Nonparametric Teaching for Multiple Learners

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

November 16, 2023

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

Overview

1. **What is Machine Teaching?**

2. **Multi-learner nonparametric teaching**
   2.1 Teaching Settings
   2.2 Vanilla Multi-learner Teaching
   2.3 Communicated Multi-learner Teaching

3. **Experiments and Results**

What is Machine Teaching?

Machine teaching (MT) [14, 15] considers the problem of how to design the most effective teaching set, typically with the smallest amount of (teaching) examples possible, to facilitate rapid learning of the target models by learners based on these examples.

It can be thought of as an inverse of machine learning, in the sense that the learner is to learn models on a given dataset, while the teacher is to seek a (minimal) dataset from a target model.

Depending on how teachers and learners interact with each other, MT can be carried out in either

- batch fashion [14, 10, 6, 12] which focuses on single-round interaction, that is, the most representative and effective teaching dataset are designed to be fed to the learner in one shot, or
- iterative fashion [7, 8, 9] where an iterative teacher would feed examples based on learners' status (current learnt models) round by round.

Multi-learner nonparametric teaching

Previous nonparametric teaching algorithms [13] merely focus on the single-learner setting (*i.e.*, teaching a scalar-valued target model or function to a single learner). To empower them to fulfill the practical needs of complex tasks, we introduce a more comprehensive task called **Multi-learner Nonparametric Teaching** (MINT). In MINT, the teacher aims to instruct multiple learners, with each learner focusing on learning a scalar-valued target model.

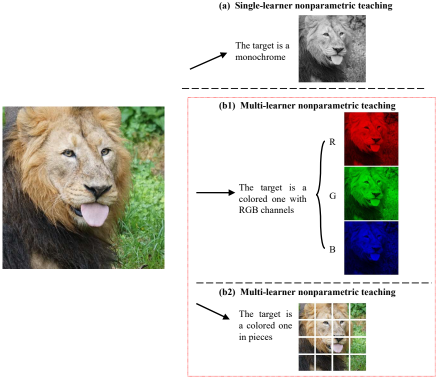


Figure: Comparison between the single-learner teaching and MINT.

Cont.

Main Contribution:

- By analyzing general vector-valued RKHS [2, 11, 1], we study the **multi-learner nonparametric teaching** (MINT), where the teacher selects examples based on a vector-valued target function (each component of the vector-valued function is a scalar-valued function for a single learner), such that multiple learners can learn their own target models simultaneously.
- By enabling the communication among multiple learners, learners can update themselves with a linear combination of current learnt functions of all learners [4, 3]. We study a communicated MINT where the teacher not only selects examples but also injects the guidance of communication.
- Under mild assumptions, we characterize the efficiency of our multi-learner generalization of nonparametric teaching. More importantly, we also empirically demonstrate its efficiency.

Teaching Settings

Vector-valued Functional Optimization: We define multi-learner noparametric teaching as a vector-valued functional minimization over the collection of potential teaching sequences $\mathbb{D}$ in the vector-valued reproducing kernel Hilbert space:

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

where $\mathcal{M}$ denotes a discrepancy measure, $\mathrm{len}(\boldsymbol{\mathcal{D}})$, which is regularized by a constant λ, is the length of the teaching sequence $\boldsymbol{\mathcal{D}}$, and $\mathcal{A}$ represents the learning algorithm of learners. Specifically, $\mathcal{A}$ is taken as $\hat{\boldsymbol{f}}^* = \underset{\boldsymbol{f}\in\mathcal{H}^d}{\arg\min}\, \mathbb{E}_{(\boldsymbol{x},\boldsymbol{y})}\left[\mathcal{L}(\boldsymbol{f}(\boldsymbol{x}), \boldsymbol{y})\right]$, where $(\boldsymbol{x}, \boldsymbol{y}) \in \mathcal{X}^d \times \mathcal{Y}^d$ and $(\boldsymbol{x}, \boldsymbol{y}) \sim [\mathbb{Q}_i(x_i, y_i)]^d$. Evaluated at an example vector $(\boldsymbol{x}, \boldsymbol{y}) = [(x_{i,j_i}, y_{i,j_i})]^d$ with the example index $j_i \in \mathbb{N}_k$, the multi-learner convex loss $\mathcal{L}$ therein is $\mathcal{L}(\boldsymbol{f}(\boldsymbol{x}), \boldsymbol{y}) = \sum_{i=1}^d \mathcal{L}_i(f_i(x_{i,j_i}), y_{i,j_i}) = E_{\boldsymbol{x}}\left[[\mathcal{L}_i(f_i, y_{i,j_i})]^d\right]$, where $\mathcal{L}_i$ is the convex loss for i-th learner.

