SUPPLEMENTARY MATERIAL - AL2: PROGRESSIVE ACTIVATION LOSS FOR LEARNING GENERAL REPRESENTATIONS IN CLASSIFICATION NEURAL NETWORKS

Majed El Helou *Frederike Dümbgen* *Sabine Süsstrunk*

School of Computer and Communication Sciences, EPFL, Switzerland.

ABSTRACT

In this supplementary material, we present the details of the neural network architecture and training settings used in all our experiments. This holds for all experiments presented in the main paper as well as in this supplementary material.

We also show the summary results of all of our 96 experiments (test accuracy, training cross-entropy loss, and regularization loss), sampled at 100 epoch intervals. We analyze these results for each of the benchmark datasets, namely MNIST, Fashion-MNIST and CIFAR10, and underline global observations we make throughout the entire experiment set.

1. ARCHITECTURE AND TRAINING DETAILS

Our network architecture is a VGG-like architecture, which we use to dissect the effects of the different regularizers. The details of the architecture are as follows:

- 2D convolution (20 channels, kernel of size 5, stride of 1)
- ReLU activation
- 2D MaxPool with window of size 2
- 2D convolution (50 channels, kernel of size 5, stride of 1)
- optional Batch Normalization (c.f. experiments)
- ReLU activation
- 2D MaxPool with window of size 2
- optional Dropout (c.f. experiments)
- Linear layer mapping to 500 neurons
- ReLU activation
- Linear layer mapping to the N classes
- log SoftMax.

We use PyTorch [1] libraries and train all networks with stochastic gradient decent (batch size 64, learning rate 10^{-3}, momentum 0.5) in all of our experiments. We also take care of using the same random initialization weights for all our trained networks, to reduce if not remove any potential effect that their variance might have on the results.

Corrupt labels are created with what is known as symmetric noise. In other words, to corrupt a label we randomly select one incorrect label (with uniform probability) and replace the old label by the new label. The corrupt assignment of labels is preserved throughout all epochs, and the training data is shuffled at the end of every epoch.

2. DETAILED RESULTS ANALYSIS

The results of training the neural network on MNIST, Fashion-MNIST and CIFAR10, each with label corruption of 75, 50, 25 and 0% are all collected in this supplementary material for selected epochs ranging from 100 to 700. In each experiment, we train a bare version of the convolutional VGG-like network, a version with batch normalization [2], with dropout [3] and one with weight decay [4], each of which is then re-trained with AL2, always starting from the same network weight initialization. These experiments result in a total of 96 neural networks trained until 700 epochs each, and we analyze their results in what follows. We begin by the MNIST dataset, followed by Fashion-MNIST then CIFAR10, in increasing degree of difficulty of the datasets.

MNIST is a relatively easy dataset for neural networks to overfit, given a large enough number of parameters [5]. We can see that for any percentage of corrupt labels in the training dataset, using AL2 significantly improves the generalization of the network assessed at the final epoch, by 60 percentage points in the most extreme case (75% corrupt labels, with weight decay, Table 1).

Fashion-MNIST is more challenging compared to MNIST. We see again that in the most extreme case we improve by up to almost 60 percentage points (Table 5). We notice as well (notably with 50% and 25% corrupt labels, Tables 6 and 7 respectively) that the performance of the network with our method is hardly affected by the baseline. In other words, irrespective of whether the network is bare, or has batch normalization, dropout or weight decay, the performance is practically the same when AL2 is used in the training. We see, however, that the best performance this time is with dropout, closely followed by weight decay with AL2 training.

CIFAR10 is also a more challenging dataset, and this is directly reflected in the experimental results without label corruption. We note here again the improvement resulting from using AL2 during the training, with the most extreme case being an improvement by about 30 percentage points (when training with 50% corrupt labels, with weight decay and also with the bare neural network). The best final results with any degree of corruption are always with AL2 training (whether with weight decay or with batch normalization or even with the bare network).

Global observations:

