Sample-wise Label Confidence Incorporation for Learning with Noisy Labels: Supplementary Material

Chanho Ahn Kikyung Kim Ji-won Baek Jongin Lim Seungju Han
Samsung Advanced Institute of Technology (SAIT), Korea

A. Proof of Theorem 1

Lemma 1. *Given the monotonically increasing function $I : \mathbb{R} \to \mathbb{R}^+$, the mean of a sequence is always less than or equal to the mean weighted by I, as follows:*

$$\frac{1}{n}\sum_{i=1}^{n} x_i \leq \frac{1}{\sum_{i=1}^{n} I(x_i)} \sum_{i=1}^{n} I(x_i) \cdot x_i.$$

Proof. This can be trivially proved by mathematical induction on the length of the given sequence. □

Corollary 1. *In the case where the function, h, which maps the robust loss value to label confidence is a monotonically decreasing function, the weighted loss function with negative label confidence incorporation has the following lower bound:*

$$\begin{aligned}\sum_{(x,y)\in\mathcal{D}} (1 - P((x,y) \in \mathcal{D}_{ce}))\mathcal{L}(f(x;\theta),y) &= \sum_{(x,y)\in\mathcal{D}} (1 - h(\mathcal{L}(f(x;\theta),y))\mathcal{L}(f(x;\theta),y) \\ &\geq \frac{\sum_{(x,y)\in\mathcal{D}}(1 - P((x,y) \in \mathcal{D}_{ce}))}{|\mathcal{D}|} \cdot \sum_{(x,y)\in\mathcal{D}} \mathcal{L}(f(x;\theta),y).\end{aligned}$$

The negative label confidence is denoted as $1 - P((x,y) \in \mathcal{D}_{ce})$. Since the negative label confidence is a monotonically increasing function, based on Lemma 1, the mean loss function can serve as the lower bound for the weighted loss function.

Lemma 2. *Let's suppose that the loss function of the classification problem is defined as a function of the prediction probability for the corresponding label as follows: $\mathcal{L}(f(x;\theta),y) := g(f(x;\theta)_y)$. For $\alpha > 0$ which satisfies $\lim_{p\to 1} \nabla_p(g(p) + \alpha \log p) < 0$, there exists $\tau < 1$ that satisfies the following proposition:*

$$h(g(p)) = 0, \ \textit{for } p < \tau \ \Rightarrow \sum_{(x,y)\in\mathcal{D}} P((x,y) \in \mathcal{D}_{ce}) \cdot \mathcal{L}(f(x;\theta),y) \geq \sum_{(x,y)\in\mathcal{D}} \alpha \cdot P((x,y) \in \mathcal{D}_{ce}) \cdot CE(f(x;\theta),y).$$

Proof. Without loss of generality, the noise-robust loss function has a minimum value 0 when the prediction probability of the corresponding label is 1. For simplicity, let's use an auxiliary variable p to represent the difference between the noise-robust loss and the cross-entropy loss as follows:

$$\mathcal{L}(f(x;\theta),y) - \alpha \cdot \mathrm{CE}(f(x;\theta),y) = g(p) + \alpha \log(p) := r(p),$$

where $p = f(x;\theta)_y$. The difference function $r(\cdot)$ satisfies the following propositions based on our assumptions: $r(1) = 0$, and $\lim_{p\to 1} r'(p) < 0$. This implies that there exists $\tau < 1$ satisfying $r(p) \geq 0$ for $p \in [\tau, 1]$. Assume that $\mathrm{P}((x,y) \in \mathcal{D}_{ce}) = h(g(f(x;\theta)_y)) = 0$ for $f(x;\theta)_y < \tau$, then the following inequality holds:

$$\begin{aligned}&\sum_{(x,y)\in\mathcal{D}} \mathrm{P}((x,y) \in \mathcal{D}_{ce}) \cdot (\mathcal{L}(f(x;\theta),y) - \alpha \cdot \mathrm{CE}(f(x;\theta),y)) \\ &\quad = \sum_{(x,y)\in\mathcal{D}} \mathbb{I}(f(x;\theta)_y \geq \tau) \cdot \mathrm{P}((x,y) \in \mathcal{D}_{ce}) \cdot (\mathcal{L}(f(x;\theta),y) - \alpha \cdot \mathrm{CE}(f(x;\theta),y)) \geq 0\end{aligned}$$

