http://lamda.nju.edu.cn

Towards Understanding Deep Learning from Noisy Labels with Small-Loss Criterion

Xian-Jin Gui, Wei Wang and Zhang-Hao Tian
National Key Laboratory for Novel Software Technology
Nanjing University, China

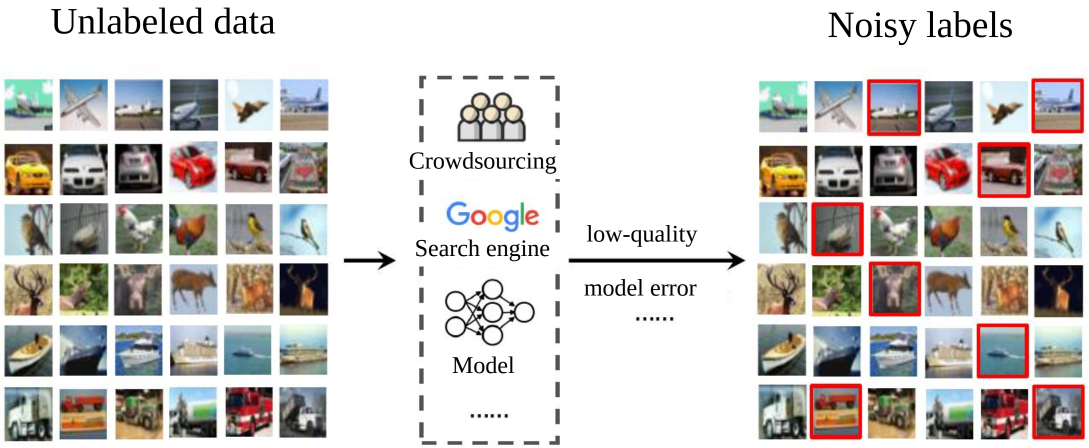
{guixj, wangw, tianzh}@lamda.nju.edu.cn

Background

Deep learning relies on large-scale data to achieve success.

In real-world applications, labels are usually collected from non-experts such as crowdsourcing.

1.2 million, Crowdsourcing, 2.5 years



- ✓ Quick
- ✓ Cheap
- ✓ Easy access
- ✓ ……

However, these annotation means will unavoidably incur many noisy labels.

The performance of models may be severely hurt if these noisy labels are blindly used.

Related work

- Noise-robust loss functions
  - mean absolute error [Ghosh *et al.*, AAAI’17]
  - information-theoretic loss [Xu *et al.*, NeurIPS’19]
- Loss correction
  - auxiliary network [Jacob *et al.*, ICLR’17]
  - unbiased loss term based on T [Patrini *et al.*, CVPR’17]
- Label correction
  - pseudo labels [Ma *et al.*, ICML’18]
  - joint optimization [Yi *et al.*, CVPR’19]
- Sample selection
  - selecting a part of clean data based on small-loss criterion [Han *et al.*, NeurIPS’18; Jiang *et al.*, ICML’18; Wei *et al.*, CVPR’20; Yu *et al.*, ICML’19]

Sample selection strategy with the small-loss criterion has been widely used.

Related work

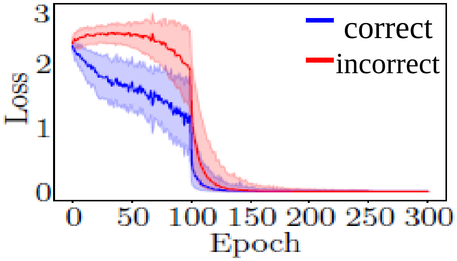
Small-loss criterion：

- select a part of examples with small loss as low-risk examples
- train models with the selected low risk examples

The small-loss criterion has been widely used and achieved prominent performance.

Experimental phenomena [Zhang *et al.*, ICLR'17, Aprit *et al.*, ICML'18]：

In the training process, the examples with correct labels tend to have smaller loss than that with incorrect labels.



But there are few theoretical analyses to explain why the small-loss criterion works.

Our contribution

- We theoretically explain why the widely-used small-loss criterion works.
- Based on the explanation, we reformalize the vanilla small-loss criterion to select examples.
- We also carefully set the selected number for each class to alleviate class imbalance in the sample selection process.
- We introduce semi-supervised learning to further exploit the unselected examples.

Preliminaries

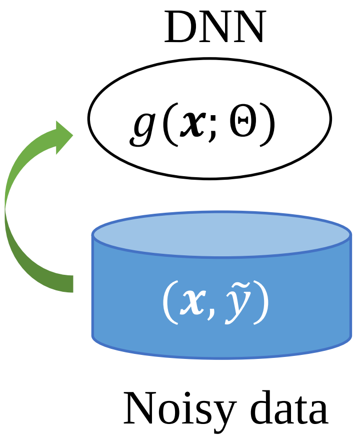
Small-loss criterion：

- select a part of examples with small loss as low-risk examples
- train models with the selected low-risk examples

- Neural network： $g(\boldsymbol{x};\Theta):\mathcal{X}\rightarrow\mathbb{R}^c$, with output $[\hat{p}_1(\boldsymbol{x}),\ldots,\hat{p}_c(\boldsymbol{x})]^\top\in\mathbb{R}^c$

  where $\hat{p}_i(\boldsymbol{x})=\dfrac{\exp\left(\boldsymbol{w}_i^\top\phi(\boldsymbol{x};\boldsymbol{\theta})\right)}{\sum_{j=1}^{c}\exp\left(\boldsymbol{w}_j^\top\phi(\boldsymbol{x};\boldsymbol{\theta})\right)}$