- **Without AL2.** In the absence of AL2 during training, the best performance was attained by the network trained with dropout.
- **Activation magnitude.** One very interesting phenomenon we found with dropout, compared with the bare, the batch normalization, and the weight decay networks, is that dropout decreases the magnitude of the activations on the feature representation layer. For example with 75% corrupt labels, dropout decreases our regularization loss down from 119.10 to 1.92 (MNIST, Table 1), from 17.36 to 0.13 (Fashion-MNIST, Table 5), and from 87.98 to 3.78 (CIFAR10, Table 9). This means that dropout also indirectly minimizes the regularization loss that we use explicitly in our training. When AL2 training is used, the regularization loss which is explicitly minimized is multiple orders of magnitude smaller than that of other networks.
- **Weight decay.** The regularization loss with the weight decay network (which we tweaked to find the weight parameter achieving the best performance of weight decay, before AL2 training) was much larger than that with other regularizers, and close to that of the bare network. This counter-intuitive phenomenon illustrates the difference between decaying weight values and regularizing the activation values of a certain intermediate feature representation. In fact, the former almost did not affect the latter (compare weight decay RL to bare RL, relative to the RL of other networks, in any of the result tables below).
- **Training cross-entropy.** 1) We note that, especially with the bare network, there is little difficulty in dropping the training cross-entropy loss and doing very well on the training dataset classification, and this is consistently observed throughout our different experiments. 2) The networks trained with AL2 have larger training cross-entropy loss, which is important (although not sufficient) for improving performance with corrupt labels. Also interestingly, we notice that when training with 0% corrupt labels, the bare network trained with AL2 has larger training cross-entropy loss yet performs similarly (MNIST and Fashion-MNIST) or even better ($+3.89$ percentage points on CIFAR10, Table 12) than the bare network trained normally. This further strengthens our generalization findings made with the corrupt-label experiments.

3. REFERENCES

[1] A. Paszke, S. Gross, S. Chintala, G. Chanan, E. Yang, Z. DeVito, Z. Lin, A. Desmaison, L. Antiga, and A. Lerer, "Automatic differentiation in PyTorch," in *NeurIPS Workshop*, 2017.

[2] S. Ioffe and C. Szegedy, "Batch normalization: Accelerating deep network training by reducing internal covariate shift," in *ICML*, 2015, pp. 448–456.

[3] N. Srivastava, G. Hinton, A. Krizhevsky, I. Sutskever, and R. Salakhutdinov, "Dropout: A simple way to prevent neural networks from overfitting," *The Journal of Machine Learning Research*, vol. 15, no. 1, pp. 1929–1958, 2014.

[4] A. Krogh and J. A. Hertz, "A simple weight decay can improve generalization," in *NeurIPS*, 1992, pp. 950–957.

[5] C. Zhang, S. Bengio, M. Hardt, B. Recht, and O. Vinyals, "Understanding deep learning requires rethinking generalization," in *ICLR*, 2017.

Different metrics evaluated across training epochs (without/with AL2)								
Baseline	Metric	epoch=100	epoch=200	epoch=300	epoch=400	epoch=500	epoch=600	epoch=700
Bare	TA	84.20/95.25	45.30/94.92	25.25/93.07	23.83/88.76	26.07/79.64	26.45/75.88	25.84/68.46
	$\mathcal{L}_c$	2.15/2.22	1.78/2.19	0.89/2.15	0.19/2.11	0.04/2.08	0.01/2.07	0.00/2.08
	$\mathcal{L}_r$	3.20/0.24	10.93/0.10	26.12/0.06	54.42/0.03	74.49/0.02	103.26/0.01	119.10/0.00
BN [2]	TA	74.72/95.47	36.65/94.48	26.72/90.20	25.97/85.34	25.88/83.02	25.60/81.53	25.55/81.16
	$\mathcal{L}_c$	2.07/2.22	1.48/2.19	0.30/2.15	0.04/2.12	0.01/2.11	0.01/2.12	0.01/2.14
	$\mathcal{L}_r$	0.84/0.24	2.35/0.10	6.46/0.06	9.25/0.03	10.40/0.01	11.06/0.01	11.51/0.00
DO [3]	TA	96.13/94.43	96.47/95.03	95.93/95.03	92.74/94.79	81.96/92.15	68.12/92.69	55.39/91.70
	$\mathcal{L}_c$	2.22/2.23	2.20/2.22	2.17/2.20	2.13/2.20	2.05/2.20	1.94/2.21	1.79/2.23
	$\mathcal{L}_r$	0.26/0.24	0.30/0.09	0.41/0.04	0.61/0.02	1.00/0.01	1.50/0.00	1.92/0.00
WD [4]	TA	88.91/95.21	50.87/95.47	27.98/95.17	27.66/94.03	25.14/91.42	28.05/89.81	25.57/86.98
	$\mathcal{L}_c$	2.16/2.22	1.87/2.20	1.06/2.18	0.32/2.16	0.07/2.16	0.04/2.17	0.02/2.19
	$\mathcal{L}_r$	2.94/0.23	10.52/0.09	26.04/0.05	53.65/0.02	81.53/0.01	84.64/0.00	107.80/0.00

Table 1. Test accuracy (TA), training cross-entropy loss $\mathcal{L}_c$, and our training regularization loss $\mathcal{L}_r$ which is shown 100 times larger with AL2 for readability. We evaluate all metrics at different epochs and with different baselines, without/with AL2. The networks are trained on the MNIST dataset with 75% corrupt labels.

