Learning with Noisy Labels via Sparse Regularization: Supplementary Materials

A. Proof of Theorems

Theorem 1. *In a multi-class classification problem, $\forall L \in \mathcal{L}$, L is noise-tolerant under symmetric label noise if $\eta < 1 - \frac{1}{k}$ and $f : \mathcal{X} \to \mathcal{P}_{\mathbf{v}}$, where $\mathbf{v}$ is a fixed vector, i.e.,*

$$\arg\min_{f:\mathcal{X}\to\mathcal{P}_{\mathbf{v}}} R_L(f) = \arg\min_{f:\mathcal{X}\to\mathcal{P}_{\mathbf{v}}} R_L^{\eta}(f). \tag{1}$$

Proof. For symmetric label noise, we have

$$\begin{aligned} R_L^{\eta}(f) &= \mathbb{E}_{\mathbf{x},y}\left[(1-\eta)L(f(\mathbf{x}),y) + \frac{\eta}{k-1}\sum_{i\neq y} L(f(\mathbf{x}),i)\right] \\ &= \mathbb{E}_{\mathbf{x},y}\left[(1-\eta)L(f(\mathbf{x}),y) + \frac{\eta}{k-1}(\sum_{i=1}^{k} L(f(\mathbf{x}),i) - L(f(\mathbf{x}),y)\right] \\ &= (1-\eta)R_L(f) + \frac{\eta}{k-1}(\sum_{i=1}^{k} L(\mathbf{v},i) - R_L(f)) \\ &= (1-\frac{\eta k}{k-1})R_L(f) + \frac{\eta}{k-1}\sum_{i=1}^{k} L(\mathbf{v},i) \end{aligned}$$

since $1 - \frac{\eta k}{k-1} > 0$ and $\sum_{i=1}^{k} L(\mathbf{v},i)$ is a constant, then f^* minimizes $R^{\eta}(f)$ if and only if f^* minimizes $R(f)$. □

Theorem 2. *In a multi-class classification problem, we let $f : \mathcal{X} \to \mathcal{P}_{\mathbf{v}}$, where $\mathbf{v}$ is a fixed vector. If $R_L(f^*) = 0$, $\forall f : \mathcal{X} \to \mathcal{P}_{\mathbf{v}}$, $\forall L \in \mathcal{L}$ and $0 \leq L \leq \frac{C}{k-1}$, L is noise-tolerant under asymmetric or class-conditional noise when $\eta_{\mathbf{y},i} < 1 - \eta_y$ with $\sum_{k\neq y} \eta_{y,i} = \eta_y$, $\forall \mathbf{x}$.*

Proof. For asymmetric or class-conditional noise, we have

$$\begin{aligned} R_L^{\eta}(f) &= \mathbb{E}_{\mathbf{x},y}(1-\eta_y)L(f(\mathbf{x}),y) + \mathbb{E}_{\mathbf{x},y}\sum_{i\neq y}\eta_{y,i}L(f(\mathbf{x}),i) \\ &= \mathbb{E}_{\mathbf{x},y}(1-\eta_y)(C - \sum_{i\neq y} L(f(\mathbf{x}),i)) + \mathbb{E}_{\mathbf{x},y}\sum_{i\neq y}\eta_{y,i}L(f(\mathbf{x}),i) \\ &= C\mathbb{E}_{\mathbf{x},y}(1-\eta_y) - \mathbb{E}_{\mathbf{x},y}\sum_{i\neq y}(1-\eta_y-\eta_{y,i})L(f(\mathbf{x}),i) \end{aligned} \tag{2}$$

Let f_{η}^* and f^* be the minimizer of $R_L^{\eta}(f)$ and $R_L(f)$ when $f : \mathcal{X} \to \mathcal{P}_{\mathbf{v}}$, respectively. We have $R_L^{\eta}(f_{\eta}^*) - R_L^{\eta}(f^*) \leq 0$ and hence derive that

$$\mathbb{E}_{\mathbf{x},y}\sum_{i\neq y}(1-\eta_y-\eta_{y,i})(L(f(\mathbf{x}),i) - L(f(\mathbf{x}),i)) \leq 0 \tag{3}$$

Since we are given $R_L(f^*) = 0$, we have $L(f^*(\mathbf{x}),y) = 0$. Given the condition on L in the theorem, this implies $L(f^*(\mathbf{x}),i) = C/(k-1)$, $i \neq y$. As per the assumption on noise in the theorem, $1 - \eta_y - \eta_{y,i} > 0$. Also, L has to

satisft $L(f_\eta^*(\mathbf{x}), i) \leq C/(k-1), \forall i$. Thus for Eq. 3 to hold, it must be the case that $L(f_\eta^*(\mathbf{x}), i) = C/(k-1)$, which implies $L(f_\eta^*(\mathbf{x}), y) = 0$. Thus, the minimizer of true risk is also a minimizer of risk under noisy case. □

