A Second-Order Approach to Learning with Instance-Dependent Label Noise

Zhaowei Zhu†, Tongliang Liu, and Yang Liu†

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

§The University of Sydney, tongliang.liu@sydney.edu.au

Paper & Code:

Motivating Example

Class-dependent label noise (CDN): $\forall X: \mathbb{P}(\widetilde{Y}|Y^*, X) = \mathbb{P}(\widetilde{Y}|Y^*)$.

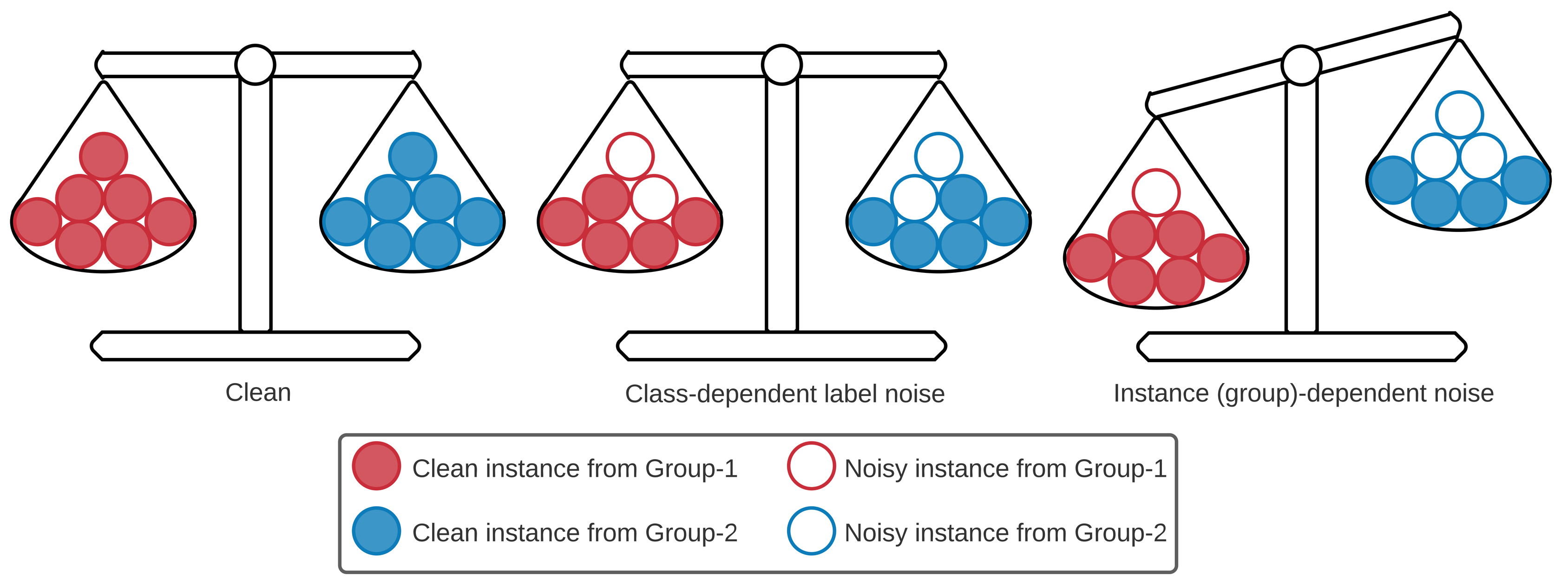
Instance-dependent label noise (IDN): $\exists X: \mathbb{P}(\widetilde{Y}|Y^*, X) \neq \mathbb{P}(\widetilde{Y}|Y^*)$

Example:

Two groups of instances. Intra-group: CDN; Inter-group: IDN.

Empirical Risk Minimization (ERM) of instances from two groups:

$$\text{Loss} = \sum_{i \in \text{Group-1}} \text{Loss}_i + \sum_{j \in \text{Group-2}} \text{Loss}_j$$



Intuition: Compare the weights of group 1 with group 2, we find:

Clean: no noise ⇒

equal #instances contribute to clean loss ⇒ equal weights in ERM

CDN: equal noise ⇒

equal #instances contribute to clean loss ⇒ equal weights in ERM

IDN: Group 2: larger noise ⇒

less #instances contribute to clean loss ⇒ smaller weights in ERM

Problems & Solutions (Overview)

One-sentence summary:

We use covariance to compensate for the "imbalances" caused by IDN such that the challenging IDN can be transformed to a easier CDN one.