Different metrics evaluated across training epochs (without/with AL2)								
Baseline	Metric	epoch=100	epoch=200	epoch=300	epoch=400	epoch=500	epoch=600	epoch=700
Bare	TA	96.28/98.14	73.52/97.63	51.41/95.92	54.69/94.10	53.48/92.14	54.27/92.82	53.62/93.50
	$\mathcal{L}_c$	1.70/1.81	1.25/1.75	0.52/1.68	0.12/1.62	0.03/1.60	0.01/1.62	0.00/1.65
	$\mathcal{L}_r$	7.38/0.44	20.34/0.18	42.51/0.11	72.29/0.06	99.61/0.03	134.51/0.01	155.10/0.00
BN [2]	TA	94.77/97.90	75.20/97.35	59.02/95.86	55.07/94.48	55.77/94.47	55.56/94.46	55.11/95.22
	$\mathcal{L}_c$	1.64/1.81	1.12/1.74	0.25/1.69	0.04/1.66	0.01/1.64	0.01/1.66	0.01/1.69
	$\mathcal{L}_r$	0.95/0.46	2.39/0.19	5.90/0.10	8.33/0.05	9.40/0.02	10.01/0.01	10.44/0.00
DO [3]	TA	98.38/97.97	98.73/97.97	98.71/97.92	98.36/97.76	96.86/97.64	94.17/97.61	89.06/97.49
	$\mathcal{L}_c$	1.82/1.84	1.79/1.82	1.75/1.80	1.70/1.78	1.62/1.79	1.52/1.80	1.41/1.83
	$\mathcal{L}_r$	0.64/0.48	0.75/0.16	0.97/0.08	1.33/0.04	1.75/0.02	2.21/0.01	2.58/0.00
WD [4]	TA	96.76/98.13	77.36/98.03	51.16/97.07	56.43/96.39	51.59/95.56	56.47/95.77	53.69/96.43
	$\mathcal{L}_c$	1.71/1.81	1.32/1.77	0.62/1.72	0.20/1.68	0.07/1.67	0.04/1.69	0.02/1.73
	$\mathcal{L}_r$	7.18/0.45	20.26/0.18	42.29/0.10	75.60/0.05	85.77/0.02	100.01/0.01	130.50/0.00

Table 2. Test accuracy (TA), training cross-entropy loss $\mathcal{L}_c$, and our training regularization loss $\mathcal{L}_r$ which is shown 100 times larger with AL2 for readability. We evaluate all metrics at different epochs and with different baselines, without/with AL2. The networks are trained on the MNIST dataset with 50% corrupt labels.

Different metrics evaluated across training epochs (without/with AL2)								
Baseline	Metric	epoch=100	epoch=200	epoch=300	epoch=400	epoch=500	epoch=600	epoch=700
Bare	TA	98.39/98.60	88.99/98.29	77.87/97.97	81.15/97.68	81.22/97.38	80.04/96.79	79.90/97.46
	$\mathcal{L}_c$	1.04/1.16	0.68/1.10	0.24/1.04	0.06/1.00	0.02/0.99	0.01/1.01	0.00/1.05
	$\mathcal{L}_r$	11.75/0.60	30.75/0.22	63.08/0.12	96.88/0.06	100.71/0.03	146.50/0.01	171.46/0.01
BN [2]	TA	97.90/98.40	90.03/98.36	81.39/98.05	79.55/97.68	80.03/97.78	79.76/97.68	79.57/97.80
	$\mathcal{L}_c$	1.02/1.15	0.67/1.09	0.20/1.05	0.04/1.03	0.01/1.03	0.01/1.05	0.00/1.08
	$\mathcal{L}_r$	0.98/0.63	2.09/0.24	4.34/0.11	6.21/0.05	7.12/0.02	7.64/0.01	8.00/0.00
DO [3]	TA	98.86/98.40	99.12/98.40	99.20/98.42	99.00/98.40	98.70/98.41	98.10/98.39	97.13/98.30
	$\mathcal{L}_c$	1.17/1.20	1.13/1.17	1.09/1.15	1.04/1.15	0.99/1.16	0.92/1.18	0.84/1.21
	$\mathcal{L}_r$	1.19/0.67	1.39/0.22	1.79/0.10	2.32/0.05	2.87/0.02	3.40/0.01	3.84/0.00
WD [4]	TA	98.51/98.63	90.12/98.38	78.80/98.29	80.91/98.02	80.42/97.97	81.08/97.65	80.00/97.96
	$\mathcal{L}_c$	1.05/1.16	0.74/1.12	0.31/1.07	0.09/1.05	0.04/1.05	0.02/1.07	0.02/1.11
	$\mathcal{L}_r$	11.37/0.61	30.28/0.23	61.39/0.12	101.03/0.06	130.50/0.03	139.38/0.01	143.93/0.01

Table 3. Test accuracy (TA), training cross-entropy loss $\mathcal{L}_c$, and our training regularization loss $\mathcal{L}_r$ which is shown 100 times larger with AL2 for readability. We evaluate all metrics at different epochs and with different baselines, without/with AL2. The networks are trained on the MNIST dataset with 25% corrupt labels.

