Iterative Machine Teaching

Weiyang Liu, Bo Dai, Ahmad Humayun, Charlene Tay,
Yu Chen, Linda B. Smith, Jim Rehg, Le Song

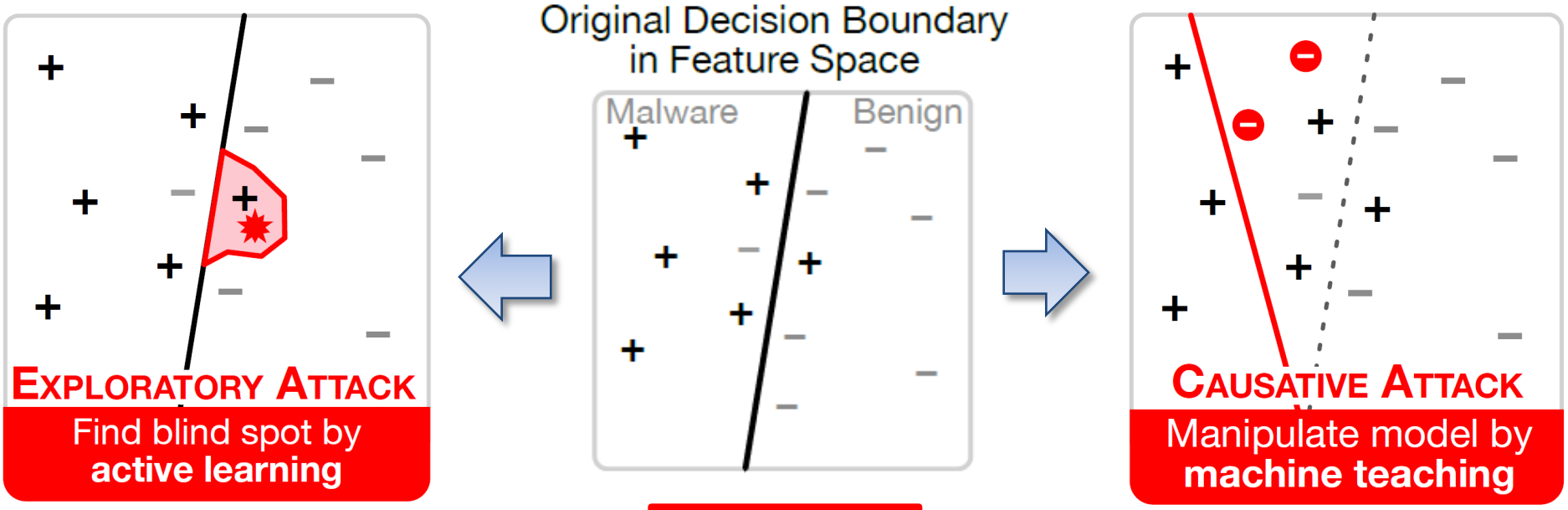
College of Computing

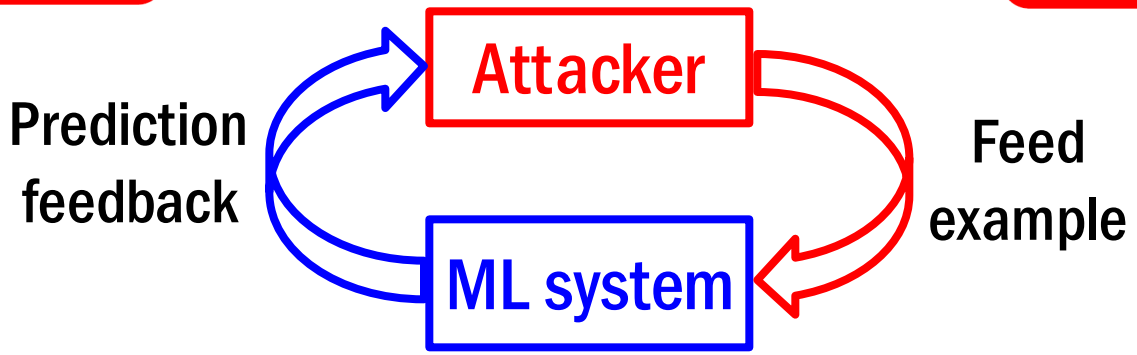
Georgia Institute of Technology

Attack with data

An ML defender system is driven by the data fed to it

- Exploratory attack: probe the decision boundary with data
- Causative attack: manipulate the decision boundary with data





Causative attack $\Leftrightarrow$ machine teaching

Machine teaching: the teacher designs labeled data intelligently, such that the learner can learns the teacher model with these data

Causative attack: the attacker generate malicious data, such that the ML system are steered towards a particular model with these data

