# A. The Proof for Theorem 1

In this paper, we use the similar proof skills in [6] to prove Theorem 1. The details are presented as follows. Our method inherits the cross-update paradigm [1]. That is, we exploit two networks, where they select possible clean examples for the peer network. Although two networks have different initialization and diverse outputs, the outputs cannot be totally different. In each epoch, there are a fixed number of examples that are selected for the network training. We denote the fixed number as $n_t$. The sets including examples for training of $f_1$ and $f_2$ are denoted as $\mathcal{S}_1$ and $\mathcal{S}_2$. Here, we analyze the composition of $\mathcal{S}_1$, since the analysis of $\mathcal{S}_2$ is the same. We omit the index of $\mathcal{S}$ for simplicity. During training, we suppose that $\mathcal{S} = \sigma_s \cup \sigma_l$. For the example in $\sigma_s$, the loss $\mathcal{L}_{\text{D}}$ is small. While, for the example in $\sigma_l$, the loss $\mathcal{L}_{\text{D}}$ is large. At its most extreme, we can measure the magnitude of the loss from whether it means different predictions [9]. This division means that, for $f_1$, the important information provided by $f_2$ exists in $\sigma_l$.

We first analyze the divergence between any $f_2 \in \mathcal{F}_2$ and $\mathcal{S}$. The divergence between $f_2 \in \mathcal{F}_2$ and the target concept $c$ is $d(f_2, c)$ for any $\mathbf{x} \in \sigma_s$. For any $\mathbf{x} \in \sigma_l$, the important information for generalization is provided by $f_1$. The divergence between any $f_2 \in \mathcal{F}_2$ and $f_1$ is $d(f_2, f_1)$. Let $X_1, \ldots, X_{n_t}$ be random variables taking on values in $[0, 1]$, which correspond the divergence between the outputs of $f_2$ and its assigned-label. We then have

$$\begin{aligned}\mathbb{E}[X] &= \mathbb{E}[\sum_{j=1}^{n_t} X_j] = n_s d(f_2, c) + n_l d(f_2, f_1) \\ &= n_s \mathcal{L}_{\text{D}}(f_2, c) + n_l \mathcal{L}_{\text{D}}(f_2, f_1)\end{aligned} \tag{1}$$

Then, we analyze the divergence between the target concept $c$ and $\mathcal{S}$. Assume that the classification loss is normalized. Let $X'_1, \ldots, X'_{n_t}$ be random variables taking on values in $[0, 1]$, which correspond the divergence between the outputs of the target concept $c$ and $\mathcal{S}$. Similar to Eq. (1), we have

$$\mathbb{E}[X'] = \mathbb{E}[\sum_{j=1}^{n_t} X'_j] = n_l d(c, f_1) = n_l \mathcal{L}_{\text{C1}}. \tag{2}$$

The empirical risk minimization is used in this paper. Therefore, the algorithm will search out the classifier that has a small divergence from $\mathcal{S}$. If there is a classifier whose loss is no larger than $\zeta_2$ with probability at least $1 - \delta$, $\mathcal{S}$ should guarantee that the classifier whose loss is larger than $\zeta_2$ has a smaller divergence with $\mathcal{S}$ than the target concept $c$ with probability no larger than $\delta$. Therefore, for $f_2 \in \mathcal{F}_2$, if

$$\mathcal{L}_{\text{D}}(f_2, c) > \zeta_2 = \frac{\mathcal{L}^0_{\text{C2}} \sqrt{n_s^2 + n_s n_l}}{n_s} - \frac{n_l \Phi}{n_s},$$

we have

$$\begin{aligned}&\mathbb{E}[X] - \mathbb{E}[X'] \\ &= n_s \mathcal{L}_{\text{D}}(f_2, c) + n_l \mathcal{L}_{\text{D}}(f_2, f_1) - n_l \mathcal{L}_{\text{C1}} \\ &> n_s \zeta_2 + n_l \mathcal{L}_{\text{D}}(f_2, f_1) - n_l \mathcal{L}_{\text{C1}} \\ &= \mathcal{L}^0_{\text{C2}} \sqrt{n_s^2 + n_s n_l}.\end{aligned} \tag{3}$$

If $f_2$ further minimizes the empirical risk on $\mathcal{S}$, $X \leq X'$. Considering that there are at most $|\mathcal{F}_2| - 1$ classifiers whose losses are larger than $\zeta_2$. According to Hoeffding bounds [3], we have

$$\begin{aligned}&p\left(X' \geq \mathbb{E}[X'] + \frac{\mathcal{L}^0_{\text{C2}} \sqrt{n_s^2 + n_s n_l}}{2}\right) \\ &\leq \exp\left(-\frac{(\mathcal{L}^0_{\text{C2}} \sqrt{n_s^2 + n_s n_l})^2}{2(n_s + n_l)}\right) \quad \text{and}\end{aligned} \tag{4}$$

$$\begin{aligned}&p\left(X \leq \mathbb{E}[X'] + \frac{\mathcal{L}^0_{\text{C2}} \sqrt{n_s^2 + n_s n_l}}{2}\right) \\ &\leq \exp\left(-\frac{(\mathcal{L}^0_{\text{C2}} \sqrt{n_s^2 + n_s n_l})^2}{2(n_s + n_l)}\right).\end{aligned} \tag{5}$$

