Table 4. CNN on MNIST and CIFAR-10/100.

Input	28×28 Gray Image 32×32 Color Image
Block 1	Conv(3×3, 128)-BN-LReLU Conv(3×3, 128)-BN-LReLU Conv(3×3, 128)-BN-LReLU MaxPool(2×2, stride = 2) Dropout(p = 0.25)
Block 2	Conv(3×3, 256)-BN-LReLU Conv(3×3, 256)-BN-LReLU Conv(3×3, 256)-BN-LReLU MaxPool(2×2, stride = 2) Dropout(p = 0.25)
Block 3	Conv(3×3, 512)-BN-LReLU Conv(3×3, 256)-BN-LReLU Conv(3×3, 128)-BN-LReLU GlobalAvgPool(128)
Score	Linear(128, 10 or 100)

Table 5. CNN for open-set noise.

Input	32×32 Color Image
Block 1	Conv(3×3, 64)-BN-LReLU Conv(3×3, 64)-BN-LReLU MaxPool(2×2)
Block 2	Conv(3×3, 128)-BN-LReLU Conv(3×3, 128)-BN-LReLU MaxPool(2×2)
Block 3	Conv(3×3, 196)-BN-LReLU Conv(3×3, 196)-BN-LReLU MaxPool(2×2)
Score	Linear(256, 10)

Table 6. 1D CNN on NEWS.

Input	Sequence of Tokens
Embed	300D GloVe
Block 1	Conv(3, 100)-ReLU GlobalMaxPool(100)
Score	Linear(100, 7)

Table 7. MLP on NEWS.

Input	Sequence of Tokens
Embed	300D GloVe
Block 1	Linear(300, 300)-Softsign Linear(300, 300)-Softsign
Score	Linear(300, 2)

A. Neural Network Architectures for Benchmark Datasets

Table 4 describes the 9-layer CNN (Laine & Aila, 2017; Miyato et al., 2019) used on MNIST and CIFAR-10/100. In fact, it has 9 convolutional layers but 19 trainable layers. Table 5 describes the CNN used on CIFAR-10 under open-set noise. It has 6 convolutional layers but 13 trainable layers. Furthermore, the 1D CNN on NEWS for SET1 methods and MLP on NEWS for SET2 methods are given in Tables 6 and 7 respectively. Here, BN stands for *batch normalization* layers (Ioffe & Szegedy, 2015); LReLU stands for *Leaky ReLU* (Xu et al., 2015), a special case of *Parametric ReLU* (He et al., 2015); GloVe stands for *global vectors* for word representation (Pennington et al., 2014); Softsign is an activation function which looks very similar to Tanh (Glorot & Bengio, 2010).

Note that the 9-layer CNN is a standard and common practice in weakly supervised learning, including but not limited to semi-supervised learning (e.g., Laine & Aila, 2017; Miyato et al., 2019) and noisy-label learning (e.g., Han et al., 2018b). Actually, this CNN was not born in those areas—it came from Salimans et al. (2016) where it served as the discriminator of GANs on CIFAR-10. We decided to use this CNN, because then the experimental results are directly comparable with previous papers in the same area, and it would be crystal clear where the proposed methods stand in the area.

That being said, SIGUA can definitely achieve better performance if given better models. In order to demonstrate this, let us take ResNet-18 (He et al., 2016), the smallest ResNet in *torch vision model zoo* for example. The experimental results are shown in Table 8. For each noise, we selected the better one among SIGUA_{SL} and SIGUA_{BC}, and replaced the model with ResNet-18. We can see from Table 8 that the improvements were very great by training bigger ResNet-18.

Table 8. Average test accuracy (in %) over the last ten epochs on CIFAR-10.

	SIGUA_{SL} under symmetry-20%	SIGUA_{SL} under symmetry-50%	SIGUA_{BC} under pair-45%
9-layer CNN	84.05	77.12	81.82
ResNet-18	89.41	81.96	89.56
Absolute Acc Increase	5.36	4.84	7.74
Relative Err Reduction	33.61	21.15	42.57

B. More Experiments

Due to the limited space, the experiments on CIFAR-100 and NEWS are moved here. The setup of CIFAR-100 is similar to CIFAR-10, but the momentum is 0.5 and lr is divided by 10 every 30 epochs for SET2 methods. The setup of NEWS is similar to other three datasets, except $\mathfrak{O}$ is AdaGrad (Duchi et al., 2011) that automatically decays lr every mini-batch.

Figure 4 shows the accuracy curves of the three methods in SET1 (CIFAR-100 in the top and NEWS in the bottom). The trend in Figure 4 is similar to the trend in Figure 2 that SIGUA_{SL} either stopped or alleviated the decrease in Standard and

