# Supplementary Material of “Sample Selection with Uncertainty of Losses for Learning with Noisy Labels”

**Anonymous Author(s)**
Affiliation
Address
email

## A Proof of Theoretical Results

### A.1 Proof of Theorem 1

For the circumstance with soft truncation, $\tilde{\mu}_z = \frac{1}{n}\sum_{i=1}^{n}\psi(z_i)$. As suggested in [1], we can exploit $\tilde{\mu}_z^-$ and $\tilde{\mu}_z^+$ such that

$$\tilde{\mu}_z^- \leq \tilde{\mu}_z \leq \tilde{\mu}_z^+, \tag{1}$$

to derive a bound for $\tilde{\mu}_z$. For some positive real parameter $\alpha$, we define

$$r(\tilde{\mu}_z) = \sum_{i=1}^{n}\psi\left[\alpha(z_i - \tilde{\mu}_z)\right] = 0. \tag{2}$$

Let us introduce the quantity

$$r(\theta) = \frac{1}{\alpha n}\sum_{i=1}^{n}\psi\left[\alpha(z_i - \theta)\right]. \tag{3}$$

With the exponential moment inequality [5] and the C$_r$ inequality [9], we have

$$\begin{aligned}\exp\{\alpha n r(\theta)\} &\leq \left\{1 + \alpha(\mu_z - \theta) + \alpha^2[\sigma^2 + (\mu_z - \theta)^2]\right\}^n \\ &\leq \exp\{n\alpha(\mu_z - \theta) + n\alpha^2[\sigma^2 + (\mu_z - \theta)^2]\}.\end{aligned} \tag{4}$$

In the same way,

$$\exp\{-\alpha n r(\theta)\} \leq \exp\{-n\alpha(\mu_z - \theta) + n\alpha^2[\sigma^2 + (\mu_z - \theta)^2]\}. \tag{5}$$

If we define for any $\mu_s \in \mathbb{R}$ the bounds

$$B_-(\theta) = \mu_z - \theta - \alpha[\sigma^2 + (\mu_z - \theta)^2] - \frac{\log(\epsilon^{-1})}{\alpha n} \tag{6}$$

and

$$B_+(\theta) = \mu_z - \theta + \alpha[\sigma^2 + (\mu_z - \theta)^2] + \frac{\log(\epsilon^{-1})}{\alpha n}. \tag{7}$$

From [2] (Lemma 2.2), we obtain that

$$P(r(\theta) > B_-(\theta)) \geq 1 - \epsilon \quad \text{and} \quad P(r(\theta) < B_+(\theta)) \geq 1 - \epsilon. \tag{8}$$

Let $\tilde{\mu}_z^-$ be the largest solution of the quadratic equation $B_-(\theta)$ and $\tilde{\mu}_z^+$ be the smallest solution of the quadratic equation $B_+(\theta)$. Also, to guarantee the solution of the quadratic equation, we assume

$$4\alpha^2\sigma^2 + \frac{4\log(\epsilon^{-1})}{n} \leq 1. \tag{9}$$

Submitted to 35th Conference on Neural Information Processing Systems (NeurIPS 2021). Do not distribute.

From [2] (Theorem 2.6), we then have

$$\tilde{\mu}_z^- \geq \mu_z - \frac{\alpha\sigma^2 + \frac{\log(\epsilon^{-1})}{\alpha n}}{\alpha - 1}, \tag{10}$$

and

$$\tilde{\mu}_z^+ \leq \mu_z + \frac{\alpha\sigma^2 + \frac{\log(\epsilon^{-1})}{\alpha n}}{\alpha - 1}. \tag{11}$$

With probability at least 1-2$\epsilon$, we have $\tilde{\mu}_z^- \leq \tilde{\mu}_z \leq \tilde{\mu}_z^+$. We can choose $\alpha = \frac{n}{\sigma^2}$. Then we have

$$|\tilde{\mu}_z - \mu_z| \leq \frac{\sigma^2(n + \frac{\sigma^2 \log(\epsilon^{-1})}{n^2})}{n - \sigma^2}, \tag{12}$$

which holds with probability at least 1-2$\epsilon$.