As $n_s \geq \frac{2}{(\mathcal{L}^0_{\text{C2}})^2} \log \frac{2|\mathcal{F}_2|}{\delta}$, we get

$$\begin{aligned}&\exp\left(-\frac{(\mathcal{L}^0_{\text{C2}} \sqrt{n_s^2 + n_s n_l})^2}{2(n_s + n_l)}\right) \\ &= \exp\left(-\frac{(\mathcal{L}^0_{\text{C2}})^2 (n_s^2 + n_s n_l)}{2(n_s + n_l)}\right) \\ &= \exp\left(-\frac{(\mathcal{L}^0_{\text{C2}})^2 n_s}{2}\right) \\ &\leq \frac{\delta}{2|\mathcal{F}_2|} \leq \frac{\delta}{2}.\end{aligned} \tag{6}$$

Therefore,

$$p\left(X' \geq \mathbb{E}[X'] + \frac{\mathcal{L}^0_{\text{C2}} \sqrt{n_s^2 + n_s n_l}}{2}\right) \leq \frac{\delta}{2}, \tag{7}$$

$$p\left(X \leq \mathbb{E}[X'] + \frac{\mathcal{L}^0_{\text{C2}} \sqrt{n_s^2 + n_s n_l}}{2}\right) \leq \frac{\delta}{2} \tag{8}$$

will hold, which guarantee that the classifier whose loss is larger than $\zeta_2$ has a smaller divergence with $\mathcal{S}$ than the target concept $c$ with probability no larger than $\delta$. Thus, we have that $p(\mathcal{L}_{\text{C2}} < \zeta_2) \geq 1 - \delta$ holds. Similarly, $p(\mathcal{L}_{\text{C1}} < \zeta_1) \geq 1 - \delta$ holds.

# B. Supplementary Experimental Settings

## B.1. Details of Used Datasets

The statistics of used datasets are shown in Table 1.

Table 1. The summary of simulated noisy datasets used in the experiments.

| Dataset | type | # of training | # of testing | # of class | size |
|---|---|---|---|---|---|
| *MNIST* | image | 60,000 | 10,000 | 10 | 28×28×1 |
| *F-MNIST* | image | 60,000 | 10,000 | 10 | 28×28×1 |
| *SVHN* | image | 73,257 | 26,032 | 10 | 32×32×3 |
| *CIFAR-10* | image | 50,000 | 10,000 | 10 | 32×32×3 |
| *CIFAR-100* | image | 50,000 | 10,000 | 100 | 32×32×3 |
| *NEWS* | text | 11,314 | 7,532 | 20 | 300-D |

### B.2. Details of Noise Generation

**Class-balanced cases.** Here, we introduce the details of generating different types of noisy labels. We mainly follow the settings in [10]. The details are described as follows:

◇ Instance-independent noise

- Symmetric noise.: We flip clean labels in each class *uniformly* to incorrect labels of other classes.

- Pairflip noise: We flip clean labels in each class to its *adjacent* class.

- Tridiagonal noise: the noise corresponds to a spectral of classes where adjacent classes are easier to be mutually mislabeled, which can be implemented by *two consecutive pair flipping* transformations in the opposite direction.

We corrupt clean datasets manually by the label transition matrix $T$, where $T_{ij} = p(\tilde{\mathbf{y}} = \mathbf{e}_j | \mathbf{y} = \mathbf{e}_i)$, given that noisy $\tilde{\mathbf{y}}$ is flipped from clean $\mathbf{y}$. When the noise rate is set to $\epsilon$, the transition matrices for the above three types of label noise are shown in (9), (10), and (11).

◇ Instance-dependent noise

- Instance noise: We consider that the probability that an instance is mislabeled depends on its *features/instances*. The generation of such a kind of noise follows the procedure in [11, 8, 4].

$$\text{Sym. } \epsilon\text{:} \quad T = \begin{bmatrix} 1-\epsilon & \frac{\epsilon}{c-1} & \dots & \frac{\epsilon}{c-1} & \frac{\epsilon}{c-1} \\ \frac{\epsilon}{c-1} & 1-\epsilon & \frac{\epsilon}{c-1} & \dots & \frac{\epsilon}{c-1} \\ \vdots & & \ddots & & \vdots \\ \frac{\epsilon}{c-1} & \dots & \frac{\epsilon}{c-1} & 1-\epsilon & \frac{\epsilon}{c-1} \\ \frac{\epsilon}{c-1} & \frac{\epsilon}{c-1} & \dots & \frac{\epsilon}{c-1} & 1-\epsilon \end{bmatrix}_{c\times c} . \tag{9}$$

$$\text{Pair. } \epsilon\text{:} \quad T = \begin{bmatrix} 1-\epsilon & \epsilon & \dots & 0 & 0 \\ 0 & 1-\epsilon & \epsilon & \dots & 0 \\ \vdots & & \ddots & & \vdots \\ 0 & \dots & 0 & 1-\epsilon & \epsilon \\ \epsilon & 0 & \dots & 0 & 1-\epsilon \end{bmatrix}_{c\times c} . \tag{10}$$

$$\text{Trid. } \epsilon\text{:} \quad T = \begin{bmatrix} 1-\epsilon & \frac{\epsilon}{2} & \dots & 0 & \frac{\epsilon}{2} \\ \frac{\epsilon}{2} & 1-\epsilon & \frac{\epsilon}{2} & \dots & 0 \\ \vdots & & \ddots & & \vdots \\ 0 & \dots & \frac{\epsilon}{2} & 1-\epsilon & \frac{\epsilon}{2} \\ \frac{\epsilon}{2} & 0 & \dots & \frac{\epsilon}{2} & 1-\epsilon \end{bmatrix}_{c\times c} . \tag{11}$$

