Learning with Instance-Dependent Label Noise: A Sample Sieve Approach

UNIVERSITY OF CALIFORNIA SANTA CRUZ

Tencent YouTu Lab

Hao Cheng*§, Zhaowei Zhu*†, (*Equal contributions)

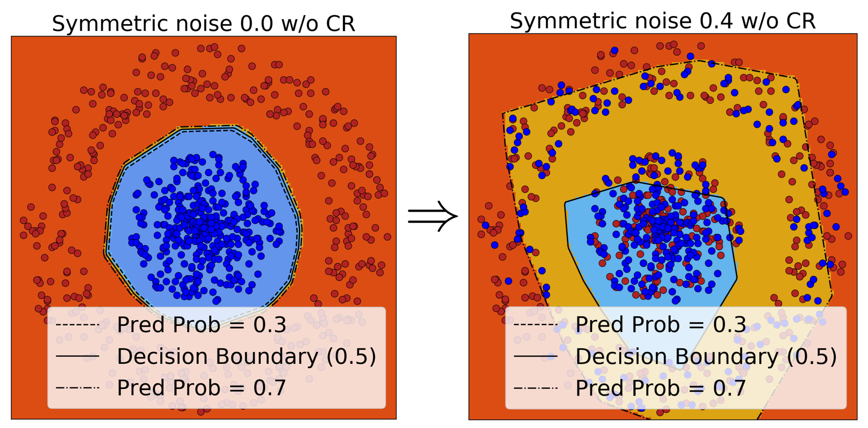
Xingyu Li†, Yifei Gong§, Xing Sun§, and Yang Liu†

†University of California, Santa Cruz, {zwzhu,xli279,yangliu}@ucsc.edu

§Tencent YouTu Lab, {louischeng,yifeigong,winfredsun}@tencent.com

Paper & Code:

Motivation



Observation: Label noise reduces the confidence of predictions

Our idea: Encourage **confident** prediction to remove corrupted examples

Problems & Solutions (Overview)

One-sentence summary: A dynamic sample sieve with theoretical guarantees to avoid overfitting to instance-dependent label noise.

Problems:

1. Label noise $(X, \widetilde{Y}) \rightarrow$ Wrong correlation patterns
2. Expensive human-efforts to reduce label noise

Challenges:

1. Unknown noise rates $\mathbb{P}(\widetilde{Y}|Y, X)$
2. Instance-dependent label noise $\mathbb{P}(\widetilde{Y}|Y, X) \neq \mathbb{P}(\widetilde{Y}|Y)$, while most existing works [1-3] assume feature independency: $\mathbb{P}(\widetilde{Y}|Y, X) = \mathbb{P}(\widetilde{Y}|Y)$
3. Loss-correction/reweighting [1-3]: Hard to estimate $\mathbb{P}(\widetilde{Y}|Y, X), \forall X$

Solutions: COnfidence Regularized Sample Sieve (CORES^2)

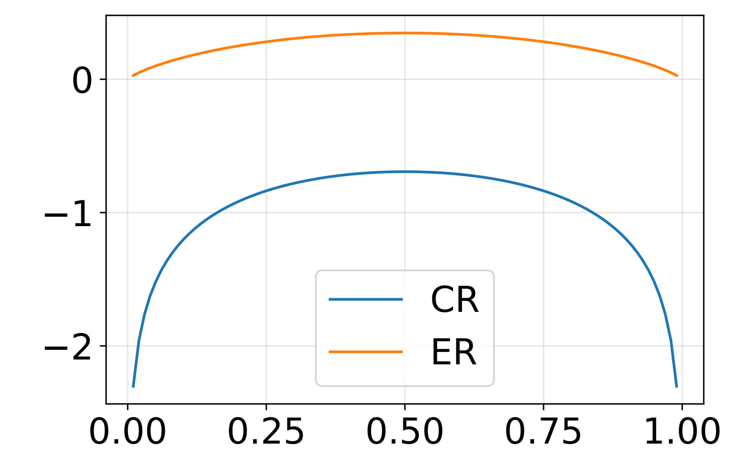
1. Confidence regularizer (learn clean distributions) - *CR*
2. Sample sieve (separate clean/corrupted examples) - *CORES^2*
3. Regular training (sieved clean examples) + Consistency training (features of sieved corrupted examples) - *CORES^{2*}*