We exploit the lower bound and let $\epsilon = \frac{1}{2t}$. Then we have

$$\ell_s^\star = \tilde{\mu}_s - \frac{\sigma^2(t + \frac{\sigma^2 \log(2t)}{t^2})}{n_t - \sigma^2}, \tag{13}$$

where $n_t$ denotes the number of times that the example was selected in the time intervals.

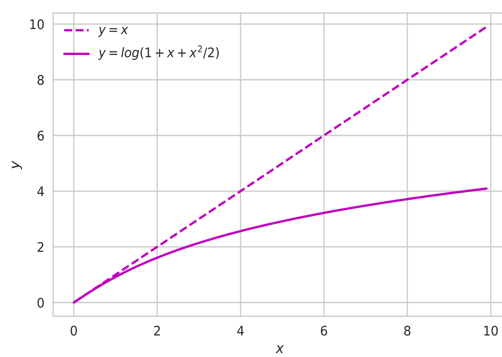



Figure 1: The illustration of the influence function for the soft estimator.

Here, we provide the graph of the used influence function for the soft estimator, which explains the mechanism of the function $y = \log(1 + x + x^2/2)$ more clearly. The illustration is presented in Figure 1. As can be seen, when $x$ is large and may be an outlier, the influence function can reduce its negative impact for mean estimation. Therefore, we exploit such an influence function for robust mean estimation, which brings better classification performance.

### A.2 Proof of Theorem 2

**Lemma 1** ([11]). *Let $Z_n = \{z_1, \ldots, z_n\}$ be a (not necessarily time homogeneous) Markov chain with mean $\mu_z$, taking values in a Polish state space $\Lambda_1 \times \ldots \times \Lambda_n$, with a mixing time $\tau(\upsilon)$ (for $0 \leq \upsilon \leq 1$). Let*

$$\tau_{\min} = \inf_{0 \leq \upsilon < 1} \tau(\upsilon) \cdot \left(\frac{2 - \upsilon}{1 - \upsilon}\right)^2. \tag{14}$$

*For some $\eta \in \mathbb{R}^+$, suppose that $f : \Lambda \to \mathbb{R}$ satisfies the following inequality:*

$$f(a) - f(b) \leq \sum_{i=1}^{n} \eta \mathbb{1}[a_i \neq b_i], \tag{15}$$

*for every $a, b \in \Lambda$. Then for any $\epsilon \geq 0$, we have*

$$P(|f(Z_n) - \mathbb{E}f(Z_n)| \geq \epsilon) \leq 2\exp\left(\frac{-2\epsilon^2}{\eta^2 \tau_{\min}}\right). \tag{16}$$

The detailed definition of the mixing time for the Markov chain can be found in [11, 12]. Let $f$ be the mean function. Following the prior work on mean estimation [7, 4, 3, 10], without loss of generality, we assume $\mu_z = 0$ for the underlying true distribution, and $|z_i|$ is upper bounded by $Z$. Then we can set $\eta$ to $4Z/n$ for Eq. (15). Combining the above analyses, we can revise Eq. (16) as follows:

$$P\left(\left|\frac{1}{n}\sum_{i=1}^{n} z_i\right| \geq \frac{2Z}{n}\sqrt{2\tau_{\min}\log\frac{2}{\epsilon_1}}\right) \leq \epsilon_1, \tag{17}$$

and

$$P\left(\max_{i\in[n]}|z_i| \geq \frac{2Z}{n}\sqrt{2\tau_{\min}\log\frac{2n}{\epsilon_2}}\right) \leq \epsilon_2, \tag{18}$$

for $\epsilon_1 > 0$ and $\epsilon_2 > 0$. If we remove the potential outliers $Z_{n_o}$ from $Z_n$. Therefore, we have

