NEURAL INFORMATION PROCESSING SYSTEMS

NeurIPS 2021

Iterative Teaching by Label Synthesis

Weiyang Liu* Zhen Liu* Hanchen Wang*

Liam Paull Bernhard Schölkopf Adrian Weller

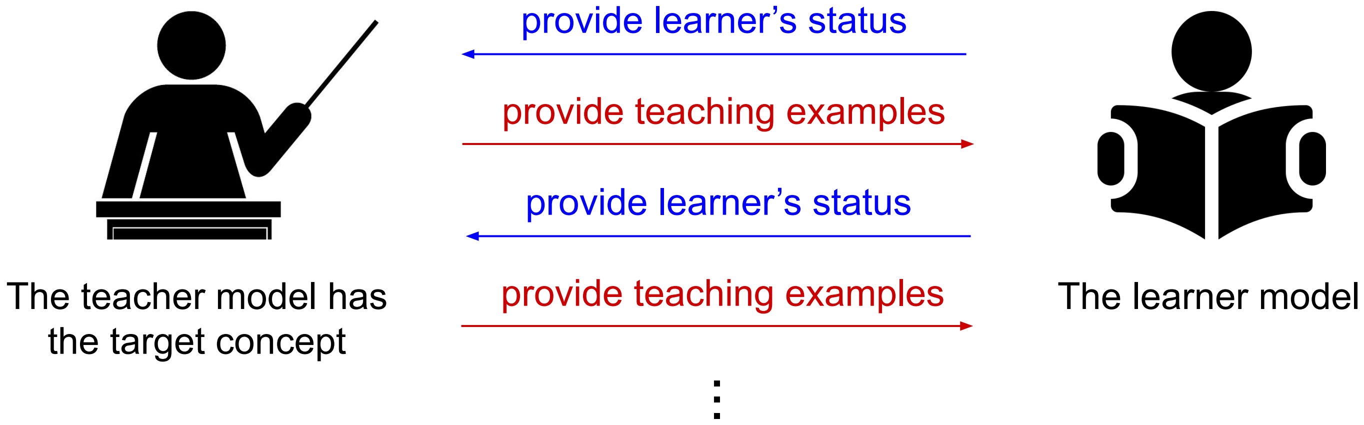
UNIVERSITY OF CAMBRIDGE

Max Planck Institute for Intelligent Systems

Mila

Iterative Machine Teaching (IMT)

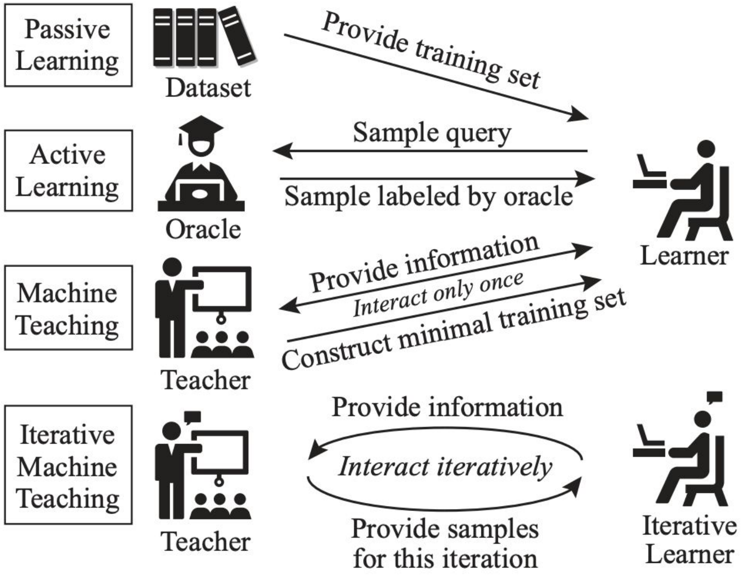
- How it works?
  - The learner is some machine learning model that aims to learn a set of target parameters.
  - The communication between the teacher and learner is contrained to be examples.



until the learner learns the target concept

Iterative Machine Teaching (IMT)

- Comparison to other machine learning paradigms



The teacher and learner interact with each other iteratively!