Different metrics evaluated across training epochs (without/with AL2)								
Baseline	Metric	epoch=100	epoch=200	epoch=300	epoch=400	epoch=500	epoch=600	epoch=700
Bare	TA	99.01/98.89	99.03/99.09	98.99/99.22	99.05/99.19	98.99/99.14	99.02/99.14	99.03/99.04
	$\mathcal{L}_c$	0.01/0.06	0.00/0.04	0.00/0.04	0.00/0.05	0.00/0.06	0.00/0.07	0.00/0.10
	$\mathcal{L}_r$	48.74/0.87	77.32/0.26	97.68/0.11	111.15/0.05	120.98/0.02	128.66/0.01	134.94/0.00
BN [2]	TA	98.98/99.00	99.02/99.07	99.02/99.12	99.03/99.07	99.00/99.05	99.01/98.89	99.05/98.97
	$\mathcal{L}_c$	0.01/0.06	0.00/0.04	0.00/0.04	0.00/0.05	0.00/0.06	0.00/0.08	0.00/0.10
	$\mathcal{L}_r$	1.03/0.89	1.19/0.27	1.28/0.11	1.33/0.05	1.37/0.02	1.40/0.01	1.42/0.00
DO [3]	TA	99.16/98.80	99.30/99.10	99.39/99.20	99.37/99.04	99.37/99.14	99.44/99.06	99.40/98.88
	$\mathcal{L}_c$	0.04/0.08	0.02/0.06	0.02/0.06	0.01/0.07	0.01/0.09	0.01/0.11	0.01/0.14
	$\mathcal{L}_r$	7.12/1.06	7.79/0.31	8.57/0.13	9.36/0.06	10.12/0.03	10.70/0.01	11.42/0.01
WD [4]	TA	99.00/98.85	99.09/99.08	99.04/99.19	99.06/99.13	99.07/99.11	99.09/99.06	99.10/98.98
	$\mathcal{L}_c$	0.01/0.06	0.00/0.05	0.00/0.05	0.00/0.06	0.00/0.07	0.00/0.10	0.00/0.13
	$\mathcal{L}_r$	43.92/0.91	61.93/0.29	72.14/0.12	77.32/0.06	80.77/0.03	83.47/0.01	85.94/0.01

Table 4. Test accuracy (TA), training cross-entropy loss $\mathcal{L}_c$, and our training regularization loss $\mathcal{L}_r$ which is shown 100 times larger with AL2 for readability. We evaluate all metrics at different epochs and with different baselines, without/with AL2. The networks are trained on the MNIST dataset with 0% corrupt labels.

Different metrics evaluated across training epochs (without/with AL2)								
Baseline	Metric	epoch=100	epoch=200	epoch=300	epoch=400	epoch=500	epoch=600	epoch=700
Bare	TA	80.37/75.91	80.23/77.92	60.66/78.78	41.21/78.81	24.88/77.77	27.28/77.38	19.97/73.59
	$\mathcal{L}_c$	2.22/2.24	2.18/2.23	2.04/2.21	1.76/2.20	1.30/2.19	0.74/2.19	0.29/2.20
	$\mathcal{L}_r$	0.76/0.17	1.38/0.06	2.54/0.03	4.17/0.02	6.91/0.01	11.00/0.00	17.36/0.00
BN [2]	TA	62.85/78.65	36.58/79.36	25.31/78.00	24.44/75.58	24.06/74.49	23.95/74.36	24.44/74.55
	$\mathcal{L}_c$	2.06/2.23	1.49/2.21	0.39/2.19	0.04/2.17	0.02/2.16	0.01/2.17	0.01/2.18
	$\mathcal{L}_r$	0.80/0.21	2.23/0.09	6.11/0.05	9.14/0.02	10.38/0.01	11.07/0.00	11.55/0.00
DO [3]	TA	77.29/74.38	80.91/76.61	83.03/77.48	83.85/77.79	83.56/77.67	82.56/77.39	79.79/72.72
	$\mathcal{L}_c$	2.24/2.25	2.23/2.24	2.22/2.23	2.20/2.23	2.19/2.23	2.16/2.24	2.13/2.25
	$\mathcal{L}_r$	0.11/0.18	0.09/0.06	0.09/0.03	0.09/0.01	0.10/0.01	0.11/0.00	0.13/0.00
WD [4]	TA	79.68/75.53	82.09/77.46	74.00/78.66	51.65/78.92	27.18/79.19	27.28/78.89	18.77/77.40
	$\mathcal{L}_c$	2.23/2.24	2.20/2.23	2.12/2.23	1.95/2.22	1.64/2.23	1.19/2.24	0.70/2.25
	$\mathcal{L}_r$	0.73/0.17	1.23/0.06	2.25/0.03	3.71/0.01	5.91/0.01	8.99/0.00	13.38/0.00