**Class-imbalanced cases.** In this paper, we consider two types of ways for building *imbalanced noisy* datasets. The first one is asymmetric noise, which is injected into four datasets, *i.e.*, *MNIST*, *F-MNIST*, *SVHN*, and *CIFAR-10*. For *MNIST*, flipping 2→7, 3→8, 5↔6. For *F-MNIST*, flipping TSHIRT→SHIRT, PULLOVER→COAT, SANDALS→SNEAKER. For *SVHN*, flipping 2→7, 3→8, 5↔6. For *CIFAR-10*, flipping TRUCK→AUTOMOBILE, BIRD→AIRPLANE, DEER→HORSE, CAT↔DOG. As some flip processes (*e.g.*, 2→7, but not 2↔7) are *not bidirectional*, the simulated noisy datasets are imbalanced accordingly.

## C. Supplementary Experimental Results

### C.1. Results on Balanced Noisy Datasets

In the main paper, we have provided experimental results on simulated *CIFAR-10* and *NEWS*. Here, we provide results on the other four balanced noisy datasets, which are shown in Table 2. Besides, before this, for the symmetric noise, we set the noise rate to 20% and 40% respectively to verify the effectiveness of our method. Here, we increase the noise levels to 50%, 60%, and 70% to further support our claims. Experiments are conducted on MNIST and F-MNIST. The experimental results in Tables 3 support our claims well.

### C.2. Results on Imbalanced Noisy Datasets

**Experiments on noisy long-tailed *CIFAR-100*.** We provide the experiments on noisy imbalanced *CIFAR-100*. Note that, it is somewhat complex to consider the visual similarity of classes in *CIFAR-100*, since there are a large number of classes. Therefore, we focuses on noisy long-tailed cases. The asymmetric noise injected into *CIFAR-100* is bulit by: the 100 classes are grouped into 20 super-classes, and each has 5 sub-classes. Each class is then flipped into the next within the same super-class. In addition, long-tailed *CIFAR-100* is built similarly to *MNIST* and *SVHN*, resulting in the *L-CIFAR-100-1* and *L-CIFAR-100-2* datasets. The other settings are the same as the experiments on noisy balanced *CIFAR-100*. The results are provided in Table 4, which verify the effectiveness of the proposed method. Note that the reported test accuracy of all methods is relatively low. It is because *CIFAR-100* is challenging,

Table 2. Mean and standard deviations of test accuracy (%) on five balanced noisy datasets with different noise levels. The test accuracy is calculated over the last ten epochs. The results are reported over five trials. The best result and second best result in each case are highlighted in red and blue respectively.

