To Smooth or Not?
When Label Smoothing Meets Noisy Labels

Jiaheng Wei, Hangyu Liu, Tongliang Liu,

Gang Niu, Masashi Sugiyama, and Yang Liu

UNIVERSITY OF CALIFORNIA SANTA CRUZ

BROWN

THE UNIVERSITY OF SYDNEY

RIKEN

AIP Center for Advanced Intelligence Project

東京大学 THE UNIVERSITY OF TOKYO

Correspondence to yangliu@ucsc.edu

Content

- Background
- Motivation
- Main Contributions
- Takeaways

Main contributions

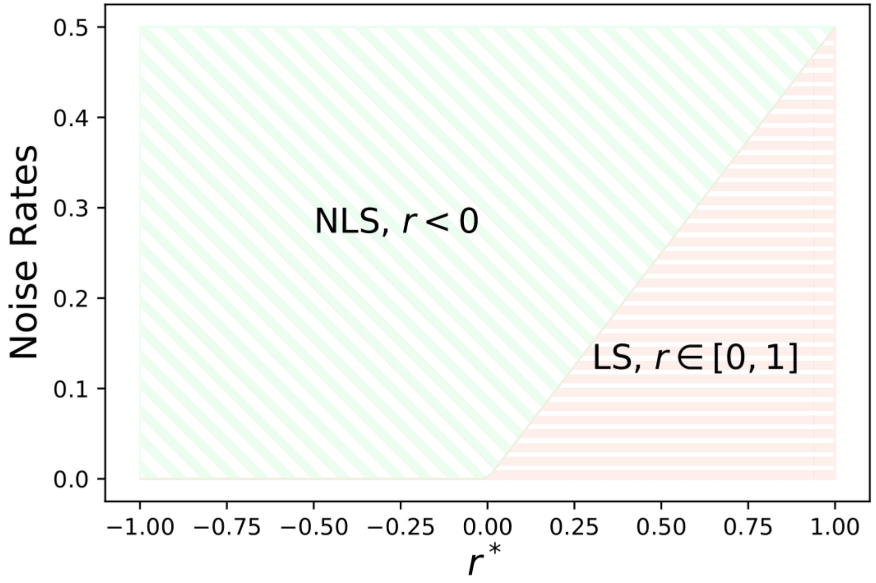
$$\mathbf{y}_i^{\mathrm{GLS},r} := (1-r)\cdot \mathbf{y}_i + \frac{r}{K}\cdot \mathbf{1},$$

We explore generalized label smoothing, where r could go negative (NLS):

1. NLS is beneficial when the label noise rate is high.

2. Build theoretical connections between NLS and existing robust methods.

3. We give empirical significances of the overlooked NLS.



The preferences between NLS, LS in binary classification task.

Background

Generalized label smoothing

Generalized label smoothing ($r < 1$)

$$\mathbf{y}_i^{\mathrm{GLS},r} := (1-r)\cdot \mathbf{y}_i + \frac{r}{K}\cdot \mathbf{1},$$

$\mathbf{y}_i$: the one-hot label of sample x_i; $\mathbf{1} = [1, 1, \dots, 1]^{\mathrm{T}}$: the all one vector; K: # of classes.

- *Hard label*: $r = 0$
  - i.e., $K = 3$, $\mathbf{y}_i^{\mathrm{GLS},r} = [0, 1, 0]^{\mathrm{T}}$;
  - Three elements indicate: class dog (1st), cat (2nd), deer (3rd), respectively.

Extended label distribution

Generalized label smoothing ($r < 1$)

$$\mathbf{y}_i^{\mathrm{GLS},r} := (1-r)\cdot\mathbf{y}_i + \frac{r}{K}\cdot\mathbf{1},$$

$\mathbf{y}_i$: the one-hot label of sample x_i; $\mathbf{1} = [1, 1, \dots, 1]^{\mathrm{T}}$: the all one vector; K: # of classes.

- *(Positive) label smoothing*: $0 < r < 1$
  - i.e., $r = 0.3 \rightarrow \mathbf{y}_i^{\mathrm{GLS},r} = [0.1, 0.8, 0.1]^{\mathrm{T}}$;
- *Negative label smoothing*: $r < 0$
  - i.e., $r = -0.3 \rightarrow \mathbf{y}_i^{\mathrm{GLS},r} = [-0.1, 1.2, -0.1]^{\mathrm{T}}$.