- Loss functions：
  - 0-1 loss: $\ell_{01}(f(\boldsymbol{x}),\tilde{y})=\mathbb{I}[f(\boldsymbol{x})\neq\tilde{y}]$
  - Cross-entropy loss: $\ell_{CE}(g(\boldsymbol{x};\Theta),\tilde{y})=-\log(\hat{p}_{\tilde{y}}(\boldsymbol{x}))$



- Optimization objective:

$$\Theta^*=\arg\min_{\Theta}\mathbb{E}_{(\boldsymbol{x},\tilde{y})}[\ell_{CE}(g(\boldsymbol{x};\Theta),\tilde{y})]. \qquad (1)$$

Preliminaries

For observed samples $(\boldsymbol{x}, \tilde{y})$, its label $\tilde{y}$ may be different from its true label y

- Feature dependent noise: $p(\tilde{y}|\boldsymbol{x}, y)$
  - $\tilde{y}$ depend on both y and $\boldsymbol{x}$

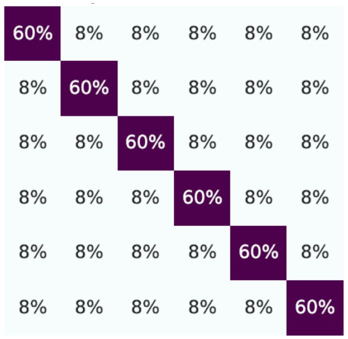
- Feature independent noise: $p(\tilde{y}|y)$
- $\tilde{y}$ only dependent on the true label y
- $p(\tilde{y}|\boldsymbol{x}, y) = p(\tilde{y}|y)$
- clean data $(\boldsymbol{x}, y)$ → noisy data $(\boldsymbol{x}, \tilde{y})$

$$p(\boldsymbol{x}, y) \quad \rightarrow \quad p(\boldsymbol{x}, \tilde{y}) = \sum_{i=1}^{c} \boxed{p(\tilde{y}|y=i)} \cdot p(\boldsymbol{x}, y)$$

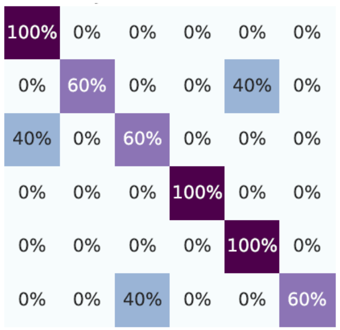
The noise transition matrix T

Noise transition matrix T

$$T_{ij} \overset{\text{def}}{=} p(\tilde{y} = j | y = i)$$



60%	8%	8%	8%	8%	8%
8%	60%	8%	8%	8%	8%
8%	8%	60%	8%	8%	8%
8%	8%	8%	60%	8%	8%
8%	8%	8%	8%	60%	8%
8%	8%	8%	8%	8%	60%



100%	0%	0%	0%	0%	0%
0%	60%	0%	0%	40%	0%
40%	0%	60%	0%	0%	0%
0%	0%	0%	100%	0%	0%
0%	0%	0%	0%	100%	0%
0%	0%	40%	0%	0%	60%

row-diagonally dominant: $T_{ii} > T_{ij}, \ \ \forall\, i, \forall\, j \neq i$

column-diagonally dominant: $T_{ii} > T_{ji}, \ \ \forall\, i, \forall\, j \neq i$

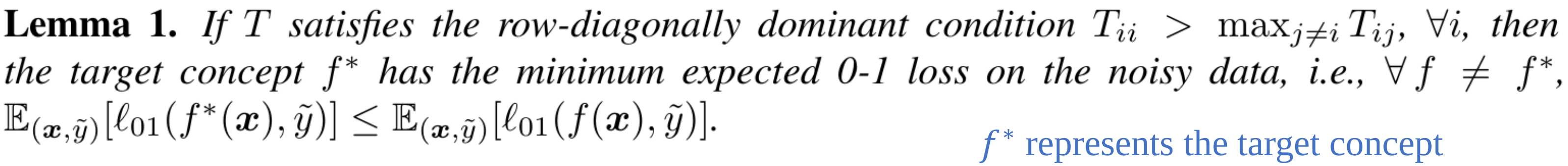
Our Work

How to answer the following questions:

Q1: Why and when does the small-loss criterion work?

Q2: What condition should the noise transition matrix T satisfy?

For the 0-1 loss function, we give Lemma 1:

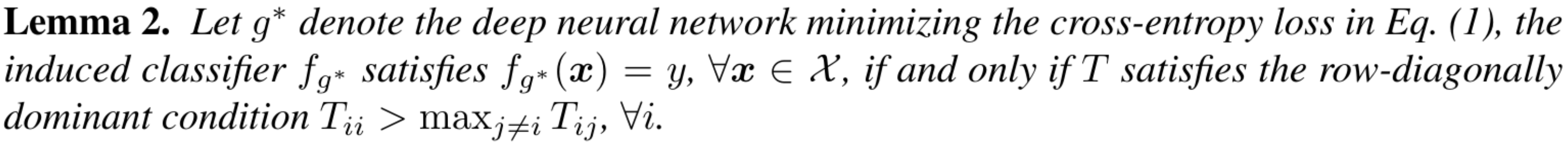


Lemma 1. *If T satisfies the row-diagonally dominant condition $T_{ii} > \max_{j \neq i} T_{ij}, \forall i$, then the target concept f^* has the minimum expected 0-1 loss on the noisy data, i.e., $\forall\, f \neq f^*$, $\mathbb{E}_{(\boldsymbol{x},\tilde{y})}[\ell_{01}(f^*(\boldsymbol{x}),\tilde{y})] \leq \mathbb{E}_{(\boldsymbol{x},\tilde{y})}[\ell_{01}(f(\boldsymbol{x}),\tilde{y})]$.*

f^* represents the target concept

row-diagonally dominant → f^* has the minimum expected 0-1 loss on the noisy data