Confidence Regularizer

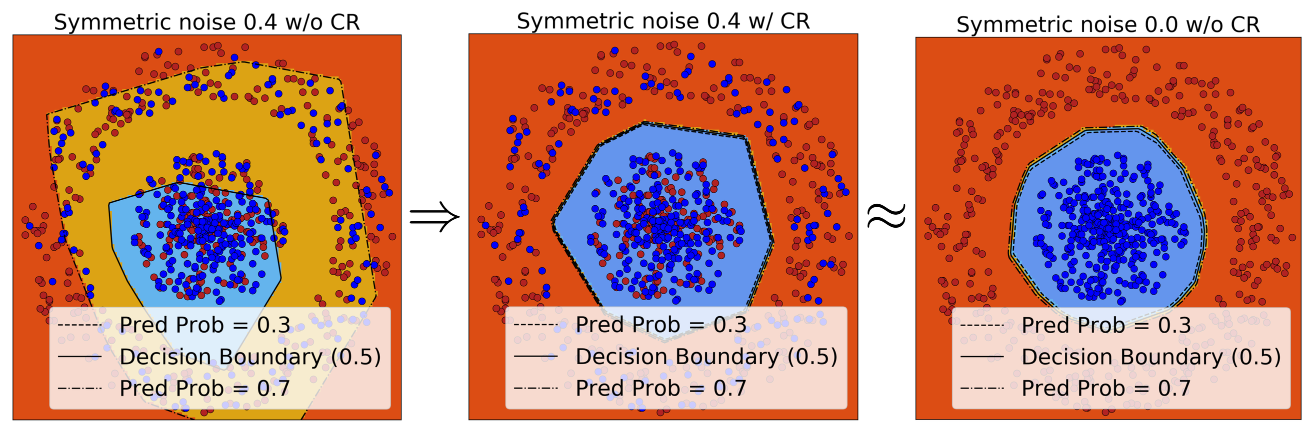
Definition: $\ell_{\text{CR}}(f(x_n)) := -\beta \cdot \mathbb{E}_{\mathcal{D}_{\widetilde{Y}|\widetilde{D}}}[\ell(f(x_n), \widetilde{Y})]$

Binary Example $\{0, 1\}$:

- Cross-Entropy loss
- $\mathbb{P}(\widetilde{Y} = 0) = \mathbb{P}(\widetilde{Y} = 1) = \frac{1}{2}$
- $p := f_{x_n}[0],\ \beta = 1$
- $\ell_{\text{CR}}(f(x_n)) = \frac{1}{2}(\ln p + \ln(1-p))$
- Confident predictions give small loss: $p \approx 0$ or $p \approx 1 \rightarrow \ell_{\text{CR}}(f(x_n)) \rightarrow -\infty$
- Unconfident predictions give large loss $\rightarrow p \approx 0.5 \rightarrow \ell_{\text{CR}}(f(x_n)) \rightarrow$ maximum



Comparison to entropy regularizer ER: $\ell_{\text{ER}}(f(x_n)) = -\frac{1}{2}(p \ln p + (1-p)\ln(1-p))$



CR helps: 1. Make *confident* predictions; 2. Learn *clean* distributions

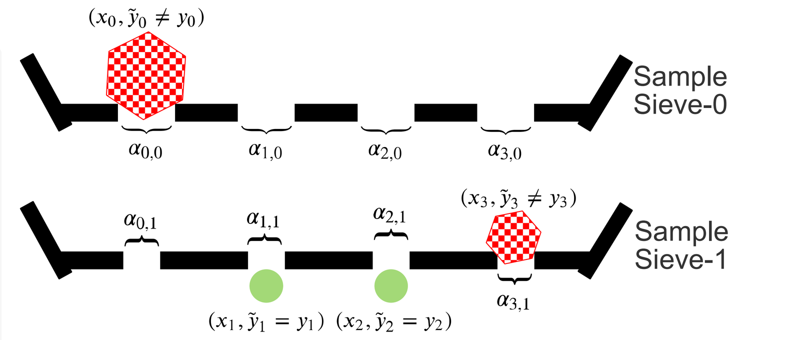
Dynamic Sample Sieve

Confidence Regularized Sample Sieve

$$\min_{\substack{f \in \mathcal{F},\\ \boldsymbol{v} \in \{0,1\}^N}} \sum_{n \in [N]} v_n \left[\ell(f(x_n), \tilde{y}_n) + \ell_{\text{CR}}(f(x_n)) - \alpha_n\right]$$

$$\text{s.t.} \quad \ell_{\text{CR}}(f(x_n)) := -\beta \cdot \mathbb{E}_{\mathcal{D}_{\widetilde{Y}|\widetilde{D}}} \ell(f(x_n), \widetilde{Y}),$$

$$\alpha_n := \frac{1}{K} \sum_{\tilde{y} \in [K]} \ell(\bar{f}(x_n), \tilde{y}) + \ell_{\text{CR}}(\bar{f}(x_n)).$$



Green circles: clean examples

Red hexagons: corrupted examples

- $v_n \in \{0, 1\}$: whether example n is clean ($v_n = 1$) or not ($v_n = 0$);
- α_n: aperture of a sieve, controls which example should be sieved out;
- $\bar{f}$: copy of f and does not contribute to the back-propagation.