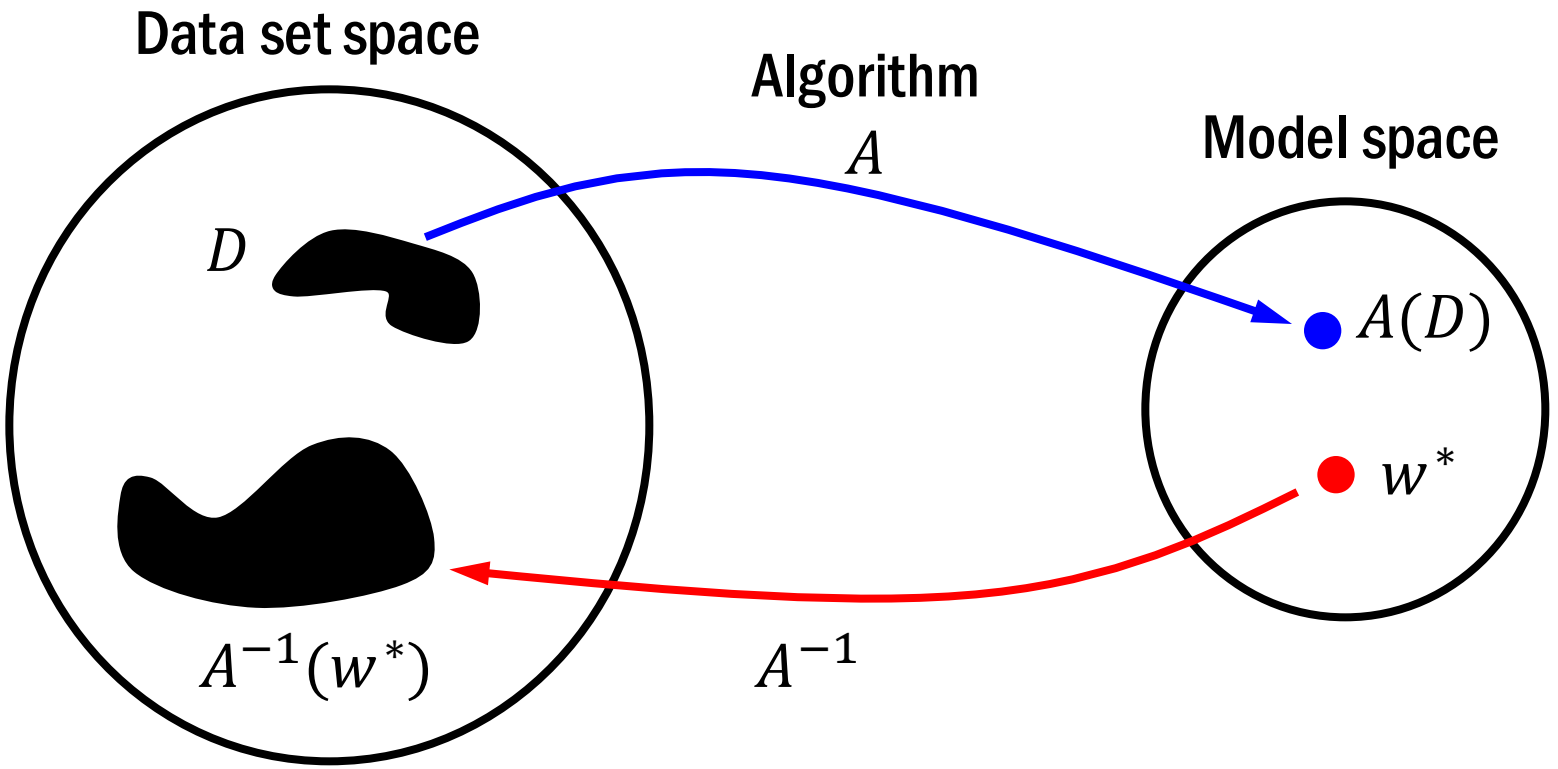
Evaluation of robustness

- **Teaching dimension**: number of examples to steer to ϵ error
- **Key results**: potentially exponentially smaller sample size
- **Implication**: need new algorithms with large teaching dimension

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



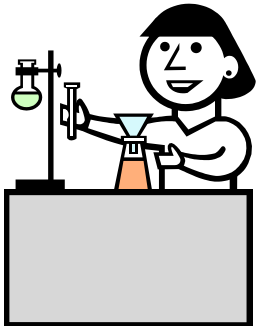
Batch machine teaching

Data Source

Unlabeled Data

Learning algorithm (estimator)

Selected data points



Expert / Teacher

ML system

outputs a classifier/regressor

Attacker

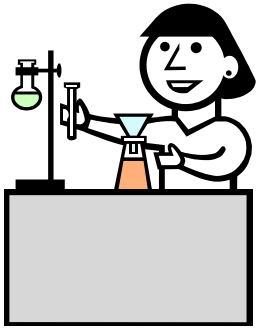
Iterative machine teaching

Data Source

Unlabeled Data

Learning algorithm (estimator)

Expert / Teacher



Provide feedback

Send one sample and its label

Provide feedback

Send one sample and its label

⋮

ML system

Attacker

outputs a classifier/regressor

Teaching speed is limited by the student

- Student ⇔ An iterative algorithm for learning a model parameter

- Depending on the type of student algorithms
  - The best example to teach will be different
  - The teaching speed (dimension) will be different

Stochastic gradient descent:
Initialize w^0
For $t = 1 \dots T$
obtain (x, y) **from teacher**

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

loss function $\ell(\cdot,\cdot)$
feature representation x
linear model $\langle w, x \rangle$
learning rate η_t
Goal: teach w^*

Teaching dimension? Best examples?