Furthermore, for the cross-entropy loss, we give Lemma 2:

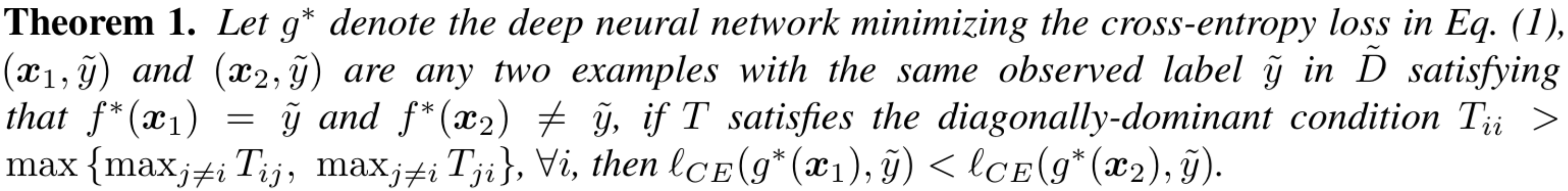


Lemma 2. *Let g^* denote the deep neural network minimizing the cross-entropy loss in Eq. (1), the induced classifier f_{g^*} satisfies $f_{g^*}(\boldsymbol{x}) = y, \forall \boldsymbol{x} \in \mathcal{X}$, if and only if T satisfies the row-diagonally dominant condition $T_{ii} > \max_{j \neq i} T_{ij}, \forall i$.*

row-diagonally dominant → good classifiers could be learned by minimizing the expected cross-entropy loss on the noisy data.

Our work

For the small-loss criterion, we give Theorem 1:



Theorem 1. *Let g^* denote the deep neural network minimizing the cross-entropy loss in Eq. (1), $(\boldsymbol{x}_1, \tilde{y})$ and $(\boldsymbol{x}_2, \tilde{y})$ are any two examples with the same observed label $\tilde{y}$ in $\tilde{D}$ satisfying that $f^*(\boldsymbol{x}_1) = \tilde{y}$ and $f^*(\boldsymbol{x}_2) \neq \tilde{y}$, if T satisfies the diagonally-dominant condition $T_{ii} > \max\{\max_{j \neq i} T_{ij},\ \max_{j \neq i} T_{ji}\}, \forall i$, then $\ell_{CE}(g^*(\boldsymbol{x}_1), \tilde{y}) < \ell_{CE}(g^*(\boldsymbol{x}_2), \tilde{y})$.*

diagonally-dominant → for g^*, examples with correct labels have smaller loss than that with incorrect labels

Theorem 1 implies that if T satisfies the diagonally-dominant condition：

- for the examples with the same observed label, the correct examples have smaller loss than the incorrect ones..
- single epoch’s loss value may not be reliable for sample selection.

Theorem 1 only focuses on the g^*which minimizes the expected cross-entropy loss.

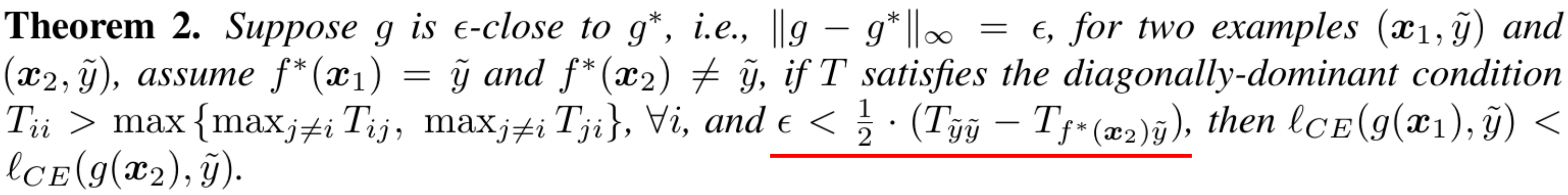
In practice, for a warmed-up neural network g, whether the small-loss criterion still works?

Our work

The small-loss criterion in practice：

1. warm up the model g on the whole noisy dataset with some epochs
2. then select small-loss examples and use them to update models

For this process, we have：



Theorem 2. *Suppose g is ϵ-close to g^*, i.e., $\|g - g^*\|_\infty = \epsilon$, for two examples $(\boldsymbol{x}_1, \tilde{y})$ and $(\boldsymbol{x}_2, \tilde{y})$, assume $f^*(\boldsymbol{x}_1) = \tilde{y}$ and $f^*(\boldsymbol{x}_2) \neq \tilde{y}$, if T satisfies the diagonally-dominant condition $T_{ii} > \max\{\max_{j \neq i} T_{ij}, \max_{j \neq i} T_{ji}\}, \forall i$, and $\underline{\epsilon < \frac{1}{2} \cdot (T_{\tilde{y}\tilde{y}} - T_{f^*(\boldsymbol{x}_2)\tilde{y}})}$, then $\ell_{CE}(g(\boldsymbol{x}_1), \tilde{y}) < \ell_{CE}(g(\boldsymbol{x}_2), \tilde{y})$.*