| | Noise type | Sym. | | Pair. | | Trid. | | Ins. | |
|---|---|---|---|---|---|---|---|---|---|
| | Setting | 20% | 40% | 20% | 40% | 20% | 40% | 20% | 40% |
| *MNIST* | APL | 98.76±0.06 | 94.92±0.31 | 98.66±0.10 | 68.44±2.95 | 98.93±0.04 | 76.44±3.04 | 97.63±0.73 | 87.90±1.94 |
| | CDR | 94.77±0.17 | 92.16±0.73 | 93.25±0.90 | 71.02±3.89 | 94.06±0.92 | 70.28±4.01 | 93.17±0.96 | 77.45±3.04 |
| | MentorNet | 95.04±0.03 | 92.08±0.42 | 93.19±0.17 | 90.93±1.54 | 96.42±0.09 | 93.28±1.37 | 94.65±0.73 | 90.11±1.26 |
| | SIGUA | 92.31±1.10 | 91.88±0.92 | 93.77±1.40 | 86.22±1.75 | 94.92±0.83 | 83.46±2.98 | 92.90±1.82 | 86.34±3.51 |
| | Co-teaching | 97.53±0.12 | 95.62±0.30 | 96.05±0.96 | 94.16±1.37 | 98.05±0.06 | 96.18±0.85 | 97.96±0.09 | 95.02±0.39 |
| | Decoupling | 98.39±0.08 | 81.56±0.72 | 97.82±0.31 | 66.48±0.78 | 98.33±0.11 | 74.55±0.97 | 98.05±0.30 | 71.87±1.24 |
| | Co-teaching+ | 98.25±0.13 | 92.63±0.34 | 97.30±0.16 | 92.00±0.31 | 98.00±0.16 | 93.06±0.24 | 96.83±0.28 | 89.99±0.37 |
| | JoCor | 98.42±0.14 | 98.04±0.07 | 98.01±0.19 | 96.85±0.43 | 98.45±0.17 | 96.98±0.25 | 98.62±0.06 | 96.07±0.31 |
| | CoDis | 98.80±0.04 | 98.33±0.09 | 98.28±0.12 | 95.39±1.24 | 98.93±0.04 | 97.17±0.14 | 98.40±0.15 | 96.12±0.96 |
| *F-MNIST* | APL | 91.73±0.20 | 89.06±0.41 | 90.22±0.80 | 78.54±4.33 | 90.84±0.22 | 86.53±0.76 | 90.96±0.77 | 85.55±2.86 |
| | CDR | 85.62±0.96 | 71.83±1.37 | 85.72±0.65 | 69.07±2.31 | 86.75±1.19 | 73.63±2.82 | 85.92±1.43 | 73.14±3.12 |
| | MentorNet | 90.37±0.17 | 86.53±0.65 | 87.92±0.18 | 83.70±0.49 | 88.74±0.33 | 85.63±0.59 | 87.52±0.15 | 83.27±1.42 |
| | SIGUA | 87.64±1.29 | 87.23±0.32 | 69.59±5.75 | 68.93±2.80 | 79.97±3.23 | 76.14±4.24 | 79.97±3.23 | 76.14±4.24 |
| | Co-teaching | 91.48±0.10 | 88.80±0.29 | 90.77±0.23 | 86.91±0.71 | 91.24±0.11 | 89.18±0.36 | 90.60±0.12 | 87.90±0.45 |
| | Decoupling | 88.89±0.47 | 70.45±0.62 | 87.03±0.32 | 60.12±0.23 | 88.42±0.37 | 65.98±1.05 | 87.16±0.77 | 63.48±0.88 |
| | Co-teaching+ | 89.95±0.18 | 83.73±0.44 | 88.33±0.45 | 71.76±1.57 | 89.68±0.41 | 79.47±0.92 | 88.64±0.26 | 75.40±2.40 |
| | JoCor | 91.97±0.13 | 89.96±0.19 | 91.52±0.24 | 87.40±0.58 | 92.01±0.17 | 89.42±0.33 | 91.43±0.71 | 87.59±0.94 |
| | CoDis | 92.21±0.17 | 90.49±0.24 | 91.66±0.31 | 87.07±0.51 | 92.19±0.30 | 88.70±0.94 | 91.48±0.52 | 88.04±0.58 |
| *SVHN* | APL | 89.05±0.43 | 83.51±3.03 | 89.29±1.23 | 68.07±4.98 | 90.88±1.31 | 80.86±2.28 | 90.21±0.52 | 72.75±4.25 |
| | CDR | 83.45±1.23 | 61.99±1.42 | 82.72±0.76 | 59.76±1.06 | 83.42±0.88 | 63.19±1.22 | 82.11±0.27 | 60.05±1.39 |
| | MentorNet | 93.18±0.26 | 92.02±0.24 | 92.78±0.25 | 81.05±0.37 | 92.99±0.16 | 90.16±0.16 | 92.21±0.27 | 87.60±0.79 |
| | SIGUA | 92.31±0.32 | 89.73±0.34 | 75.88±2.43 | 72.21±3.61 | 82.94±2.06 | 78.14±4.25 | 77.29±7.68 | 76.40±3.85 |
| | Co-teaching | 93.61±0.11 | 91.89±0.25 | 93.53±0.20 | 90.37±0.49 | 93.62±0.19 | 90.65±0.43 | 93.13±0.36 | 89.99±0.65 |
| | Decoupling | 88.46±0.19 | 65.22±3.74 | 87.80±0.83 | 63.02±3.28 | 89.04±0.61 | 66.73±0.64 | 87.25±0.93 | 62.06±1.34 |
| | Co-teaching+ | 90.31±0.30 | 87.60±0.54 | 89.85±0.37 | 69.17±1.58 | 90.31±0.31 | 80.15±0.92 | 88.43±0.55 | 70.16±3.00 |
| | JoCor | 93.70±0.20 | 92.16±0.26 | 93.54±0.43 | 90.73±0.17 | 93.74±0.12 | 90.97±0.39 | 93.32±0.42 | 89.37±0.56 |
| | CoDis | 93.75±0.17 | 92.22±0.42 | 93.54±0.26 | 91.29±0.33 | 93.65±0.14 | 90.75±0.27 | 93.42±1.02 | 90.15±1.29 |
| *CIFAR-100* | APL | 27.36±0.56 | 22.30±1.31 | 27.51±0.82 | 19.56±0.89 | 28.07±1.43 | 21.07±0.62 | 26.96±0.63 | 18.80±1.99 |
| | CDR | 31.42±0.74 | 25.77±0.63 | 32.88±0.65 | 23.35±1.62 | 33.04±1.05 | 26.74±2.86 | 32.26±0.94 | 21.77±2.16 |
| | MentorNet | 43.15±0.42 | 37.62±0.89 | 40.06±0.37 | 27.17±0.92 | 42.20±0.30 | 31.74±0.88 | 40.54±0.69 | 33.09±1.53 |
| | SIGUA | 42.03±0.33 | 40.53±0.49 | 36.48±0.47 | 26.73±0.33 | 39.21±0.40 | 32.69±0.36 | 39.19±0.32 | 33.51±0.43 |
| | Co-teaching | 45.17±0.25 | 40.95±0.52 | 42.50±0.39 | 30.07±0.17 | 44.41±0.41 | 34.96±0.35 | 42.23±0.52 | 35.87±1.47 |
| | Decoupling | 31.53±0.28 | 19.09±0.29 | 35.85±0.35 | 25.36±0.38 | 35.01±0.12 | 24.72±0.47 | 33.46±0.51 | 22.53±0.58 |
| | Co-teaching+ | 35.89±0.70 | 24.95±0.96 | 36.16±0.40 | 24.76±0.46 | 36.85±0.61 | 26.06±0.30 | 36.19±0.57 | 25.89±0.37 |
| | JoCor | 45.93±0.21 | 41.56±0.57 | 42.12±0.35 | 30.12±0.65 | 44.98±0.27 | 34.23±1.13 | 44.28±0.59 | 35.60±0.99 |
| | CoDis | 45.19±0.31 | 41.53±0.88 | 42.63±0.10 | 30.58±0.30 | 45.42±0.88 | 35.35±0.98 | 44.25±0.26 | 36.49±0.73 |