$$\begin{aligned}\left|\frac{1}{n-n_o}\sum_{z_i\in Z_n\setminus Z_{n_o}} - \mu_z\right| &= \frac{1}{n-n_o}\left|\sum_{z_i\in Z_n} - \sum_{z_i\in Z_{n_o}}\right| \\ &\leq \frac{1}{n-n_o}\left(\left|\sum_{z_i\in Z_n}\right| + \left|\sum_{z_i\in Z_{n_o}}\right|\right) \\ &\leq \frac{1}{n-n_o}\left(\left|\sum_{z_i\in Z_n}\right| + n_o\max_{i\in[n]}|z_i|\right) \\ &\leq \frac{1}{n-n_o}\left(2Z\sqrt{2\tau_{\min}\log\frac{2}{\epsilon_1}} + \frac{2Zn_o}{n}\sqrt{2\tau_{\min}\log\frac{2n}{\epsilon_2}}\right),\end{aligned} \tag{19}$$

which holds with probability at least $1 - \epsilon_1 - \epsilon_2$.

For our task, we exploit the concentration inequality. Let $\epsilon_1 = \epsilon_2 = \frac{1}{2t}$, and the losses be bounded by $L$. Next we can obtain

$$\begin{aligned}|\tilde{\mu}_h - \mu| &\leq \frac{2L}{t-t_o}\left(\sqrt{2\tau_{\min}\log(4t)} + \frac{t_o}{t}\sqrt{4\tau_{\min}\log(4t)}\right) \\ &= \frac{2\sqrt{2\tau_{\min}}L(t+\sqrt{2}t_o)}{(t-t_o)t}\sqrt{\log(4t)}\end{aligned} \tag{20}$$

with the probability at least $1 - \frac{1}{t}$. In practice, it is easy to identify the value of $L$. For example, we can training deep networks on noisy datasets to observe the loss distributions. Then, we exploit the lower bound such that

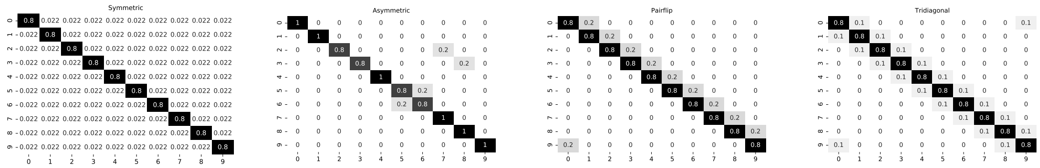
$$\ell_h^\star = \tilde{\mu}_h - \frac{2\sqrt{2\tau_{\min}}L(t+\sqrt{2}t_o)}{(t-t_o)\sqrt{t}}\sqrt{\frac{\log(4t)}{n_t}} \tag{21}$$

for sample selection.

# B Complementary Experimental Analyses

| | # of training | # of testing | # of class | size |
|---|---|---|---|---|
| *MNIST* | 60,000 | 10,000 | 10 | 28×28×1 |
| *F-MNIST* | 60,000 | 10,000 | 10 | 28×28×1 |
| *CIFAR-10* | 50,000 | 10,000 | 10 | 32×32×3 |
| *CIFAR-100* | 50,000 | 10,000 | 100 | 32×32×3 |

Table 1: Summary of synthetic datasets used in the experiments.




(a) (b) (c) (d)

Figure 2: Synthetic class-dependent transition matrices used in our experiments on *MNIST*. The noise rate is set to 20%.

### B.1 The Details of Datasets and Generating Noisy Labels

For the details of datasets, the important statistics of the used datasets are summarized in Table 1.

For the details of generating noisy labels, we exploit both *class-dependent* and *instance-dependent label noise* which include five types of synthetic label noise to verify the effectiveness of the proposed method. Here, we describe the details of the noise setting as follows:

(1). Class-dependent label noise:

• Symmetric noise: this kind of label noise is generated by flipping labels in each class uniformly to incorrect labels of other classes.

• Asymmetic noise : this kind of label noise is generated by flipping labels within a set of similar classes. In this paper, for *MNIST*, flipping 2→7, 3→8, 5↔6. For *F-MNIST*, flipping TSHIRT→SHIRT, PULLOVER→COAT, SANDALS→SNEAKER. For *CIFAR-10*, flipping TRUCK→AUTOMOBILE, BIRD→AIRPLANE, DEER→HORSE, CAT↔DOG. For *CIFAR-100*, 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.

• Pairflip noise: the noise flips 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, unlike the unidirectional pair flipping. It can be implemented by two consecutive pair flipping transformations in the opposite direction.