Existing problems that motivate our method

- Classic IMT algorithms need to traverse the entire dataset to obtain the teaching examples for the learner.
    - Computationally expensive and not scalable to large datasets!
- Classic IMT algorithms typically restrict the teaching to example selection.
    - Low teaching capacity!
- Classic IMT algorithms solve a combinatorial problem by nature.
    - Nontrivial to learn a parameterized teaching policy!

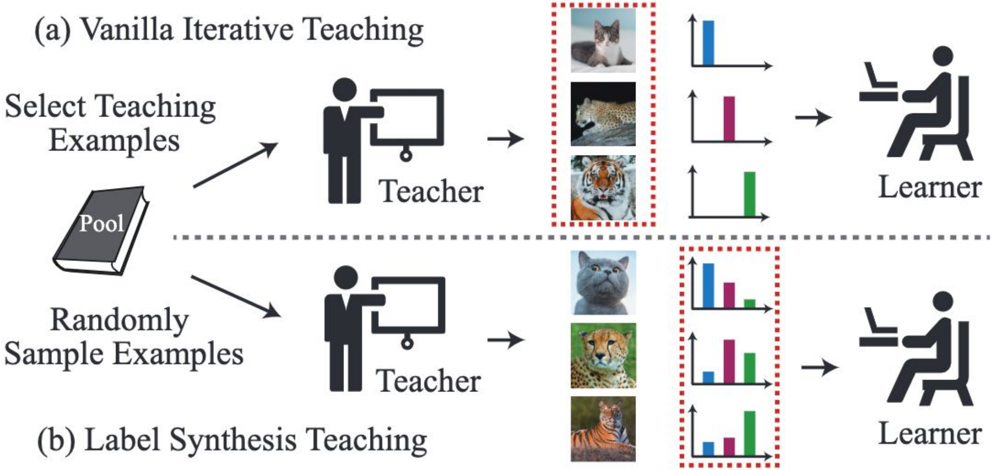
Existing problems that motivate our method

Classic IMT algorithms

- Need to traverse the entire dataset to obtain the teaching examples for the learner
    - Computationally expensive and not scalable to large datasets!
- Solve a combinatorial problem by nature
    - Nontrivial to learn a parameterized teaching policy!
- Typically restrict the form of teaching to example selection

Our Approach: Label Synthesis Teaching (LAST)

- We aim to avoid the traverse of the entire dataset by teaching the learner through label synthesis instead of example selection.



The red dottedframes indicate the teacher's efforts.

Why label synthesis?

- Effectively avoids traversing the dataset.
- Yields a computational cost that is independent of the dataset size.
- Provides a unified framework to think about label smoothing, knowledge distillation, etc.
- Has the same convergence speed-up guarantees as IMT.

Theorem 1 (Exponential teachability). *Assume that the learner loss function ℓ_i has the property of interpolation, L_i-Lipschitz, and convexity. And f is order-1 μ strongly convex. Then LAST can achieve ET with $g(y)=c_1\|\boldsymbol{w}^t-\boldsymbol{w}^*\|$,* i.e., $\mathbb{E}\{\|\boldsymbol{w}^T-\boldsymbol{w}^*\|^2\}\leq(1-c_1\eta_t\bar{\mu}+c_1^2\eta_t^2L_{\max})^T\left\|\boldsymbol{w}^0-\boldsymbol{w}^*\right\|^2$ *where* $L_{\max}=\max_i L_i$ *and* $\bar{\mu}=\sum_i\mu_i/n$. *It implies that* $(\log\frac{1}{c_2})^{-1}\log(\frac{1}{\epsilon})$ *samples are needed to achieve* $\mathbb{E}\{\|\boldsymbol{w}^T-\boldsymbol{w}^*\|^2\}\leq\epsilon$. $c_2=1-c_1\eta_t\bar{\mu}+c_1^2\eta_t^2L_{\max}$ *and* c_1 *is adjusted such that* $0<c_1\eta_t<\bar{\mu}/L_{\max}$.