Intuition from SGD updates

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

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

$$\min_{x,y \in \mathcal{X} \times \mathcal{Y}} \left\| w^t - \eta_t \frac{\partial \ell(\langle w^t, x \rangle, y)}{\partial w^t} - w^* \right\|_2^2$$

- SGD randomly selects x, y, which may hurt the performance.
- Select optimal samples $x, y \in \mathcal{X} \times \mathcal{Y}$ will lead to minimizing distance to optimal solution, therefore, accelerating learning speed.
- Extremely, in a rich candidate training set, we can select the optimal sample x, y which reduces $\|w^{t+1} - w^*\|_2^2$ to 0, i.e., learning with one sample.

Intuition from SGD updates

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

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

- $T_1(x, y|w^t)$: 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

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

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

Student side:

For $t = 1, \ldots, T$

Receive training samples from teacher

Update the model by

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

Teacher side:

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

For $t = 1, \ldots, T$

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

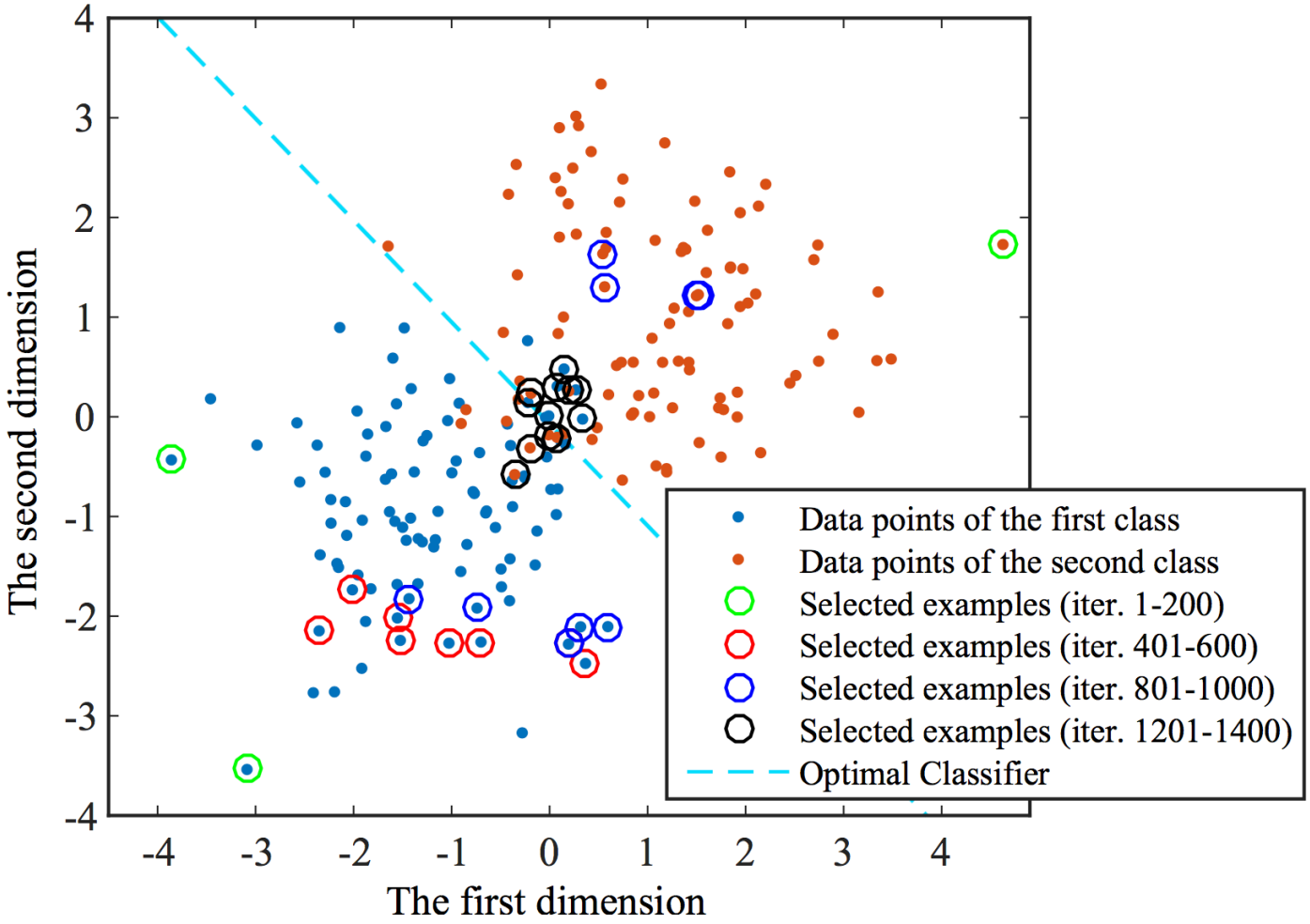
Trade-off between sample difficulty and sample usefulness

- Consider fixed learning rate η_t
  - $\frac{\partial \ell(\langle w^t, x \rangle, y)}{\partial w^t}$ is large at the beginning stage $\Longrightarrow T_1\ (x, y|w^t)$ is dominated
  - $\frac{\partial \ell(\langle w^t, x \rangle, y)}{\partial w^t}$ is small when w^t approaches $w^* \Longrightarrow T_2\ (x, y|w^t)$ is dominated

- Pick easy example first and difficult later, gradually focusing on difficult examples.
  - Connected to curriculum learning and boosting.