Theorem 2 implies that if the model g is not far away from g^*(ϵ is not too large):

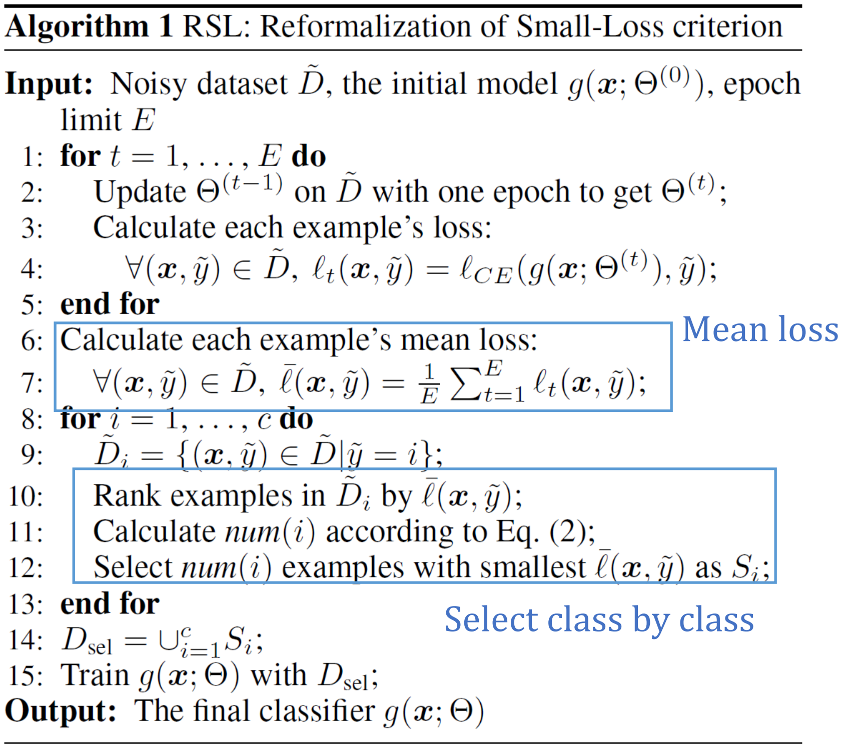
for the examples with the same observed labels, the correct examples still have smaller loss than the incorrect ones.

This explains why the small-loss criterion works in practice.

Our work

Based on the theoretical analysis, we reformalize the vanilla small-loss criterion:

- use the *mean loss* of each example along the training process to select samples
- select the examples with small mean loss *class by class*



```
Algorithm 1 RSL: Reformalization of Small-Loss criterion
Input: Noisy dataset $\tilde{D}$, the initial model $g(\boldsymbol{x}; \Theta^{(0)})$, epoch limit $E$
for $t = 1, \ldots, E$ do
    Update $\Theta^{(t-1)}$ on $\tilde{D}$ with one epoch to get $\Theta^{(t)}$;
    Calculate each example's loss:
        $\forall (\boldsymbol{x}, \tilde{y}) \in \tilde{D},\ \ell_t(\boldsymbol{x}, \tilde{y}) = \ell_{CE}(g(\boldsymbol{x}; \Theta^{(t)}), \tilde{y})$;
end for
Calculate each example's mean loss:
    $\forall (\boldsymbol{x}, \tilde{y}) \in \tilde{D},\ \bar{\ell}(\boldsymbol{x}, \tilde{y}) = \frac{1}{E}\sum_{t=1}^{E} \ell_t(\boldsymbol{x}, \tilde{y})$;
for $i = 1, \ldots, c$ do
    $\tilde{D}_i = \{(\boldsymbol{x}, \tilde{y}) \in \tilde{D} | \tilde{y} = i\}$;
    Rank examples in $\tilde{D}_i$ by $\bar{\ell}(\boldsymbol{x}, \tilde{y})$;
    Calculate $num(i)$ according to Eq. (2);
    Select $num(i)$ examples with smallest $\bar{\ell}(\boldsymbol{x}, \tilde{y})$ as $S_i$;
end for
$D_{\text{sel}} = \cup_{i=1}^{c} S_i$;
Train $g(\boldsymbol{x}; \Theta)$ with $D_{\text{sel}}$;
Output: The final classifier $g(\boldsymbol{x}; \Theta)$
```

Mean loss

Select class by class

Selection number $num(i)$:

- first introduce parameter $\beta \geq 0$ to make $prop(i)$ less than $1 - \eta_i$:

$$prop(i) = max\{1 - (1 + \beta)\eta_i, (1 - \beta)(1 - \eta_i)\}$$

Issue: $[prop(1) \cdot n_1, \cdots prop(c) \cdot n_c]$ may seriously deviate from the true class distribution $[p_1, \cdots, p_c]$.

- set the selected data as $[p_1 \cdot m, \ldots, p_c \cdot m]$ to obey $[p_1, \cdots, p_c]$:

$$m = \min_{1 \leq i \leq c} \{prop(i) \cdot n_i / p_i\}$$

Issue: too many useful data may be wasted.