Theorem 3. *In a multi-class classification problem, if the loss function L satisfies* $|\sum_{i=1}^{k}(L(\mathbf{u}_1, i) - L(\mathbf{u}_2, i))| \leq \delta$ *when* $\|\mathbf{u}_1 - \mathbf{u}_2\|_2 \leq \varepsilon$*, and* $\delta \to 0$ *as* $\varepsilon \to 0$*, then for symmetric label noise satisfying* $\eta < 1 - \frac{1}{k}$*, the risk bound can be expressed as*

$$R_L(f_\eta^*) - R_L(f^*) \leq 2c\delta,$$

where $c = \frac{\eta}{(1-\eta)k-1}$*,* f_η^* *and* f^* *denote the minimizer of* $R_L^\eta(f)$ *and* $R_L(f)$ *when* $f \in \mathcal{H}_{\mathbf{v},\varepsilon}$*, respectively.*

Proof. For symmetric label noise, we have

$$\begin{aligned} R_L^\eta(f^*) &= \mathbb{E}_{\mathbf{x},y}\big[(1-\eta)L(f^*(\mathbf{x}), y) + \frac{\eta}{k-1}\sum_{i\neq y} L(f^*(\mathbf{x}), i)\big] \\ &= (1 - \frac{\eta k}{k-1})R_L(f^*) + \frac{\eta}{k-1}\mathbb{E}_{\mathbf{x},y}\left[\sum_{i=1}^{k} L(f^*(\mathbf{x}), i)\right] \\ &= (1 - \frac{\eta k}{k-1})R_L(f^*) + \frac{\eta}{k-1}\sum_{i=1}^{k} L(\mathbf{v}, i) + \frac{\eta}{k-1}\delta_1 \end{aligned} \tag{4}$$

where $\delta_1 = \mathbb{E}_{\mathbf{x},y}[\sum_{i=1}^{k} L(f(\mathbf{x}), i) - \sum_{i=1}^{k} L(\mathbf{v}, i)]$. On the other hand, $f^* \in \mathcal{H}_{\mathbf{v},\varepsilon}$, *i.e.*, $\|f^*(\mathbf{x}) - \mathbf{v}\|_2 \leq \varepsilon$, so we have $|\sum_{i=1}^{k} L(f(\mathbf{x}), i) - \sum_{i=1}^{k} L(\mathbf{v}, i)| \leq \delta$. This means that $\delta_1 \in [-\delta, \delta]$. Similarly, we can obtain

$$R_L^\eta(f_\eta^*) = (1 - \frac{\eta k}{k-1})R_L(f_\eta^*) + \frac{\eta}{k-1}\sum_{i=1}^{k} L(\mathbf{v}, i) + \frac{\eta}{k-1}\delta_2 \tag{5}$$

Since $f_\eta^* = \arg\min_{f\in\mathcal{H}_{\mathbf{v},\varepsilon}} R_L^\eta(f)$, and $f^* = \arg\min_{f\in\mathcal{H}_{\mathbf{v},\varepsilon}} R_L(f)$, we have

$$\begin{aligned} 0 &\geq R_L^\eta(f_\eta^*) - R_L^\eta(f^*) \\ &= (1 - \frac{\eta k}{k-1})(R_L(f_\eta^*) - R_L(f^*)) + \frac{\eta}{k-1}(\delta_2 - \delta_1) \\ \Rightarrow R_L(f_\eta^*) - R_L(f^*) &\leq \frac{\eta}{(1-\eta)k-1}(\delta_1 - \delta_2) \leq \frac{2\eta\delta}{(1-\eta)k-1} \end{aligned} \tag{6}$$

where we have used the fact that $1 - \frac{\eta k}{k-1} > 0$, and $\delta_2 - \delta_1 \leq 2\delta$ holds for $\delta_1, \delta_2 \in [-\delta, \delta]$. □

B. Experiments

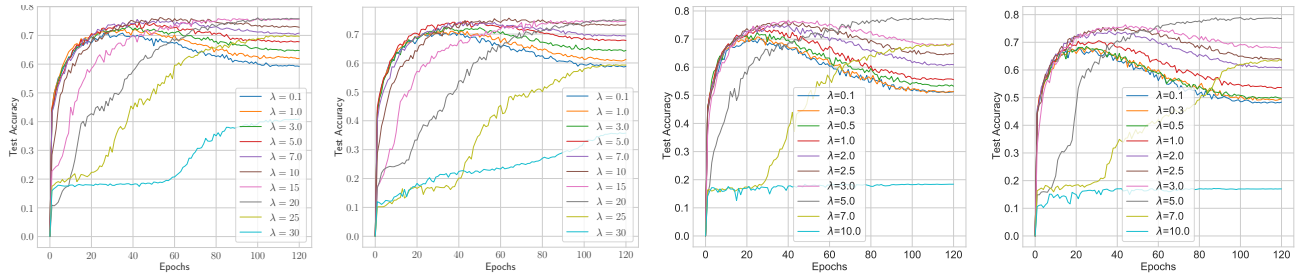
In this section, we provide the experimental details.