Table 3. Mean and standard deviations of test accuracy (%) on *MNIST* and *F-MNIST* with high noise levels over the last ten epochs. The best result and second best result in each case are highlighted in red and blue respectively.

| | Method/Noise | Sym. 50% | Sym. 60% | Sym. 70% |
|---|---|---|---|---|
| *MNIST* | APL | 84.97±2.97 | 75.68±1.22 | 70.11±0.52 |
| | CDR | 76.85±2.46 | 57.22±1.92 | 54.22±0.94 |
| | MentorNet | 91.14±0.17 | 90.11±0.37 | 88.72±0.46 |
| | SIGUA | 91.35±2.62 | 88.62±1.93 | 86.08±6.04 |
| | Co-teaching | 95.60±0.38 | 95.44±0.30 | 94.11±0.38 |
| | Decoupling | 80.22±0.33 | 78.36±2.16 | 74.63±1.66 |
| | Co-teaching+ | 92.30±0.55 | 90.77±0.41 | 86.52±0.89 |
| | JoCor | 97.14±0.10 | 96.47±0.46 | 95.01±0.29 |
| | CoDis | 97.10±0.04 | 96.62±0.13 | 95.39±0.26 |
| *F-MNIST* | APL | 76.80±3.21 | 72.77±4.37 | 68.39±7.17 |
| | CDR | 53.41±1.81 | 45.82±2.77 | 41.33±3.69 |
| | MentorNet | 86.51±0.11 | 85.91±0.44 | 83.27±0.55 |
| | SIGUA | 83.39±3.29 | 79.36±4.54 | 72.14±4.28 |
| | Co-teaching | 88.72±0.14 | 87.92±0.34 | 85.92±0.72 |
| | Decoupling | 66.12±2.37 | 63.77±0.94 | 57.68±0.49 |
| | Co-teaching+ | 83.25±0.35 | 80.92±0.65 | 77.52±0.73 |
| | JoCor | 89.16±0.27 | 87.93±0.61 | 86.99±0.92 |
| | CoDis | 89.63±0.30 | 88.24±0.40 | 87.15±0.85 |

Table 4. Mean and standard deviations of test accuracy (%) on noisy long-tailed *CIFAR-100* with different noise levels. The test accuracy is calculated over the last ten epochs. The results are reported over five trials. The best result and second best result in each case are highlighted in red and blue respectively.

| | Noise type | *L-CIFAR-100-1* | | | | *L-CIFAR-100-2* | | | |
|---|---|---|---|---|---|---|---|---|---|
| | Setting | Asym. 20% | Asym. 30% | Asym. 40% | Asym. 45% | Asym. 20% | Asym. 30% | Asym. 40% | Asym. 45% |
| *L-CIFAR-100* | APL | 20.93±0.55 | 17.43±1.11 | 13.09±0.92 | 10.50±2.27 | 22.19±0.74 | 17.22±0.25 | 12.06±1.09 | 10.95±0.88 |
| | CDR | 30.22±0.65 | 23.06±1.07 | 18.77±1.90 | 13.79±3.54 | 23.19±1.13 | 18.15±1.10 | 14.22±0.14 | 13.52±0.96 |
| | MentorNet | 33.66±0.73 | 28.57±0.75 | 21.98±1.07 | 18.32±0.63 | 27.27±0.65 | 24.47±1.30 | 19.81±1.07 | 16.88±0.99 |
| | SIGUA | 24.83±0.41 | 21.41±0.18 | 16.71±0.65 | 13.76±0.33 | 20.83±0.40 | 19.51±0.59 | 13.30±0.61 | 11.61±0.59 |
| | Co-teaching | 34.30±0.70 | 29.88±0.44 | 24.40±0.50 | 20.39±0.65 | 32.25±0.47 | 26.94±0.69 | 20.14±1.08 | 18.77±0.67 |
| | Decoupling | 28.69±0.67 | 24.16±0.33 | 19.79±0.40 | 17.73±0.31 | 25.90±0.58 | 21.93±0.33 | 17.98±0.30 | 16.26±0.20 |
| | Co-teaching+ | 28.10±0.31 | 23.50±0.54 | 18.78±0.46 | 16.52±0.26 | 25.56±0.56 | 21.55±0.51 | 17.18±0.61 | 15.09±0.38 |
| | JoCor | 35.38±0.71 | 28.27±0.65 | 20.73±0.92 | 18.66±0.65 | 29.46±0.97 | 25.07±0.44 | 19.07±0.70 | 16.26±0.52 |
| | CoDis | 34.92±0.45 | 31.14±0.34 | 24.75±0.73 | 21.26±0.75 | 33.15±0.13 | 28.11±0.16 | 22.41±0.60 | 20.23±0.55 |

