Nonparametric Teaching for Multiple Learners

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

Machine Teaching

Machine teaching (MT) 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 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 where an iterative teacher would feed examples based on learners' status (current learnt models) round by round, such that the learner can converge to a target model within fewer rounds.

Motivation

Previous nonparametric teaching algorithms 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.

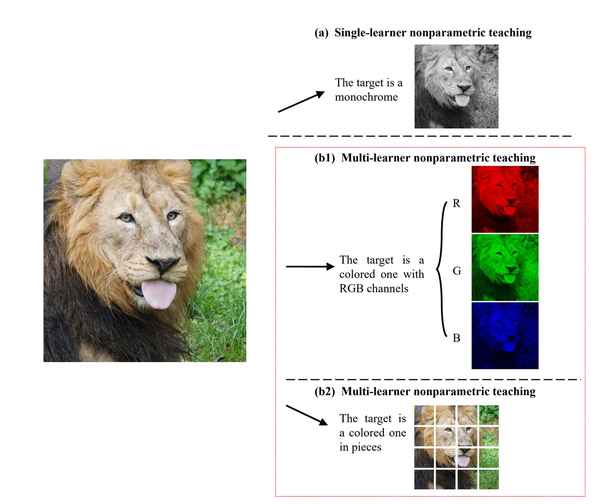


Figure: Comparison between the single-learner teaching and MINT.

Main Contribution:

- By analyzing general vector-valued RKHS, we study the **multi-learner nonparametric teaching** (MINT), where the teacher selects examples based on a vector-valued target function (each component of it is a scalar-valued one for a single learner) such that multiple learners can learn its components simultaneously in a fast speed.
- Allowing the communication across multiple learners, that is, learners are allowed to carry out linear combination on current learnt functions of all learners, we investigate a communicated MINT where the teacher not only selects examples but also constructs a matrix as the guide of communication in each iteration.
- Under a mild assumption, we theoretically prove the efficiency of our multi-learner generalization of nonparametric teaching. We also empirically demonstrate its applicability and efficiency in extensive multi-learner experiments.

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:

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

where $\mathcal{M}$ denotes a discrepancy measure, 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 $\hat{\boldsymbol{f}}^* = \operatorname*{arg\,min}_{\boldsymbol{f}\in\mathcal{H}^d} \mathbb{E}_{(\boldsymbol{x},\boldsymbol{y})}[\mathcal{L}(\boldsymbol{f}(\boldsymbol{x}),\boldsymbol{y})]$, 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.

Vanilla Multi-learner Teaching

Lemma 1 (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:

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

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

Theorem 2 (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 $\mathcal{O}\left(2(\mathbb{E}_{\boldsymbol{x}\sim[\mathbb{P}_i(x_i)]^d}\left[\mathcal{L}(\boldsymbol{f}^0)\right] - \epsilon)/(d\dot{\eta}\min_{i\in\mathbb{N}_d}\left(m_{i,t-1}(\mu_i) + m''_{i,t-1}(\mu_i)\sigma_i^2/2\right))\right)$ iterations to reduce the multi-learner loss $\mathcal{L}$ to a sufficiently small value and to reach a stationary point in terms of $\mathcal{L}$.

Lemma 3 (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 $\mathcal{L}$:

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

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

Theorem 4 (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, and given a small constant $\epsilon > 0$ it would need around $\mathcal{O}\left(2(\mathbb{E}_{\boldsymbol{x}\sim[\mathbb{P}_i(x_i)]^d}\left[\mathcal{L}(\boldsymbol{f}^0)\right] - \epsilon)/(d\dot{\eta}\min_{i\in\mathbb{N}_d} m_{i,t-1}(x_i^{t-1*}))\right)$ iterations to decrease the multi-learner loss $\mathcal{L}$ to a sufficiently small value and to reach a stationary point in terms of $\mathcal{L}$.

Communicated Multi-learner Teaching

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 $0 \leq \mathcal{L}(\boldsymbol{f}^t) - \mathcal{L}(A^t\boldsymbol{f}^t) \leq 2L_\mathcal{L}\|\boldsymbol{f}^t - \boldsymbol{f}^*\|_{\mathcal{H}^d}$.