Datasets. We verify the effectiveness of our method on benchmark datasets, including MNIST [3], CIFAR-10/-100 [2] with synthetic label noise.

Since MNIST, CIFAR-10 and CIFAR-100 are clean, following previous works [7, 5], we experiment with two types of label noise: symmetric (uniform) noise and asymmetric (class-conditional) noise. For symmetric noise, we corrupt the training labels by flipping the labels in each class randomly to incorrect labels in other classes with flip probability $\eta \in \{0.2, 0.4, 0.6, 0.8\}$. For asymmetric noise, we flip the labels within a specific set of classes, for example, for MNIST, flipping $2 \to 7$, $7 \to 1$, $5 \leftrightarrow 6$, and $3 \to 8$; for CIFAR-10, flipping TRUCK $\to$ AUTOMOBILE, BIRD $\to$ AIRPLANE, DEER $\to$ HORSE, and CAT $\leftrightarrow$ DOG; for CIFAR-100, the 100 classes are grouped into 20 super-classes with each has 5 sub-classes, and each class are flipped within the same super-classes into the next.

Baselines. We experiment with the following state-of-the art methods, and two effective loss functions CE and Focal Loss (FL) [4] for classification. Moreover, we add the proposed sparse regularization mechanism to CE, FL and GCE, i.e., CE+SR, FL+SR and GCE+SR. All the implementations and experiments are based on PyTorch.

- GCE [8]. The Generalized Cross Entropy (GCE) is defined as $L_{GCE}(\mathbf{u}, i) = (1 - u_i^q)/q$ $(0 < q \leq 1)$.



(a) CE+SR ($\tau = 1.0$) (b) FL+SR ($\tau = 1.0$) (c) CE+SR ($\tau = 0.5$) (d) FL+SR ($\tau = 0.5$)

Figure 1. Test accuracy curve of different λ on CIFAR-10 with 0.6 symmetric label noise.

- SCE [7]. The Symmetric Cross Entropy (SCE) can be regarded as a weighted loss of CE and RCE (scaled MAE): $L_{SCE}(\mathbf{u}, i) = \alpha L_{CE}(\mathbf{u}, i) + \beta L_{RCE}(\mathbf{u}, i)$.
- NLNL [1]. NLNL improves robustness with a complementary label.
- APL [5]. The Active Passive Loss (APL) was proposed to combine a robust active loss and a robust passive loss, i.e., $L_{APL} = \alpha L_{Active} + \beta L_{Passive}$.

Network Structure and Training Details. Following the setting in [5], we use a 4-layer CNN for MNIST, an 8-layer CNN for CIFAR-10 and a ResNet-34 for CIFAR-100. The networks are trained for 50, 120, 200 epochs for MNIST, CIFAR-10, CIFAR-100, respectively. For all the training, we use SGD optimizer with momentum 0.9 and cosine learning rate annealing. Weight decay is set as 1×10^{-3}, 1×10^{-4}, 1×10^{-5} for MNIST, CIFAR-10, CIFAR-100, respectively. The initial learning rate is set to 0.01 for MNIST/CIFAR-10 and 0.1 for CIFAR-100. Batch size is set to 128. Typical data augmentations including random width/height shift and horizontal flip are applied.

Parameter Settings. We set the parameters which match their original papers for all baseline methods. Specifically, for FL, we set $\gamma = 0.3$. For GCE, we set $q = 0.7$. For SCE, we set $A = -3$, and $\alpha = 0.01$, $\beta = 1$ for MNIST, $\alpha = 0.1$, $\beta = 1$ for CIFAR-10, $\alpha = 6$, $\beta = 0.1$ for CIFAR-100. For APL (NCE+MAE), we set $\alpha = 1$, $\beta = 100$ for MNIST, $\alpha, \beta = 1$ for CIFAR-10, and $\alpha = 10$, $\beta = 0.1$ for CIFAR-100. For our sparse regularization, we set $(\tau, p, \lambda_0, \rho, r) = (0.1, 0.1, 4, 2, 5)$ for MNIST, $(0.5, 0.1, 1.1, 1.03, 1)$ for CIFAR-10, and $(0.5, 0.01, \cdot, 1.02, 1)$ for CIFAR-100. Otherwise, on CIFAR-100, we set λ_0 to 4 and 10 for symmetric and asymmetric label noise, respectively.