Cont.

We investigate MINT in the gray-box setting, which is equivalent to the one considered in [13]. To facilitate the theoretical analysis, we adopt some moderate assumptions regarding $\mathcal{L}_i$ and kernels, which align with those made in [13].

Assumption 1

Each loss $\mathcal{L}_i(f_i), i \in \mathbb{N}_d$ is $L_{\mathcal{L}_i}$-Lipschitz smooth, *i.e.*, $\forall f_i, f'_i \in \mathcal{H}, x_i \in \mathcal{X}$ and $i \in \mathbb{N}_d$

$$\left|E_{x_i}\left[\nabla_f \mathcal{L}_i(f_i)\right] - E_{x_i}\left[\nabla_f \mathcal{L}_i(f'_i)\right]\right| \leq L_{\mathcal{L}_i}\left|E_{x_i}\left[f_i\right] - E_{x_i}\left[f'_i\right]\right|,$$

where $L_{\mathcal{L}_i} \geq 0$ is a constant. To simplify the notation, we assume that $L_{\mathcal{L}_i} = L_{\mathcal{L}}$ for all $i \in \mathbb{N}_d$.

Assumption 2

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

Vanilla Multi-learner Teaching

In tackling MINT, we begin by examining a basic scenario in which multiple learners concurrently learns corresponding components of a vector-valued target function without communication between them [5, 3].

Lemma 3 (Sufficient Descent for multi-learner RFT)

Suppose there are d learners, and the example mean for each learner is $\mu_i = \mathbb{E}_{x_i \sim \mathbb{P}_i(x_i)}(x_i) < \infty$, and the variance $\sigma_i^2 = \mathbb{E}_{x_i \sim \mathbb{P}_i(x_i)}(x_i - \mu_i)^2 < \infty, i \in \mathbb{N}_d$. Under the Lipschitz smooth and bounded kernel assumptions, if $\eta_i^t \leq \frac{1}{2L_{\mathcal{L}} \cdot M_K}$ for all $i \in \mathbb{N}_d$, then RFT teachers can, on average, reduce the multi-learner loss $\mathcal{L}(\boldsymbol{f})$ by:

$$\mathbb{E}_{\boldsymbol{x} \sim [\mathbb{P}_i(x_i)]^d} \left[\mathcal{L}(\boldsymbol{f}^{t+1}) - \mathcal{L}(\boldsymbol{f}^t) \right] \leq -\frac{\tilde{\eta}^t}{2} \sum_{i=1}^{d} (m_{i,t}(\mu_i) + \frac{m''_{i,t}(\mu_i)}{2} \sigma_i^2), \tag{2}$$

where $\tilde{\eta}^t = \min_{i \in \mathbb{N}_d} \eta_i^t$ and $m_{i,t}(\dot{x}) := E_{\dot{x}}[(\nabla_f \mathcal{L}_i(f)|_{f=f_i^t})^2]$.

Cont.

Theorem 4 (Convergence for multi-learner RFT)

Suppose the vector-valued model for multiple learners is initialized with $\boldsymbol{f}^0 \in \mathcal{H}^d$ and returns $\boldsymbol{f}^t \in \mathcal{H}^d$ after t iterations, we have the upper bound of $\min_{i\in\mathbb{N}_d}\left(m_{i,t}(\mu_i) + m''_{i,t}(\mu_i)\sigma_i^2/2\right)$ w.r.t. t:

$$\min_{i\in\mathbb{N}_d}\left(m_{i,t-1}(\mu_i) + m''_{i,t-1}(\mu_i)\sigma_i^2/2\right) \leq 2\mathbb{E}_{\boldsymbol{x}\sim[\mathbb{P}_i(x_i)]^d}\left[\mathcal{L}(\boldsymbol{f}^0)\right]/(d\dot{\eta}t), \tag{3}$$