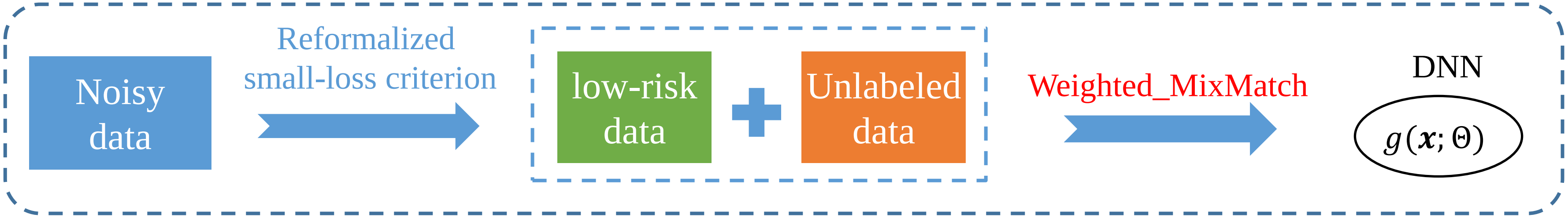
- additionally introduce parameter $\gamma \geq 1$:

$$num(i) = \min\{\gamma \cdot p_i \times m, prop(i) \times n_i\}$$

Our work

The overall process:

- identify low-risk examples with the reformalized small-loss criterion
- treat low-risk examples as clean labeled data, and the rest as unlabeled data
- simultaneously exploit clean and unlabeled data with Weighted_MixMatch



Noisy data

Reformalized small-loss criterion

low-risk data

Unlabeled data

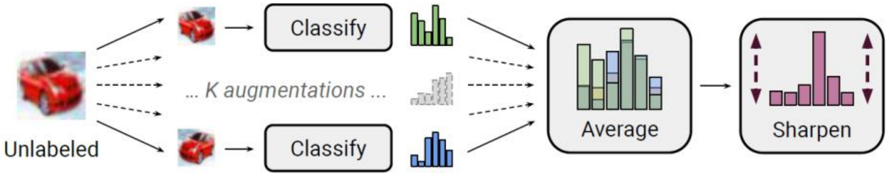
Weighted_MixMatch

DNN

$g(\boldsymbol{x};\Theta)$

Weighted_MixMatch

consistency regularization



Issue: the low-risk examples may still have label noise

Solution: reweigh the low-risk examples

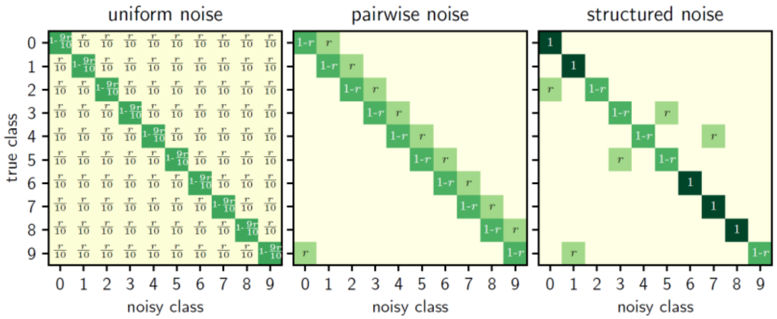
$$w(\boldsymbol{x},\tilde{y}) = \exp\Big(-\kappa\frac{\bar{\ell}(\boldsymbol{x},\tilde{y})-\ell_*(i)}{\ell^*(i)-\ell_*(i)}\Big)$$

Experimental setups

Datasets:

- Noisy CIFAR-10: uniform/pairwise/structured noise
- Noisy CIFAR-100: uniform/pairwise noise
- WebVision [Li *et al.*, ECCV’17]:
  - 50 classes, 2.4 million pictures
  - noise rate is about 20%

noise transition matrix



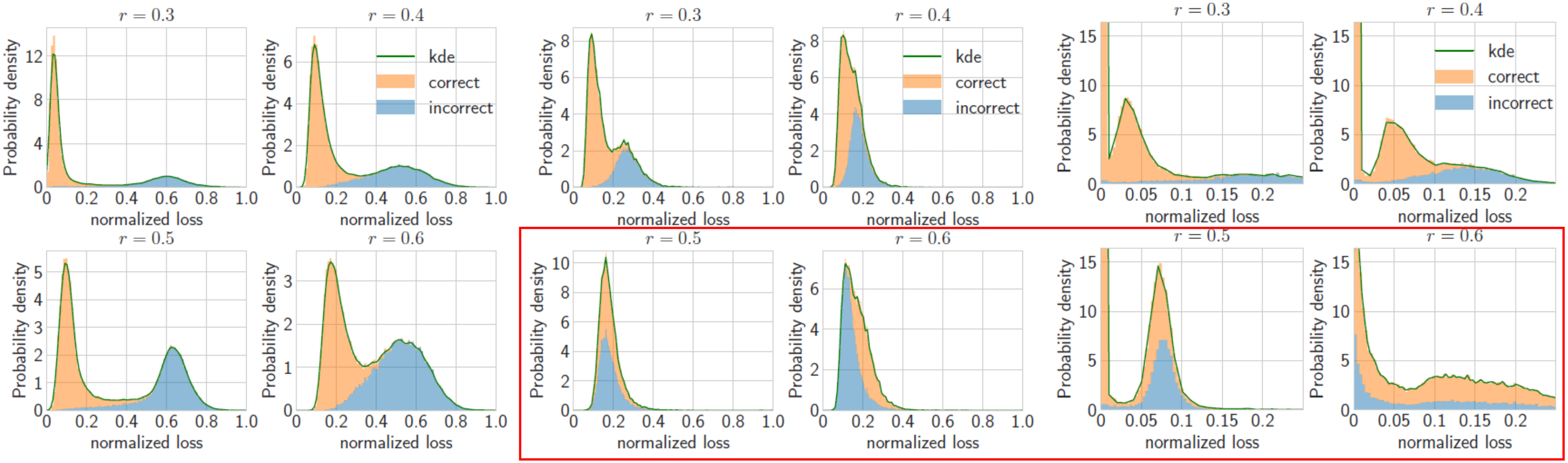
Three types of label noise on CIFAR-10

Baselines:

- Methods based on sample selection:
  - Co-teaching [Han *et al.*, NeurIPS’18]
  - Co-teaching+ [Yu *et al.*, ICML’19]
  - INCV [Chen *et al.*, ICML’19]
  - JoCoR [Wei *et al.*, CVPR’20]
- Methods based on robust loss function:
  - Truncated $\mathcal{L}_q$ [Zhang *et al.*, NeurIPS’18]
  - $\mathcal{L}_{DMI}$ [Xu *et al.*, NeurIPS’19]

Experimental results

The diagonally-dominant condition is necessary for small-loss criterion



(a) uniform label noise (b) pairwise label noise (c) structured label noise

When the diagonally-dominant condition is not satisfied, many incorrect examples (blue) may even have smaller loss than correct ones (yellow), see subfigure (b) $r = 0.5$ and $r = 0.6$, (c) $r = 0.5$ and $r = 0.6$.

Experimental results

The loss of correct examples is smaller than the loss of incorrect ones.

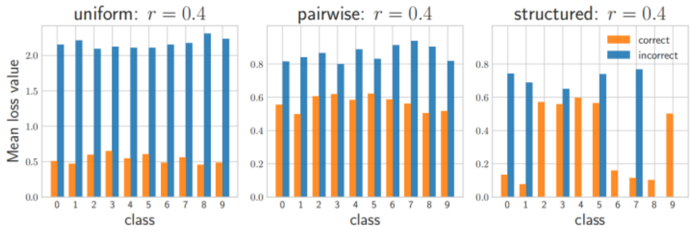


Figure 3: Mean values of the mean loss of correct examples and incorrect ones for each class. For structured noise ($r = 0.4$), some classes do not have label noise.

The mean loss is more stable than single epoch's loss.

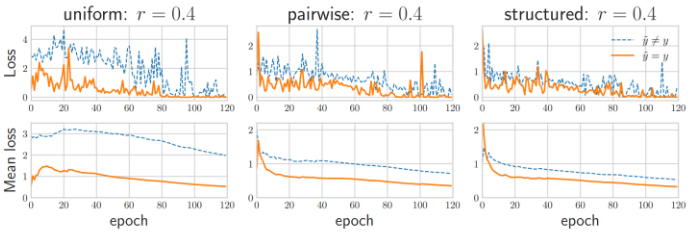
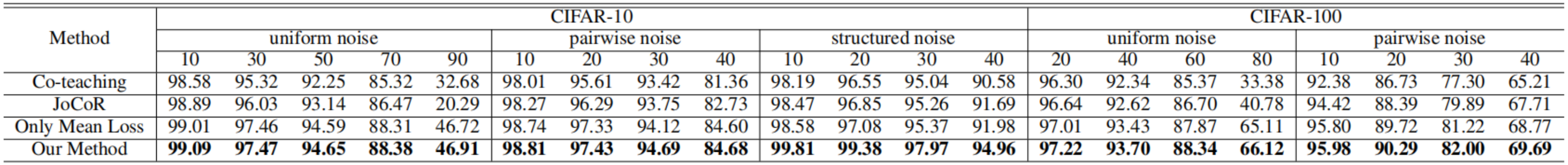


Figure 4: Each epoch's loss and the cumulative mean loss for randomly chosen one pair of correct example and incorrect example. Additional figures of other pairs can be found in Appendix B.

The precision of the examples selected by our method is higher.

Table 1: The precision (%) of the selected data with different methods on CIFAR-10 and CIFAR-100. "Only Mean Loss" represents using mean loss but not selecting examples class by class. "Our Method" represents using mean loss and selecting examples class by class.



Method	CIFAR-10													CIFAR-100							
	uniform noise					pairwise noise				structured noise				uniform noise				pairwise noise			
	10	30	50	70	90	10	20	30	40	10	20	30	40	20	40	60	80	10	20	30	40
Co-teaching	98.58	95.32	92.25	85.32	32.68	98.01	95.61	93.42	81.36	98.19	96.55	95.04	90.58	96.30	92.34	85.37	33.38	92.38	86.73	77.30	65.21
JoCoR	98.89	96.03	93.14	86.47	20.29	98.27	96.29	93.75	82.73	98.47	96.85	95.26	91.69	96.64	92.62	86.70	40.78	94.42	88.39	79.89	67.71
Only Mean Loss	99.01	97.46	94.59	88.31	46.72	98.74	97.33	94.12	84.60	98.58	97.08	95.37	91.98	97.01	93.43	87.87	65.11	95.80	89.72	81.22	68.77
Our Method	**99.09**	**97.47**	**94.65**	**88.38**	**46.91**	**98.81**	**97.43**	**94.69**	**84.68**	**99.81**	**99.38**	**97.97**	**94.96**	**97.22**	**93.70**	**88.34**	**66.12**	**95.98**	**90.29**	**82.00**	**69.69**

