Towards Black-box Iterative Machine Teaching

Weiyang Liu, Bo Dai, Xingguo Li, Zhen Liu, Jim Rehg, Le Song

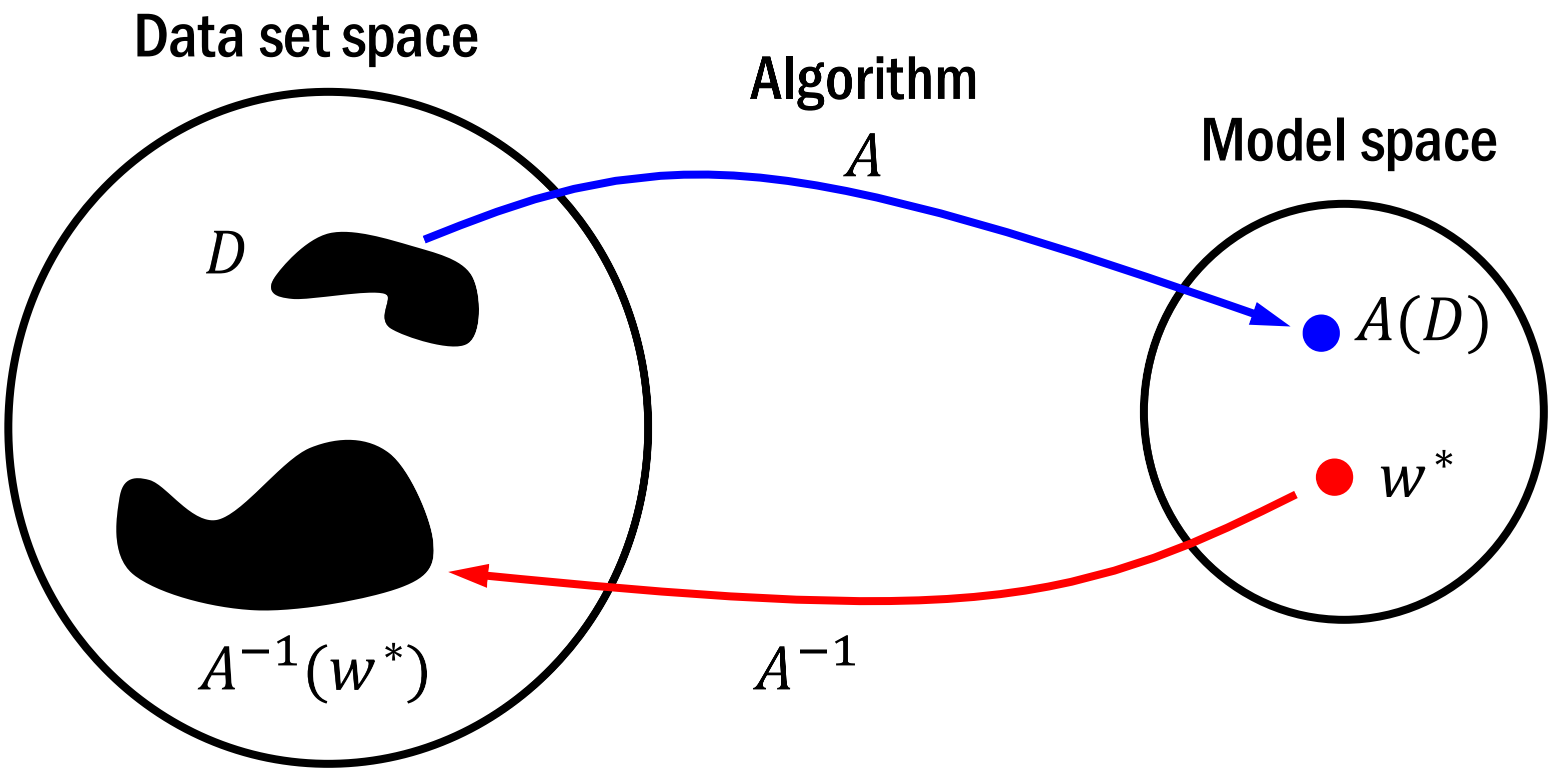
College of Computing

Georgia Institute of Technology