What do negative labels really mean?

The cross-entropy loss ℓ, model prediction logit on a sample $\mathbf{f}(x_i)$, i.e., $[0.2, 0.6, 0.2]^{\mathrm{T}}$

- Evaluate on hard label: $\mathbf{y}_i^{\mathrm{GLS},r} = [0, 1, 0]^{\mathrm{T}}$
  - $\ell = -\log(0.6);$
- Evaluate on positive label: $\mathbf{y}_i^{\mathrm{GLS},r} = [0.1, 0.8, 0.1]^{\mathrm{T}}$
  - $\ell = -0.1 * \log(0.2) - 0.8 * \log(0.6) - 0.1 * \log(0.2);$
- Evaluate on negative label: $\mathbf{y}_i^{\mathrm{GLS},r} = [-0.1, 1.2, -0.1]^{\mathrm{T}}$
  - $\ell = 0.1 * \log(0.2) - 1.2 * \log(0.6) + 0.1 * \log(0.2);$
  - High confidence on irrelevant class is **punished**!

Negative labels encourage confident predictions

Evaluate on negative label: $\mathbf{y}_i^{\mathrm{GLS},r} = [-0.1, 1.2, -0.1]^{\mathrm{T}}$

- Unconfident model prediction logit
  - i.e., $\mathbf{f}(x_i) = [0.2, 0.6, 0.2]^{\mathrm{T}}$;
  - $\ell = 0.1 * \log(0.2) - 1.2 * \log(0.6) + 0.1 * \log(0.2) = 0.13$;
- Confident model prediction logit
  - i.e., $\mathbf{f}(x_i) = [0, 1, 0]^{\mathrm{T}}$;
  - $\ell = -1.2 * \log(1) = 0$;

Model is encouraged to give confident predictions.

Similar designs w.r.t. negative labels

In the binary setting ($y_i \in \{0, 1\}$), the loss on $\left(x_i, \mathbf{y}_i^{\mathrm{GLS},r}\right)$ is:

$$\ell\left(\mathbf{f}(x_i), \mathbf{y}_i^{\mathrm{GLS},r}\right) = \left(1 - \frac{r}{2}\right) \ell(\mathbf{f}(x_i), y_i) - \frac{|r|}{2}\ \ell(\mathbf{f}(x_i), 1 - y_i),$$

where y_i is the label of sample x_i.

In label-noise learning:

- *Backward Loss Correction* [Natarajan et al. 13, Partini et al. 17]
  - $\ell_{\mathbf{BLC}}(\mathbf{f}(x_i), y_i) = c_1\ \ell(\mathbf{f}(x_i), y_i) - c_2\ \ell(\mathbf{f}(x_i), 1 - y_i)$, for some $c_1, c_2 > 0$;
- *Peer Loss* [Liu & Guo, 20]
  - $\ell_{\mathbf{PL}}(\mathbf{f}(x_i), y_i) = \ell(\mathbf{f}(x_i), y_i) - \ell\left(\mathbf{f}(x_i), y_{\mathbf{rand},i}\right)$;
  - $\mathrm{P}\left(y_{\mathbf{rand},i} = y_i\right) = \mathrm{P}(y_i)$, random sampling.

What are Noisy Labels?

X: Feature; Y: Clean Label; $\tilde{Y}$: Noisy Label;

Noise transition matrix: $T_{i,j}(X) = \mathrm{P}(\tilde{Y} = j | Y = i, X)$.

