Co-teaching: Robust Training of Deep Neural Networks with Noisy Labels

Bo Han*[1,2] Quanming Yao*[3] Xingrui Yu[1] Gang Niu[2]
Miao Xu[2] Weihua Hu[4] Ivor W. Tsang[1] Masashi Sugiyama[2,5]

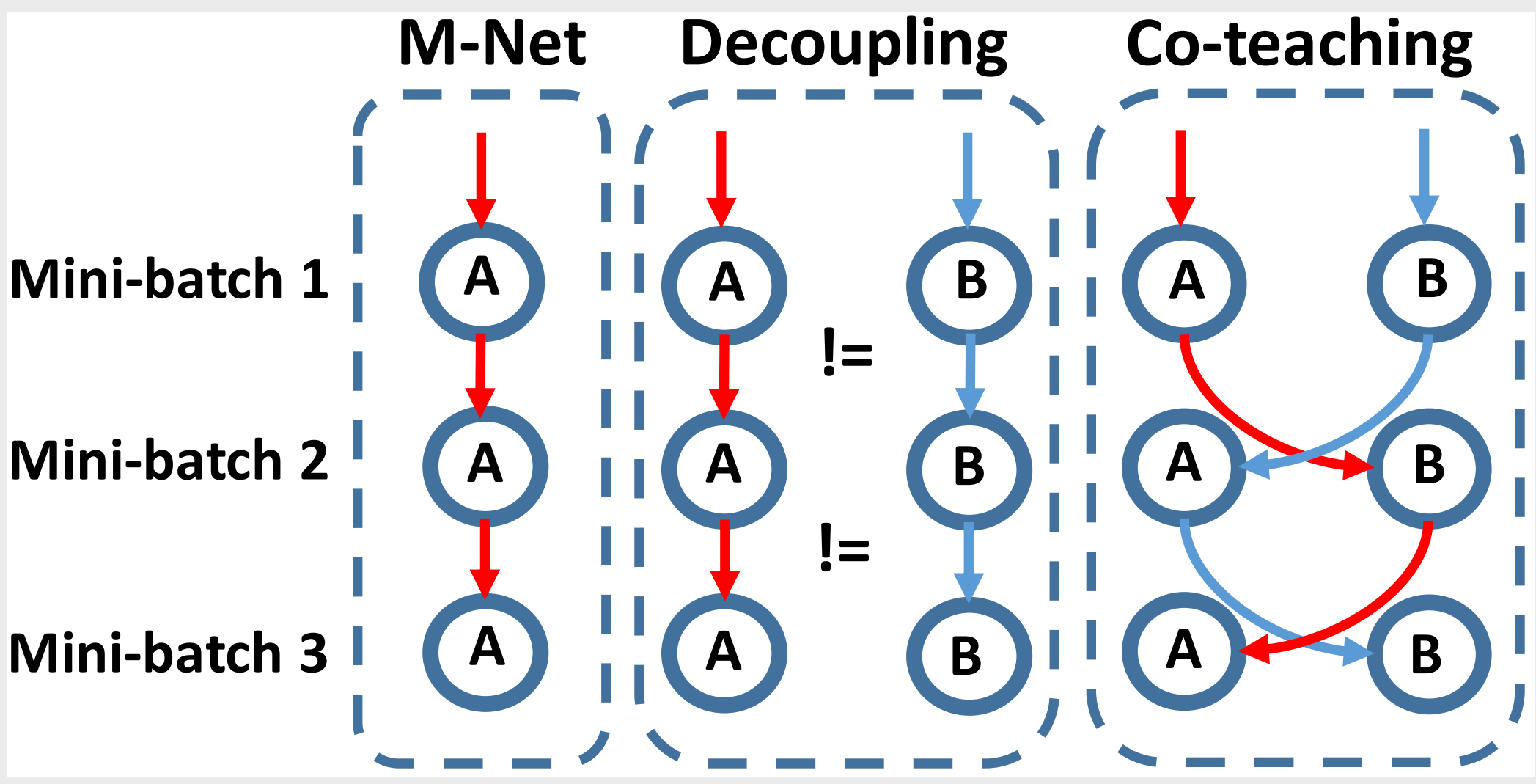
[1]CAI, University of Technology Sydney [2]AIP, RIKEN [3]4Paradigm Inc. [4]Stanford University [5]The University of Tokyo

Overview

TL;DR: We train **two** networks, and each network samples its **small-loss instances** as the useful knowledge to update the parameters of its **peer network**.