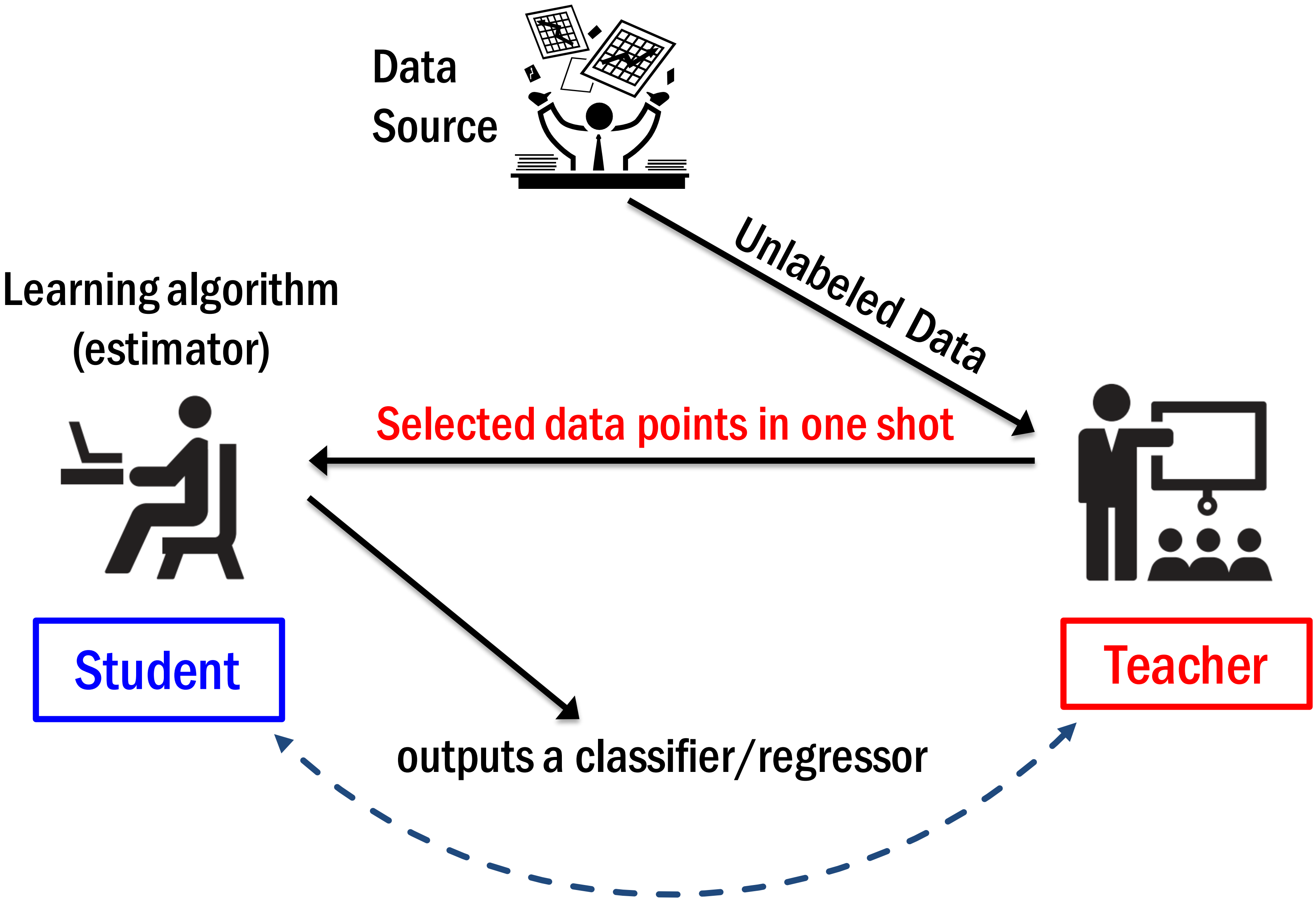
Machine teaching

An inverse problem to machine learning

- Teacher has a target model w^*, and knows learner's algorithm A
- Find the smallest data set (teaching dimension) to steer A