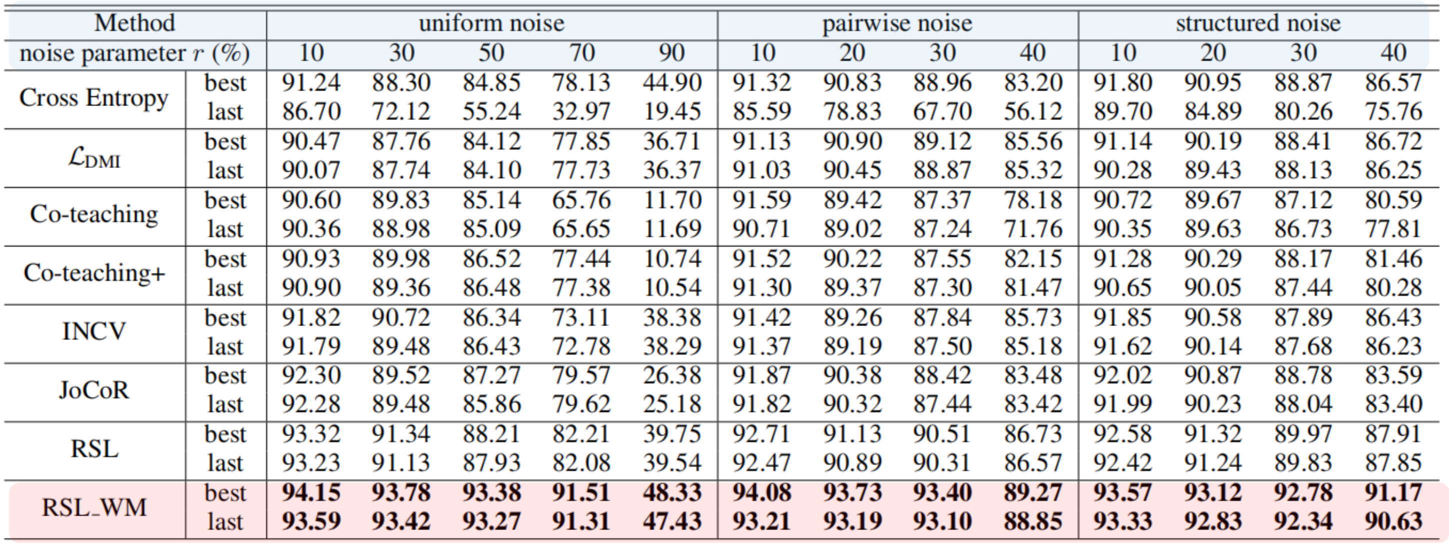
Ablation study: the precision of the selected examples on CIFAR-10/100 datasets

Experimental results

Our method achieves better performance compared with all baselines.

Table 2: The accuracy (%) results on CIFAR-10, where "best" means the test accuracy of the epoch when validation accuracy is maximum, and "last" means the test accuracy of the last epoch.

different noise setting



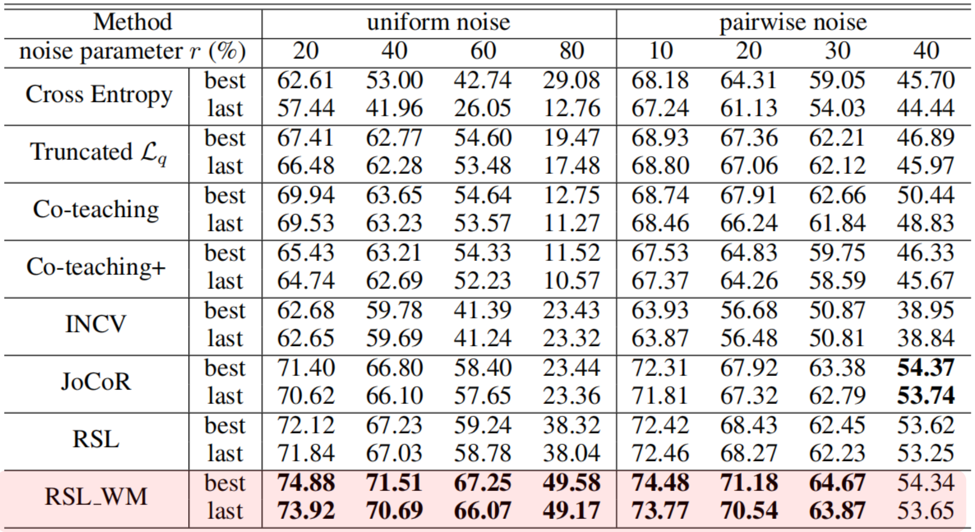
Method		uniform noise					pairwise noise				structured noise			
noise parameter r (%)		10	30	50	70	90	10	20	30	40	10	20	30	40
Cross Entropy	best	91.24	88.30	84.85	78.13	44.90	91.32	90.83	88.96	83.20	91.80	90.95	88.87	86.57
	last	86.70	72.12	55.24	32.97	19.45	85.59	78.83	67.70	56.12	89.70	84.89	80.26	75.76
$\mathcal{L}_{\text{DMI}}$	best	90.47	87.76	84.12	77.85	36.71	91.13	90.90	89.12	85.56	91.14	90.19	88.41	86.72
	last	90.07	87.74	84.10	77.73	36.37	91.03	90.45	88.87	85.32	90.28	89.43	88.13	86.25
Co-teaching	best	90.60	89.83	85.14	65.76	11.70	91.59	89.42	87.37	78.18	90.72	89.67	87.12	80.59
	last	90.36	88.98	85.09	65.65	11.69	90.71	89.02	87.24	71.76	90.35	89.63	86.73	77.81
Co-teaching+	best	90.93	89.98	86.52	77.44	10.74	91.52	90.22	87.55	82.15	91.28	90.29	88.17	81.46
	last	90.90	89.36	86.48	77.38	10.54	91.30	89.37	87.30	81.47	90.65	90.05	87.44	80.28
INCV	best	91.82	90.72	86.34	73.11	38.38	91.42	89.26	87.84	85.73	91.85	90.58	87.89	86.43
	last	91.79	89.48	86.43	72.78	38.29	91.37	89.19	87.50	85.18	91.62	90.14	87.68	86.23
JoCoR	best	92.30	89.52	87.27	79.57	26.38	91.87	90.38	88.42	83.48	92.02	90.87	88.78	83.59
	last	92.28	89.48	85.86	79.62	25.18	91.82	90.32	87.44	83.42	91.99	90.23	88.04	83.40
RSL	best	93.32	91.34	88.21	82.21	39.75	92.71	91.13	90.51	86.73	92.58	91.32	89.97	87.91
	last	93.23	91.13	87.93	82.08	39.54	92.47	90.89	90.31	86.57	92.42	91.24	89.83	87.85
RSL_WM	best	**94.15**	**93.78**	**93.38**	**91.51**	**48.33**	**94.08**	**93.73**	**93.40**	**89.27**	**93.57**	**93.12**	**92.78**	**91.17**
	last	**93.59**	**93.42**	**93.27**	**91.31**	**47.43**	**93.21**	**93.19**	**93.10**	**88.85**	**93.33**	**92.83**	**92.34**	**90.63**