Table 5. Test accuracy (TA), training cross-entropy loss $\mathcal{L}_c$, and our training regularization loss $\mathcal{L}_r$ which is shown 100 times larger with AL2 for readability. We evaluate all metrics at different epochs and with different baselines, without/with AL2. The networks are trained on the FashionMNIST dataset with 75% corrupt labels.

Different metrics evaluated across training epochs (without/with AL2)								
Baseline	Metric	epoch=100	epoch=200	epoch=300	epoch=400	epoch=500	epoch=600	epoch=700
Bare	TA	86.21/84.62	87.03/86.81	77.32/86.64	62.95/85.63	46.36/83.39	46.73/83.74	45.98/84.00
	$\mathcal{L}_c$	1.86/1.89	1.75/1.84	1.52/1.78	1.14/1.74	0.68/1.71	0.31/1.71	0.08/1.74
	$\mathcal{L}_r$	2.22/0.39	3.60/0.15	6.33/0.08	9.93/0.05	15.05/0.02	22.24/0.01	32.80/0.00
BN [2]	TA	82.21/86.60	67.69/85.92	52.87/84.05	49.68/82.92	49.73/82.93	49.32/83.51	49.30/83.78
	$\mathcal{L}_c$	1.68/1.85	1.15/1.78	0.32/1.73	0.04/1.70	0.02/1.70	0.01/1.72	0.01/1.75
	$\mathcal{L}_r$	0.89/0.50	2.28/0.22	5.61/0.11	8.25/0.05	9.39/0.02	10.03/0.01	10.47/0.00
DO [3]	TA	84.62/83.02	87.01/85.82	87.85/86.73	88.40/86.38	88.49/86.23	88.55/85.97	87.59/84.91
	$\mathcal{L}_c$	1.91/1.92	1.88/1.89	1.85/1.86	1.83/1.85	1.80/1.85	1.76/1.87	1.70/1.89
	$\mathcal{L}_r$	0.25/0.42	0.25/0.15	0.27/0.07	0.31/0.04	0.36/0.02	0.46/0.01	0.60/0.00
WD [4]	TA	86.16/84.29	87.63/86.58	82.13/87.11	69.27/86.66	48.44/85.91	51.95/86.17	45.95/85.67
	$\mathcal{L}_c$	1.86/1.90	1.78/1.85	1.62/1.82	1.33/1.79	0.92/1.78	0.54/1.79	0.27/1.82
	$\mathcal{L}_r$	2.24/0.39	3.35/0.15	5.80/0.08	9.20/0.04	13.73/0.02	19.36/0.01	26.64/0.00

Table 6. Test accuracy (TA), training cross-entropy loss $\mathcal{L}_c$, and our training regularization loss $\mathcal{L}_r$ which is shown 100 times larger with AL2 for readability. We evaluate all metrics at different epochs and with different baselines, without/with AL2. The networks are trained on the FashionMNIST dataset with 50% corrupt labels.

Different metrics evaluated across training epochs (without/with AL2)								
Baseline	Metric	epoch=100	epoch=200	epoch=300	epoch=400	epoch=500	epoch=600	epoch=700
Bare	TA	88.28/87.45	88.20/88.50	86.14/88.49	81.39/88.33	73.53/87.87	70.35/87.44	71.62/87.38
	$\mathcal{L}_c$	1.25/1.29	1.12/1.22	0.91/1.16	0.62/1.13	0.33/1.11	0.14/1.13	0.03/1.16
	$\mathcal{L}_r$	3.92/0.57	6.15/0.21	9.96/0.11	14.93/0.06	21.67/0.03	30.71/0.01	43.06/0.00
BN [2]	TA	86.96/88.56	80.56/87.92	74.99/87.42	72.88/87.36	73.46/87.24	72.37/87.61	72.49/86.49
	$\mathcal{L}_c$	1.08/1.24	0.73/1.16	0.24/1.11	0.04/1.10	0.01/1.10	0.01/1.13	0.01/1.17
	$\mathcal{L}_r$	0.94/0.72	2.00/0.28	4.28/0.13	6.33/0.06	7.29/0.03	7.83/0.01	8.19/0.00
DO [3]	TA	86.80/86.36	88.81/87.88	89.61/88.99	90.04/88.44	90.19/88.63	90.26/88.29	90.11/87.69
	$\mathcal{L}_c$	1.34/1.35	1.29/1.30	1.25/1.27	1.22/1.27	1.18/1.27	1.14/1.29	1.09/1.32
	$\mathcal{L}_r$	0.46/0.63	0.46/0.22	0.50/0.10	0.56/0.05	0.66/0.02	0.80/0.01	0.96/0.00
WD [4]	TA	88.11/87.30	88.41/88.66	87.76/89.15	83.85/88.93	75.06/88.55	73.79/88.28	70.54/88.12
	$\mathcal{L}_c$	1.26/1.30	1.15/1.24	0.98/1.19	0.75/1.17	0.48/1.17	0.27/1.19	0.13/1.22
	$\mathcal{L}_r$	3.91/0.58	5.83/0.22	9.30/0.11	13.89/0.06	19.53/0.03	26.25/0.01	33.99/0.01