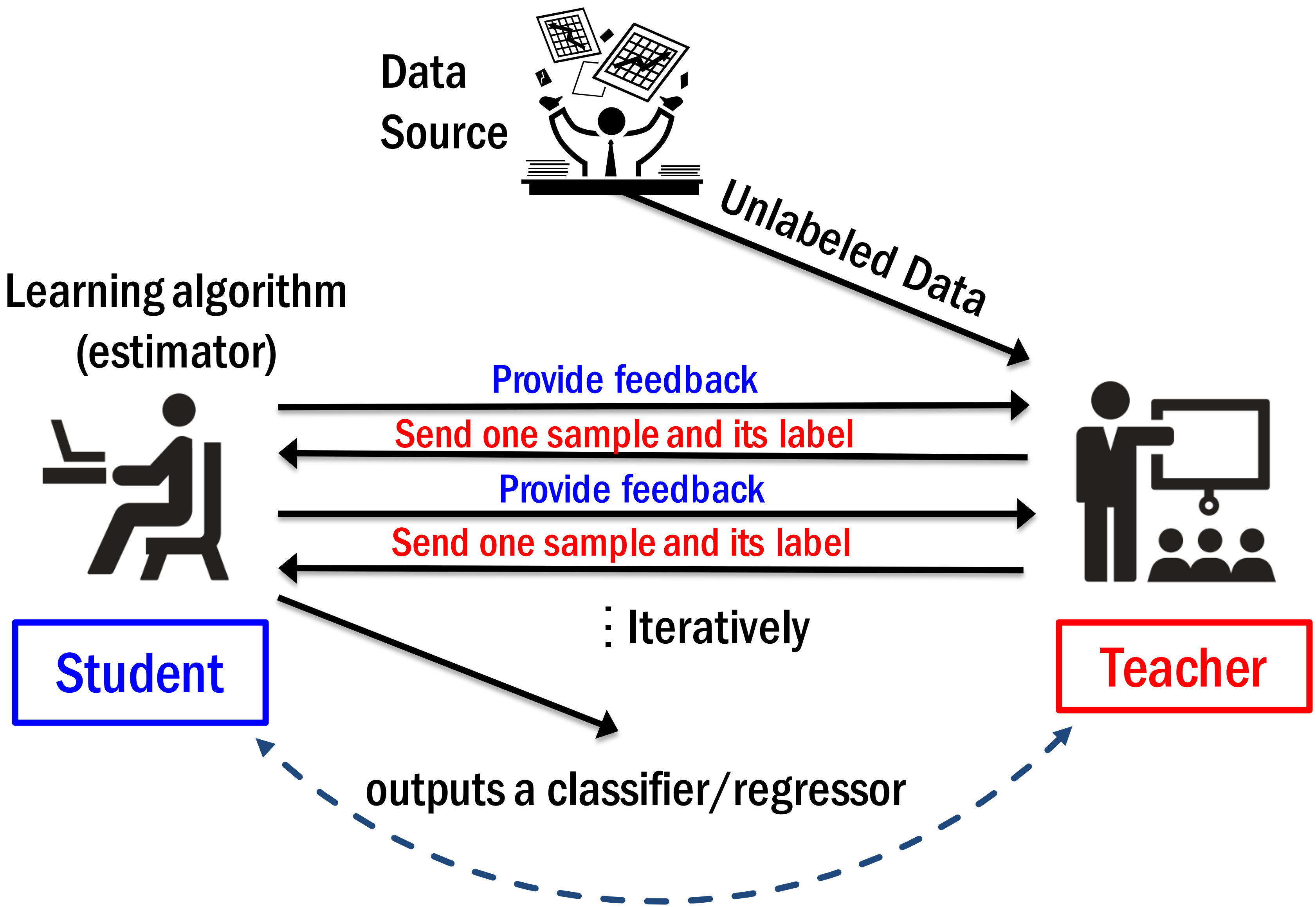


Batch machine teaching



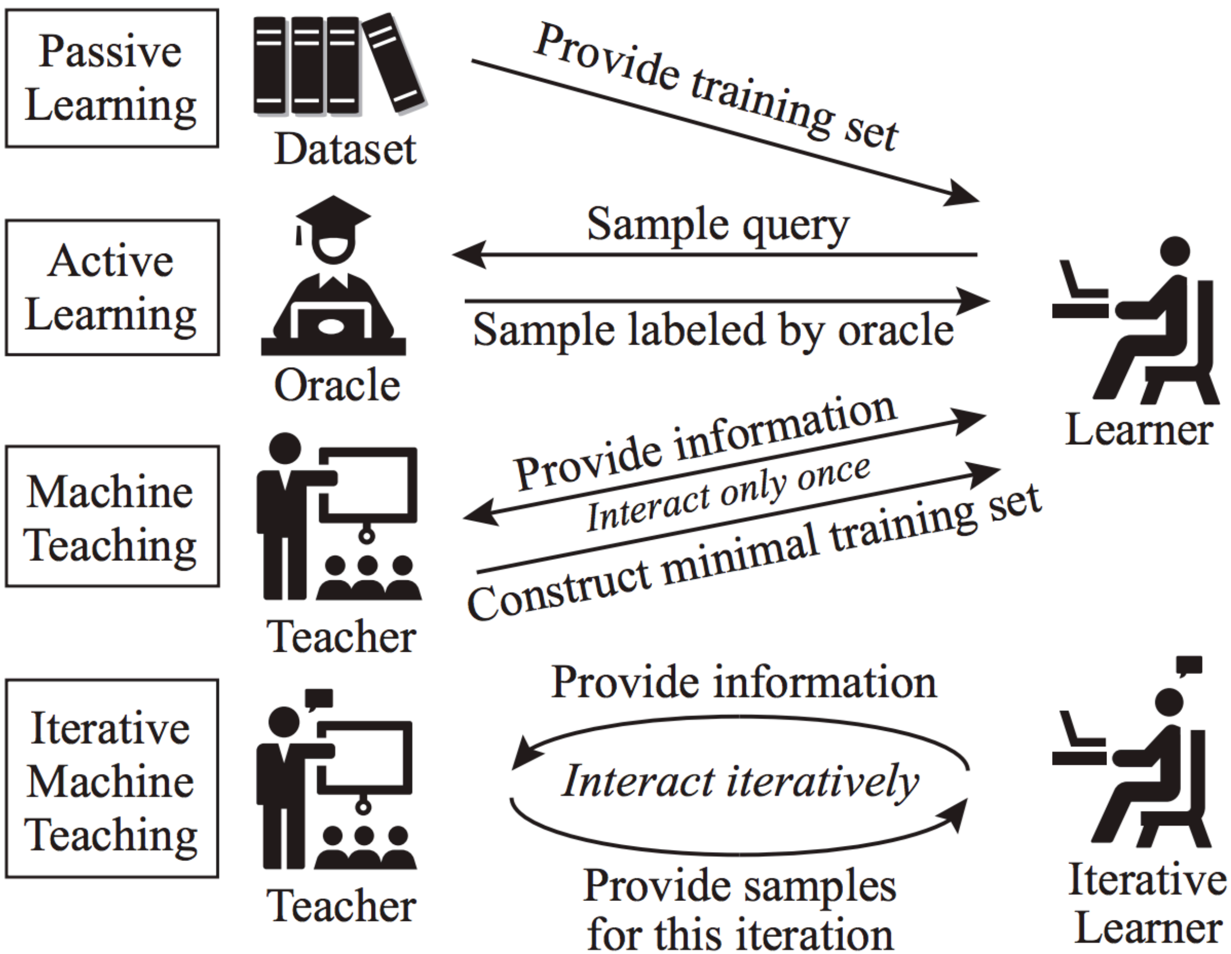
The teacher has full information about the learner