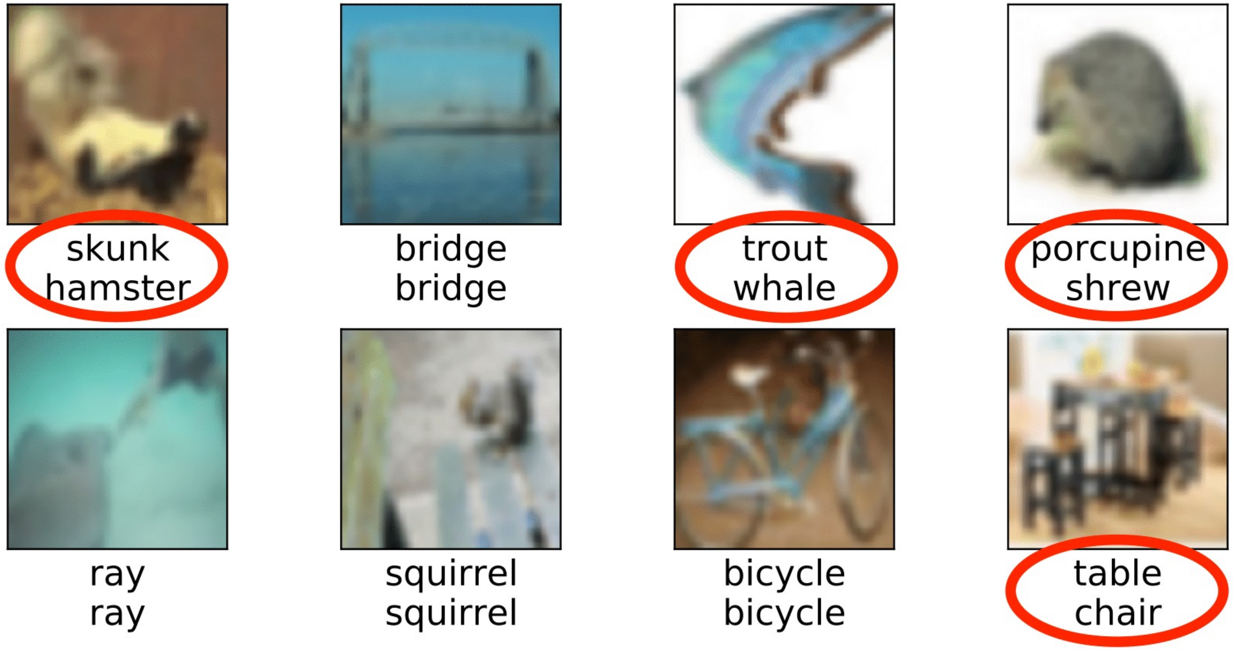


Figure 1: Human annotations for CIFAR-100 training images [Wei et al. 22].

First row in text: ground-truth labels; Second row in text: human annotations.

Motivation

Motivation: A Seemingly Conflict

[Lukasik et al. 20]

(Positive) label smoothing (LS) is beneficial when learning with noisy labels

V.S.

[Our observations]

Negative label smoothing (NLS) is closely related to several existing learning-with-noisy-label solutions

Our Contributions

Contribution 1

Address the question:

Q: Whether should we smooth labels or not, when learning with noisy labels?

or

Q: When should we prefer negative label smoothing (NLS) than positive ones (LS)?

Short answer:

A: NLS is more beneficial in the high noise regime.

Theoretical guarantees:

- Closed form of the optimal r when learning with noisy labels;
- See Theorem 3.3, 3.6.

Sketch of Contribution 1

In the risk minimization framework:

$$\min_{f\in\mathcal{F}} \mathbb{E}_{(X,\widetilde{Y})\sim\widetilde{\mathcal{D}}}\left[\ell(\mathbf{f}(X), \widetilde{Y}^{\mathrm{GLS},r})\right], \tag{1}$$

where $X, \widetilde{Y}, \widetilde{Y}^{\mathrm{GLS},r}$ denote the variable of sample, label, and smoothed label.

We bridge the gap between (1) and (2) by giving the closed form of r in (1):

$$\min_{f\in\mathcal{F}} \mathbb{E}_{(X,Y)\sim\mathcal{D}}\left[\ell(\mathbf{f}(X), Y^*)\right], \tag{2}$$

where $Y^* = Y^{\mathrm{GLS},\,r^*}$, for some optimal r^* on the clean data.

Sketch of Contribution 1

For $i \neq j$, if $T_{i,j}(X) = \mathrm{P}\left(\tilde{Y} = j \middle| Y = i, X\right) = \frac{\epsilon}{K-1}$,

we have: $r_{\text{opt}} = \frac{(K-1)\cdot r^* - K\cdot\epsilon}{(K-1) - K\cdot\epsilon}$.

- **Low noise** ($\epsilon \leq \frac{(K-1)\cdot r^*}{K}$): NLS is worse.
- **High noise** ($\epsilon > \frac{(K-1)\cdot r^*}{K}$): NLS is better.