where $\mathbb{I}$ is an indicator function. □

Lemma 2 becomes meaningless if τ is close to 1, but τ may not be close to 1. For example, in the case of a generalized cross entropy loss [2] with a hyper-parameter of 0.5 ($g(p) = (1 - p^{0.5})/0.5$), $\tau \approx 0.2128$ for $\alpha = 0.5$. The gradient of the cross-entropy loss increases rapidly as the predictive probability value for the corresponding label is closer to 0. Therefore, the cross-entropy loss cannot always be smaller than the noise-robust loss containing underfitting issue. This is why the threshold condition exists in the proposed label confidence. Actually, Lemma 2 assumes τ forcing the cross-entropy loss to be smaller than the noise-robust loss in the range of consideration. Consequently, using Corollary 1 and Lemma 2, we can prove our theorem in the paper as follows:

Theorem 1. *Let us assume that $\mathcal{L}(f(x;\theta), y) := g(f(x;\theta)_y)$, where $g : [0,1] \to \mathbb{R}^+$. Given $\alpha > 0$ such that $\lim_{p\to 1} \nabla_p(g(p) + \alpha \log p) < 0$, there exists a value of $\tau < 1$ that satisfies the following inequality:*

$$\mathcal{R}_{\mathcal{L}}(\theta;\mathcal{D}) \geq \frac{n - |\mathcal{D}_{ce}|}{n}\mathcal{R}_{\mathcal{L}}(\theta;\mathcal{D}) + \alpha\frac{|\mathcal{D}_{ce}|}{n}\mathcal{R}_{CE}(\theta;\mathcal{D}_{ce}),$$

where n is the number of samples in $\mathcal{D}$ and $h(\cdot)$ is a monotonically decreasing function which satisfies $h(l) = 0$ for $l > g(\tau)$.

Proof. For the τ defined in Lemma 2, the following inequalities hold:

$$\begin{aligned}
n \cdot \text{LHS} = n \cdot \mathcal{R}_{\mathcal{L}}(\theta;\mathcal{D}) &= \sum_{(x,y)\in\mathcal{D}} \mathcal{L}(f(x;\theta), y) \\
&= \sum_{(x,y)\in\mathcal{D}} \text{P}((x,y) \in \mathcal{D}_{ce}) \cdot \mathcal{L}(f(x;\theta), y) + (1 - \text{P}((x,y) \in \mathcal{D}_{ce})) \cdot \mathcal{L}(f(x;\theta), y) \\
&\geq \sum_{(x,y)\in\mathcal{D}} \text{P}((x,y) \in \mathcal{D}_{ce}) \cdot \mathcal{L}(f(x;\theta), y) + \frac{\sum_{(x,y)\in\mathcal{D}} 1 - \text{P}((x,y) \in \mathcal{D}_{ce})}{n} \cdot \sum_{(x,y)\in\mathcal{D}} \mathcal{L}(f(x;\theta), y) \\
&\geq \sum_{(x,y)\in\mathcal{D}} \alpha \cdot \text{P}((x,y) \in \mathcal{D}_{ce}) \cdot \text{CE}(f(x;\theta), y) + (n - |\mathcal{D}_{ce}|) \cdot \frac{1}{n} \cdot \sum_{(x,y)\in\mathcal{D}} \mathcal{L}(f(x;\theta), y) \\
&= \alpha \cdot |\mathcal{D}_{ce}| \cdot \mathcal{R}_{\text{CE}}(\theta;\mathcal{D}_{ce}) + (n - |\mathcal{D}_{ce}|) \cdot \mathcal{R}_{\mathcal{L}}(\theta;\mathcal{D}) = n \cdot \text{RHS}.
\end{aligned}$$

The third line of the proof has validity by Corollary 1 and the fourth line holds by Lemma 2. □

τ can be simply calculated when the base noise-robust loss and α are determined. In addition, theoretically the proposed method requires satisfying the strict thresholding based on the τ, but the soft-thresholding which makes the algorithm simple works strongly in our experiments.