Toy Experiments

- Teaching Linear Learner with Gaussian Training data
- Visualization of selected samples: easy ones first and gradually difficult ones.



Qualitatively result

- Define the Teaching Volume as

$$TV(w) = \max_{x,y \in X \times Y} -\eta_t{}^2 T_1(x, y|w) + 2\eta_t T_2(x, y|w)$$

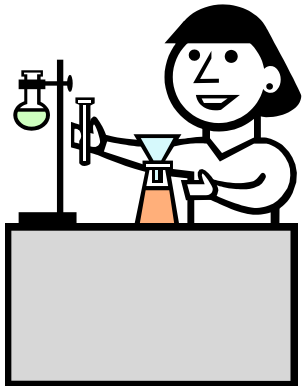
- Teaching monotonicity:

$$\|w_1 - w^*\|_2^2 - TV(w_1) \leq \|w_2 - w^*\|_2^2 - TV(w_2)$$

for any w_1, w_2 satisfying $\|w_1 - w^*\|_2^2 \leq \|w_2 - w^*\|_2^2$

- Under the teaching monotonicity condition, the teacher algorithm can always converge not slower than teacher selecting random examples

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:

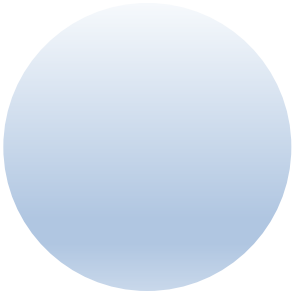
$$\mathcal{X} = \{x | \ \|x\| \leq R, x = \sum_{i=1}^{m} \alpha_i x_i, \ x_i \in \mathcal{D}\},$$

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

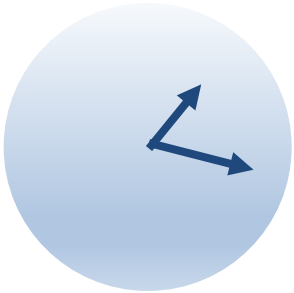
- Rescaled Pool-based teaching:

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

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



Synthesis



Combination

Rescaled Pool

Theoretical results

- Exponential speedup learning procedure
  - 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.

Exponential rate: the student can learn an ϵ-approximation of w^* with $\mathcal{O}\left(C \log\frac{1}{\epsilon}\right)$ samples.

Different training sets will affect the constant C in the learning rate.

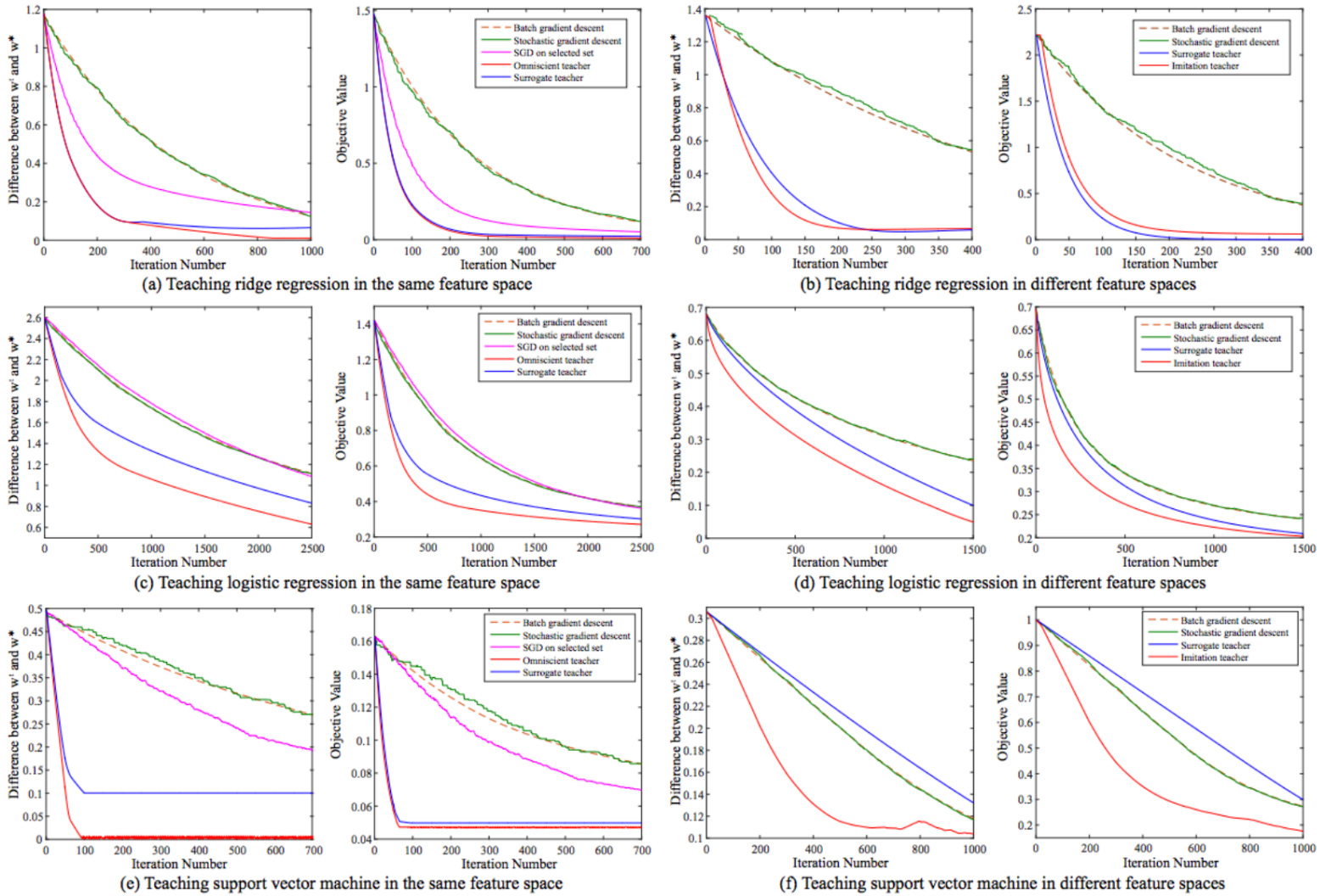
Commonly used loss functions for regression: square loss, absolute loss
Commonly used loss functions for classification: hinge loss, logistic loss