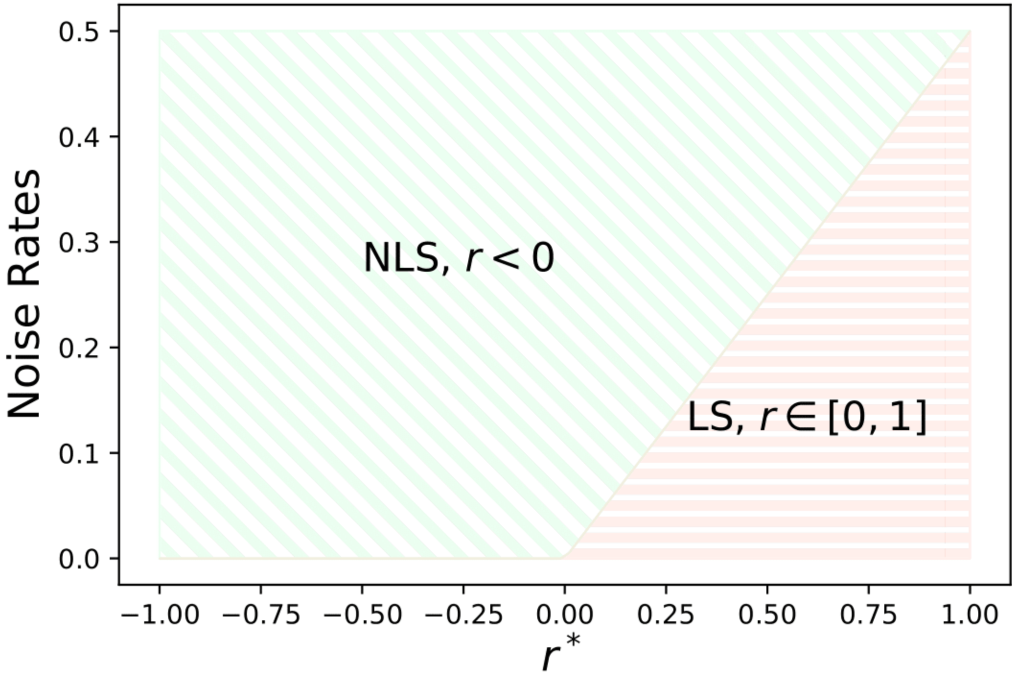


Figure 2: The preferences between NLS, LS in binary classification task.

Empirical verification of contribution 1

Table 1: Test accuracies of GLS on clean and noisy UCI datasets with best two (possibly tied) smooth rates (green: NLS; red: LS).

Smooth Rate	Twonorm					Splice				
	$e_i = 0$	$e_i = 0.1$	$e_i = 0.2$	$e_i = 0.3$	$e_i = 0.4$	$e_i = 0$	$e_i = 0.1$	$e_i = 0.2$	$e_i = 0.3$	$e_i = 0.4$
$r = 0.8$	0.990	0.990	0.986	0.982	0.968	0.980	0.946	0.919	0.856	0.760
$r = 0.6$	0.990	0.989	0.987	0.981	0.972	0.978	0.939	0.913	0.869	0.778
$r = 0.4$	0.990	0.990	0.987	0.983	0.971	0.978	0.948	0.922	0.885	0.797
$r = 0.2$	0.990	0.989	0.986	0.985	0.969	0.978	0.948	0.919	0.878	0.800
$r = 0.0$	0.990	0.989	0.987	0.985	0.973	0.976	0.948	0.926	0.876	0.806
$r = -0.4$	0.986	0.988	0.988	0.986	0.972	0.961	0.956	0.928	0.880	0.817
$r = -0.6$	0.986	0.988	0.987	0.984	0.974	0.961	0.956	0.926	0.880	0.819
$r = -1.0$	0.986	0.986	0.988	0.985	0.977	0.956	0.954	0.932	0.889	0.819
$r = -2.0$	0.986	0.986	0.986	0.986	0.978	0.952	0.946	0.935	0.898	0.830
$r = -4.0$	0.986	0.986	0.986	0.986	0.983	0.946	0.943	0.939	0.911	0.830
$r = -8.0$	0.986	0.986	0.986	0.985	0.986	0.943	0.946	0.939	0.915	0.845
$r_{\text{opt}} =$	[0.0, 0.8]	[0.4, 0.8]	[-1.0, -0.4]	[-4.0, -0.4]	-8.0	[0.0, 0.8]	[-0.6, -0.4]	[-8.0, -4.0]	-8.0	-8.0

Empirical verification of contribution 1

Table 2: Test accuracies (mean $\pm$ std) of GLS on synthetic noisy CIFAR datasets. Best two smooth rates for each synthetic noise setting are highlighted for each ϵ (green: NLS; red: LS).