B. Detailed loss function

In our paper, we describe the final loss as a combination of robust loss, weighted cross-entropy loss with label confidence incorporation, penalty loss for the distance between two models and augmentation-invariant regularizer. Specifically, we can formulate the final loss function as follows:

$$\sum_{(x,y)\in\mathcal{D}} \left[\frac{1}{n}\mathcal{L}(f(x;\theta), y) + \alpha\frac{h(f(x;\theta)_y)}{n - |\mathcal{D}_{ce}|}\text{CE}(f(x;\theta^*), \tilde{y}) + \rho \cdot \text{JSD}(f(x';\theta^*), f(x'';\theta^*))\right] + \lambda||\theta - \theta^*||_F^2,$$
$$s.t.\ \ h(f(x;\theta)_y) = \sigma(0.5 \cdot (-\mathcal{L}(f(x;\theta), y) + \mu + m)),$$

where JSD represents Jensen-Shannon Divergence between two predictions which is the augmentation-invariant regularizer, $||\cdot||_F^2$ denotes the Frobenius norm, x' and x'' are two different images generated by applying random transformation to the input image x, and $\tilde{y}$ is the label which combines the ground-truth label and the label expected by the noise-robust model. While the paper presents a milestone to combine the noise-robust loss and the cross-entropy loss, the proposed framework includes four hyper-parameters (α, ρ, λ, and m).

To simplify the application of our algorithm, we fix two hyperparameters which do not significantly affect performance. According to the findings in the paper (Section 4.1), the proposed framework shows consistent performance even when the influence of cross-entropy loss changes. Based on this observation, we fix two hyperparameters that contribute to the influence of the cross-entropy loss: α and m. Considering α and threshold τ independently, as α increases, the effect of

cross-entropy also increases. Similarly, as m increases, more samples are trained by cross-entropy and the effect of cross-entropy increases. Since m is a variable that directly controls the criteria in our soft-thresholding, it is required to verify whether it can be determined freely from α. Fortunately, we can replace α with 1 when α and a certain threshold satisfy the theoretical conditions, because the following inequality holds: $\mathcal{L}(f(x;\theta),y) + \alpha W \cdot \mathrm{CE}(f(x;\theta^*),y) \geq \alpha \cdot (\mathcal{L}(f(x;\theta),y) + W \cdot \mathrm{CE}(f(x;\theta^*),y))$. However, note that, as the threshold increases, the approximation of the proposed framework becomes inaccurate, increasing the gap between the upper and lower bounds. It was experimentally confirmed in the paper that the variable m that affects the threshold size does not have significant effect on the overall performance, and we fix the value at 0.05 which is the central value of the effective settings in the experiments (Section 4.1).

C. Hyperparameters of our method in the experimental scenarios

Table 1. **Hyperparameters of the proposed method in our experimental scenarios.** 'NR' denotes the symmetric noise ratio.

Hyper-parameter	CIFAR-100 (NR $\geq$ 60 %)	CIFAR (others)	mini-WebVision	Clothing1M
Batch size	128	128	64	64
Learning rate	0.2	0.1	0.1	0.001
Weight decay	2e-5	1e-4	1e-4	1e-4
λ	1e-4	1e-3	1e-4	1e-4
ρ	10	5	5	10

In Table 1, we report the hyperparameters of the proposed method for the experiments represented in the paper. Hyperparameters include the basic elements of general deep learning model (batch size, learning rate, and weight decay). The specific hyperparameters of our method include weighting factors for the augmentation-invariant regularization, ρ, and the penalty of the parameter distance between two models, λ. λ was selected from 1, 5, and 10 times of the weight decay and ρ was selected from 1, 5, and 10. For the hyperparameter of the baseline noise-robust loss (GCE [2]) used by the proposed method, we followed the same value in [1] for the CIFAR experiments. For mini-WebVision and Clothing1M experiments, 0.5 was used for the hyperparameter of GCE.