More generally, the Lipschitz smooth and strongly convex loss function are exponential teachable

Experiments

- Teaching Linear Learner with Gaussian Training data

  Faster Convergence than SGD!



(a) Teaching ridge regression in the same feature space

(b) Teaching ridge regression in different feature spaces

(c) Teaching logistic regression in the same feature space

(d) Teaching logistic regression in different feature spaces

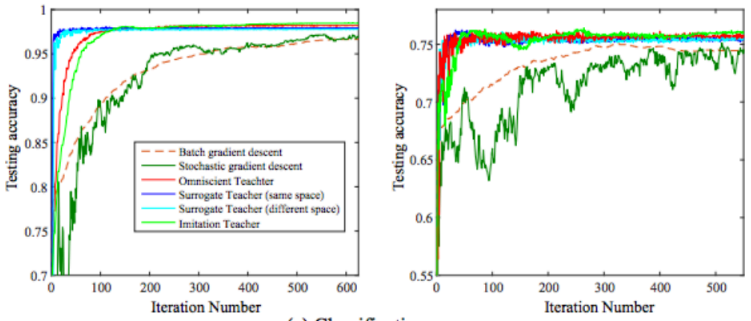
(e) Teaching support vector machine in the same feature space

(f) Teaching support vector machine in different feature spaces

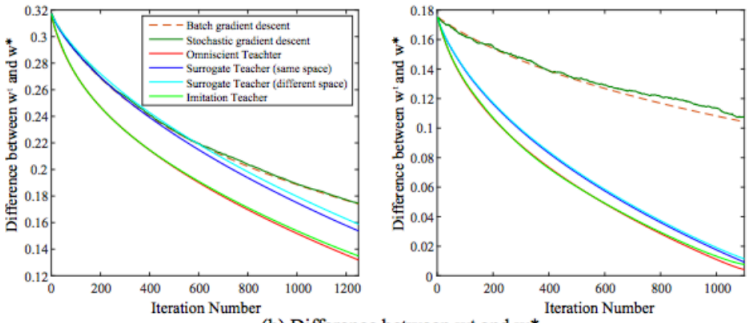
Experiments

- Teaching Linear Learner in MNIST dataset

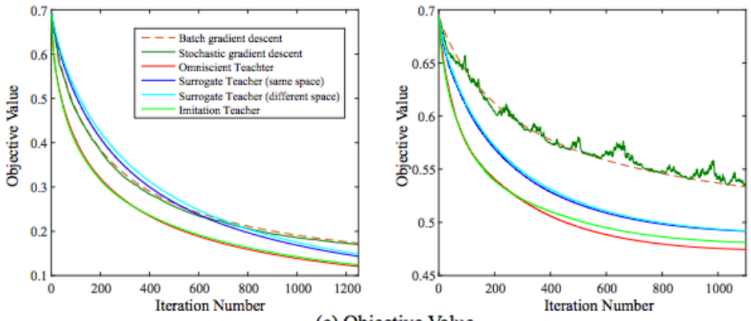
Faster Convergence than SGD!



(a) Classification accuracy



(b) Difference between w^t and w^*



(c) Objective Value

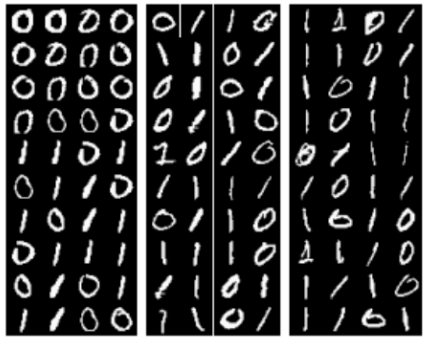
0/1, 3/5 binary classification

Visualization of selected samples

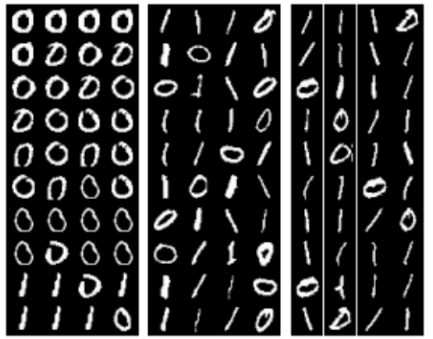
From easy examples to difficult examples