- **Noisy labels** are corrupted from ground-truth labels, which degenerates the robustness of learning models.
- **Deep neural networks** have the high capacity to fit any noisy labels. The solutions are as follows.
  - ⋄ Noise **transition** matrix estimation. E.g., F-correction.
  - ⋄ **Regularization**. E.g., VAT and Mean teacher.
  - ⋄ Training on **selected** samples. E.g., MentorNet.
- We present a new paradigm called **Co-teaching** combating with extremely noisy labels.
  - ⋄ We train **two** networks simultaneously.
  - ⋄ In each mini-batch data, each network **filters** noisy instances based on memorization effects.
  - ⋄ It teaches the **remaining** instances to its **peer** network for updating the parameters.
- Empirical results on **MNIST, CIFAR-10** demonstrate that the robustness of deep learning models trained by Co-teaching approach is superior than that of SOTA methods.

Motivation



Co-teaching Algorithm

```
for $T = 1, 2, \dots, T_{\max}$ do
    Shuffle training set $\mathcal{D}$;                //noisy dataset
    for $N = 1, \dots, N_{\max}$ do
        Fetch mini-batch $\bar{\mathcal{D}}$ from $\mathcal{D}$;
        Obtain $\bar{\mathcal{D}}_f = \arg\min_{\mathcal{D}':|\mathcal{D}'|\geq R(T)|\bar{\mathcal{D}}|} \ell(f, \mathcal{D}')$; //sample $R(T)\%$ small-loss instances
        Obtain $\bar{\mathcal{D}}_g = \arg\min_{\mathcal{D}':|\mathcal{D}'|\geq R(T)|\bar{\mathcal{D}}|} \ell(g, \mathcal{D}')$; //sample $R(T)\%$ small-loss instances
        Update $w_f = w_f - \eta\nabla\ell(f, \bar{\mathcal{D}}_g)$;
        Update $w_g = w_g - \eta\nabla\ell(g, \bar{\mathcal{D}}_f)$;
    end
    Update $R(T) = 1 - \min\left\{\frac{T}{T_k}\tau, \tau\right\}$;
end
```

Two Important Questions

Q1. Why can sampling **small-loss instances** based on $R(T)$ help us find clean instances?
Q2. Why do need two networks and **cross-update** the parameters?

QR Code

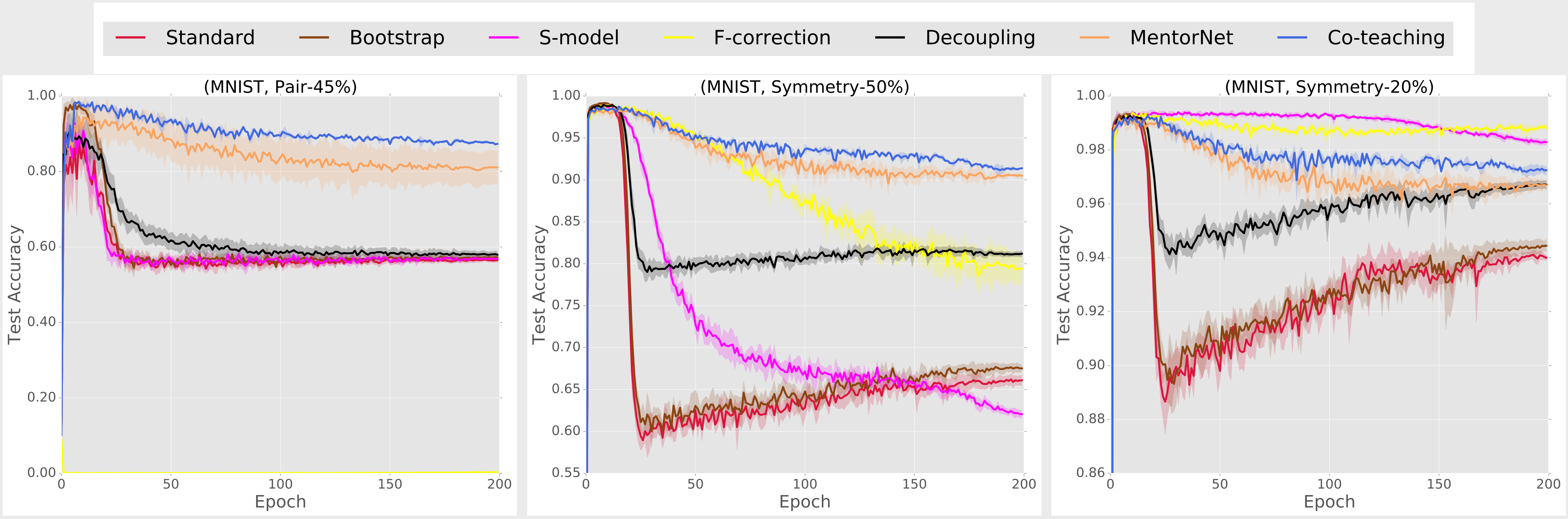
Co-teaching with State-of-the-Art Methods

