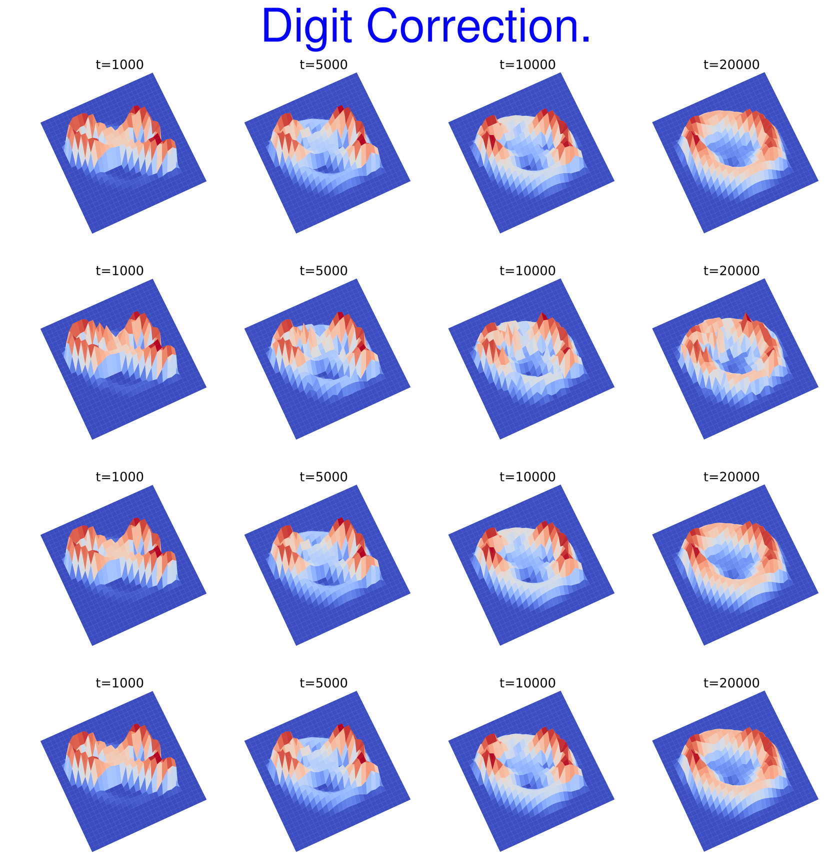


2D Classification.

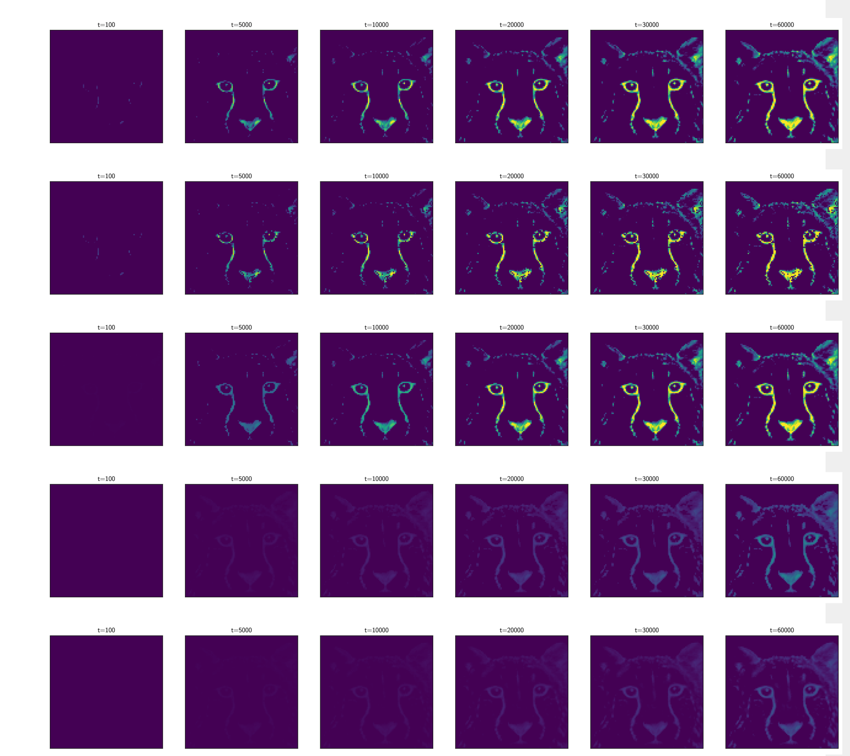


Real-world data.

Digit Correction.



Cheetah Impartation.



Sketch for Missing Person Report.

[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

Mailto: xiaofengcao@jlu.edu.cn