As for the parameter settings for Webvision, we use the suggested $q = 0.7$ for GCE, $A = -4$, $\alpha = 10$, $\beta = 1$ for SCE, while for APL, we set $\alpha = 50$, $\eta = 0.1$. For our CE+SR and FL+SR, we set $\tau = 0.5$, $p = 0.01$, $\lambda_0 = 2$, $\rho = 1.02$ and $f = 1$.

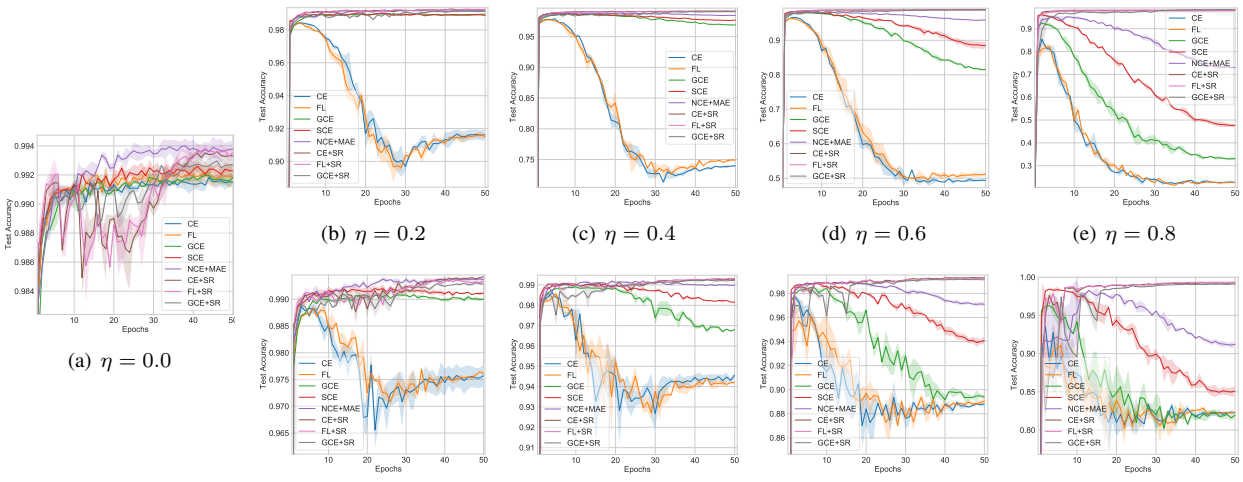
More experiments about hyperparameter selection. We offer more experimental results on selecting different λ on CIFAR-10 with 0.6 symmetric label noise. We adjust τ from 1.0 to 0.5. The results are shown in Fig. 1. We found that the output sharpening can benefit the sparse regularization. We can achieve the similar robustness result of $\lambda = 20$ ($\tau = 1$) by setting $\lambda = 5$ and $\tau = 0.5$, which demonstrates that the output sharpening also plays the role of sparse regularization when using ℓ_p-norm. Moreover, smaller λ can help maintain the fitting ability of the model with classification loss $L(f(\mathbf{x}), y)$ (*i.e.*, learning efficiently while keeping robustness). As a evidence, the eventual accuracy ($\tau = 0.5$, $\lambda = 5$) is higher than the experiments with $\tau = 1.0$.

More results of Comparison study. Fig. 2 shows test accuracy vs. epochs on MNIST. As can be observed, the commonly-used loss functions CE and FL suffer from significant overfitting in all noisy cases. The state-of-the-art methods GCE, SCE and APL show non-trival effectiveness of mitigating label noise, but the effects are crippled when meeting hard label noise. On the contrary, our proposed SR-enhanced methods CE+SR, FL+SR and GCE+SR perform better robustness and more efficiency . Fig. 3 shows test accuracy vs. epochs on CIFAR-10. The results are similar to MNIST, our SR-enhanced methods keep robust and achieve the best accuracy in most cases. Fig. 4 shows test accuracy vs. epochs on CIFAR-100. Our methods are of better fitting ability than commonly-used losses in the clean case, while the state-of-the-art GCE, SCE and APL encounter a little underfitting. For 0.2 and 0.4 symmetric label noise, our methods perform the best test accuracy. Interestingly, for all asymmetric label noise, our methods perform overfitting at the beginning, but they later mitigate label noise and outperform other methods.