(2). Instance-dependent label noise:

• Instance noise: the noise is quite realistic, where the probability that an instance is mislabeled depends on its features. We generate this type of label noise to validate the effectiveness of the proposed method as did in [13].

We use synthetic noisy *MNIST* as an example and plot the noise transition matrices in Figure 2. The noise rate is set to 20%.

### B.2 Comparison with Other Types of Baselines

As we focus on the sample selection approach in learning with noisy labels, in the main paper (Section 3.1), we fairly compare our methods with the baselines which also focus on sample selection. Here, we evaluate other types of baselines. We exploit APL [8] and CDR [14], which add implicit regularization from different perspectives. The experiments are conducted on *MNIST* and *F-MNIST*. Other experimental settings are the same as those in the main paper. The experimental results are provided in Table 2 and 3, which show that the proposed methods can outperform them with respect to classification performance.

### B.3 Experiments on Synthetic *CIFAR-100*

For *CIFAR-100*, we use a 7-layer CNN structure from [17, 16]. Other experimental settings are the same as those in the experiments on *MNIST*, *F-MNIST*, and *CIFAR-10*. The results are provided in Table 4. We can see the proposed method outperforms all the baselines.

| Noise type | Sym. | | Asym. | | Pair. | | Trid. | | Ins. | |
|---|---|---|---|---|---|---|---|---|---|---|
| Method/Noise ratio | 20% | 40% | 20% | 40% | 20% | 40% | 20% | 40% | 20% | 40% |
| APL | 98.76 ±0.06 | 94.92 ±0.31 | 98.63 ±0.05 | 88.65 ±1.72 | 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 | 96.73 ±0.19 | 91.05 ±0.76 | 93.25 ±0.90 | 71.02 ±3.89 | 94.06 ±0.92 | 70.28 ±4.01 | 93.17 ±0.96 | 77.45 ±3.04 |

Table 2: Test accuracy (%) on *MNIST* over the last ten epochs.

| Noise type | Sym. | | Asym. | | Pair. | | Trid. | | Ins. | |
|---|---|---|---|---|---|---|---|---|---|---|
| Method/Noise ratio | 20% | 40% | 20% | 40% | 20% | 40% | 20% | 40% | 20% | 40% |
| APL | 91.73 ±0.20 | 89.06 ±0.41 | 90.13 ±0.17 | 80.34 ±0.63 | 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 | 89.78 ±0.41 | 79.05 ±1.39 | 85.72 ±0.65 | 69.07 ±2.31 | 86.75 ±1.19 | 73.63 ±2.82 | 85.92 ±1.43 | 73.14 ±3.12 |

Table 3: Test accuracy on *F-MNIST* over the last ten epochs.

| Noise type | Sym. | | Asym. | | Pair. | | Trid. | | Ins. | |
|---|---|---|---|---|---|---|---|---|---|---|
| Method/Noise ratio | 20% | 40% | 20% | 40% | 20% | 40% | 20% | 40% | 20% | 40% |
| S2E | 44.59 ±0.32 | 25.78 ±5.44 | 42.18 ±1.73 | 26.81 ±2.25 | 42.99 ±1.54 | 26.96 ±2.48 | 43.16 ±0.93 | 27.72 ±3.56 | 43.13 ±0.67 | 27.12 ±3.86 |
| MentorNet | 43.15 ±0.42 | 37.62 ±0.89 | 41.03 ±0.22 | 28.27 ±0.41 | 40.06 ±0.37 | 27.17 ±0.92 | 42.20 ±0.30 | 31.74 ±0.88 | 40.54 ±0.69 | 33.09 ±1.53 |
| Co-teaching | 45.17 ±0.25 | 40.95 ±0.52 | 42.76 ±0.34 | 30.27 ±0.33 | 42.50 ±0.39 | 30.07 ±0.17 | 44.41 ±0.41 | 34.96 ±0.35 | 42.23 ±0.52 | 35.87 ±1.47 |
| SIGUA | 42.03 ±0.33 | 40.53 ±0.49 | 36.67 ±0.25 | 26.71 ±0.42 | 36.48 ±0.37 | 26.73 ±0.33 | 39.21 ±0.40 | 32.69 ±0.36 | 39.19 ±0.32 | 33.51 ±0.43 |
| JoCor | 45.93 ±0.21 | 41.56 ±0.57 | 42.89 ±0.37 | 29.19 ±1.42 | 42.12 ±0.35 | 30.12 ±0.65 | 44.98 ±0.27 | 34.23 ±1.13 | 44.28 ±0.59 | 35.60 ±0.99 |
| CNLCU-S | **46.09 ±0.29** | **42.11 ±0.70** | **43.06 ±0.28** | **30.47 ±0.37** | **43.08 ±0.92** | **30.33 ±0.74** | **45.19 ±0.90** | **35.49 ±1.30** | **44.80 ±0.70** | **36.23 ±0.49** |
| CNLCU-H | **46.27 ±0.38** | **42.05 ±0.87** | **43.21 ±0.93** | **30.55 ±0.72** | **43.25 ±0.75** | **30.79 ±0.86** | **45.02 ±1.06** | **35.24 ±0.93** | **45.02 ±1.07** | **36.17 ±1.54** |