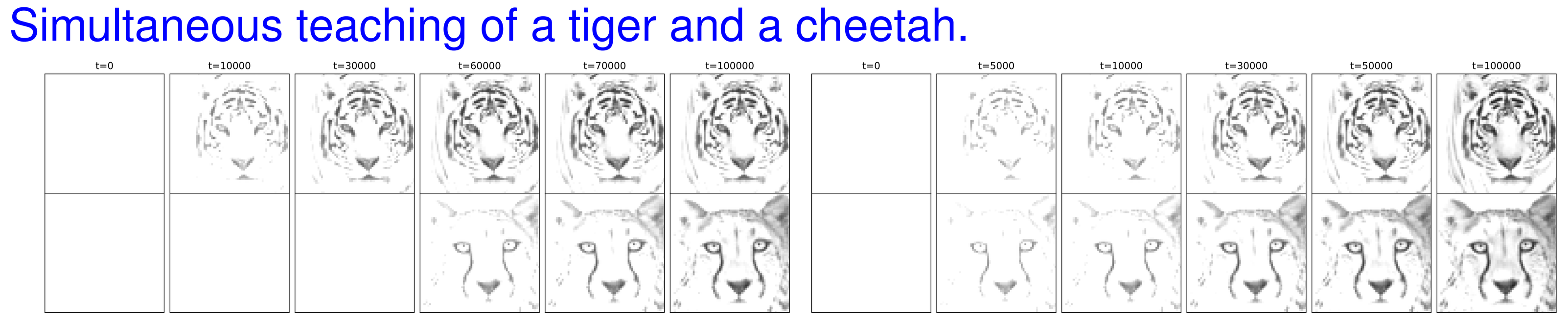
Theorem 7 Suppose the communication in the t-th iteration of multiple learners is denoted by the matrix A^t and returns $\boldsymbol{f}_{A^t}^{t+1} \in \mathcal{H}^d$, for both RFT and GFT we have:

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

Experiments and Results

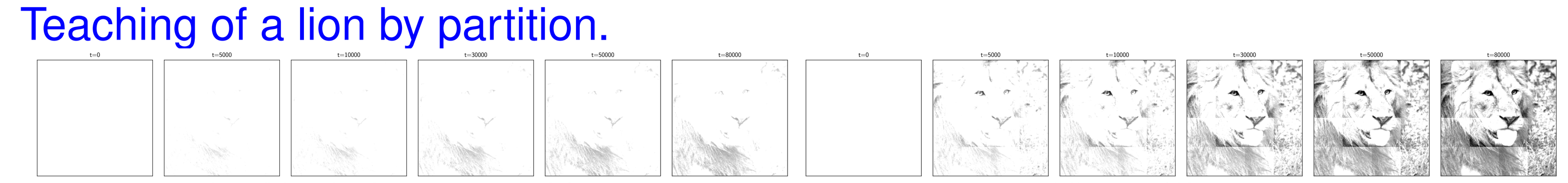
- **MINT in gray scale.**

Simultaneous teaching of a tiger and a cheetah.



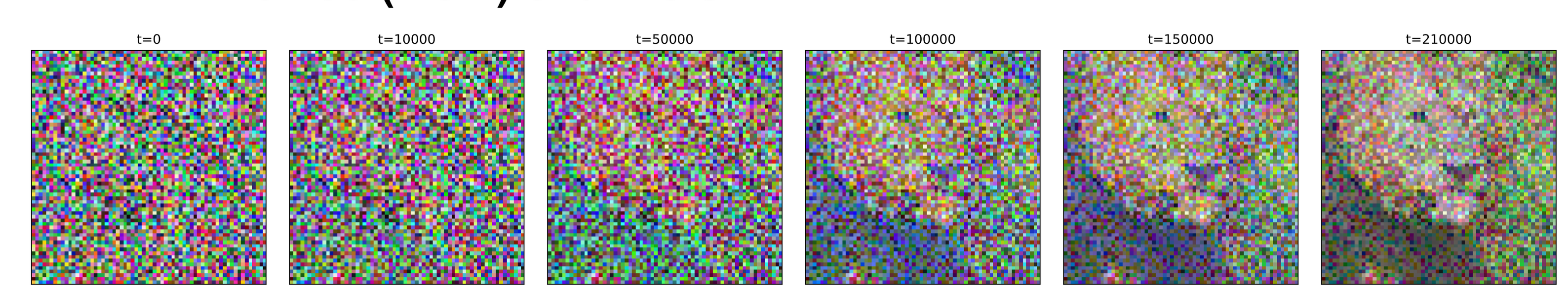
(a) Single-learner teaching (b) Vanilla MINT