More results of visualizations. More visualizations of representations on different datasets are shown in Fig. 5, 6 and 7. As can be seen, the representations learned by the proposed sparse regularization (SR)-enhanced methods are more discriminative than those learned by original losses, which are with more separated and clearly bound margins.

Table 1. Test accuracies (%) of different methods on benchmark datasets with clean or asymmetric label noise ($\eta \in [0.1, 0.2, 0.3, 0.4]$). The results (mean±std) are reported over 3 random runs and the top 3 best results are **boldfaced**.

Datasets	Methods	Asymmetric Noise Rate (η) 0.1	0.2	0.3	0.4
MNIST	CE	97.57 ± 0.22	94.56 ± 0.22	88.81 ± 0.10	82.27 ± 0.40
	FL	97.58 ± 0.09	94.25 ± 0.15	89.09 ± 0.25	82.13 ± 0.49
	GCE	99.01 ± 0.04	96.69 ± 0.12	89.12 ± 0.24	81.51 ± 0.19
	SCE	99.14 ± 0.04	98.03 ± 0.05	93.68 ± 0.43	85.36 ± 0.17
	NLNL	98.63 ± 0.06	98.35 ± 0.01	97.51 ± 0.15	95.84 ± 0.26
	APL	**99.32 ± 0.09**	98.89 ± 0.04	96.93 ± 0.17	91.45 ± 0.40
	CE+SR	**99.42 ± 0.02**	**99.27 ± 0.06**	**99.24 ± 0.08**	**99.23 ± 0.07**
	FL+SR	**99.34 ± 0.05**	**99.31 ± 0.02**	**99.23 ± 0.02**	**99.36 ± 0.05**
	GCE+SR	99.28 ± 0.06	**99.22 ± 0.02**	**99.13 ± 0.05**	**99.09 ± 0.02**
CIFAR-10	CE	87.55 ± 0.14	83.32 ± 0.12	79.32 ± 0.59	74.67 ± 0.38
	FL	86.43 ± 0.30	83.37 ± 0.07	79.33 ± 0.08	74.28 ± 0.44
	GCE	88.33 ± 0.05	85.93 ± 0.23	80.88 ± 0.38	74.29 ± 0.43
	SCE	**89.77 ± 0.11**	86.20 ± 0.37	81.38 ± 0.35	75.16 ± 0.39
	NLNL	88.54 ± 0.25	84.74 ± 0.08	81.26± 0.43	76.97 ± 0.52
	APL	88.31 ± 0.20	86.50 ± 0.31	83.34 ± 0.39	77.14 ± 0.33
	CE+SR	**89.08 ± 0.08**	**87.70 ± 0.19**	**85.63 ± 0.07**	**79.29 ± 0.20**
	FL+SR	88.68 ± 0.23	**87.56 ± 0.29**	**85.10 ± 0.23**	**79.07 ± 0.50**
	GCE+SR	**89.20 ± 0.23**	**87.55 ± 0.08**	**84.69 ± 0.46**	**79.01 ± 0.18**
CIFAR-100	CE	64.85 ± 0.37	58.11 ± 0.32	50.68 ± 0.55	40.17 ± 1.31
	FL	64.78 ± 0.50	58.05 ± 0.42	51.15 ± 0.84	41.18 ± 0.68
	GCE	63.01 ± 1.01	59.35 ± 1.10	53.83 ± 0.64	40.91 ± 0.57
	NLNL	59.55 ± 1.22	50.19 ± 0.56	42.81 ± 1.13	35.10 ± 0.20
	SCE	64.26 ± 0.43	58.16 ± 0.73	50.98 ± 0.33	41.54 ± 0.52
	APL	66.48 ± 0.12	62.80 ± 0.05	56.74 ± 0.53	42.61 ± 0.24
	CE+SR	**68.96 ± 0.22**	**64.79 ± 0.01**	**59.09 ± 2.10**	**49.51 ± 0.59**
	FL+SR	**68.96 ± 0.17**	**64.61 ± 0.67**	**58.94 ± 0.33**	**46.94 ± 1.68**
	GCE+SR	**69.27 ± 0.31**	**64.35 ± 0.78**	**57.22 ± 0.80**	**49.51 ± 1.31**