Table 7. Test accuracy (TA), training cross-entropy loss $\mathcal{L}_c$, and our training regularization loss $\mathcal{L}_r$ which is shown 100 times larger with AL2 for readability. We evaluate all metrics at different epochs and with different baselines, without/with AL2. The networks are trained on the FashionMNIST dataset with 25% corrupt labels.

Different metrics evaluated across training epochs (without/with AL2)								
Baseline	Metric	epoch=100	epoch=200	epoch=300	epoch=400	epoch=500	epoch=600	epoch=700
Bare	TA	89.50/89.36	90.58/90.61	90.31/90.79	90.97/90.64	90.94/90.92	91.09/90.59	90.82/90.25
	$\mathcal{L}_c$	0.22/0.27	0.13/0.19	0.07/0.16	0.03/0.15	0.01/0.16	0.00/0.18	0.00/0.21
	$\mathcal{L}_r$	10.71/0.93	14.12/0.31	20.06/0.14	27.64/0.06	36.08/0.03	44.31/0.01	51.06/0.01
BN [2]	TA	90.00/90.38	89.07/90.64	88.94/90.63	88.93/90.44	88.88/90.75	88.85/90.31	88.88/90.18
	$\mathcal{L}_c$	0.11/0.22	0.03/0.16	0.01/0.15	0.00/0.15	0.00/0.17	0.00/0.19	0.00/0.23
	$\mathcal{L}_r$	1.14/1.02	1.60/0.34	1.92/0.14	2.09/0.06	2.19/0.03	2.26/0.01	2.31/0.01
DO [3]	TA	88.83/88.74	90.52/90.22	91.21/90.50	91.72/90.75	92.01/90.62	92.19/90.17	92.22/90.04
	$\mathcal{L}_c$	0.32/0.34	0.26/0.27	0.22/0.25	0.19/0.25	0.17/0.27	0.15/0.30	0.14/0.33
	$\mathcal{L}_r$	1.66/1.05	1.32/0.34	1.31/0.15	1.36/0.07	1.44/0.03	1.58/0.01	1.69/0.01
WD [4]	TA	89.45/89.26	90.51/90.53	90.58/90.78	91.08/90.72	91.08/90.63	91.01/90.76	90.92/90.18
	$\mathcal{L}_c$	0.23/0.27	0.15/0.20	0.10/0.18	0.06/0.18	0.04/0.19	0.02/0.22	0.02/0.27
	$\mathcal{L}_r$	10.35/0.95	12.87/0.33	16.91/0.15	21.37/0.07	25.74/0.03	29.56/0.01	32.45/0.01

Table 8. Test accuracy (TA), training cross-entropy loss $\mathcal{L}_c$, and our training regularization loss $\mathcal{L}_r$ which is shown 100 times larger with AL2 for readability. We evaluate all metrics at different epochs and with different baselines, without/with AL2. The networks are trained on the FashionMNIST dataset with 0% corrupt labels.