Teaching of a lion by partition.

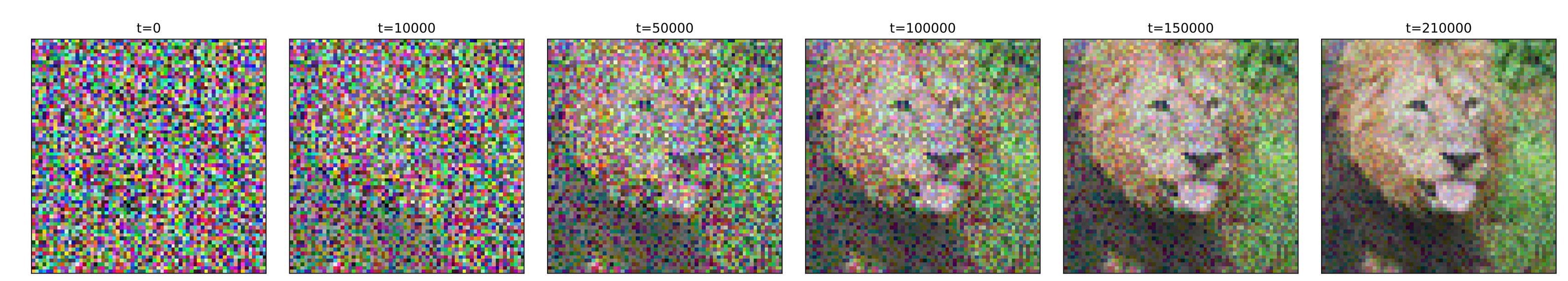


(c) Single-learner teaching (d) Vanilla MINT

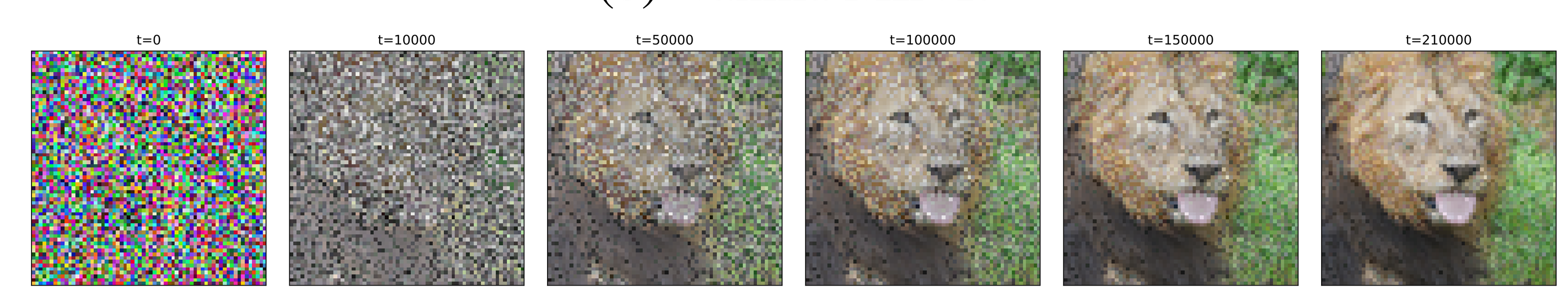
- **MINT in three (RGB) channels.**



(a) Single-learner teaching.



(b) Vanilla MINT.



(c) Communicated MINT.

[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