Table 4: Test accuracy (%) on *CIFAR-100* over the last ten epochs. The best two results are in bold.

### B.4 Experiments for Ablation Study

We conduct the ablation study to analyze the sensitivity of the length of time intervals. The results are shown in Figure. 3 and 4. As we can seen, the proposed method, i.e., CNLCU-S and CNLCU-H are robust to the choices of hyperparameters.

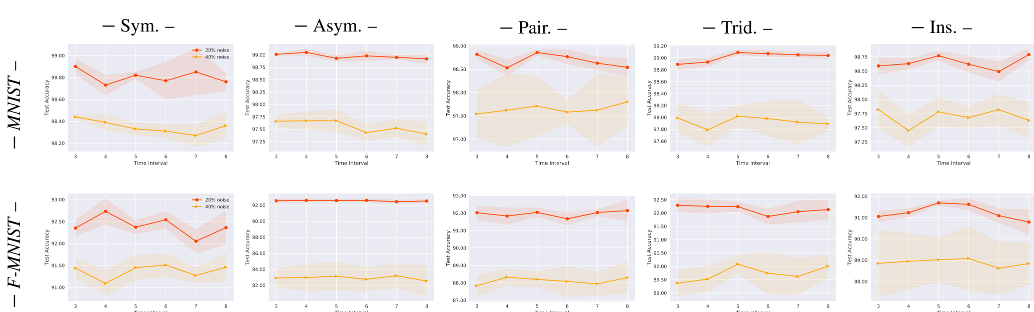



Figure 3: Illustrations of the hyperparameter sensitivity for the proposed CNLCU-S. The error bar for standard deviation in each figure has been shaded.

Note that in this paper, we concern uncertainty from *two aspects*, i.e., the uncertainty about small-loss examples and the uncertainty about large-loss examples. Here, we conduct ablation study to show the effect of removing different components to provide insights into what makes the proposed methods successful. The experiments are conducted on *MNIST* and *F-MNIST*. Other experimental settings

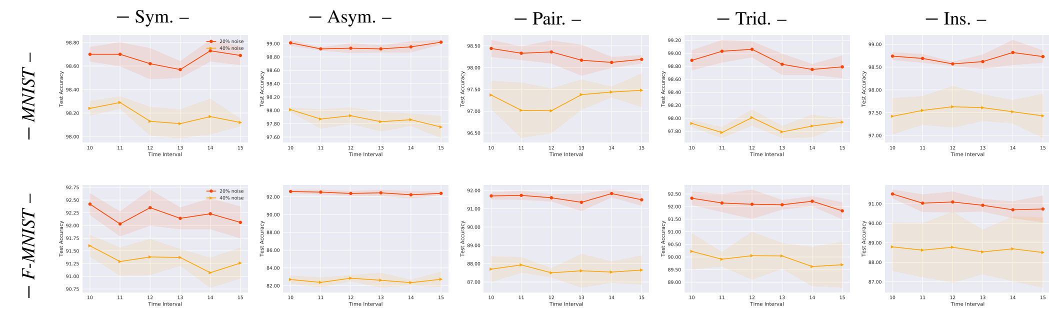



Figure 4: Illustrations of the hyperparameter sensitivity for the proposed CNLCU-H. The error bar for standard deviation in each figure has been shaded.

are the same as those in the main paper (Section 3.1). Note that we employ two networks to teach each other following [6]. Therefore, when we do not consider uncertainty in sample selection, the proposed methods will *reduce to* the baseline Co-teaching [6].