- **"large class":** can deal with a large number of classes;
- **"heavy noise":** can combat the heavy noise, i.e., high noise rates;
- **"flexibility":** not need combine with specific network architecture;
- **"no pre-train":** can be trained from scratch, i.e, Decoupling needs 5000 iterations to pre-train two networks first, then switches to training with the "Update by Disagreement" rule.

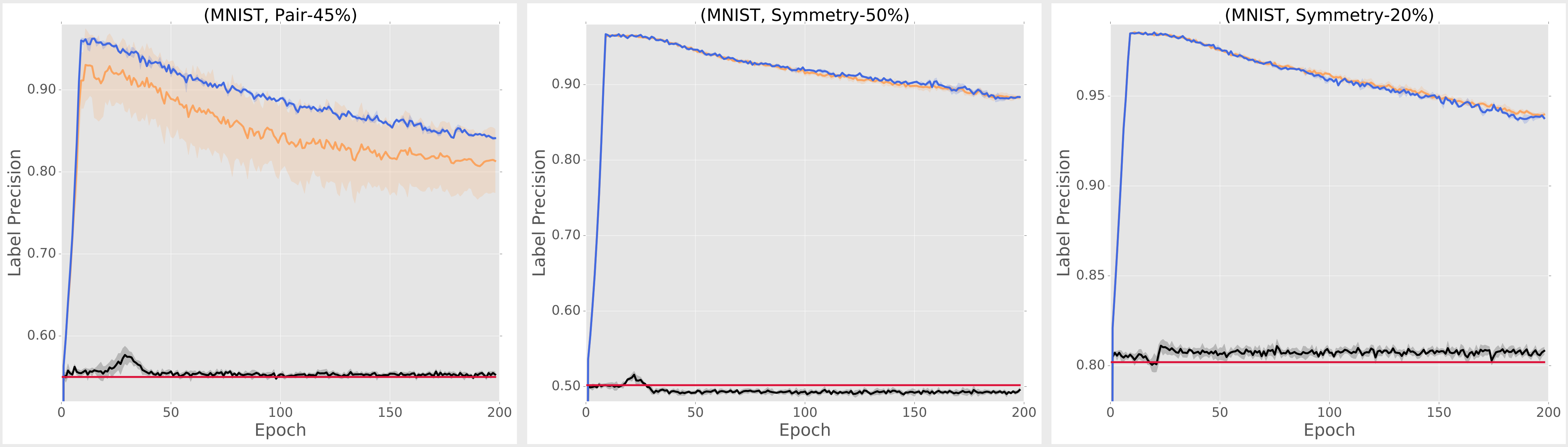
	Bootstrap	S-model	F-correction	Decoupling	**MentorNet**	**Co-teaching**
large class	×	×	×	✓	✓	✓
heavy noise	×	×	×	×	✓	✓
flexibility	×	×	✓	✓	✓	✓
no pre-train	✓	×	×	×	✓	✓

Results on MNIST

- Test accuracy vs number of epochs on **MNIST** dataset.