Two LAST variants

- LAST is solving the following optimizaiton in general (d is some discrepancy measure)

$$\min_{\{\boldsymbol{y}^1,\cdots,\boldsymbol{y}^T\}} d(\boldsymbol{w}^T, \boldsymbol{w}^*)$$

learner's parameters after T steps target parameters

- Two ways of approximating the solution
    - One-step approximation with greedy LAST:

$$\min_{\boldsymbol{y}^1} d(\boldsymbol{w}^1, \boldsymbol{w}^*)$$

    - Multi-step approximation with parameterized LAST:

$$\min_{\{\boldsymbol{y}^1,\cdots,\boldsymbol{y}^T\}} d(\boldsymbol{w}^T, \boldsymbol{w}^*) \longrightarrow \min_{\boldsymbol{\theta}} \left\| \boldsymbol{w}^T(\boldsymbol{\theta}) - \boldsymbol{w}^* \right\|_2^2$$

Greedy LAST

- Intuition: synthesize the label that leads to the maximum discrepancy reduction.
- Step 1: randomly select an example from the dataset.
- Step 2: generate the label of the selected example with

$$\min_{\boldsymbol{y}} \left\{ \left\| \boldsymbol{w}^t - \eta_t \frac{\partial \ell(\boldsymbol{x}, \boldsymbol{y} | \boldsymbol{w}^t)}{\partial \boldsymbol{w}^t} - \boldsymbol{w}^* \right\|_2^2 \right\}$$

original gradients

current learner's parameters

target parameters

- Step 3: update the learner with gradients using the synthesized label

$$\boldsymbol{w}^t = \boldsymbol{w}^{t-1} - \eta_t \frac{\partial \ell(\boldsymbol{x}^t, \boldsymbol{y}^t | \boldsymbol{w}^t)}{\partial \boldsymbol{w}^t}$$

Parameterized LAST

- Intuition: use a parameterized teaching policy and learn it end-to-end by (1) *unrolling multi-step gradient updates* or (2) *policy gradients*.
- Nested Optimization: solve the following optimization by performing gradient descent on theta.

$$\min_{\boldsymbol{\theta}} \left\| \boldsymbol{w}^T(\boldsymbol{\theta}) - \boldsymbol{w}^* \right\|_2^2$$

$$\text{s.t.} \;\; \boldsymbol{w}^T(\boldsymbol{\theta}) = \arg\min_{\boldsymbol{w}} \mathbb{E}_{\{\boldsymbol{x},\tilde{\boldsymbol{y}}\}} \left\{ \ell(\boldsymbol{x}, \pi_{\boldsymbol{\theta}}(\boldsymbol{x}, \tilde{\boldsymbol{y}}, \boldsymbol{w}^t, \boldsymbol{w}^*) | \boldsymbol{w}) \right\}$$

Parameterized LAST

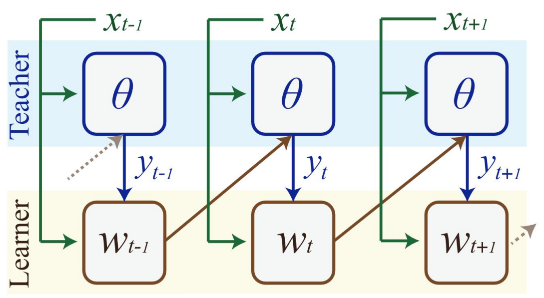
- Intuition: use a parameterized teaching policy and learn it end-to-end.
- Nested Optimization: solve the following optimization by performing gradient descent on theta.

$$\min_{\boldsymbol{\theta}} \left\| \boldsymbol{w}^T(\boldsymbol{\theta}) - \boldsymbol{w}^* \right\|_2^2$$

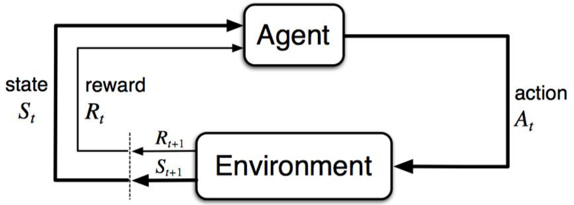
$$\text{s.t. } \boldsymbol{w}^T(\boldsymbol{\theta}) = \arg\min_{\boldsymbol{w}} \mathbb{E}_{\{\boldsymbol{x}, \tilde{\boldsymbol{y}}\}} \left\{ \ell(\boldsymbol{x}, \pi_{\boldsymbol{\theta}}(\boldsymbol{x}, \tilde{\boldsymbol{y}}, \boldsymbol{w}^t, \boldsymbol{w}^*) | \boldsymbol{w}) \right\}$$

Ways to learn the parameterized LAST

- Unrolling the parameterized teaching policy into multi-step gradient updates.