To study the effect of concerning uncertainty about small-loss examples, we remove the concerns about large-loss examples, i.e., the network is not encouraged to choose the less selected examples for updates. We express such a setting as "**with**o**u**t **c**oncerning about **l**arge-loss examples" (abbreviated as *w/o cl*). To study the effect of concerning uncertainty about large-loss examples, we remove the concerns about small-loss examples, i.e., we only exploit the predictions of the current network. We express such a setting as "**with**o**u**t **c**oncerning about **s**mall-loss examples" (abbreviated as *w/o cs*). Besides, we express the setting which directly uses non-robust mean as Co-teaching-M.

The experimental results of ablation study are provided in Table 5 and 6. As can be seen, both aspects of uncertainty concerns can improve the robustness of models. Therefore, combining two uncertainty concerns, we can better combat noisy labels. In addition, robust mean estimation is superior to the non-robust mean in learning with noisy labels.

| Noise type | Sym. | | Asym. | | Pair. | | Trid. | | Ins. | |
|---|---|---|---|---|---|---|---|---|---|---|
| Method/Noise ratio | 20% | 40% | 20% | 40% | 20% | 40% | 20% | 40% | 20% | 40% |
| CNLCU-S | 98.82 ±0.03 | 98.31 ±0.05 | 98.93 ±0.06 | 97.67 ±0.22 | 98.86 ±0.06 | 97.71 ±0.64 | 99.09 ±0.04 | 98.02 ±0.17 | 98.77 ±0.08 | 97.78 ±0.25 |
| CNLCU-S *w/o cl* | 98.02 ±0.08 | 96.83 ±0.29 | 98.50 ±0.04 | 96.25 ±0.13 | 98.22 ±0.13 | 96.08 ±0.75 | 98.64 ±0.31 | 97.25 ±0.24 | 98.17 ±0.20 | 97.13 ±0.40 |
| CNLCU-S *w/o cs* | 98.15 ±0.20 | 97.12 ±0.22 | 98.36 ±0.07 | 96.39 ±0.48 | 98.04 ±0.24 | 96.12 ±0.68 | 98.74 ±0.05 | 97.30 ±0.52 | 98.11 ±0.15 | 97.32 ±0.43 |
| CNLCU-H | 98.70 ±0.06 | 98.24 ±0.06 | 99.01 ±0.04 | 98.01 ±0.03 | 98.44 ±0.19 | 97.37 ±0.32 | 98.89 ±0.15 | 97.92 ±0.05 | 98.74 ±0.16 | 97.42 ±0.39 |
| CNLCU-H *w/o cl* | 98.06 ±0.13 | 96.92 ±0.23 | 98.39 ±0.04 | 96.51 ±0.57 | 97.04 ±0.87 | 95.62 ±0.93 | 98.33 ±0.47 | 97.41 ±0.92 | 98.01 ±0.20 | 96.15 ±0.28 |
| CNLCU-H *w/o cs* | 98.19 ±0.22 | 97.05 ±0.49 | 98.76 ±0.59 | 97.17 ±0.60 | 97.26 ±1.19 | 96.31 ±0.25 | 98.29 ±0.17 | 97.65 ±0.92 | 98.34 ±0.36 | 96.49 ±0.48 |
| Co-teaching-M | 97.72 ±0.08 | 97.78 ±0.32 | 98.27 ±0.03 | 95.42 ±0.42 | 96.22 ±0.10 | 95.01 ±0.65 | 97.92 ±0.14 | 96.64 ±0.77 | 98.02 ±0.04 | 96.03 ±0.57 |
| Co-teaching | 97.53 ±0.12 | 95.62 ±0.30 | 98.25 ±0.08 | 95.08 ±0.43 | 96.05 ±0.96 | 94.16 ±1.37 | 98.05 ±0.06 | 96.18 ±0.85 | 97.96 ±0.09 | 95.02 ±0.39 |

Table 5: Test accuracy (%) on *MNIST* over last ten epochs.