where $0 < \dot{\eta} = \min_{l\in\{0\}\bigcup\mathbb{N}_{t-1}} \tilde{\eta}^l \leq 1/(2L_{\mathcal{L}} \cdot M_K)$, and given a small constant $\epsilon > 0$ it would take approximately $\mathcal{O}(2\mathbb{E}_{\boldsymbol{x}\sim[\mathbb{P}_i(x_i)]^d}\left[\mathcal{L}\left(\boldsymbol{f}^0\right)\right]/(d\dot{\eta}\epsilon))$ iterations to reach a stationary point.

Cont.

NEURAL INFORMATION PROCESSING SYSTEMS

Lemma 5 (Sufficient Descent for multi-learner GFT)

Under the same assumption, if $\eta_i^t \leq \frac{1}{2L_{\mathcal{L}} \cdot M_K}$ for all $i \in \mathbb{N}_d$, the GFT teachers can achieve a greater reduction in the multi-learner loss $\mathcal{L}$:

$$\mathbb{E}_{\boldsymbol{x} \sim [\mathbb{P}_i(x_i)]^d} \left[\mathcal{L}(\boldsymbol{f}^{t+1}) - \mathcal{L}(\boldsymbol{f}^t) \right] \leq -\frac{\tilde{\eta}^t}{2} \sum_{i=1}^{d} m_{i,t}(x_i^{t*}), \quad (4)$$

where $\tilde{\eta}^t$ and $m_{i,t}(\cdot)$ retain their previous meaning.

Cont.

Theorem 6 (Convergence for multi-learner GFT)

Suppose the vector-valued model for multiple learners is initialized with $\boldsymbol{f}^0 \in \mathcal{H}^d$ and returns $\boldsymbol{f}^t \in \mathcal{H}^d$ after t iterations, we have the upper bound of $\min_{i \in \mathbb{N}_d} m_{i,t}(x_i^{t*})$ w.r.t. t:

$$\min_{i \in \mathbb{N}_d} m_{i,t-1}(x_i^{t-1*}) \leq \frac{2}{d\dot{\eta}t} \mathbb{E}_{\boldsymbol{x} \sim [\mathbb{P}_i(x_i)]^d} \left[\mathcal{L}(\boldsymbol{f}^0)\right] + \frac{1}{d} \sum_{l=0}^{t-1} \sum_{i=1}^{d} \left(\|x_i^{l*} - \mu_i\|_2 \right), \tag{5}$$

where $\dot{\eta}$ has the same definition as before.

Communicated Multi-learner Teaching

An infant would integrate previously learnt knowledge to grasp a new target concept, such as comprehending what a zebra is by combining the learnt ideas of horses and black-and-white stripes. Such an efficient paradigm motivates us to explore the communicated MINT, which enables the communication between learners.

Proposition 5

If the proximity between $\boldsymbol{f}^t$ and $\boldsymbol{f}^*$ is sufficiently close, meaning that $\|\boldsymbol{f}^t - \boldsymbol{f}^*\|_{\mathcal{H}^d} \leq \epsilon$ where ϵ is a tiny positive constant, then A^t equals the identity matrix I_d.

Lemma 6

Under Lipschitz smooth assumption, the communication across learners will result in a reduction of the multi-learner convex loss $\mathcal{L}$ by
$0 \leq \mathcal{L}(\boldsymbol{f}^t) - \mathcal{L}(A^t \boldsymbol{f}^t) \leq 2L_{\mathcal{L}}\|\boldsymbol{f}^t - \boldsymbol{f}^*\|_{\mathcal{H}^d}$.

Cont.