Table 5. Mean and standard deviations of test accuracy (%) on two class-imbalanced noisy datasets with different noise levels. **ResNet-34 is used**. The test accuracy is calculated over the last ten epochs. The results are reported over five trials. The best result and second best result in each case are highlighted in red and blue respectively.

| | Noise type | Asym. 20% | Asym. 30% | Asym. 40% | Asym. 45% |
|---|---|---|---|---|---|
| *SVHN* | APL | 92.09±0.15 | 87.30±0.32 | 78.64±0.50 | 65.76±1.46 |
| | CDR | 91.06±1.09 | 85.73±1.95 | 75.44±2.32 | 63.77±2.79 |
| | MentorNet | 92.15±0.53 | 86.71±0.26 | 76.05±4.40 | 60.80±3.50 |
| | SIGUA | 85.49±0.91 | 77.65±0.69 | 50.80±2.77 | 48.75±3.80 |
| | Co-teaching | 95.43±0.08 | 93.95±0.20 | 91.03±0.46 | 88.20±3.17 |
| | Decoupling | 92.17±0.52 | 85.17±0.93 | 82.19±0.77 | 77.83±1.62 |
| | Co-teaching+ | 93.03±1.24 | 88.97±1.07 | 85.73±1.21 | 80.29±1.31 |
| | JoCor | 93.93±0.28 | 89.12±2.65 | 70.73±4.11 | 52.59±3.61 |
| | CoDis | 95.73±0.11 | 95.44±0.19 | 94.52±0.40 | 93.92±0.37 |
| *CIFAR-10* | APL | 80.17±0.62 | 75.33±2.18 | 71.65±1.75 | 56.92±1.06 |
| | CDR | 79.36±0.58 | 76.22±0.39 | 70.44±1.06 | 53.92±1.75 |
| | MentorNet | 80.91±1.54 | 77.43±0.59 | 63.16±7.17 | 52.05±2.77 |
| | SIGUA | 77.58±0.48 | 71.20±1.35 | 60.24±2.17 | 36.82±3.99 |
| | Co-teaching | 83.14±0.26 | 81.83±0.52 | 72.13±0.76 | 55.93±3.92 |
| | Decoupling | 78.86±0.34 | 74.69±0.20 | 67.11±0.58 | 52.17±2.95 |
| | Co-teaching+ | 77.76±0.69 | 73.32±0.65 | 69.82±1.73 | 51.80±1.95 |
| | JoCor | 83.47±0.26 | 80.43±0.46 | 70.77±1.94 | 50.45±3.05 |
| | CoDis | 83.59±0.15 | 82.37±0.61 | 73.06±0.19 | 57.28±1.35 |

and we use a simple CNN as did in [9]. In addition, we do not employ other techniques, *e.g.*, data augmentations.

**Experiments with different networks.** Before this, we use a 9-layer CNN for *SVHN* and *CIFAR-10*. To show that our method is robust to network structures, we use ResNet-34 and MobileNet V2 [5] for these two datasets. The results are provided in Tables 5 and 6 respectively. We can see that with different network structures, CoDis exhibits superior robustness to multiple baselines.

**Experiments with data augmentation.** Before this, we verify the effectiveness of our method without data augmentation, as did in [2, 7]. Here, we exploit data augmentation that is commonly used. That is, we perform data augmentation by horizontal random flips and 32×32 random crops after padding 4 pixels on each side. The networks ResNet-34 and MobileNet V2 are employed for *SVHN*. The results are provided in Table 7. As can be seen, when data augmentation is used, CoDis still works well in all cases.

## C.3. Hyperparameter Sensitivity Analysis

**Analysis of $\alpha$.** It is easy to analyze the role of the used divergence strategy by comparing our method with Co-teaching. As we employ $\alpha$ to keep divergence of two deep networks, we the algorithm stability with different values of $\alpha$. The experiments are conducted with noisy datasets with symmetric noise. Implementation details are kept the same as above. The results in Figure 1 demonstrate the stability of our method with different $\alpha$. The ablation study about $\alpha$ on imbalanced noisy datasets is provided in Figure 2. We can see that in the certain value range, our method is robust to the choice of $\alpha$. The results mean that our method can be easy to apply, without sophisticated hyperparameter tuning.

**Analysis of $T_k$.** Here we exploit *MNIST*. Following the original paper of Co-teaching, we fix $T_{\max}$=200 and set $T_k$=5, 10, 15 respectively. Results are provided in Table 8. As can be seen, our method is not sensitive to varying $T_k$.

**Analysis of $T_{\max}$.** We fix $T_k$=10 and set $T_{\max}$=225, 250, 275 respectively. Results are shown in Table 9, which demonstrate the stability of CoDis to the changes of $T_{\max}$.