# C Complementary Explanation for Network Structures

Table 7 describes the 9-layer CNN [6] used on *MNIST*, *F-MNIST*, and *CIFAR-10*. Table 8 describes the 9-layer CNN [17] used on *CIFAR-100*. Here, LReLU stands for Leaky ReLU [15]. The slopes of all LReLU functions in the networks are set to 0.01. Note that that the 7/9-layer CNN is a standard and common practice in weakly supervised learning. We decided to use these CNNs, since then the experimental results are directly comparable with previous approaches in the same area, i.e., learning with noisy labels.

| Noise type | Sym. | | Asym. | | Pair. | | Trid. | | Ins. | |
|---|---|---|---|---|---|---|---|---|---|---|
| Method/Noise ratio | 20% | 40% | 20% | 40% | 20% | 40% | 20% | 40% | 20% | 40% |
| CNLCU-S | 92.37 ±0.15 | 91.45 ±0.28 | 92.57 ±0.15 | 83.14 ±1.77 | 92.04 ±0.26 | 88.20 ±0.44 | 92.24 ±0.17 | 90.08 ±0.34 | 91.69 ±0.10 | 89.02 ±1.02 |
| CNLCU-S *w/o cl* | 91.77 ±0.35 | 89.40 ±0.26 | 91.25 ±0.30 | 72.93 ±2.63 | 91.53 ±0.17 | 87.31 ±0.59 | 91.31 ±0.52 | 89.50 ±0.32 | 91.09 ±0.13 | 88.45 ±0.57 |
| CNLCU-S *w/o cs* | 91.85 ±0.33 | 90.76 ±0.28 | 91.94 ±0.09 | 80.99 ±2.74 | 91.28 ±0.20 | 87.31 ±0.72 | 91.39 ±0.07 | 89.29 ±0.51 | 90.98 ±0.43 | 88.73 ±0.62 |
| CNLCU-H | 92.42 ±0.21 | 91.60 ±0.19 | 92.60 ±0.18 | 82.69 ±0.43 | 91.70 ±0.18 | 87.70 ±0.69 | 92.33 ±0.26 | 90.22 ±0.71 | 91.50 ±0.21 | 88.79 ±1.22 |
| CNLCU-H *w/o cl* | 91.70 ±0.04 | 90.05 ±0.31 | 91.08 ±0.06 | 71.35 ±2.30 | 91.03 ±0.29 | 87.22 ±0.72 | 91.59 ±0.07 | 90.01 ±0.24 | 90.80 ±0.27 | 88.31 ±1.09 |
| CNLCU-H *w/o cs* | 91.82 ±0.13 | 90.92 ±0.42 | 92.45 ±0.25 | 80.73 ±1.63 | 91.21 ±0.17 | 87.49 ±0.32 | 92.08 ±0.13 | 89.72 ±0.24 | 91.21 ±0.38 | 88.62 ±0.73 |
| Co-teaching-M | 91.33 ±0.18 | 89.05 ±0.73 | 91.14 ±0.90 | 71.03 ±3.73 | 90.85 ±0.61 | 86.95 ±0.19 | 91.50 ±0.46 | 89.18 ±0.44 | 90.74 ±1.06 | 88.25 ±0.92 |
| Co-teaching | 91.48 ±0.10 | 88.80 ±0.29 | 91.03 ±0.14 | 68.07 ±4.58 | 90.77 ±0.23 | 86.91 ±0.71 | 91.24 ±0.11 | 89.18 ±0.36 | 90.60 ±0.12 | 87.90 ±0.45 |

Table 6: Test accuracy (%) on *F-MNIST* over last ten epochs.