(a) Omniscient Teacher

(b) Imitation Teacher



(a) Omniscient Teacher

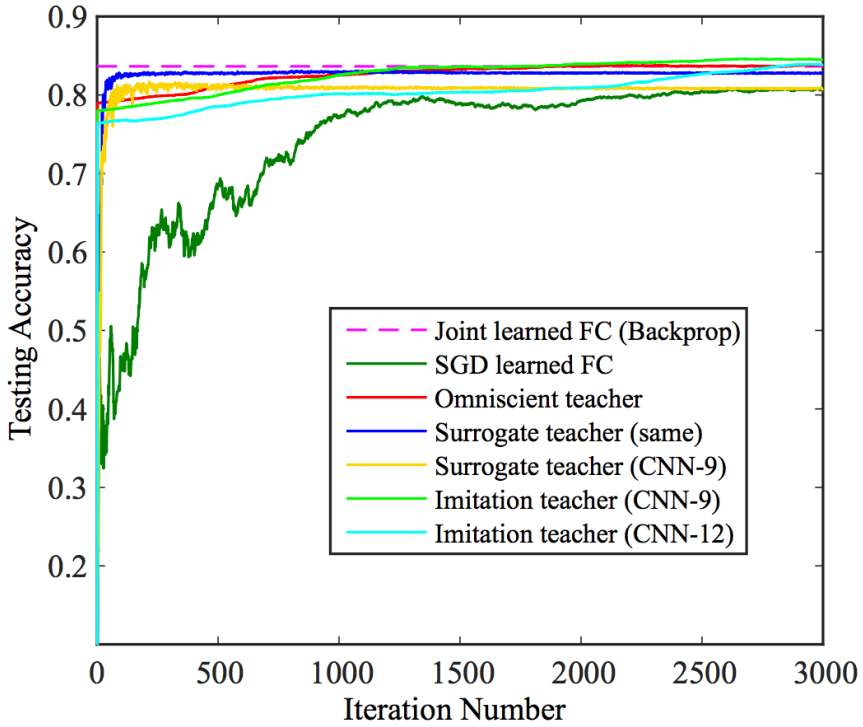


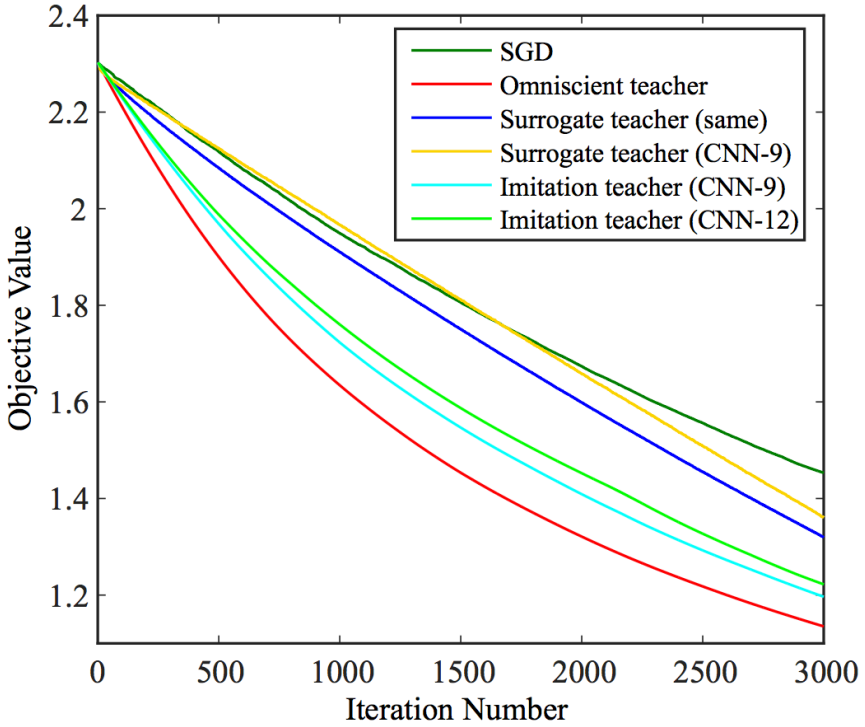
(b) Imitation Teacher

Experiments

- Teaching Linear Learner in CIFAR-10 dataset

