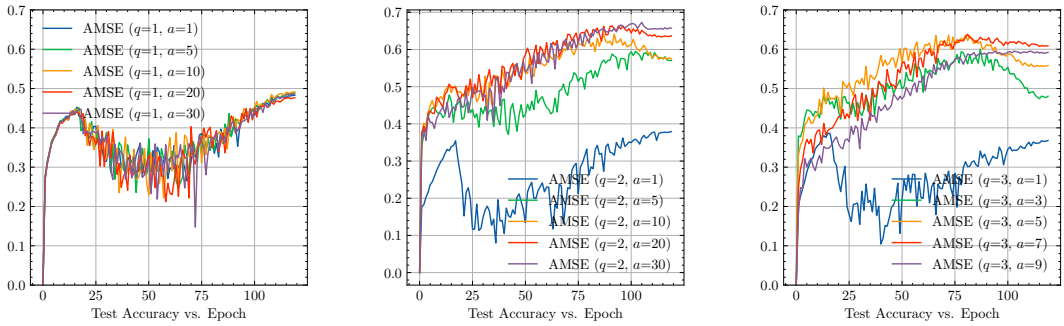
Joint Asymmetric Loss for Learning with Noisy Labels
Supplementary Materials

1. More Results

More Ablation Experiments about AMSE. We present the ablation experiments for different q and a in Figure 1. As illustrated: 1) For $q = 1$, the asymmetric condition always holds. In this case, a is a constant with zero gradient, making different choices of a equivalent. The loss is difficult to optimize, similar to MAE. 2) For $q = 2$, the asymmetric condition holds when $a \geq 9$. For the gradient, we have $\frac{\partial L(f(\mathbf{x}),y)}{\partial f(\mathbf{x})_y} = -\frac{2}{K}(a - f(\mathbf{x})_y)$, and a does not affect $\frac{\partial L(f(\mathbf{x}),y)}{\partial f(\mathbf{x})_{k\neq y}}$. As a increases, the weight of high-confidence (clean) samples in the gradient increases, while the weight of low-confidence (noisy) samples decreases. This explains why a larger a leads to better robustness. 3) For $q = 3$, the condition holds when $a \geq 4.73$. The performance of the loss is similar to $q = 2$, but it is more sensitive to the hyperparameter, as higher powers amplify the loss error. Therefore, using $q = 2$ is an appropriate choice.



(a) AMSE ($q = 1$) (b) AMSE ($q = 2$) (c) AMSE ($q = 3$)

Figure 1. Ablation experiments for AMSE on CIFAR-10 with 0.8 symmetric noise.

More Results for AGCE+MAE. For the experiment for AGCE+MAE, we use the same $a = 6, q = 1.5$ in [10], and search for $\alpha, \beta \in [1, 10]$. The complete results are presented in Table 1, while the results for $\alpha = 1, \beta = 1$ are shown in the main paper.

Table 1. Last epoch test accuracies (%) of different methods on CIFAR-10 with symmetric ($\eta \in [0.4, 0.8]$) and asymmetric ($\eta \in [0.2, 0.4]$) label noise. The results "mean±std" are reported over 3 random trials and the best results are in **bold**. † RCE actually equals a scaled MAE [5]. In order to be consistent with the original APL paper [4], we still write RCE here.

CIFAR-10	**Symmetric**		**Asymmetric**	
	0.4	0.8	0.2	0.4
MAE	82.03±3.63	44.45±6.49	77.20±4.45	57.86±1.23
NCE	69.37±0.22	41.20±1.25	72.20±0.38	65.33±0.40
AGCE	83.39±0.17	44.42±0.74	86.67±0.14	60.91±0.20
AGCE+MAE ($\alpha = 1, \beta = 1$)	85.25±0.12	44.61±5.72	78.28±4.67	57.80±2.53
AGCE+MAE ($\alpha = 1, \beta = 10$)	85.86±0.11	39.44±0.71	77.64±3.75	56.50±0.41
AGCE+MAE ($\alpha = 10, \beta = 1$)	85.71±0.29	23.36±2.85	75.43±4.16	57.55±1.83
AGCE+MAE ($\alpha = 10, \beta = 10$)	85.85±0.55	21.83±1.47	78.92±4.59	56.49±0.50
NCE+RCE†	**85.89**±0.31	**54.99**±2.13	**88.62**±0.29	**77.94**±0.21

2. Experiments

2.1. Evaluation on Benchmark Datasets

Noise Generation. We follow the approach of the previous work [8] to experiment with two types of synthetic label noise: symmetric noise and asymmetric noise. In the case of symmetric label noise, we intentionally corrupt the training labels by randomly flipping labels within each class to incorrect labels in other classes. As for asymmetric label noise, we flip the labels within a specific sets of classes: For CIFAR-10, the flips occur from TRUCK $\rightarrow$ AUTOMOBILE, BIRD $\rightarrow$ AIRPLANE, DEER $\rightarrow$ HORSE, and CAT $\leftrightarrow$ DOG. For CIFAR-100, the 100 classes are grouped into 20 super-classes, each containing 5 sub-classes, and we flip the labels within the same super-class into the next. For instance-dependent noise, we follow the approach in PDN [6] for generating label noise.

Experimental Setting. We follow the experimental settings in [4, 8, 10]: An 8-layer CNN is used for CIFAR-10 and a ResNet-34 [1, 3] for CIFAR-100. The networks are trained for 120 and 200 epochs for CIFAR-10 and CIFAR-100 with batch size 128. We use the SGD optimizer with momentum 0.9 and L1 weight decay 5×10^{-5} and 5×10^{-6} for CIFAR-10 and CIFAR-100. The learning rate is set to 0.01 for CIFAR-10 and 0.1 for CIFAR-100 with cosine annealing. Typical data augmentations including random shift and horizontal flip are applied.