Iterative machine teaching

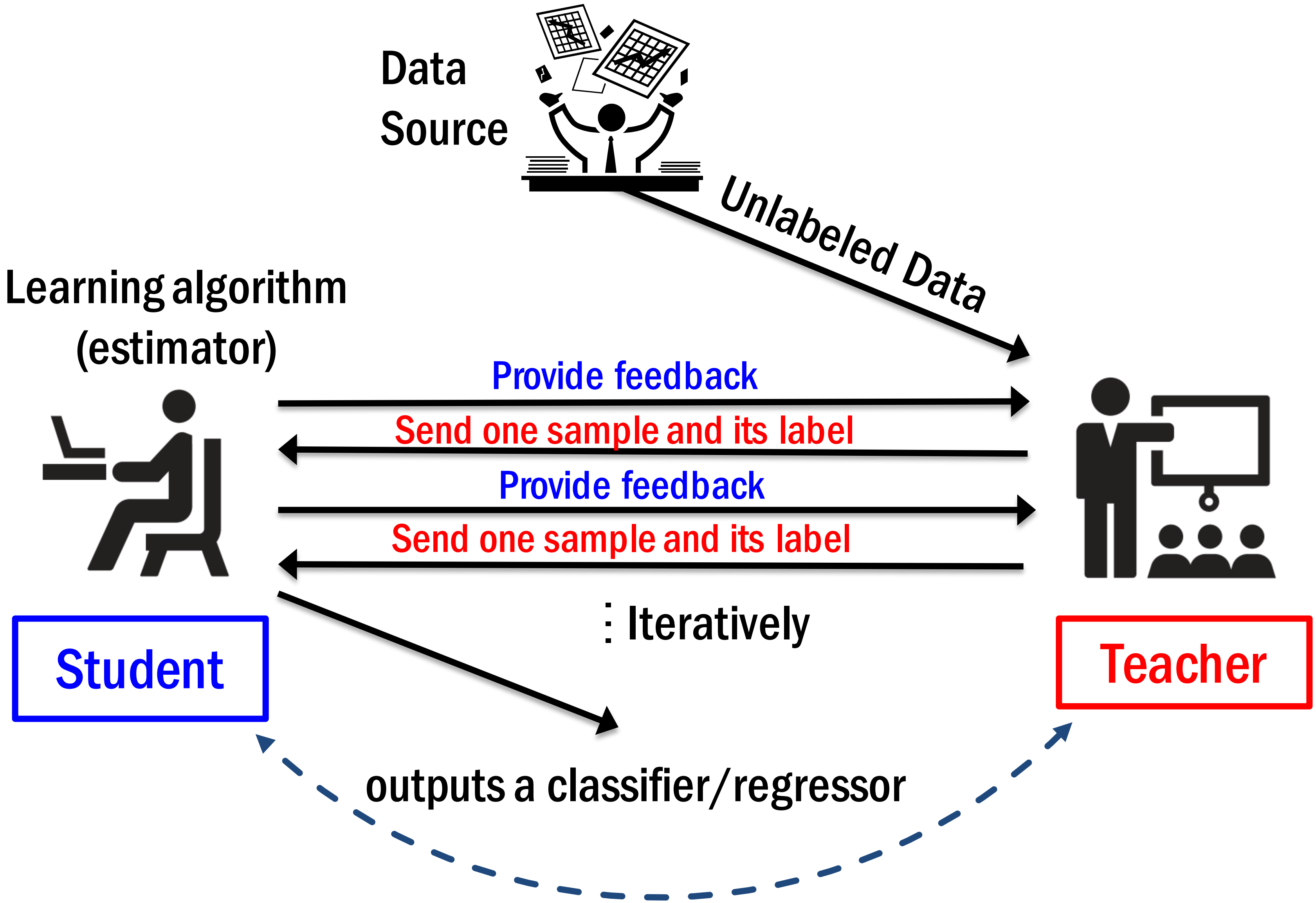


The teacher has full information about the learner

Illustrative comparison to existing learning paradigms



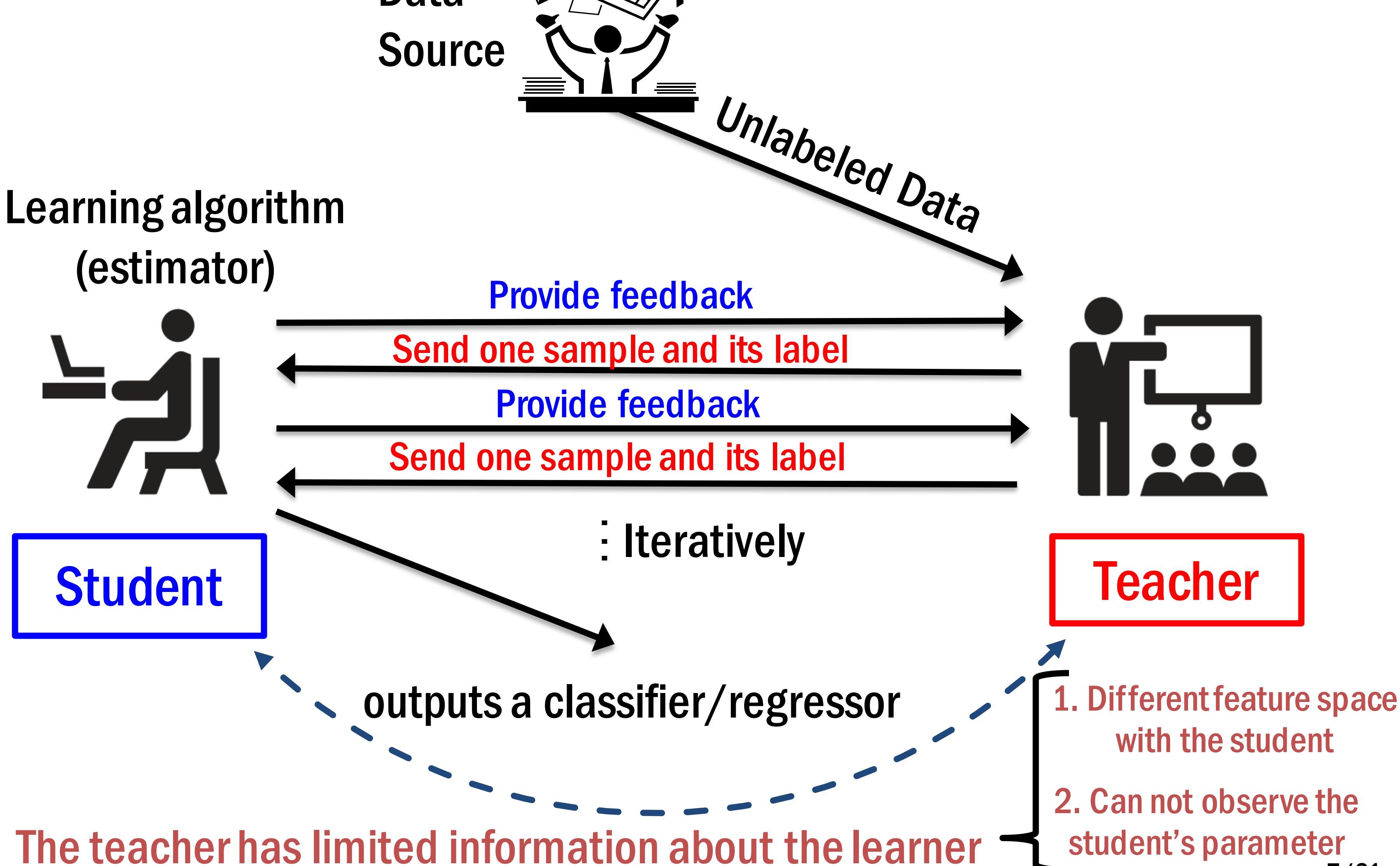
Black-box iterative machine teaching



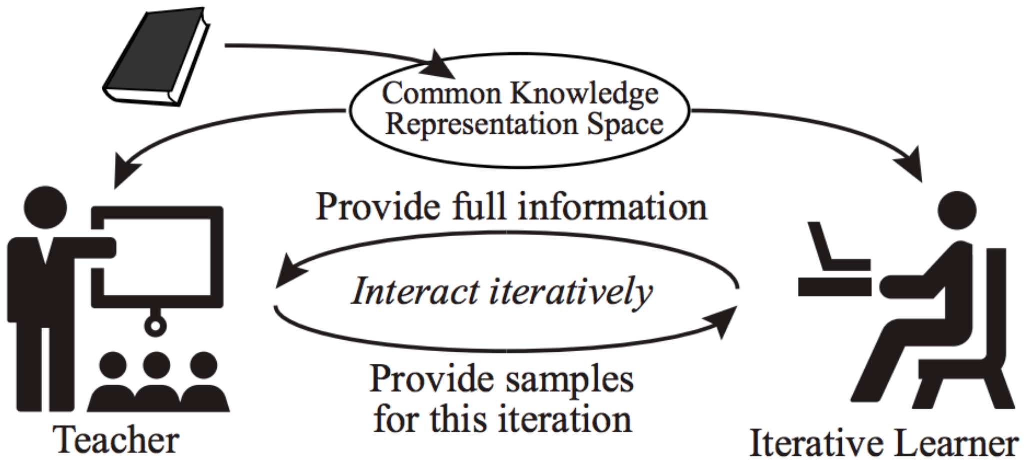
The teacher has no information about the learner

Cross-space iterative machine teaching

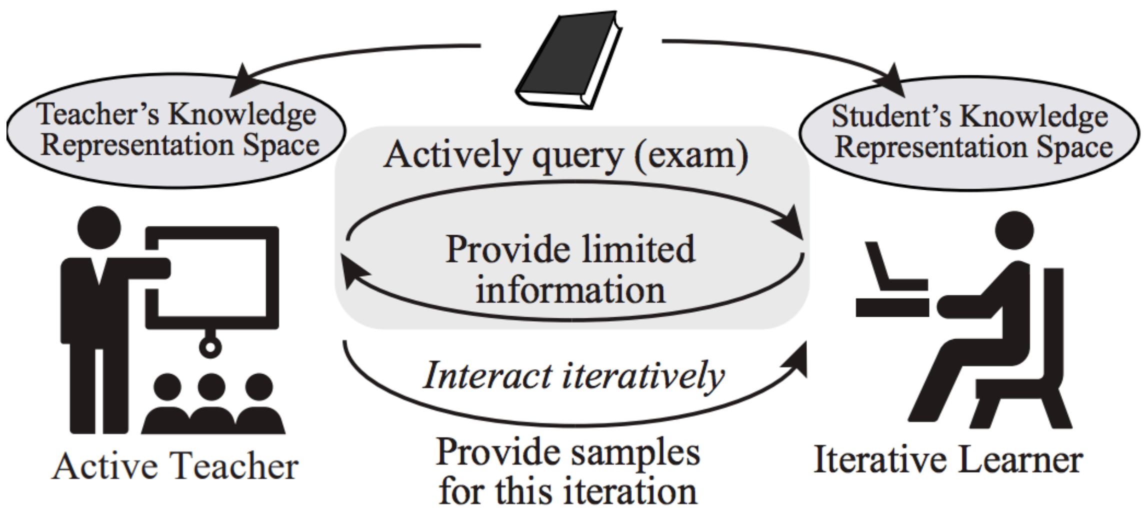
-- a tradeoff between iterative teaching and fully black-box iterative teaching



Illustrative comparison to iterative machine teaching



(a) Iterative Machine Teaching



(b) Cross-Space Machine Teaching by Active Teacher

Intuition of omniscient iterative teaching algorithm

- Consider the $t+1$-step SGD solution quality comparing to optimal model:

$$\|w^{t+1} - w^*\|_2^2 = \left\| w^t - \eta_t \frac{\partial \ell(\langle w^t, x\rangle, y)}{\partial w^t} - w^* \right\|_2^2$$