Smooth Rate	CIFAR-10 Symmetric				CIFAR-10 Asymmetric		CIFAR-100 Symmetric	
	$\varepsilon = 0.0$	$\varepsilon = 0.2$	$\varepsilon = 0.4$	$\varepsilon = 0.6$	$\varepsilon = 0.2$	$\varepsilon = 0.3$	$\varepsilon = 0.4$	$\varepsilon = 0.6$
$r = 0.8$	92.91±0.06	88.88±1.61	81.48±2.91	73.16±0.16	90.45±0.06	87.83±0.13	54.04±0.93	39.50±0.18
$r = 0.6$	92.33±0.09	87.50±1.31	82.11±0.86	73.59±0.15	90.41±0.09	87.83±0.13	52.72±0.15	40.49±0.07
$r = 0.4$	93.05±0.04	87.13±0.07	81.50±1.42	74.21±0.19	90.49±0.10	87.90±0.13	54.26±0.07	41.57±0.05
$r = 0.0$	91.44±0.16	85.08±0.86	80.42±2.29	75.34±0.13	88.32±0.24	86.27±0.32	48.03±0.29	38.11±0.14
$r = -0.4$	93.55±0.06	87.55±0.08	81.58±0.19	75.95±0.13	87.27±1.83	88.33±0.06	56.87±0.08	43.70±0.16
$r = -0.8$	92.74±0.05	88.46±0.11	81.56±0.15	76.15±0.14	86.40±1.32	87.96±0.43	57.35±0.08	44.10±0.06
$r = -1.0$	92.58±0.08	88.58±0.08	81.95±0.10	76.20±0.10	88.47±0.15	87.50±0.73	57.44±0.09	43.85±0.19
$r = -2.0$	93.30±0.03	88.78±0.09	83.64±0.15	76.11±0.07	88.66±0.17	87.27±0.70	58.10±0.08	44.88±0.11
$r = -4.0$	93.13±0.04	88.90±0.07	84.34±0.13	77.22±0.09	89.56±0.17	87.29±0.59	58.35±0.09	46.38±0.05
$r = -6.0$	93.14±0.08	88.94±0.11	84.52±0.13	77.42±0.16	89.70±0.24	87.57±0.42	57.73±0.10	46.46±0.09

Empirical verification of contribution 1

Table 3: Test accuracy comparisons on clean and symmetric noisy AGNews dataset. Highlighted numbers indicate the best performance under each ϵ.

Smooth Rate	AGNews (4 classes)				
	$\epsilon = 0$	$\epsilon = 0.1$	$\epsilon = 0.2$	$\epsilon = 0.3$	$\epsilon = 0.4$
$r = 0.4$	86.33	85.55	83.93	82.29	79.80
$r = 0.2$	87.79	86.99	85.67	83.47	81.04
$r = 0.0$	**88.20**	87.79	86.80	85.24	82.39
$r = -0.15$	85.04	**88.00**	87.47	85.83	83.09
$r = -0.2$	84.08	87.30	87.50	85.85	83.34
$r = -0.36$	81.39	84.47	**87.75**	86.14	83.62
$r = -0.4$	80.76	83.99	87.28	86.36	83.96
$r = -0.6$	77.62	80.80	84.68	**87.26**	84.37
$r = -0.67$	76.70	79.91	83.87	87.21	84.58
$r = -1.14$	72.38	74.84	78.28	82.45	**86.43**
$r = r_{\text{opt}} = \frac{(K-1)r^* - K\epsilon}{(K-1) - K\epsilon}$	**88.20**	**88.00**	**87.75**	87.21	**86.43**