## C.4. Experiments with Label Precision

We provide comparison results about the label precision. Here we compare CoDis with Co-teaching that also employs the cross-update way, where *MNIST* is used. We report results in Table 10. As can be seen, the label precision of CoDis is higher than Co-teaching, especially when train-

Table 6. Mean and standard deviations of test accuracy (%) on two class-imbalanced noisy datasets with different noise levels. **MobileNet V2 is used**. The test accuracy is calculated over the last ten epochs. The results are reported over five trials. The best result and second best result in each case are highlighted in red and blue respectively.

| | Noise type | Asym. 20% | Asym. 30% | Asym. 40% | Asym. 45% |
|---|---|---|---|---|---|
| *SVHN* | APL | 92.06±0.17 | 85.22±0.39 | 73.14±1.19 | 68.74±1.16 |
| | CDR | 91.75±1.04 | 83.15±1.28 | 67.84±2.40 | 63.33±3.64 |
| | MentorNet | 92.37±0.22 | 86.05±2.54 | 72.34±2.35 | 62.14±4.24 |
| | SIGUA | 82.17±0.34 | 70.82±3.41 | 40.77±3.96 | 40.52±4.90 |
| | Co-teaching | 95.18±0.11 | 94.11±0.19 | 84.46±3.63 | 73.47±4.16 |
| | Decoupling | 92.52±0.17 | 85.13±0.65 | 74.73±2.65 | 65.25±1.04 |
| | Co-teaching+ | 93.03±1.24 | 88.92±0.45 | 76.86±0.61 | 64.38±1.13 |
| | JoCor | 93.77±0.15 | 82.93±1.29 | 73.11±4.74 | 62.18±3.70 |
| | CoDis | 95.59±0.09 | 95.45±0.12 | 92.13±0.20 | 83.16±2.32 |
| *CIFAR-10* | APL | 80.17±0.33 | 76.33±0.62 | 68.73±2.06 | 53.92±1.50 |
| | CDR | 79.09±0.32 | 74.88±1.09 | 65.85±0.92 | 50.11±2.05 |
| | MentorNet | 79.65±0.76 | 76.45±0.30 | 65.18±1.24 | 51.24±3.06 |
| | SIGUA | 78.11±0.69 | 70.11±0.50 | 60.20±1.84 | 40.27±3.96 |
| | Co-teaching | 82.47±0.14 | 81.09±0.32 | 72.90±0.22 | 53.12±2.21 |
| | Decoupling | 77.82±0.47 | 75.22±0.59 | 67.28±1.76 | 51.28±3.55 |
| | Co-teaching+ | 78.49±0.41 | 75.10±1.27 | 66.28±1.42 | 51.75±2.87 |
| | JoCor | 82.13±0.27 | 79.74±0.24 | 65.45±6.93 | 51.80±3.32 |
| | CoDis | 82.60±0.22 | 81.25±0.27 | 73.05±0.19 | 54.50±1.06 |

Table 7. Mean and standard deviations of test accuracy (%) on noisy *SVHN* with different noise levels. **Data augmentation is employed**. The test accuracy is calculated over the last ten epochs. The results are reported over five trials. The best result and second best result in each case are highlighted in red and blue respectively.

| | Noise type | Asym. 20% | Asym. 30% | Asym. 40% | Asym. 45% |
|---|---|---|---|---|---|
| ResNet-34 | APL | 95.06±0.07 | 92.45±0.18 | 90.45±0.73 | 88.84±1.75 |
| | CDR | 93.95±0.23 | 86.24±0.66 | 82.34±3.65 | 70.54±7.34 |
| | MentorNet | 92.60±0.19 | 89.08±0.34 | 78.84±3.14 | 65.54±1.94 |
| | SIGUA | 92.15±0.62 | 86.39±1.36 | 73.68±2.44 | 65.63±5.81 |
| | Co-teaching | 95.66±0.05 | 93.87±0.24 | 91.22±0.66 | 90.33±2.33 |
| | Decoupling | 93.06±0.07 | 91.07±0.54 | 89.72±0.99 | 86.03±2.26 |
| | Co-teaching+ | 95.21±0.04 | 94.47±0.90 | 94.12±0.39 | 89.78±5.23 |
| | JoCor | 94.10±0.06 | 91.08±0.32 | 80.19±3.58 | 63.74±2.69 |
| | CoDis | 96.50±0.07 | 96.10±0.06 | 95.47±0.14 | 95.05±0.92 |
| MobileNet V2 | APL | 94.33±0.12 | 92.65±0.18 | 90.37±0.83 | 89.67±1.95 |
| | CDR | 94.02±0.67 | 91.88±0.77 | 86.73±3.77 | 75.65±6.72 |
| | MentorNet | 92.77±0.02 | 89.01±0.71 | 74.71±3.35 | 65.18±1.22 |
| | SIGUA | 89.73±2.14 | 85.77±0.91 | 70.67±3.14 | 62.63±2.69 |
| | Co-teaching | 95.59±0.11 | 94.17±0.18 | 91.88±0.34 | 84.90±5.20 |
| | Decoupling | 95.13±0.14 | 93.81±0.36 | 91.38±0.60 | 86.70±2.95 |
| | Co-teaching+ | 95.54±0.24 | 95.06±0.10 | 94.77±0.28 | 89.38±4.40 |
| | JoCor | 94.13±0.40 | 90.94±0.21 | 72.43±6.44 | 65.34±2.34 |
| | CoDis | 96.50±0.09 | 95.86±0.13 | 95.32±0.20 | 94.83±0.30 |

Table 8. The sensitivity analysis of the parameter $T_k$.