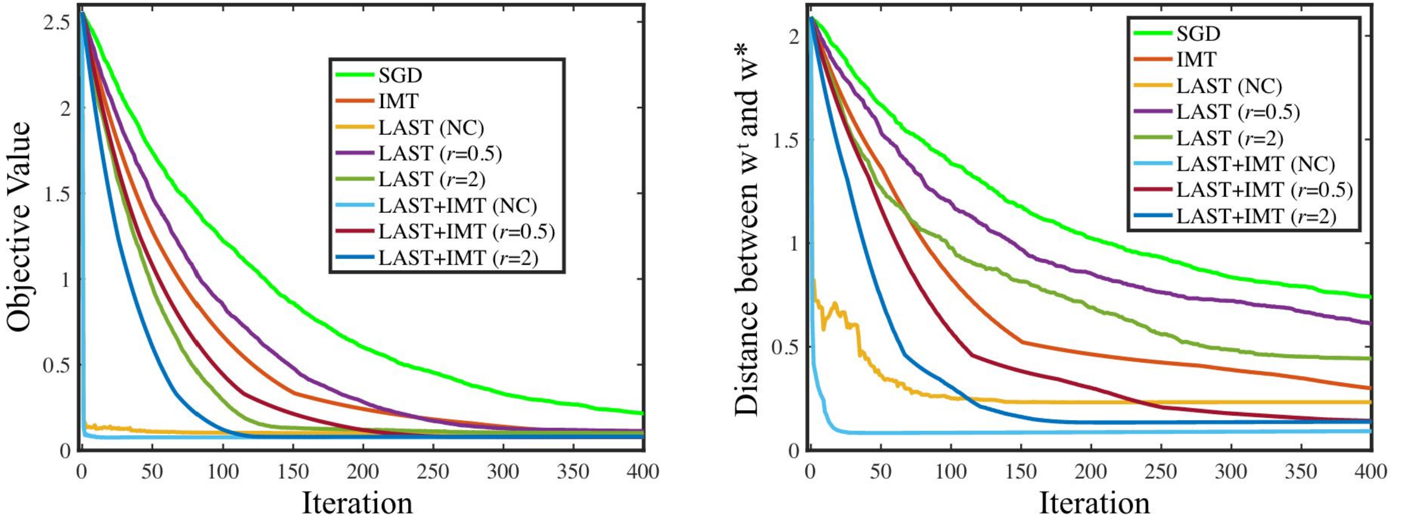


- Use the negative objective as the reward function and use policy gradients to update the teacher parameters.



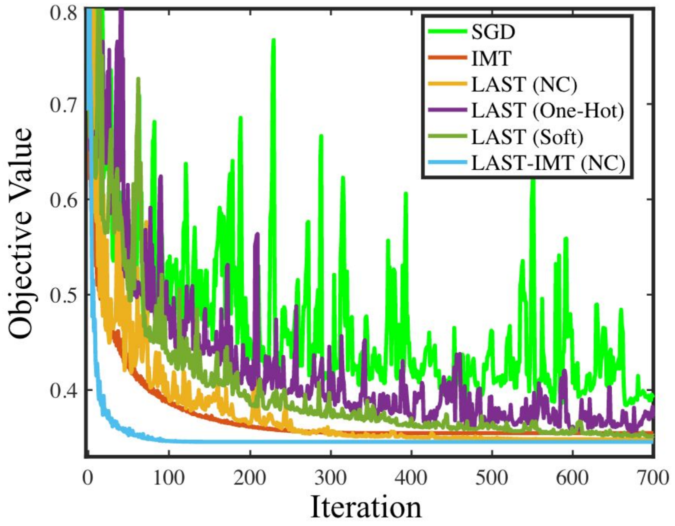
Teaching least square regression learners