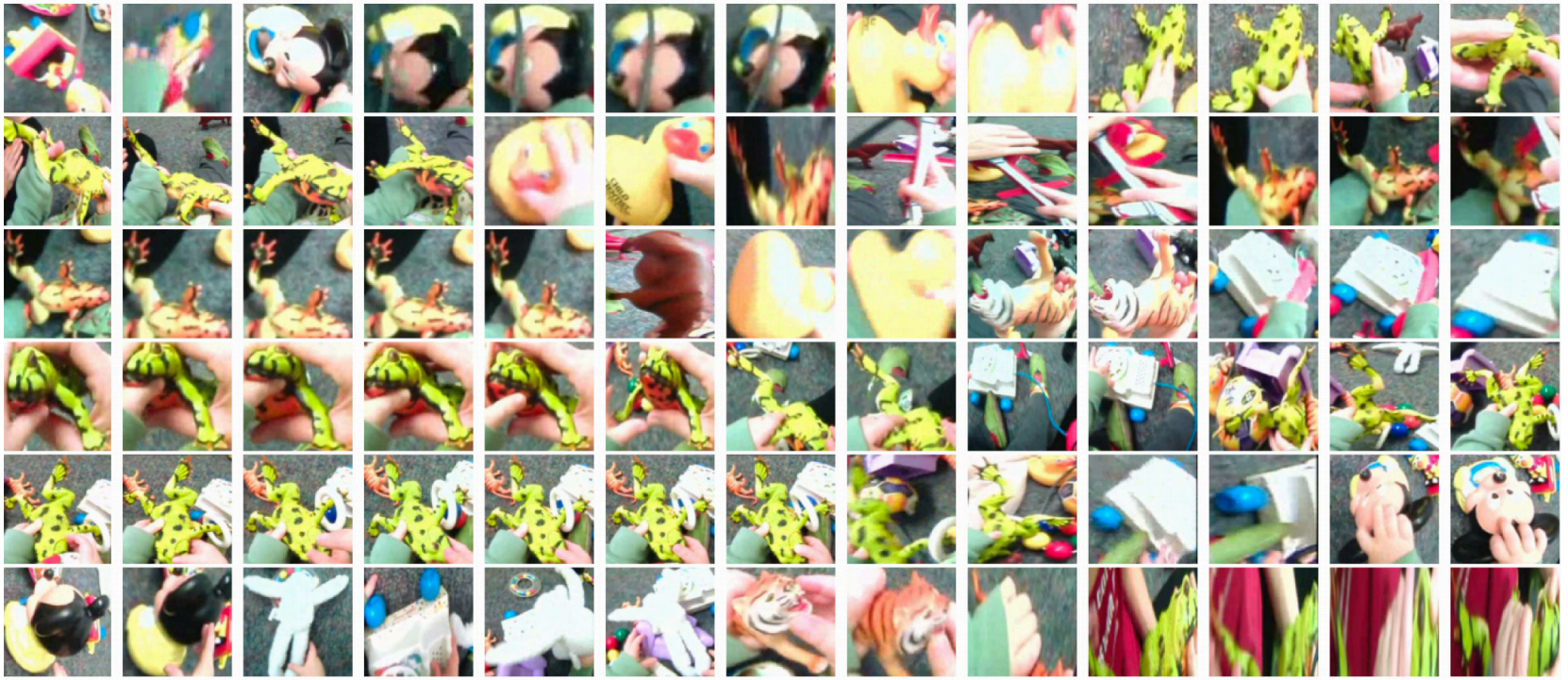
  Consistently faster Convergence than SGD!





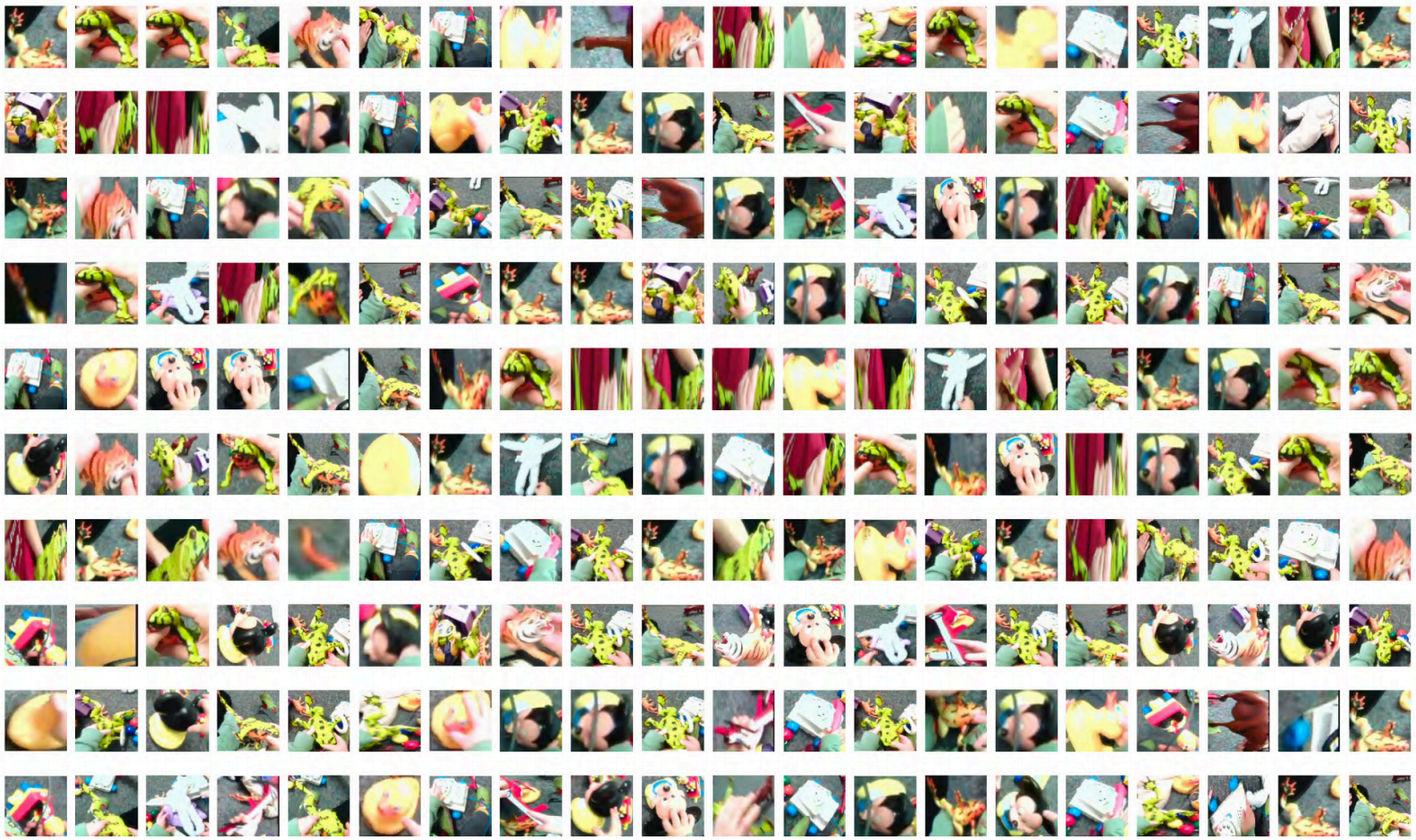
Object learning experiment on children's ego-centric visual data