Theorem 7

Suppose the communication in the t-th iteration of multiple learners is denoted by the matrix A^t and returns $\boldsymbol{f}_{A^t}^{t+1} \in \mathcal{H}^d$, for both RFT and GFT we have:

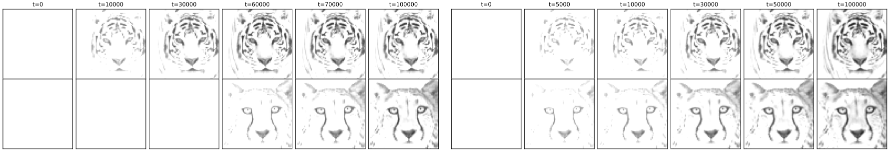
$$\mathbb{E}_{\boldsymbol{x}\sim[\mathbb{P}_i(x_i)]^d}\left[\mathcal{L}(\boldsymbol{f}_{A^t}^{t+1}) - \mathcal{L}(\boldsymbol{f}^t)\right] \leq \mathbb{E}_{\boldsymbol{x}\sim[\mathbb{P}_i(x_i)]^d}\left[\mathcal{L}(\boldsymbol{f}_{A^t}^{t+1}) - \mathcal{L}(A^t\boldsymbol{f}^t)\right] \leq 0.$$

Experiments and Results

Testing the teaching of a multi-learner (vector-valued) target model, MINT presents more satisfactory performance than repeatedly carrying out the single-learner teaching, which is consistent with our theoretical findings.

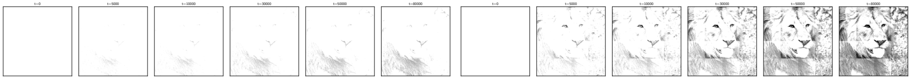
- **MINT in gray scale.**

Simultaneous teaching of a tiger and a cheetah.



(a) Single-learner teaching (b) Vanilla MINT

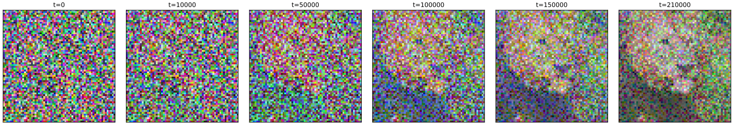
Teaching of a lion by partition.



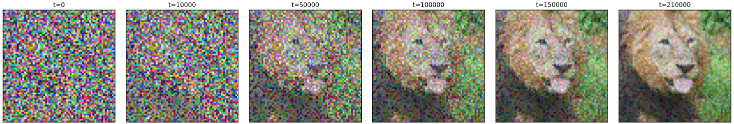
(c) Single-learner teaching (d) Vanilla MINT

Cont.

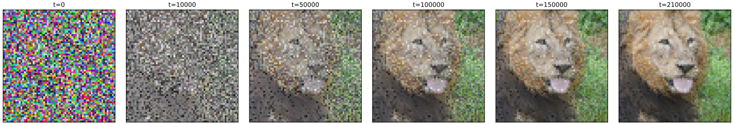
- **MINT in three (RGB) channels.**



(a) Single-learner teaching.



(b) Vanilla MINT.



(c) Communicated MINT.

Thank you for listening!

References I

[1] Mauricio A Alvarez, Lorenzo Rosasco, Neil D Lawrence, et al. Kernels for vector-valued functions: A review. Foundations and Trends® in Machine Learning, 4(3):195–266, 2012.

[2] Claudio Carmeli, Ernesto De Vito, and Alessandro Toigo. Vector valued reproducing kernel hilbert spaces of integrable functions and mercer theorem. Analysis and Applications, 4(04):377–408, 2006.

[3] Nicolò Cesa-Bianchi, Pierre Laforgue, Andrea Paudice, et al. Multitask online mirror descent. Transactions of Machine Learning Research.

[4] Peter Gehler and Sebastian Nowozin. On feature combination for multiclass object classification. In ICCV, 2009.

[5] Sham M Kakade, Shai Shalev-Shwartz, and Ambuj Tewari. Regularization techniques for learning with matrices. The Journal of Machine Learning Research, 13(1):1865–1890, 2012.

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

[7] 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.

References II

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

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

[10] 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.

[11] Ha Q Minh and Vikas Sindhwani. Vector-valued manifold regularization. In ICML, 2011.

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

[13] Chen Zhang, Xiaofeng Cao, Weiyang Liu, Ivor Tsang, and James Kwok. Nonparametric iterative machine teaching. In ICML, 2023.

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

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