Other contributions

2. Theoretical connections between NLS and existing robust methods

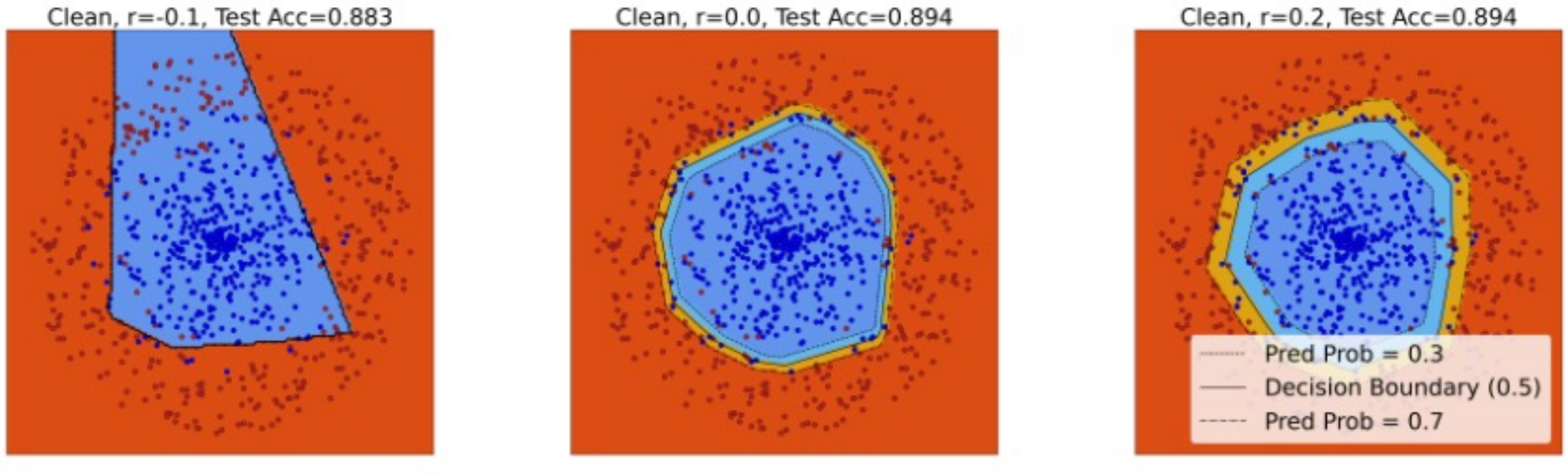
   - NLS and forward/backward loss correction [Natarajan et al. 13, Partini et al. 17]
     See Proposition 5.1, Theorem 5.2.
   - NLS and complementary loss [Ishida et al. 17]
     See Theorem 5.3.
   - NLS and peer loss functions [Liu & Guo, 20]
     See Proposition 5.4, Theorem 5.5.

3. Empirical significances of negative label smoothing

Empirical Significances

Label smoothing avoids overly model confidence (2D-synthetic data)

Left → Right: Smooth rate increases.

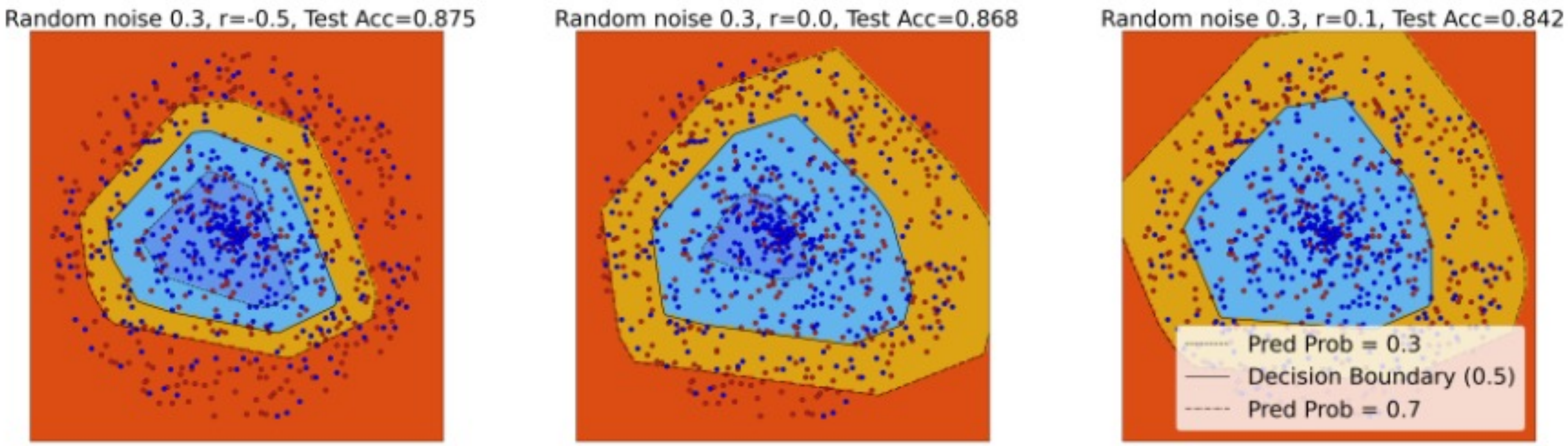


NLS (Test Acc: 0.883) CE (Test Acc: 0.894) LS (Test Acc: 0.894)

Empirical Significances

Negative label smoothing increases model confidence (2D-synthetic data)

Left → Right: Smooth rate increases.