D. Ablation studies

Table 2. **Comparison of classification accuracy (%) on noisy CIFAR-10 and CIFAR-100 datasets.** 'Single model' refers to the approach of solving the problem with just one model, while 'Two model' represents the approach where we remove the augmentation-invariant regularization and label correction from the proposed method. Similarly, 'GJS' can be interpreted as a method that adds augmentation-invariant regularization to the JS method. The best performance, excluding shaded cells, is highlighted in bold.

		no noise	symmetric noise				asymmetric noise	
Dataset	Method	0	20	40	60	80	20	40
CIFAR-10	GCE [2]	95.75	94.24	92.82	89.37	79.19	92.83	87.00
	JS [1]	**95.89**	**94.52**	93.01	89.64	76.06	92.18	87.99
	Single model	94.22	86.56	84.26	80.62	78.10	84.22	90.54
	Two models	94.98	93.71	**93.10**	**90.80**	**81.49**	**93.38**	**91.20**
	GJS [1]	95.91	95.33	93.57	91.64	79.11	93.94	89.65
	Ours	96.10	95.78	95.47	94.47	91.13	95.68	93.17
CIFAR-100	GCE [2]	77.65	75.02	71.54	65.21	49.68	72.13	51.50
	JS [1]	77.95	**75.41**	71.12	64.36	45.05	71.70	49.36
	Single model	78.07	73.54	68.30	54.52	40.57	**73.45**	58.34
	Two models	**78.50**	75.14	**71.96**	**65.56**	**52.22**	72.50	**62.45**
	GJS [1]	79.27	78.05	75.71	70.15	44.49	74.60	63.70
	Ours	79.40	78.21	75.82	71.28	61.05	77.08	68.05

In the experimental part of the main paper, one of our objectives was to rigorously assess the performance of the proposed learning model on a modular level. In pursuit of this objective, we conducted a thorough performance evaluation of two distinctive variations: the method which omits the augmentation-invariant regularizer and label correction ('Two models'), and the approach that refrains from employing the auxiliary model altogether ('Single model'). To facilitate a more

comprehensive and insightful analysis, we systematically explored a wide range of noise rates within our experimental framework. Furthermore, we extended our investigation to encompass a comparative assessment involving the GJS method [1], an augmentation-invariant regularized variant of the Jensen-Shannon loss (JS).

The performance comparison is delineated in Table 2. Notably, the optimization methodology of the proposed approach exhibits robust performance when confronted with high noise rates or asymmetric noise distributions, as evident from evaluations on both the CIFAR-10 and CIFAR-100 datasets. However, performance degradation are observed in cases where the single model is adopted. This phenomenon finds correlation with the empirical findings articulated in the paper, emphasizing that a simple combination of cross-entropy loss and noise-robust loss within a single model does not yield commensurate outcomes. A pivotal aspect lies in the distinct behavioral patterns exhibited by the two models proposed in our framework; the noise-free model and the noise-robust model. Despite parallel training dynamics on the training set, nuanced performance disparities emerge on the test set, as depicted in Figure 3 in the main paper. This nuanced discrepancy offers credence to the central premise of the paper, advocating that divergent model structures tailored to disparate objectives and characteristics prove to be more efficacious. Also, it is noteworthy that the divergence in performance between the 'Two models' approach and the proposed method occasionally exceeds the performance contrast between the JS method and the GJS method. This phenomenon can be interpreted as an efficacious response to the application of a regularizer to the cross-entropy loss, mitigating the potential perils of overfitting.

References

[1] Erik Englesson and Hossein Azizpour. Generalized jensen-shannon divergence loss for learning with noisy labels. *Neural Information Processing Systems (NIPS)*, 34:30284–30297, 2021. 3, 4

[2] Zhilu Zhang and Mert Sabuncu. Generalized cross entropy loss for training deep neural networks with noisy labels. *Neural Information Processing Systems (NIPS)*, 31, 2018. 2, 3