Different metrics evaluated across training epochs (without/with AL2)								
Baseline	Metric	epoch=100	epoch=200	epoch=300	epoch=400	epoch=500	epoch=600	epoch=700
Bare	TA	41.71/32.61	24.64/36.82	17.98/37.06	17.64/36.68	17.46/36.44	17.55/34.60	17.53/30.61
	$\mathcal{L}_c$	2.26/2.29	1.97/2.29	0.61/2.28	0.01/2.28	0.00/2.28	0.00/2.28	0.00/2.29
	$\mathcal{L}_r$	0.53/0.05	5.66/0.02	25.44/0.01	66.22/0.00	69.73/0.00	81.11/0.00	87.98/0.00
BN [2]	TA	25.32/42.40	17.77/43.08	17.68/41.63	17.73/38.35	17.78/37.18	17.80/36.42	17.59/34.36
	$\mathcal{L}_c$	1.78/2.28	0.20/2.26	0.02/2.25	0.01/2.25	0.01/2.25	0.00/2.26	0.00/2.27
	$\mathcal{L}_r$	1.31/0.13	5.05/0.06	6.98/0.03	7.68/0.01	8.09/0.01	8.37/0.00	8.59/0.00
DO [3]	TA	41.40/20.17	45.83/22.02	46.23/22.63	43.91/21.86	32.53/22.00	25.33/21.28	21.74/19.43
	$\mathcal{L}_c$	2.28/2.30	2.26/2.29	2.23/2.29	2.17/2.29	1.99/2.29	1.61/2.29	1.28/2.30
	$\mathcal{L}_r$	0.09/0.03	0.11/0.01	0.17/0.01	0.37/0.00	1.30/0.00	2.86/0.00	3.78/0.00
WD [4]	TA	41.50/30.77	27.03/33.82	17.99/34.66	17.71/33.62	17.86/27.83	17.42/25.03	17.83/24.66
	$\mathcal{L}_c$	2.26/2.29	2.05/2.29	0.81/2.29	0.02/2.29	0.01/2.29	0.01/2.29	0.01/2.30
	$\mathcal{L}_r$	0.46/0.05	5.09/0.02	23.58/0.01	69.40/0.00	69.85/0.00	83.25/0.00	79.32/0.00

Table 9. Test accuracy (TA), training cross-entropy loss $\mathcal{L}_c$, and our training regularization loss $\mathcal{L}_r$ which is shown 100 times larger with AL2 for readability. We evaluate all metrics at different epochs and with different baselines, without/with AL2. The networks are trained on the CIFAR10 dataset with 75% corrupt labels.

Different metrics evaluated across training epochs (without/with AL2)								
Baseline	Metric	epoch=100	epoch=200	epoch=300	epoch=400	epoch=500	epoch=600	epoch=700
Bare	TA	53.69/54.93	36.25/61.55	31.94/63.78	32.24/62.51	31.84/61.28	31.99/60.68	31.88/60.46
	$\mathcal{L}_c$	1.98/2.10	1.07/2.02	0.02/1.96	0.00/1.91	0.00/1.88	0.00/1.88	0.00/1.91
	$\mathcal{L}_r$	2.35/0.34	14.51/0.16	64.54/0.09	73.26/0.05	84.72/0.02	91.89/0.01	97.15/0.00
BN [2]	TA	46.31/61.83	35.88/57.19	34.69/52.38	34.71/51.13	34.46/51.49	34.39/52.79	34.53/49.53
	$\mathcal{L}_c$	1.52/1.99	0.19/1.85	0.02/1.74	0.01/1.68	0.01/1.68	0.00/1.72	0.00/1.79
	$\mathcal{L}_r$	1.35/0.66	4.62/0.34	6.41/0.18	7.08/0.08	7.47/0.04	7.74/0.01	7.95/0.01
DO [3]	TA	54.37/51.83	60.83/59.06	63.01/61.95	58.81/61.51	50.10/62.11	47.01/60.85	44.70/60.21
	$\mathcal{L}_c$	2.10/2.13	2.02/2.07	1.92/2.04	1.69/2.02	1.33/2.02	1.06/2.03	0.91/2.06
	$\mathcal{L}_r$	0.20/0.33	0.25/0.14	0.51/0.07	1.44/0.03	2.84/0.02	3.70/0.01	3.98/0.00
WD [4]	TA	54.25/54.32	37.95/61.24	31.33/63.61	32.08/63.69	31.95/63.81	32.35/62.14	32.06/62.44
	$\mathcal{L}_c$	2.00/2.10	1.17/2.03	0.05/1.99	0.01/1.96	0.01/1.95	0.01/1.97	0.02/2.01
	$\mathcal{L}_r$	2.23/0.34	13.88/0.16	49.55/0.08	75.57/0.04	69.24/0.02	79.34/0.01	59.70/0.00

Table 10. Test accuracy (TA), training cross-entropy loss $\mathcal{L}_c$, and our training regularization loss $\mathcal{L}_r$ which is shown 100 times larger with AL2 for readability. We evaluate all metrics at different epochs and with different baselines, without/with AL2. The networks are trained on the CIFAR10 dataset with 50% corrupt labels.