| Noise Setting | $T_k$=5 | $T_k$=10 | $T_k$=15 |
|---|---|---|---|
| *MNIST*+Sym. 40% | 98.25±0.14 | 98.33±0.09 | 98.33±0.05 |
| *MNIST*+Asym. 40% | 99.03±0.03 | 99.01±0.14 | 98.82±0.08 |

Table 9. The sensitivity analysis of the parameter $T_{\max}$.

| Noise Setting | $T_{\max}$=225 | $T_{\max}$=250 | $T_{\max}$=275 |
|---|---|---|---|
| *MNIST*+Sym. 40% | 98.23±0.02 | 98.19±0.04 | 98.12±0.04 |
| *MNIST*+Asym. 40% | 98.89±0.20 | 99.01±0.12 | 98.91±0.14 |

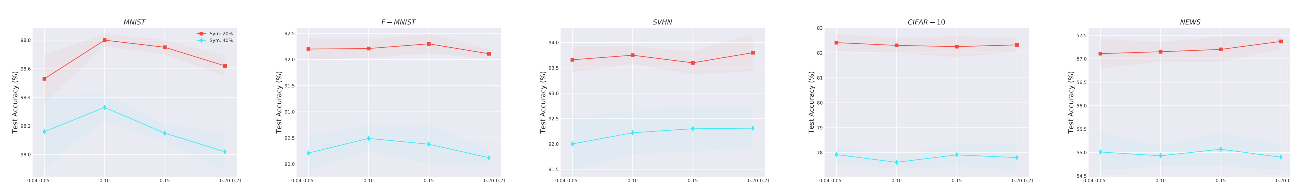


Figure 1. Illustrations of the hyperparameter sensitivity for our method. The error bar for standard deviation in each figure has been shaded.

ing data are class-imbalanced (Asym. 20%).

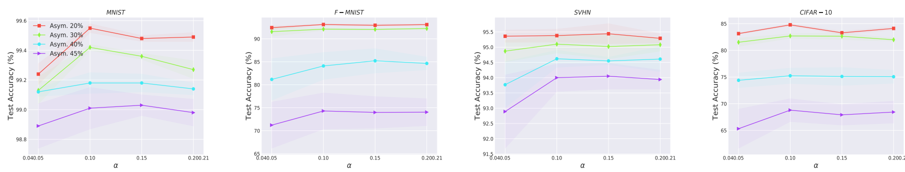


Figure 2. Illustrations of the hyperparameter sensitivity for the proposed method on four imbalanced noisy datasets. The error bar for standard deviation in each figure has been shaded.

Table 10. Comparing CoDis to Co-teaching about label precision (%) that is calculated over the last ten epochs.

| | Sym. 20% | Pair.20% | Ins. 20% | Asym.20% |
|---|---|---|---|---|
| Co-teaching | 94.62±0.24 | 92.65±0.40 | 94.66±0.08 | 94.95±0.16 |
| CoDis | **95.93±0.16** | **95.01±0.07** | **94.92±0.13** | **99.24±0.02** |

## References


[1] Avrim Blum and Tom Mitchell. Combining labeled and unlabeled data with co-training. In *ACCLT*, pages 92–100, 1998. 1

[2] Bo Han, Quanming Yao, Xingrui Yu, Gang Niu, Miao Xu, Weihua Hu, Ivor Tsang, and Masashi Sugiyama. Co-teaching: Robust training of deep neural networks with extremely noisy labels. In *NeurIPS*, pages 8527–8537, 2018. 4

[3] Wassily Hoeffding. Probability inequalities for sums of bounded random variables. In *The collected works of Wassily Hoeffding*, pages 409–426. 1994. 1

[4] Zhimeng Jiang, Kaixiong Zhou, Zirui Liu, Li Li, Rui Chen, Soo-Hyun Choi, and Xia Hu. An information fusion approach to learning with instance-dependent label noise. In *ICLR*, 2022. 2

[5] Mark Sandler, Andrew Howard, Menglong Zhu, Andrey Zhmoginov, and Liang-Chieh Chen. Mobilenetv2: Inverted residuals and linear bottlenecks. In *CVPR*, pages 4510–4520, 2018. 4

[6] Wei Wang and Zhi-Hua Zhou. Theoretical foundation of co-training and disagreement-based algorithms. *arXiv preprint arXiv:1708.04403*, 2017. 1

[7] Hongxin Wei, Lei Feng, Xiangyu Chen, and Bo An. Combating noisy labels by agreement: A joint training method with co-regularization. In *CVPR*, pages 13726–13735, 2020. 4

[8] Xiaobo Xia, Tongliang Liu, Bo Han, Nannan Wang, Mingming Gong, Haifeng Liu, Gang Niu, Dacheng Tao, and Masashi Sugiyama. Part-dependent label noise: Towards instance-dependent label noise. In *NeurIPS*, 2020. 2

[9] Xingrui Yu, Bo Han, Jiangchao Yao, Gang Niu, Ivor W Tsang, and Masashi Sugiyama. How does disagreement benefit co-teaching? In *ICML*, 2019. 1, 4

[10] Yivan Zhang, Gang Niu, and Masashi Sugiyama. Learning noise transition matrix from only noisy labels via total variation regularization. In *ICML*, 2021. 2

[11] Zhaowei Zhu, Tongliang Liu, and Yang Liu. A second-order approach to learning with instance-dependent label noise. In *CVPR*, pages 10113–10123, 2021. 2