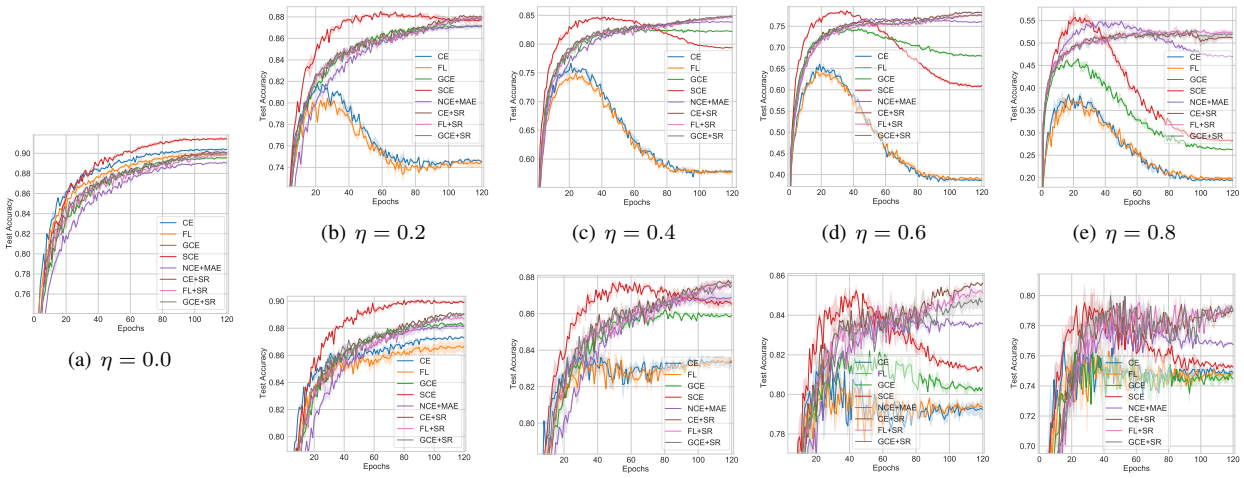
(a) $\eta = 0.0$ (b) $\eta = 0.2$ (c) $\eta = 0.4$ (d) $\eta = 0.6$ (e) $\eta = 0.8$

(f) $\eta = 0.1$ (g) $\eta = 0.2$ (h) $\eta = 0.3$ (i) $\eta = 0.4$

Figure 2. Test accuracies of different methods on MNIST with different label noise, where (a) denotes the clean case, (b-e) denote the symmetric label noise, and (f-i) denote the asymmetric label noise.

References

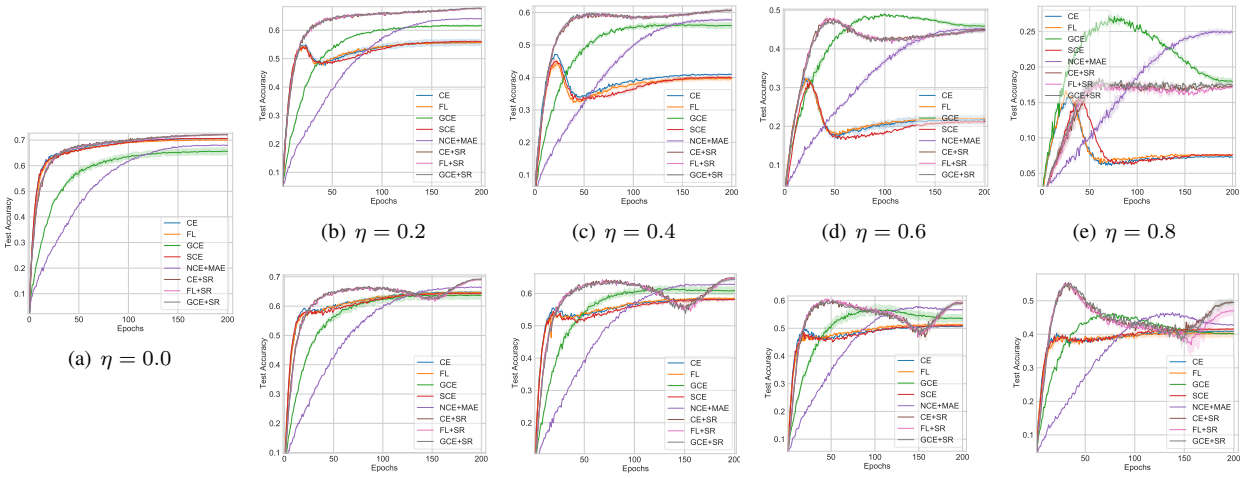
[1] Youngdong Kim, Junho Yim, Juseung Yun, and Junmo Kim. Nlnl: Negative learning for noisy labels. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 101–110, 2019.



(a) $\eta = 0.0$ (b) $\eta = 0.2$ (c) $\eta = 0.4$ (d) $\eta = 0.6$ (e) $\eta = 0.8$

(f) $\eta = 0.1$ (g) $\eta = 0.2$ (h) $\eta = 0.3$ (i) $\eta = 0.4$

Figure 3. Test accuracies of different methods on CIFAR-10 with different label noise, where (a) denotes the clean case, (b-e) denote the symmetric label noise, and (f-i) denote the asymmetric label noise.