$$= \|w^t - w^*\|_2^2 + {\eta_t}^2 \underbrace{\left\| \frac{\partial \ell(\langle w^t, x\rangle, y)}{\partial w^t} \right\|_2^2}_{T_1(x,y|w^t)} - 2\eta_t \underbrace{\left\langle w^t - w^*, \frac{\partial \ell(\langle w^t, x\rangle, y)}{\partial w^t} \right\rangle}_{T_2(x,y|w^t)}$$

- $T_1(x, y|w^t)$: characterize the difficulty of an example
  - For linear regression, $T_1(x, y|w^t) = \|\langle w^t, x\rangle - y\|_2^2$
  - For logistic regression, $T_1(x, y|w^t) = \left\| \frac{1}{1+\exp(y\langle w^t, x\rangle)} \right\|_2^2$

- $T_2(x, y|w^t)$: characterize the usefulness of an example
  - Correlation between $w^t - w^*$ and gradient caused by x, y

Omniscient iterative teaching algorithm for iterative machine teaching

$$\min_{x,y\in\mathcal{X}\times\mathcal{Y}}\|w^{t+1}-w^*\|_2^2 \Longrightarrow \min_{x,y\in\mathcal{X}\times\mathcal{Y}} {\eta_t}^2 T_1(x,y|w^t) - 2\eta_t\, T_2(x,y|w^t)$$

- The omniscient teacher knows the student's model, w, in each iteration

Student side:
For $t = 1, \dots, T$
Receive training samples from teacher
Update the model by