Theoretical Guarantee

Theorem: CORES^2 sieves out the corrupted examples:

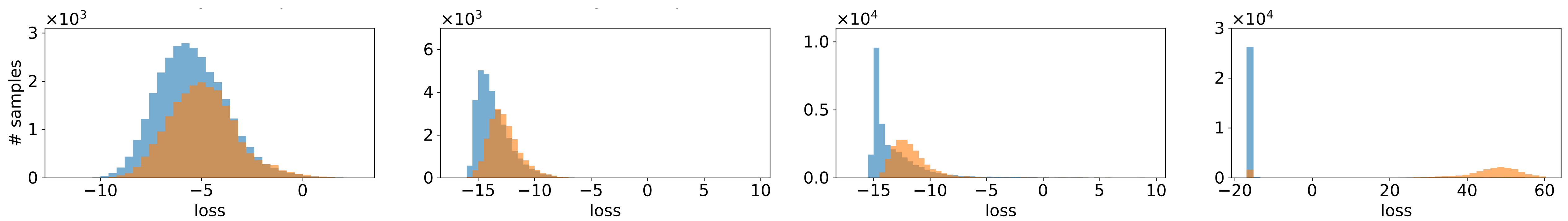
- When the model prediction on x_n is better than *random guess*, clean examples will not be wrongly identified as being corrupted
- When: $Y = Y^*$ (*clean labels are Bayes optimal*), $T_{ii}(X) - T_{ij}(X) > 0$ (*informative*), with *infinite model capacity* and *sufficiently many examples*, all the sieved clean examples are effectively clean.

Main steps of the proof:

1. **Decoupling the expected CR-regularized CE loss:** noisy loss with CR = clean loss + label shift + *noise effect* (β)
2. **CR helps learn the clean distribution:** noise effect can be *canceled* or *reversed* by proper β
3. **Proper setup of threshold** α

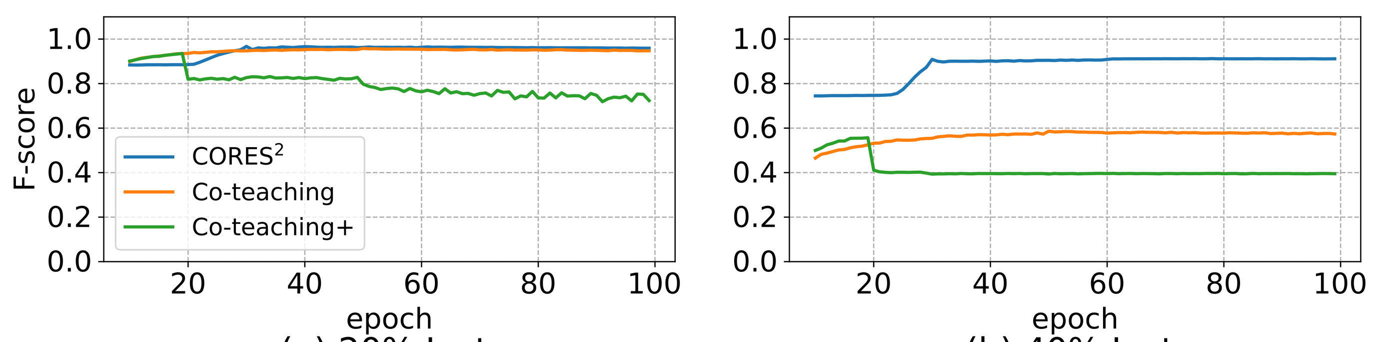
Experiments (CIFAR-10 with instance-dependent label noise)

Loss distributions: CE sieve: dynamic sample sieve without CR.



(a) CE Sieve (inst., epoch 20) (b) CE Sieve (inst., epoch 70) (c) CORES^2 (inst., epoch 20) (d) CORES^2 (inst., epoch 70)

F-scores:



(a) 20% Inst. (b) 40% Inst.

See our paper for more results of CORES^2 and CORES^{2*} on different noisy datasets

Relevant Works

[1] N. Natarajan, et al. "Learning with noisy labels." NeurIPS'13.

[2] T. Liu & D. Tao. "Classification with noisy labels by importance reweighting." TPAMI'15.

[3] G. Patrini, et al. "Making deep neural networks robust to label noise: A loss correction approach." CVPR'17.

Related other works from our lab

- Peer loss functions: learning from noisy labels without knowing noise rates, ICML'20
- CE → f-divergence: When optimizing f-divergence is robust with label noise, ICLR'21
- High-order statistics: A second-order approach to learning with instance-dependent label noise, CVPR'21 (oral)

Acknowledgement: partially supported by the National Science Foundation (NSF) under grant IIS-2007951 and the Office of Naval Research under grant N00014-20-1-22.