- Natural learning order of a child



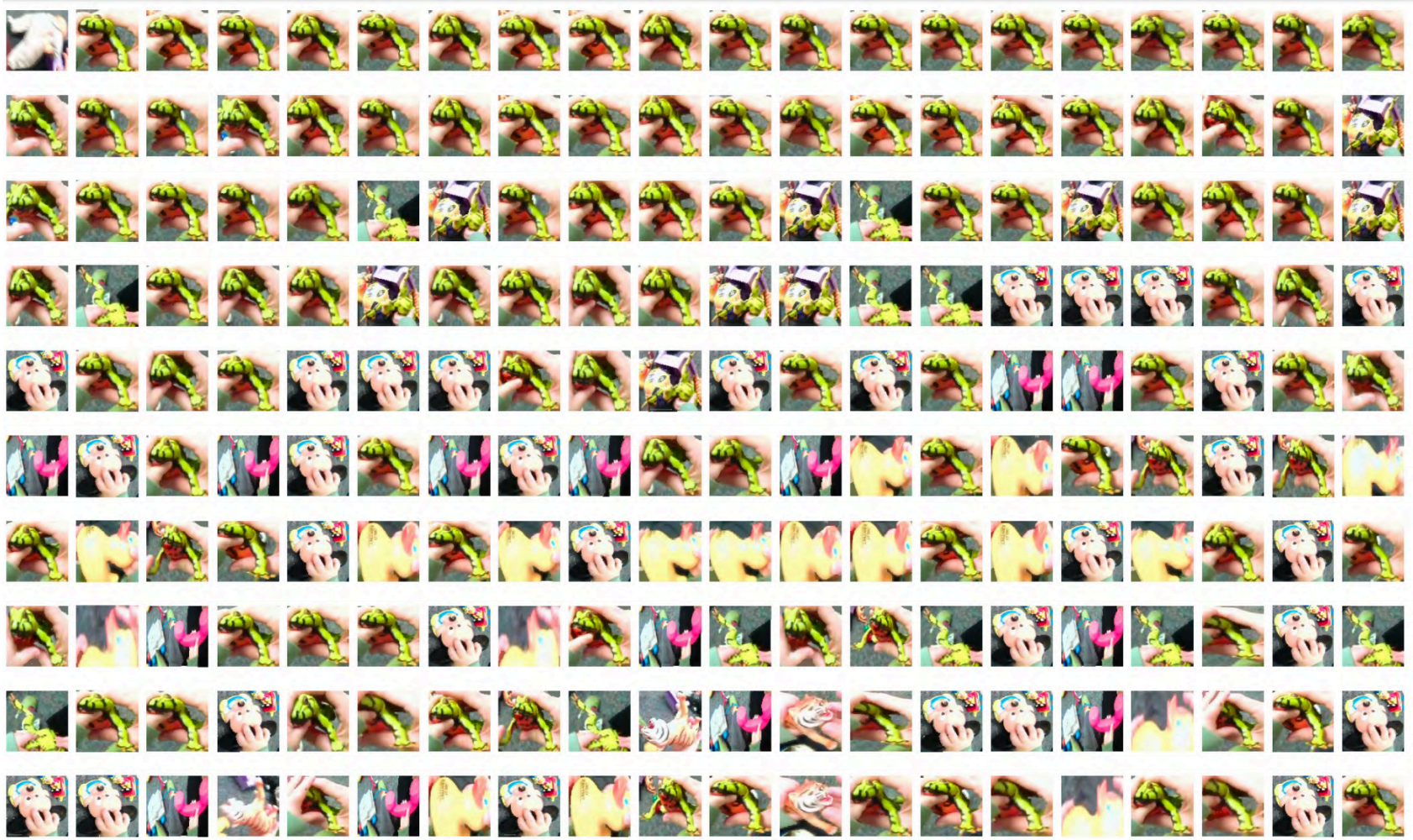
Object learning experiment on children's ego-centric visual data

- SGD learning order



Object learning experiment on children's ego-centric visual data

- Iterative teaching order



Similar to natural learning order of children!