Problems:

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

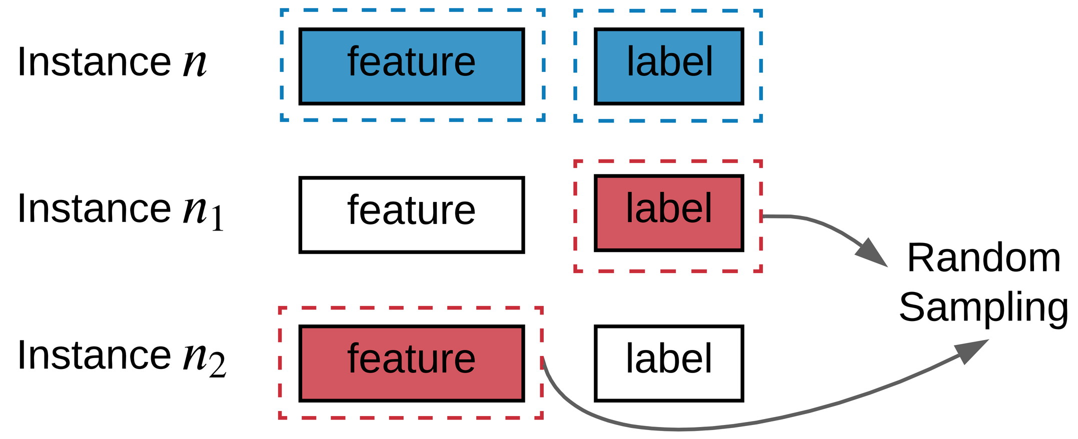
Challenges:

1. Unknown instance-dependent noise rates $\mathbb{P}(\widetilde{Y}|Y^*, X)$, while most existing works [1-5] assume feature independency: $\mathbb{P}(\widetilde{Y}|Y^*, X) = \mathbb{P}(\widetilde{Y}|Y^*)$
2. Loss-correction/reweighting [1-3]: Hard to estimate $\mathbb{P}(\widetilde{Y}|Y^*, X), \forall X$
3. IDN causes imbalances in different feature group (see Motivation)

Solutions: CAL: IDN $\xrightarrow{\text{2nd-Order}}$ CDN $\xrightarrow{\text{1st-Order}}$ Clean

Peer Loss (Use First-Order Statistics)

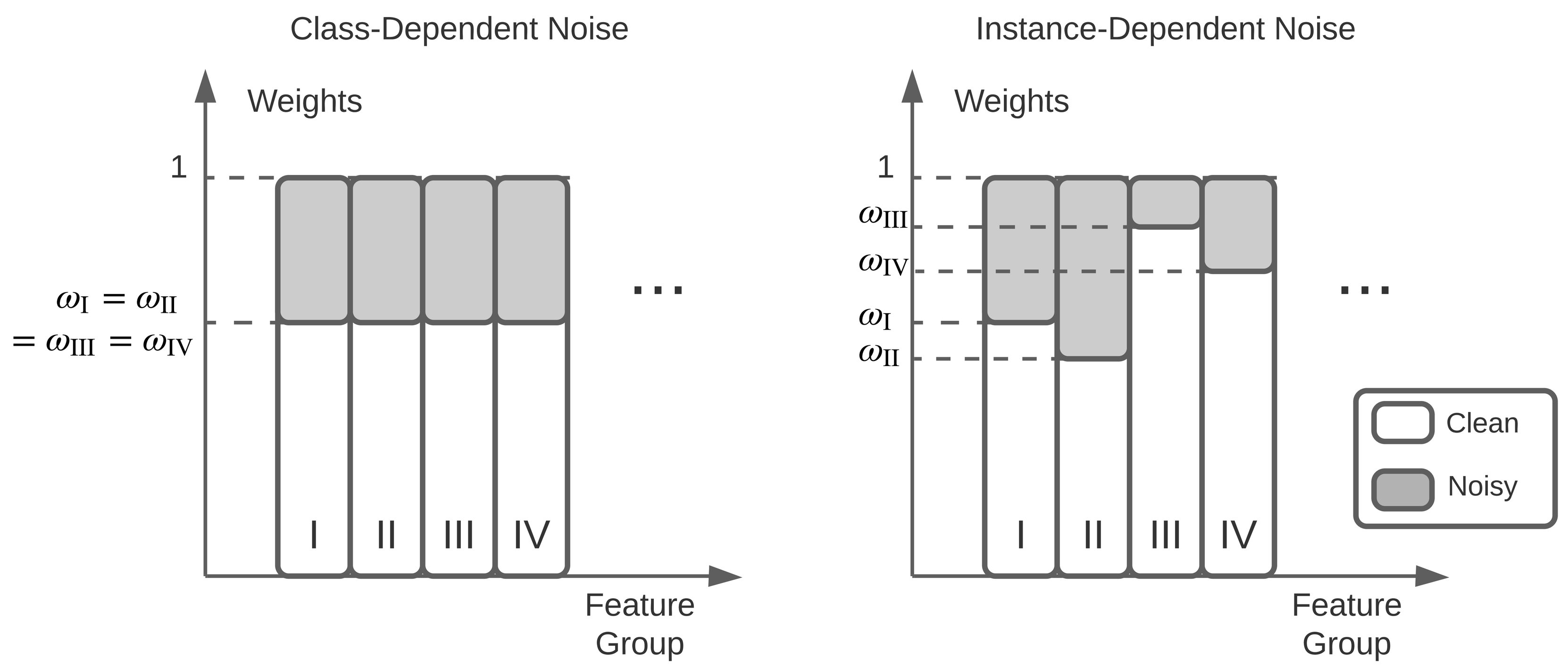
Definition: $\ell_{\text{PL}}(f(x_n), \tilde{y}_n) := \ell(f(x_n), \tilde{y}_n) - \ell(f(x_{n_1}), \tilde{y}_{n_2})$