Different metrics evaluated across training epochs (without/with AL2)								
Baseline	Metric	epoch=100	epoch=200	epoch=300	epoch=400	epoch=500	epoch=600	epoch=700
Bare	TA	59.63/64.04	45.79/68.32	47.61/67.19	47.62/66.10	47.52/65.77	47.59/65.14	47.40/65.81
	$\mathcal{L}_c$	1.43/1.65	0.28/1.50	0.01/1.40	0.00/1.33	0.00/1.31	0.00/1.33	0.00/1.40
	$\mathcal{L}_r$	4.83/0.69	29.10/0.32	73.63/0.17	72.89/0.08	82.03/0.04	88.06/0.02	92.60/0.01
BN [2]	TA	61.76/67.72	52.76/64.62	52.01/63.74	51.71/62.08	51.80/63.12	51.86/61.81	51.63/59.39
	$\mathcal{L}_c$	1.11/1.49	0.17/1.33	0.02/1.24	0.01/1.21	0.01/1.23	0.00/1.28	0.00/1.38
	$\mathcal{L}_r$	1.29/1.07	3.74/0.47	5.21/0.22	5.79/0.10	6.13/0.04	6.37/0.02	6.55/0.01
DO [3]	TA	62.72/61.13	68.15/67.70	70.29/68.90	68.43/68.48	65.54/68.66	64.32/66.68	63.34/63.98
	$\mathcal{L}_c$	1.70/1.72	1.55/1.61	1.39/1.57	1.12/1.55	0.85/1.55	0.67/1.58	0.56/1.64
	$\mathcal{L}_r$	0.29/0.66	0.38/0.27	0.84/0.13	1.90/0.06	2.97/0.03	3.64/0.01	3.97/0.01
WD [4]	TA	59.60/63.69	45.58/68.45	48.24/68.30	48.12/67.90	48.29/68.22	48.82/68.38	48.73/68.84
	$\mathcal{L}_c$	1.45/1.66	0.34/1.52	0.01/1.44	0.01/1.39	0.01/1.39	0.01/1.43	0.01/1.52
	$\mathcal{L}_r$	4.67/0.70	27.10/0.32	62.08/0.16	73.66/0.08	80.52/0.04	73.45/0.02	80.98/0.01

Table 11. Test accuracy (TA), training cross-entropy loss $\mathcal{L}_c$, and our training regularization loss $\mathcal{L}_r$ which is shown 100 times larger with AL2 for readability. We evaluate all metrics at different epochs and with different baselines, without/with AL2. The networks are trained on the CIFAR10 dataset with 25% corrupt labels.

Different metrics evaluated across training epochs (without/with AL2)								
Baseline	Metric	epoch=100	epoch=200	epoch=300	epoch=400	epoch=500	epoch=600	epoch=700
Bare	TA	66.29/69.74	66.94/73.11	66.64/73.24	66.59/73.75	66.56/72.31	66.56/70.18	66.63/70.52
	$\mathcal{L}_c$	0.52/0.80	0.02/0.58	0.00/0.49	0.00/0.46	0.00/0.48	0.00/0.54	0.00/0.63
	$\mathcal{L}_r$	8.51/1.27	39.95/0.52	58.82/0.24	67.64/0.11	73.39/0.05	77.64/0.02	81.03/0.01
BN [2]	TA	68.98/73.74	70.51/73.79	70.73/72.22	70.59/70.88	68.84/73.72	70.29/70.16	69.32/71.11
	$\mathcal{L}_c$	0.38/0.62	0.08/0.48	0.02/0.43	0.01/0.44	0.01/0.48	0.00/0.55	0.00/0.66
	$\mathcal{L}_r$	1.22/1.59	2.12/0.59	2.74/0.25	3.05/0.10	3.25/0.04	3.39/0.02	3.50/0.01
DO [3]	TA	68.65/68.31	74.17/72.49	76.48/74.08	77.13/73.91	77.08/73.27	77.55/70.06	77.64/68.91
	$\mathcal{L}_c$	0.93/0.94	0.71/0.78	0.52/0.73	0.37/0.72	0.25/0.74	0.19/0.80	0.16/0.87
	$\mathcal{L}_r$	0.40/1.23	0.59/0.47	1.17/0.21	1.99/0.09	2.80/0.04	3.37/0.02	3.84/0.01
WD [4]	TA	66.45/69.92	66.88/73.32	66.76/73.53	67.00/74.72	66.78/73.42	66.91/72.39	66.96/72.67
	$\mathcal{L}_c$	0.54/0.81	0.03/0.60	0.01/0.53	0.01/0.51	0.01/0.55	0.01/0.62	0.01/0.73
	$\mathcal{L}_r$	8.11/1.28	35.98/0.54	51.36/0.25	56.71/0.11	60.34/0.05	63.62/0.02	66.82/0.01

Table 12. Test accuracy (TA), training cross-entropy loss $\mathcal{L}_c$, and our training regularization loss $\mathcal{L}_r$ which is shown 100 times larger with AL2 for readability. We evaluate all metrics at different epochs and with different baselines, without/with AL2. The networks are trained on the CIFAR10 dataset with 0% corrupt labels.