Table 7: CNN on *MNIST*, *F-MNIST*, and *CIFAR-10*.

| CNN on *MNIST* | CNN on *F-MNIST* | CNN on *CIFAR-10* |
|---|---|---|
| 28×28 Gray Image | 28×28 Gray Image | 32×32 RGB Image |
| 3×3 conv, 128 LReLU<br>3×3 conv, 128 LReLU<br>3×3 conv, 128 LReLU | | |
| 2×2 max-pool<br>dropout, $p = 0.25$ | | |
| 3×3 conv, 256 LReLU<br>3×3 conv, 256 LReLU<br>3×3 conv, 256 LReLU | | |
| 2×2 max-pool<br>dropout, $p = 0.25$ | | |
| 3×3 conv, 512 LReLU<br>3×3 conv, 256 LReLU<br>3×3 conv, 128 LReLU | | |
| avg-pool | | |
| dense 128→10 | dense 128→10 | dense 128→10 |

Table 8: CNN on *CIFAR-100*.

| CNN on *CIFAR-100* |
|---|
| 32×32 RGB Image |
| 3×3 conv, 64 ReLU<br>3×3 conv, 64 ReLU |
| 2×2 max-pool |
| 3×3 conv, 128 ReLU<br>3×3 conv, 128 ReLU |
| 2×2 max-pool |
| 3×3 conv, 196 ReLU<br>3×3 conv, 196 ReLU |
| 2×2 max-pool |
| dense 256→100 |

# References


[1] Olivier Catoni. Challenging the empirical mean and empirical variance: a deviation study. In *Annales de l'IHP Probabilités et statistiques*, volume 48, pages 1148–1185, 2012.

[2] Peng Chen, Xinghu Jin, Xiang Li, and Lihu Xu. A generalized catoni's $m$-estimator under finite $\alpha$-th moment assumption with $\alpha \in (1, 2)$. *arXiv preprint arXiv:2010.05008*, 2020.

[3] Ilias Diakonikolas and Daniel M Kane. Recent advances in algorithmic high-dimensional robust statistics. *arXiv preprint arXiv:1911.05911*, 2019.

[4] Ilias Diakonikolas, Daniel M Kane, and Ankit Pensia. Outlier robust mean estimation with subgaussian rates via stability. *arXiv preprint arXiv:2007.15618*, 2020.

[5] Evarist Giné, Rafał Latała, and Joel Zinn. Exponential and moment inequalities for u-statistics. In *High Dimensional Probability II*, pages 13–38. Springer, 2000.

[6] 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.

[7] Liu Liu, Tianyang Li, and Constantine Caramanis. High dimensional robust m-estimation: Arbitrary corruption and heavy tails. *arXiv preprint arXiv:1901.08237*, 2019.

[8] Xingjun Ma, Hanxun Huang, Yisen Wang, Simone Romano, Sarah Erfani, and James Bailey. Normalized loss functions for deep learning with noisy labels. In *ICML*, pages 6543–6553, 2020.

[9] Mehryar Mohri, Afshin Rostamizadeh, and Ameet Talwalkar. *Foundations of Machine Learning*. MIT Press, 2018.

[10] Laura Niss and Ambuj Tewari. What you see may not be what you get: Ucb bandit algorithms robust to $\epsilon$-contamination. In *UAI*, pages 450–459, 2020.

[11] Daniel Paulin et al. Concentration inequalities for markov chains by marton couplings and spectral methods. *Electronic Journal of Probability*, 20, 2015.

[12] Gareth O Roberts, Jeffrey S Rosenthal, et al. General state space markov chains and mcmc algorithms. *Probability surveys*, 1:20–71, 2004.

[13] 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.

[14] Xiaobo Xia, Tongliang Liu, Bo Han, Chen Gong, Nannan Wang, Zongyuan Ge, and Yi Chang. Robust early-learning: Hindering the memorization of noisy labels. In *ICLR*, 2021.

[15] Bing Xu, Naiyan Wang, Tianqi Chen, and Mu Li. Empirical evaluation of rectified activations in convolutional network. *arXiv preprint arXiv:1505.00853*, 2015.

[16] Quanming Yao, Hansi Yang, Bo Han, Gang Niu, and James Tin-Yau Kwok. Searching to exploit memorization effect in learning with noisy labels. In *ICML*, pages 10789–10798, 2020.

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