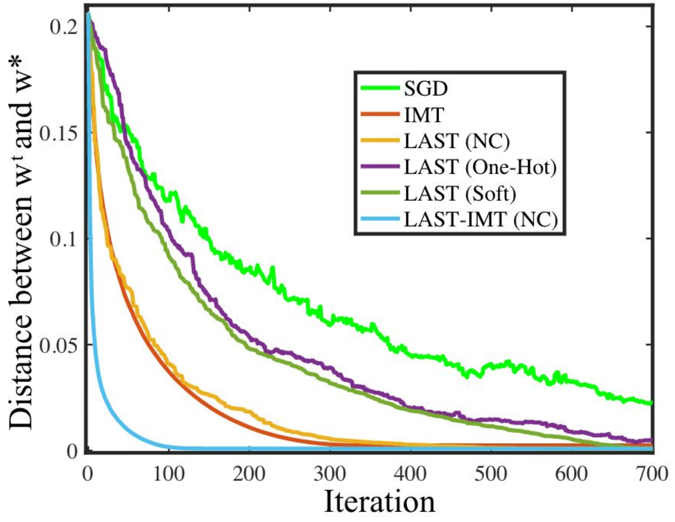
- SGD, IMT, Greedy LAST
- Greedy LAST + IMT: first use IMT to select examples and then use greedy LAST to synthesize labels.