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

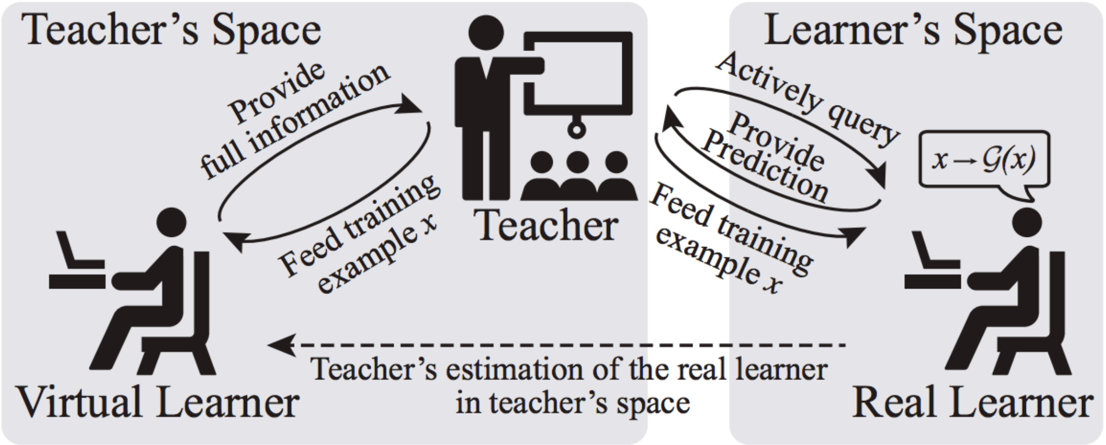
This algorithm can only be applied to the scenario that the teacher knows everything about the student.

What if the the teacher can not fully observe the student?

Teacher side:
Set up $\mathcal{X}\times\mathcal{Y}$ according to the learning setting
For $t = 1, \dots, T$
Check student's w^t
Select the training sample from $\mathcal{X}\times\mathcal{Y}$ by
$\min_{x,y\in\mathcal{X}\times\mathcal{Y}} {\eta_t}^2 T_1(x,y|w^t)$ - $2\eta_t\, T_2(x,y|w^t)$
Send the selected x, y to student

High-level intuition of the proposed active teaching

- To address the cross-space iterative teaching, we propose the active teaching algorithm:



- The key idea is that the teacher will actively query the student using some examples, and the student can provide its prediction output to the teacher. Such procedure is similar to student taking exams.
- In reality, the teacher will often make students to take exams in order to see how they have mastered the knowledge.

Active teaching algorithm for cross-space itertative teaching

- The active teacher can not observe the student's model w.

Real student side:

Initialize w^0

For $t = 1, \dots, T$

Receive training sample x from teacher

Update the model by

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

Active teacher constructs a virtual student based on the estimation of the real student, and perform omniscient teaching on the virtual one.

Virtual student side:

Estimation w^r from the active teacher

For $t = 1, \dots, T$

Use the estimated parameter to construct a virtual student which is fully observable.

Solve the omniscient teaching optimization and obtain the training sample x, similar to the omniscient teacher.

Actively query

Provide training sample x

. . .

Provide partial prediction output

Provide full information

. . .

Provide training sample x

Active Teacher side:

Set up $\mathcal{X}\times\mathcal{Y}$ according to the learning setting

For $t = 1, \dots, T$

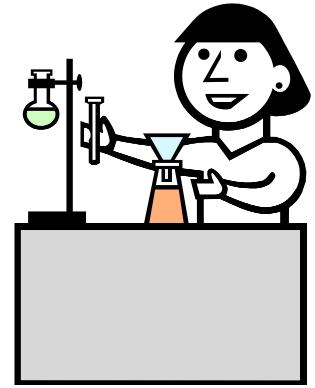
Use active queries to estimate the real student's w^t and feature space, and construct a virtual student.

Select the training sample from $\mathcal{X}\times\mathcal{Y}$ by

$$\min_{x,y\in\mathcal{X}\times\mathcal{Y}} \eta_t{}^2 T_1(x, y|w^t) - 2\eta_t\, T_2(x, y|w^t)$$

Send the selected x, y to the real student

Three ways of generating teaching examples



Regression: $\mathcal{Y} = \mathbb{R}$, Classification $\mathcal{Y} = \{-1, 1\}$

- Synthesis-based teaching:

$$\mathcal{X} = \{x \in \mathbb{R}^d, \|x\| \leq R\}$$

- Combination-based teaching:

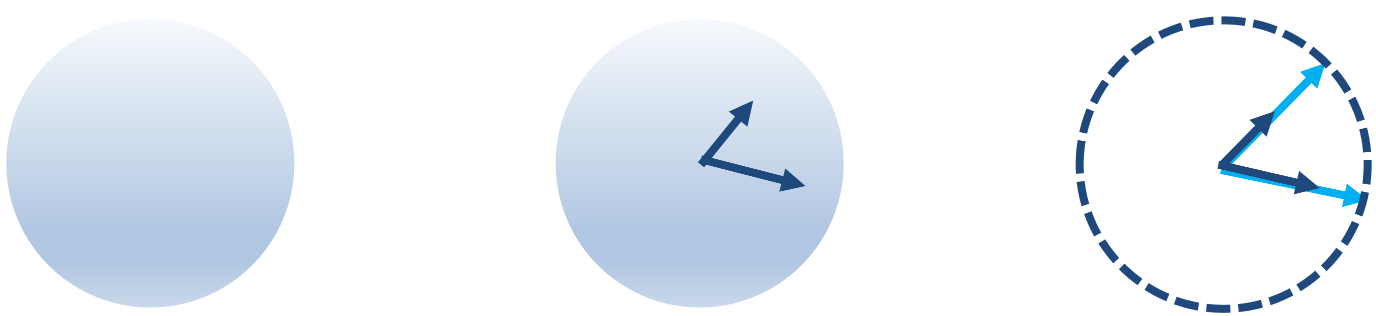
$$\mathcal{X} = \{x \mid \|x\| \leq R, x = \textstyle\sum_{i=1}^{m} \alpha_i x_i, \; x_i \in \mathcal{D}\},$$

with $\mathcal{D} = \{x_i\}_{i=1}^{m}$

- Rescaled Pool-based teaching:

$$\mathcal{X} = \{x \mid \|x\| \leq R, x = \gamma x_i, \; x_i \in \mathcal{D}\},$$

with $\mathcal{D} = \{x_i\}_{i=1}^{m}$



Synthesis Combination Rescaled Pool

Two ways of constructing the virtual student

- Exact recovery of the real student
  - The learner returns a prediction in the form of $F(\langle w, x \rangle)$. In general, if $F(\cdot)$ is an one-to-one mapping, we can exactly recover the ideal virtual learner (i.e. $G^T(w)$) in the teacher's space using the system of linear equations.
- Approximate recovery of the real student
  - In general, if $F(\cdot)$ is not an one-to-one mapping (e.g. sign function), we can only approximate the real student with some sampling complexities. Here, we use active learning to perform such approximation.