(a) $\eta = 0.0$ (b) $\eta = 0.2$ (c) $\eta = 0.4$ (d) $\eta = 0.6$ (e) $\eta = 0.8$

(f) $\eta = 0.1$ (g) $\eta = 0.2$ (h) $\eta = 0.3$ (i) $\eta = 0.4$

Figure 4. Test accuracies of different methods on CIFAR-100 with different label noise, where (a) denotes the clean case, (b-e) denote the symmetric label noise, and (f-i) denote the asymmetric label noise.

[2] A. Krizhevsky and G. Hinton. Learning multiple layers of features from tiny images. *Computer Science Department, University of Toronto, Tech. Rep*, 1, 01 2009.

[3] Y. Lecun, L. Bottou, Y. Bengio, and P. Haffner. Gradient-based learning applied to document recognition. *Proceedings of the IEEE*, 86(11):2278–2324, 1998.

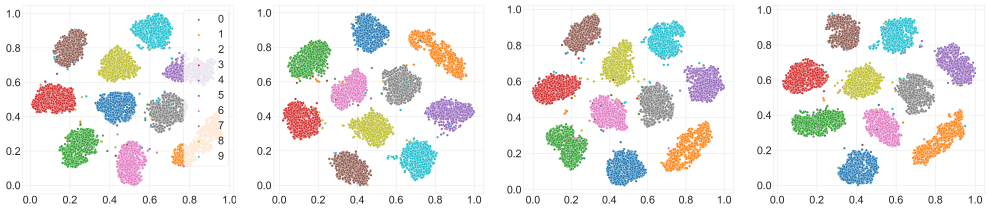
[4] Tsung-Yi Lin, Priya Goyal, Ross Girshick, Kaiming He, and Piotr Dollár. Focal loss for dense object detection. In *Proceedings of the IEEE international conference on computer vision*, pages 2980–2988, 2017.

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

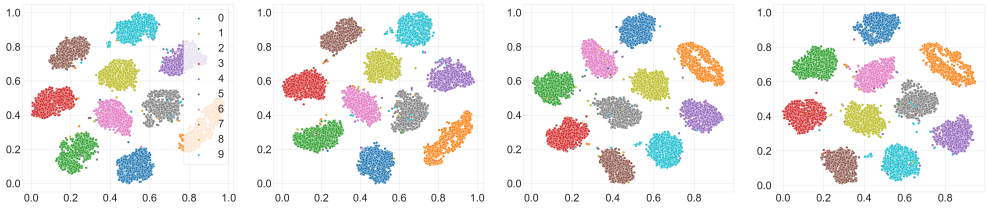
[6] Laurens Van der Maaten and Geoffrey Hinton. Visualizing data using t-sne. *Journal of machine learning research*, 9(11), 2008.

[7] Yisen Wang, Xingjun Ma, Zaiyi Chen, Yuan Luo, Jinfeng Yi, and James Bailey. Symmetric cross entropy for robust learning with noisy labels. In *IEEE International Conference on Computer Vision*, 2019.

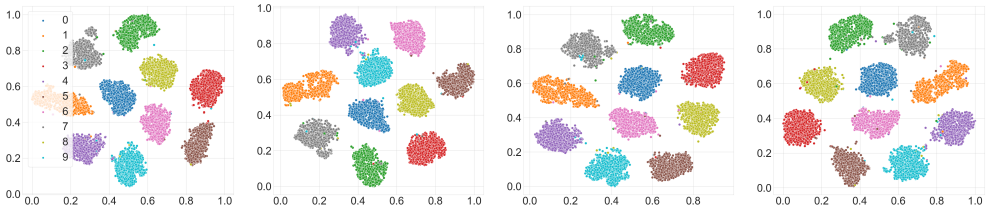
[8] Zhilu Zhang and Mert R Sabuncu. Generalized cross entropy loss for training deep neural networks with noisy labels. In *NeurIPS*, 2018.