Parameters Setting. For baselines, we use the same parameter settings in [4, 8, 10], which match their best parameters. The detailed parameters for JAL and baselines can be found in Table 2. For LT-APL [9], we take results directly from the original paper. For our method, we follow a principled strategy for parameter tuning: the range of a can be initially estimated through theoretical guidance, and then selected from $[5, 10, 20, 30]$ based on experimental results.

Table 2. Parameter settings for different methods.

Parameter	CIFAR-10	CIFAR-100	WebVision	Clothing1M
CE	-	-	-	-
FL (γ)	(0.5)	(0.5)	-	-
GCE (q)	(0.9)	(0.7)	(0.7)	(0.6)
SCE (α, β, A)	(0.1, 1, -4)	(6, 1, -4)	(10, 1, -4)	(10, 1, -4)
NCE	-	-	-	-
NCE+RCE (α, β, A)	(1, 1, -4)	(10, 0.1, -4)	(50, 0.1, -4)	(10, 1, -4)
NCE+AUL (α, β, a, p)	(1, 3, 6.3, 1.5)	(10, 0.015, 6, 3)	-	-
NCE+AGCE (α, β, a, q)	(10, 4, 6, 1.5)	(10, 0.1, 1.8, 3)	(50, 0.1, 2.5, 3)	(50, 0.1, 2.5, 3)
ANL-CE (α, β)	(5, 5)	(10, 1)	(20, 1)	(5, 0.1)
ANL-FL (α, β, γ)	(5, 5, 0.5)	(10, 1, 0.5)	(20, 1, 0.5)	(5, 0.1, 0.5)
JAL-CE (α, β, a)	(1, 1, 30)	(5, 1, 20)	(50, 1, 30)	(5, 0.1, 5)
JAL-FL (α, β, a, γ)	(1, 1, 30, 0.5)	(5, 1, 20, 0.5)	(50, 1, 30, 0.5)	(5, 0.1, 5, 0.5)

2.2. Evaluation on Real-World Datasets

Experiment Setting for WebVision / ILSVRC12. For WebVision, we use the mini setting [2], which includes the first 50 classes of the google image subset. We train a ResNet-50 using SGD for 250 epochs with initial learning rate 0.4, nesterov momentum 0.9 and weight decay 3×10^{-5} and batch size 256. The learning rate is multiplied by 0.97 after each epoch of training. All the images are resized to 224×224. Typical data augmentations including random shift, color jittering, and horizontal flip are applied. We train the model on Webvision and evaluate the trained model on the same 50 concepts on the corresponding WebVision and ILSVRC12 validation sets.

Experiment Setting for Clothing1M. For Clothing1M, we use ResNet-50 pre-trained on ImageNet similar to [7]. All the images are resized to 224×224. We use SGD with a momentum of 0.9, a weight decay of 1×10^{-3}, and batch size of 256. We train the network for 10 epochs with a learning rate of 5×10^{-3} and a decay of 0.1 at 5 epochs. Typical data augmentations including random shift and horizontal flip are applied.

References

[1] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 770–778, 2016. 2

[2] Lu Jiang, Zhengyuan Zhou, Thomas Leung, Li-Jia Li, and Li Fei-Fei. Mentornet: Learning data-driven curriculum for very deep neural networks on corrupted labels. In *International conference on machine learning*, pages 2304–2313. PMLR, 2018. 2

[3] Yann LeCun, Bernhard Boser, John S Denker, Donnie Henderson, Richard E Howard, Wayne Hubbard, and Lawrence D Jackel. Backpropagation applied to handwritten zip code recognition. *Neural computation*, 1(4):541–551, 1989. 2

[4] Xingjun Ma, Hanxun Huang, Yisen Wang, Simone Romano, Sarah Erfani, and James Bailey. Normalized loss functions for deep learning with noisy labels. In *International conference on machine learning*, pages 6543–6553. PMLR, 2020. 1, 2

[5] Yisen Wang, Xingjun Ma, Zaiyi Chen, Yuan Luo, Jinfeng Yi, and James Bailey. Symmetric cross entropy for robust learning with noisy labels. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 322–330, 2019. 1

[6] 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. *Advances in Neural Information Processing Systems*, 33:7597–7610, 2020. 2

[7] Tong Xiao, Tian Xia, Yi Yang, Chang Huang, and Xiaogang Wang. Learning from massive noisy labeled data for image classification. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 2691–2699, 2015. 2

[8] Xichen Ye, Xiaoqiang Li, Songmin Dai, Tong Liu, Yan Sun, and Weiqin Tong. Active negative loss functions for learning with noisy labels. In *Thirty-seventh Conference on Neural Information Processing Systems*, 2023. 2

[9] Shuo Zhang, Jian-Qing Li, Hamido Fujita, Yu-Wen Li, Deng-Bao Wang, Ting-Ting Zhu, Min-Ling Zhang, and Cheng-Yu Liu. Student loss: Towards the probability assumption in inaccurate supervision. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 46(6):4460–4475, 2024. 2

[10] Xiong Zhou, Xianming Liu, Junjun Jiang, Xin Gao, and Xiangyang Ji. Asymmetric loss functions for learning with noisy labels. In *International conference on machine learning*, pages 12846–12856. PMLR, 2021. 1, 2