Theoretical results

- Exponential teaching
  - Under some mild conditions, the omniscient teacher can select samples from synthesis-based, combination-based, and rescaled pool-based training set to accelerate the student learning with SGD to exponential rate under commonly used loss functions.
  - Under some assumptions, we show that the active teacher can also achieve exponential teaching.
  - The student can provide its prediction output to the teacher using the form of $F(\langle w, x \rangle)$. According to the specific form of $F(\cdot)$, we have different conclusions for its exponential teachability (i.e., the ability to achieve exponential convergence by the active teacher):

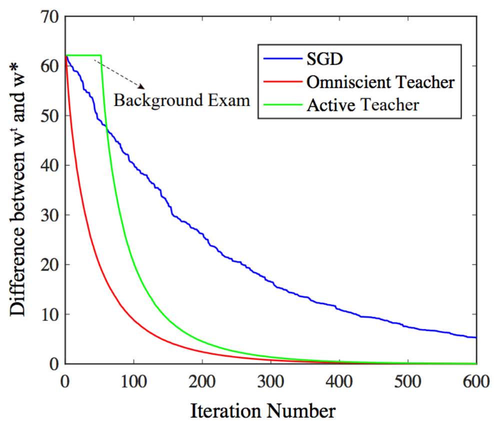
$F(\cdot)$	Synthesis teaching	Combination teaching	Rescalable pool teaching
One-to-one or hinge function	$\surd$	$\surd$	$\surd$
The other function	$\surd$	$\surd$	$\times$

✓ denotes exponentially teachable.

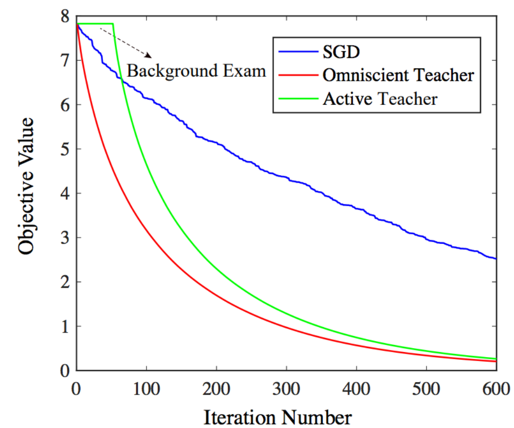
Experiments

- **Teaching Linear Learner with Gaussian Training data**
- **Exact recovery of the real student**

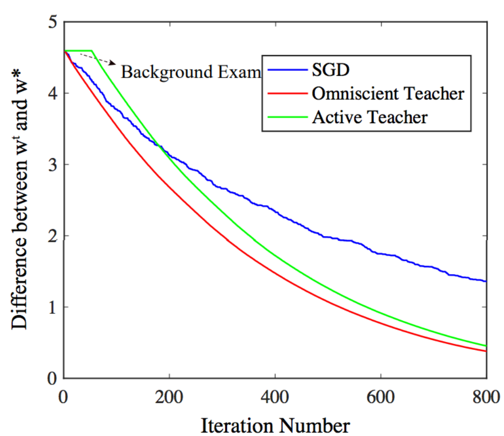
Faster Convergence than SGD and comparable to the omniscient teacher!



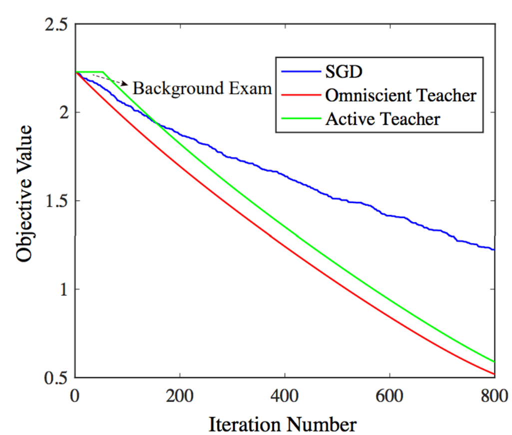
LSR



LSR



LR ($F(z)$ is sigmoid)

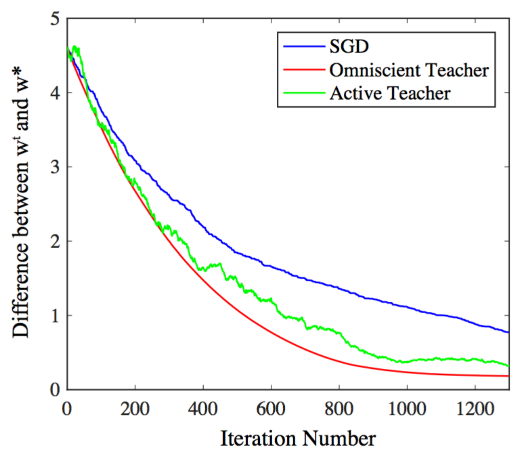


LR ($F(z)$ is sigmoid)

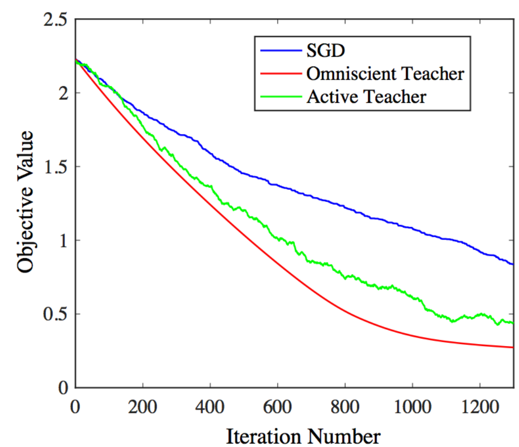
Experiments

- Teaching Linear Learner with Gaussian Training data
- Approximate recovery of the real student

Faster Convergence than SGD and comparable to the omniscient teacher!