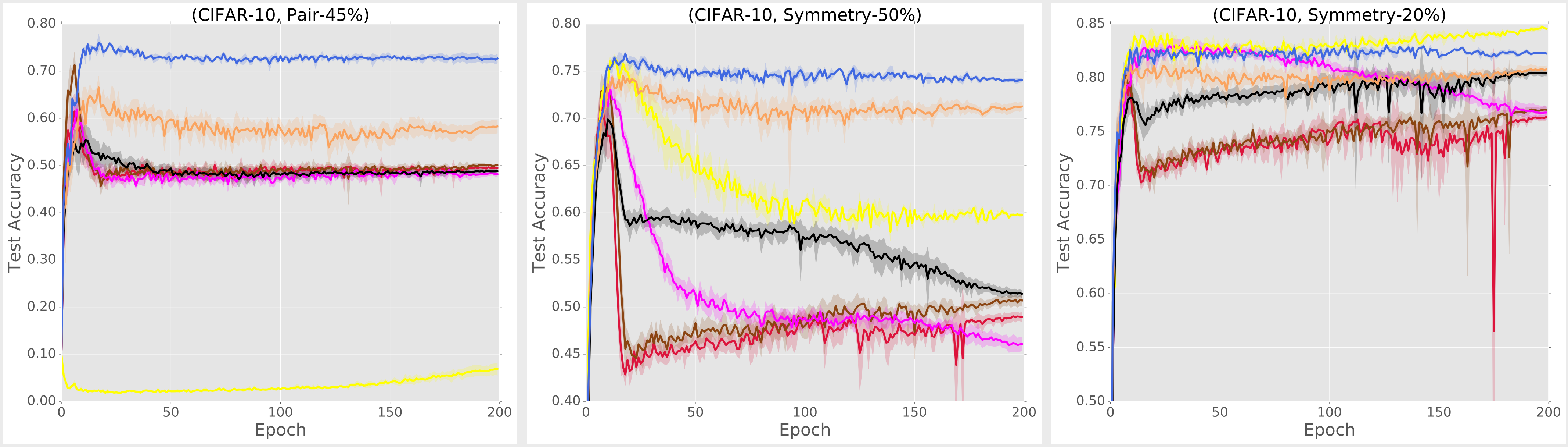


- Label precision vs number of epochs on **MNIST** dataset.

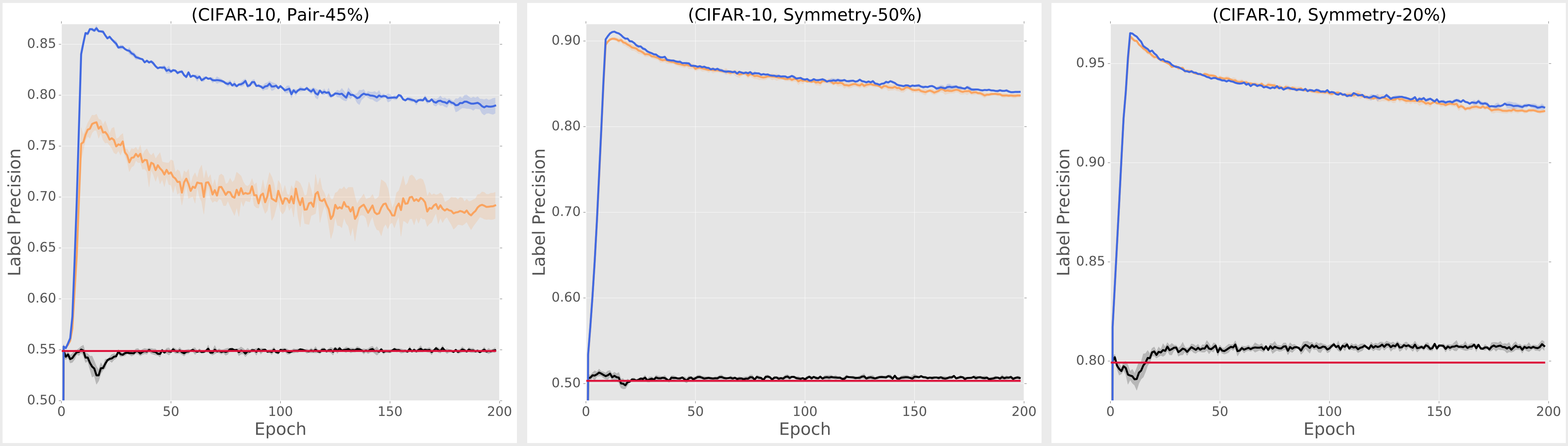


Results on CIFAR-10

- Test accuracy vs number of epochs on **CIFAR-10** dataset.



- Label precision vs number of epochs on **CIFAR-10** dataset.


(CIFAR-10, Pair-45%)
(CIFAR-10, Symmetry-50%)
(CIFAR-10, Symmetry-20%)
Label Precision
Epoch