Teaching logistic regression learners

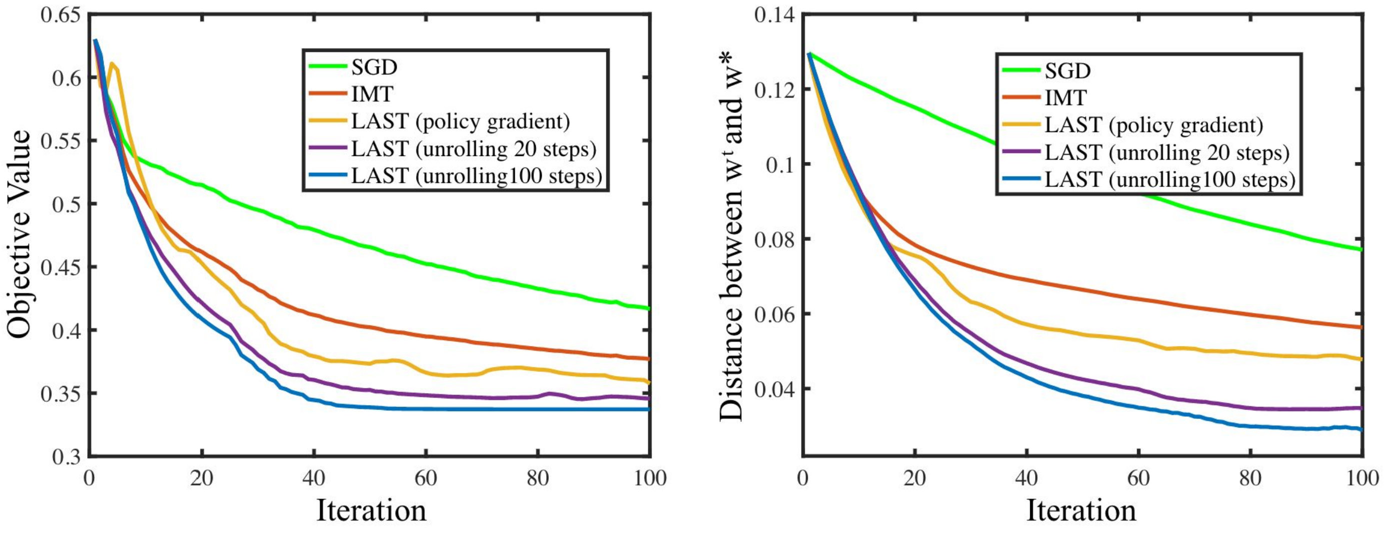
- Greedy LAST



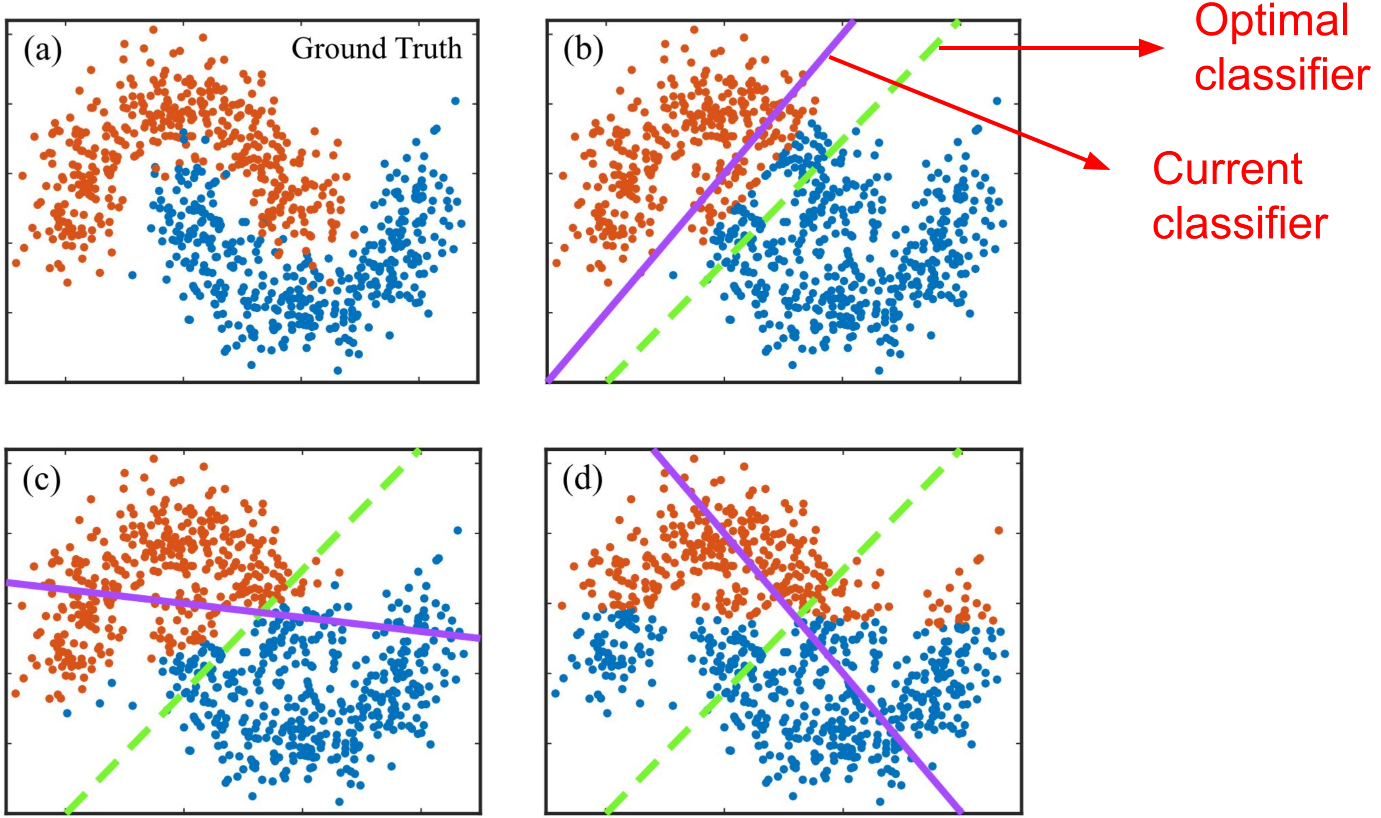


Teaching logistic regression learners

- Parameterized LAST
    - Hyperparameters and settings are slightly different from the previous experiments.



How LAST changes the ground truth label



Summary

- We propose a novel iterative teaching paradigm by label synthesis.
- Advantages of LAST
  - Scalable: applicable to large datasets
  - Flexible: applicable in various settings and well connected to existing soft label methods
  - Easy to train: the parameterized LAST is end-to-end trainable
  - Theoretically guaranteed: yielding the same convergence speed-up as IMT
  - Empirically validated: achieving comparable or better empirical convergence as IMT