LR ($F(z)$ is sign)

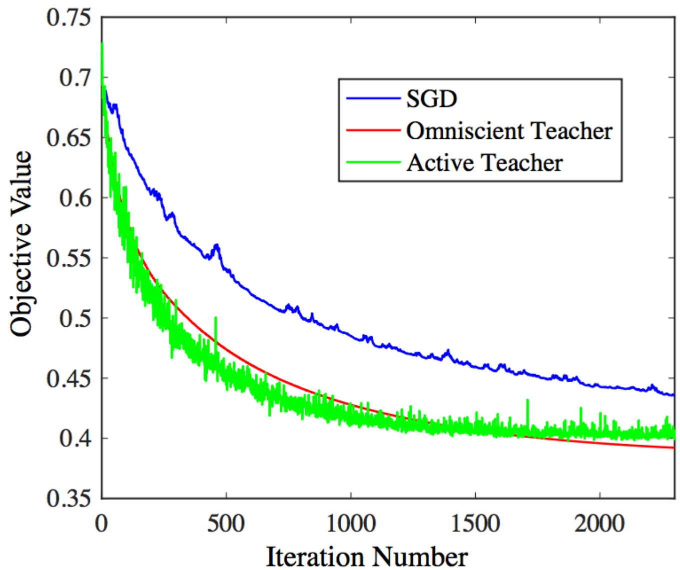


LR ($F(z)$ is sign)

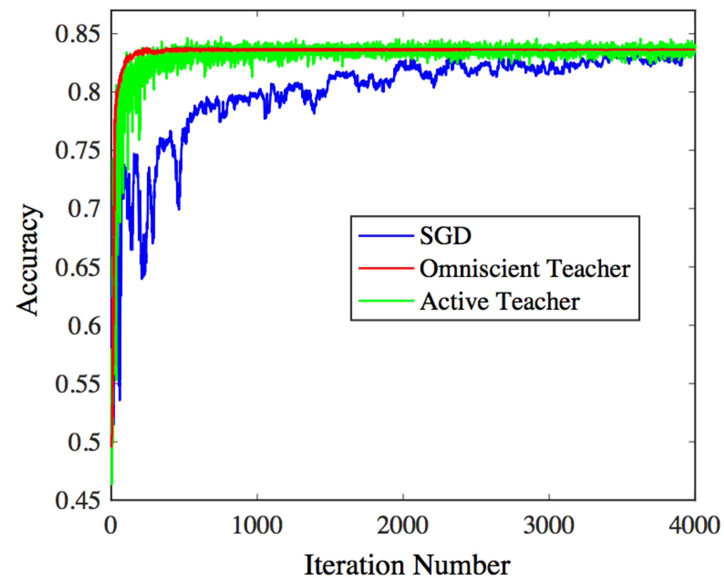
Experiments

- **Teaching Linear Learner in MNIST dataset**

  Faster Convergence than SGD!



LR ($F(z)$ is sign)



LR ($F(z)$ is sign)

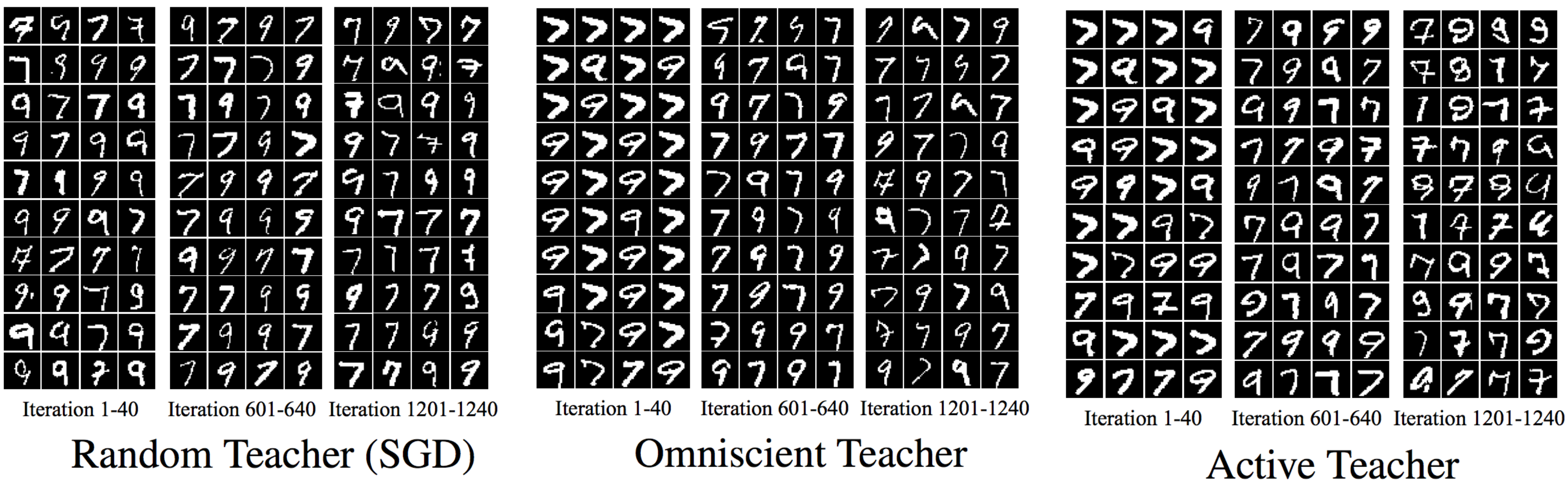
7/9 binary classification

Experiments

- Teaching Linear Learner in MNIST dataset

Visualization of selected samples

Active teacher shares similar behavior with the omniscient teacher:
selecting samples from easy examples to difficult examples



7/9 binary classification

Summary

- A step towards fully black-box iterative machine teaching: cross-space iterative machine teaching.
- A conceptually simple and well motivated teaching model: active teacher.
- Interesting connections with practical human education: the usefulness for students to take exams.