The performance on the CIFAR-10 dataset.

Experimental results

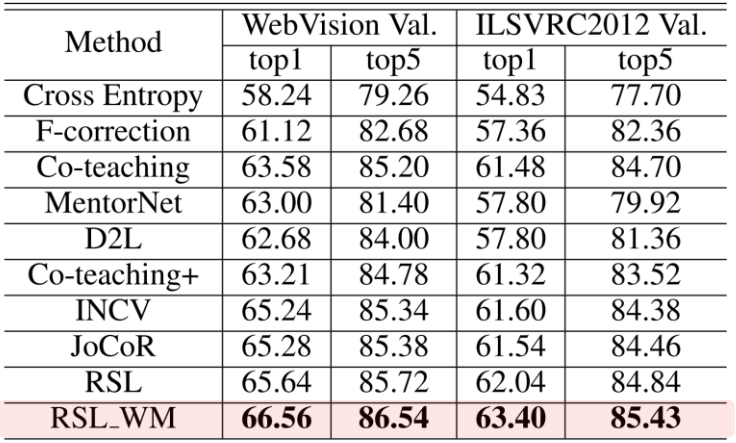
Our method achieves better performance in almost all settings compared with baselines.

Table 3: The accuracy (%) results on CIFAR-100.



Method		uniform noise				pairwise noise			
noise parameter r (%)		20	40	60	80	10	20	30	40
Cross Entropy	best	62.61	53.00	42.74	29.08	68.18	64.31	59.05	45.70
	last	57.44	41.96	26.05	12.76	67.24	61.13	54.03	44.44
Truncated $\mathcal{L}_q$	best	67.41	62.77	54.60	19.47	68.93	67.36	62.21	46.89
	last	66.48	62.28	53.48	17.48	68.80	67.06	62.12	45.97
Co-teaching	best	69.94	63.65	54.64	12.75	68.74	67.91	62.66	50.44
	last	69.53	63.23	53.57	11.27	68.46	66.24	61.84	48.83
Co-teaching+	best	65.43	63.21	54.33	11.52	67.53	64.83	59.75	46.33
	last	64.74	62.69	52.23	10.57	67.37	64.26	58.59	45.67
INCV	best	62.68	59.78	41.39	23.43	63.93	56.68	50.87	38.95
	last	62.65	59.69	41.24	23.32	63.87	56.48	50.81	38.84
JoCoR	best	71.40	66.80	58.40	23.44	72.31	67.92	63.38	**54.37**
	last	70.62	66.10	57.65	23.36	71.81	67.32	62.79	**53.74**
RSL	best	72.12	67.23	59.24	38.32	72.42	68.43	62.45	53.62
	last	71.84	67.03	58.78	38.04	72.46	68.27	62.23	53.25
RSL_WM	best	**74.88**	**71.51**	**67.25**	**49.58**	**74.48**	**71.18**	**64.67**	54.34
	last	**73.92**	**70.69**	**66.07**	**49.17**	**73.77**	**70.54**	**63.87**	53.65

Table 4: The accuracy (%) results on WebVision.



Method	WebVision Val.		ILSVRC2012 Val.	
	top1	top5	top1	top5
Cross Entropy	58.24	79.26	54.83	77.70
F-correction	61.12	82.68	57.36	82.36
Co-teaching	63.58	85.20	61.48	84.70
MentorNet	63.00	81.40	57.80	79.92
D2L	62.68	84.00	57.80	81.36
Co-teaching+	63.21	84.78	61.32	83.52
INCV	65.24	85.34	61.60	84.38
JoCoR	65.28	85.38	61.54	84.46
RSL	65.64	85.72	62.04	84.84
RSL_WM	**66.56**	**86.54**	**63.40**	**85.43**

The performance on the CIFAR-100 and WebVision datasets.

Conclusion

- We establish the connection between noisy data distribution and the small-loss criterion.

- Then we theoretically explain why the widely-used small-loss criterion works and reformalize the vanilla small-loss criterion.

- The experimental results verify our theoretical explanation and also demonstrate the effectiveness of the reformalization.

Our theoretical analysis also gives the following insights:

- the empirically diagonally-dominant condition is theoretically justified

- the loss value for examples with different labels are not comparable so the small-loss level should be determined class by class

- the warm-up stage is necessary for the small-loss criterion

Thank You!