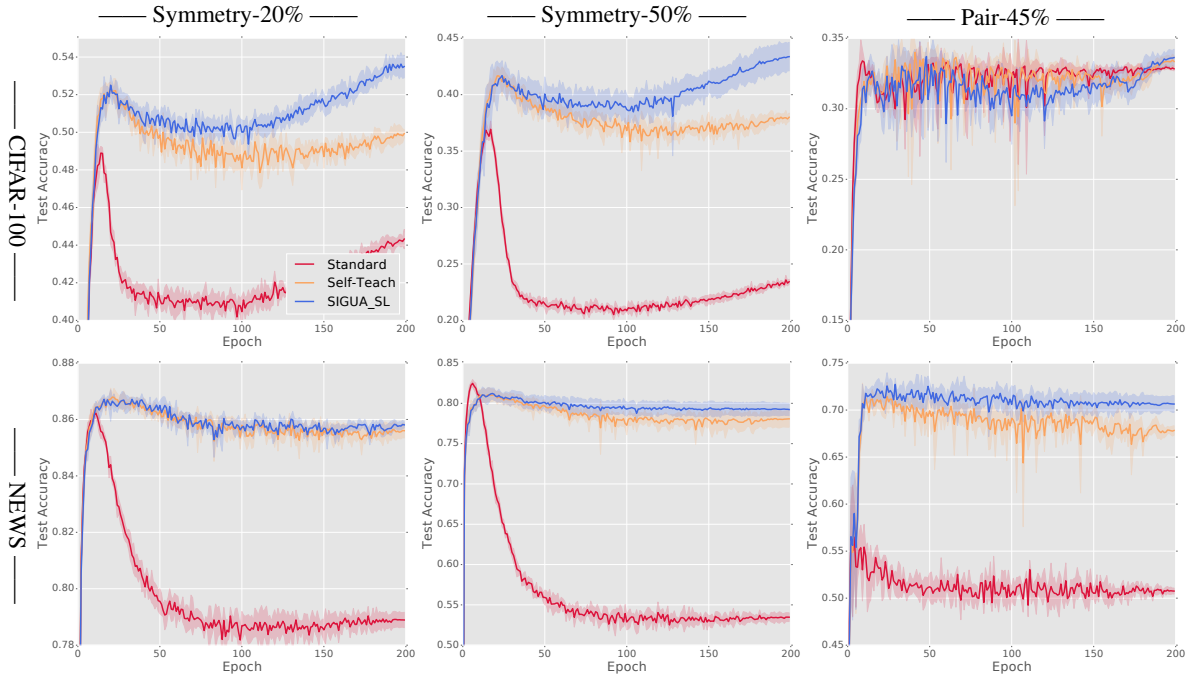

Figure 4. Accuracy curves of training deep networks using the three learning methods in SET1.

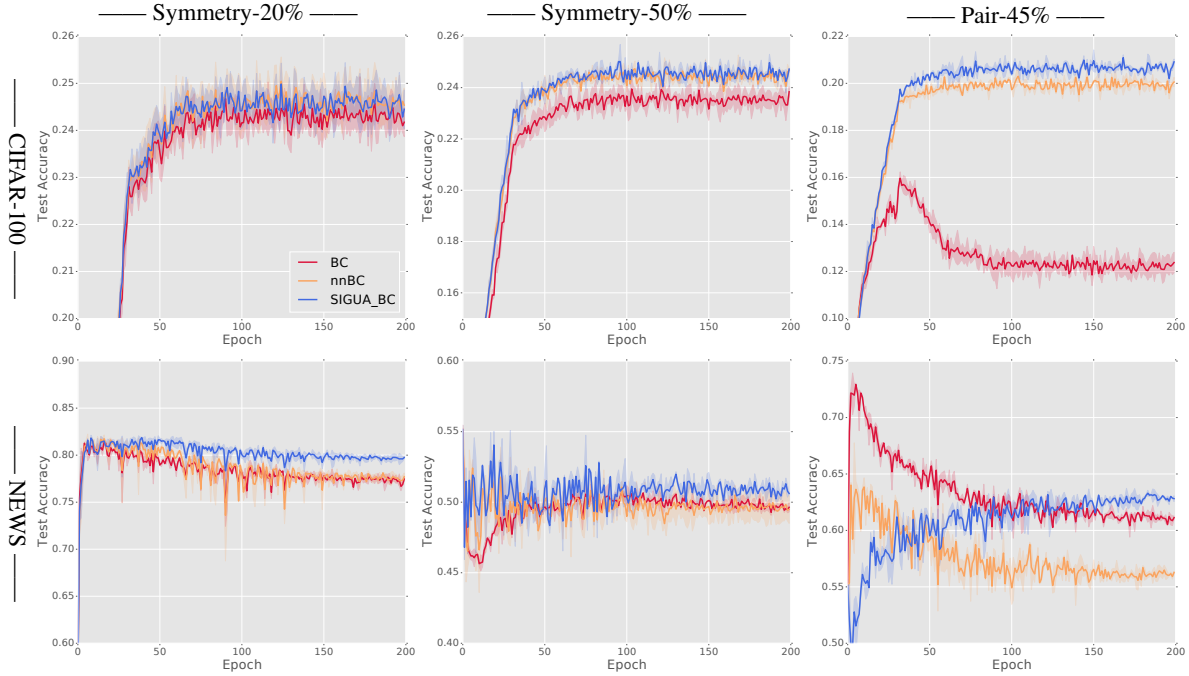

Figure 5. Accuracy curves of training deep networks using the three learning methods in SET2.

Self-Teach. Especially on CIFAR-100, after a remarkable decrease in the first half, the accuracy in the second half started to increase once more, and it eventually surpassed the best accuracy that can be obtained by early stopping. If we plot the test error rather than the test accuracy, this phenomenon is exactly an epoch-wise double descent (Nakkiran et al., 2020). Figure 5 shows the accuracy curves of the three methods in SET2 where SIGUA_{BC} still stopped or alleviated the decrease in BC and/or nnBC. The reason why BC suffered more under pair-45% may be explained by the maximum and minimum elements of T^{-1}: when $k = 10$, they are 2.101 and -1.719 under pair-45% but 2.125 and -0.125 under symmetry-50%. It is interesting on CIFAR-10, nnBC and SIGUA_{BC} under pair-45% outperformed themselves under symmetry-20%, which provides an evidence that the issue of negative losses can be fixed at least empirically.