(a) CE+SR with $\eta = 0.1$ (b) CE+SR with $\eta = 0.2$ (c) CE+SR with $\eta = 0.3$ (d) CE+SR with $\eta = 0.4$

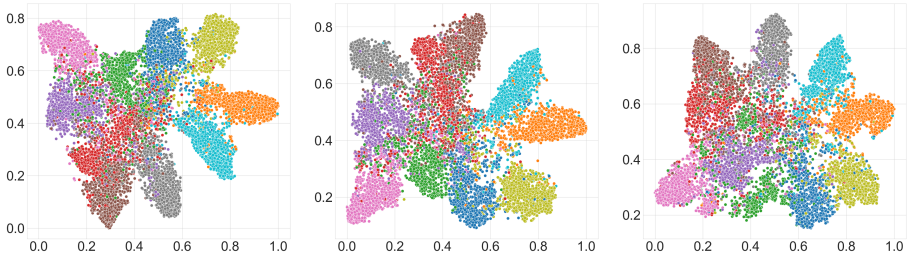


(e) FL+SR with $\eta = 0.1$ (f) FL+SR with $\eta = 0.2$ (g) FL+SR with $\eta = 0.3$ (h) FL+SR with $\eta = 0.4$

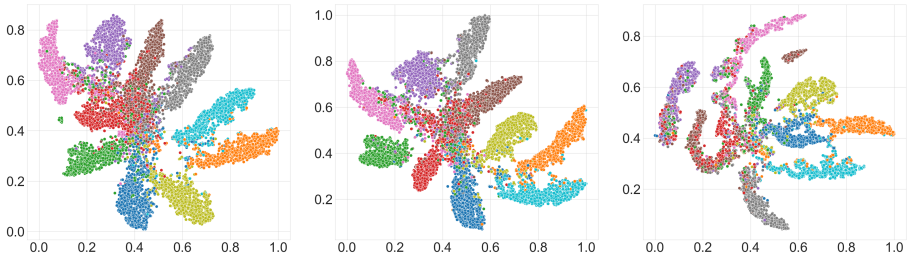


(i) GCE+SR with $\eta = 0.1$ (j) GCE+SR with $\eta = 0.2$ (k) GCE+SR with $\eta = 0.3$ (l) GCE+SR with $\eta = 0.4$

Figure 5. Features visualization for CE+SR (top) and FL+SR (bottom) on MNIST with different asymmetric label noise ($\eta \in [0.1, 0.2, 0.3, 0.4]$) by t-SNE [6] 2D embeddings at the last second full-connected layer.

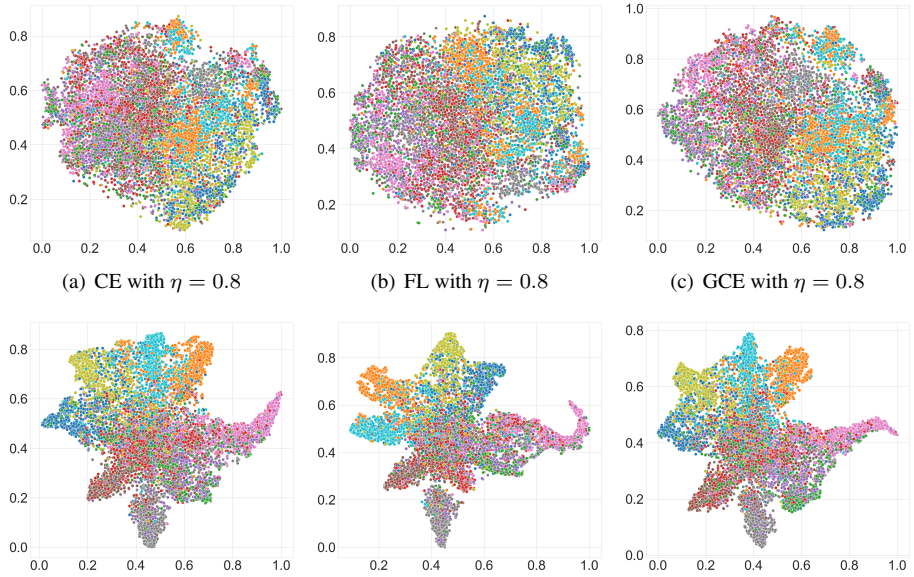


(a) CE with $\eta = 0.2$ (b) FL with $\eta = 0.2$ (c) GCE with $\eta = 0.2$



(d) CE+SR with $\eta = 0.2$ (e) FL+SR with $\eta = 0.2$ (f) GCE+SR with $\eta = 0.2$

Figure 6. Features visualization for CE (top) and CE+SR (bottom) on CIFAR10 with 0.2 symmetric label noise by t-SNE [6] 2D embeddings at the last second full-connected layer.



(a) CE with $\eta = 0.8$ (b) FL with $\eta = 0.8$ (c) GCE with $\eta = 0.8$

(d) CE+SR with $\eta = 0.8$ (e) FL+SR with $\eta = 0.8$ (f) GCE+SR with $\eta = 0.8$

Figure 7. Features visualization for CE (top) and CE+SR (bottom) on CIFAR10 with 0.8 symmetric label noise by t-SNE [6] 2D embeddings at the last second full-connected layer.