NLS (Test Acc: 0.875) CE (Test Acc: 0.868) LS (Test Acc: 0.842)

Empirical Significances

Comparisons with existing robust approaches (real-world noisy labels)

Table 5: Performance comparisons on Clothing 1M and CIFAR-N: results of baselines are obtained through the public leader-board.

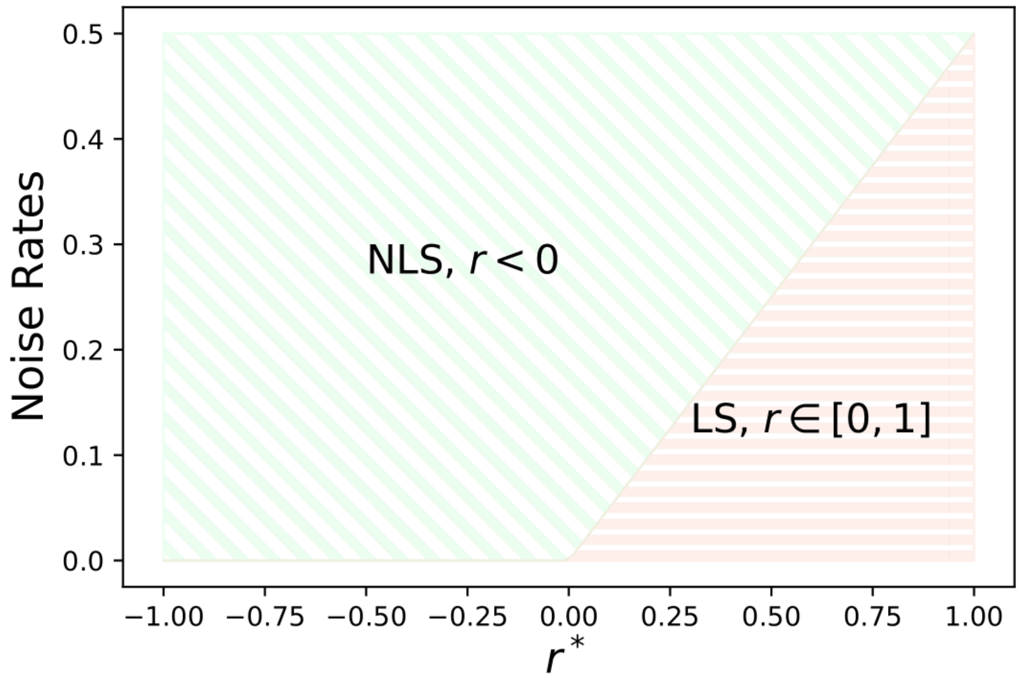
Method	Clothing 1M	CIFAR-10N Aggre	CIFAR-10N Rand1	CIFAR-10N Worse	CIFAR-100N Fine
CE	68.94	87.77	85.02	77.69	55.50
BLC	69.13	88.13	87.14	77.61	57.14
FLC	69.84	88.24	86.88	79.79	57.01
PL	72.60	90.75	89.06	82.53	57.59
F-div	73.09	91.64	89.70	82.53	57.10
LS (best)	73.44	91.57	89.80	82.76	55.84
NLS (best)	**74.24**	**91.97**	**90.29**	**82.99**	**58.59**

Takeaways

Message 1: NLS is favorable when the label noise rate is high

- LS may be beneficial
  when the label noise rate is low;

- NLS becomes more competitive
  in the high-noise regime.



Takeaways

Message 2: Interpolating existing approaches in extended label smoothing

We show, when several popular learning-with-noisy-label methods could be unified in the extended label smoothing framework.

Takeaways

Message 3: Empirical significances of the overlooked negative labels

- The nice performance of NLS on UCI synthetic noisy datasets.

- With a pre-trained model, NLS
  - works much better on synthetic noisy CIFAR datasets than CE/LS;
  - Ranks 4^{th} /33 on Clothing 1M dataset.

Paper

Negative-Label-Smoothing

SCAN ME

Thank you !

Q&A

Code

Negative-Label-Smoothing

SCAN ME