Lemma: Peer loss [4] is invariant to CDN: NoisyPL $= \omega \cdot$ CleanPL

Summary: 1) CDN $\xrightarrow{\text{Peer Loss}}$ Clean; 2) Unknown ω: Noise ↑, weight ω ↓

Insufficiency of First-Order Statistics



Summary: IDN causes weights imbalances

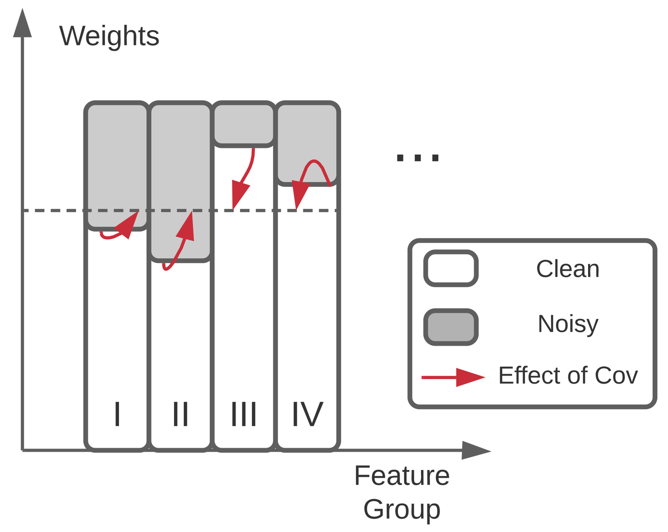
CDN: Only one unknown constant ω. *Equal* for all features.

IDN: Multiple unknown constants ω_g. *Down-weight* high-noise features.

Covariance-Assisted Learning (CAL)

Our method: Peer Loss + Covariance (requires constructing $\hat{D}$ for T):

$$\ell_{\text{CAL}}(f(x_n), \tilde{y}_n) = \ell_{\text{PL}}(f(x_n), \tilde{y}_n) - \text{Cov(Noise Trans. } T, \text{Model Pred.)}$$



Summary:

- CAL balances weights of each feature
  - High-noise (I, II): improve weights
  - Low-noise (III,IV): reduce weights
- IDN $\xrightarrow{\text{Covariance}}$ CDN $\xrightarrow{\text{Peer Loss}}$ Clean

Benefits: CAL is a "soft" correction (vs. "hard" label correction)

- Use an average term, less sensitive to estimation of each instance
- Tolerant of inaccurate $\hat{D}$

Algorithm (Sketch)

1. Construct $\hat{D}$ (unbiased estimate of $D^* \sim \mathcal{D}^*$) with sample sieve [5]
2. Estimate (unbiased) $\hat{T}$ with $\hat{D}$ (complexity $O(\text{SampleSize})$)
3. [Train DNN] Implement CAL in SGD (each point $O(1)$ complexity)

Theoretical Guarantee

Theorem:

1) With perfect covariance estimates, $\mathbb{1}_{\text{CAL}}$ is robust to IDN (induces the Bayes optimal classifier).

2) With imperfect covariance estimates, error rate can be upper bounded.

Experiments

Table: Comparison of test accuracies (%) using different methods.

Method	*Inst. CIFAR10*			*Inst. CIFAR100*		
	$\eta = 0.2$	$\eta = 0.4$	$\eta = 0.6$	$\eta = 0.2$	$\eta = 0.4$	$\eta = 0.6$
CE (Standard)	85.45±0.57	76.23±1.54	59.75±1.30	57.79±1.25	41.15±0.83	25.68±1.55
Forward T [2]	87.22±1.60	79.37±2.72	66.56±4.90	58.19±1.37	42.80±1.01	27.91±3.35
T-Revision [3]	90.04±0.46	84.11±2.47	72.18±2.47	58.00±0.36	43.83±8.42	36.07±9.73
Peer Loss [4]	89.12±0.76	83.26±0.42	74.53±1.22	61.16±0.64	47.23±1.23	31.71±2.06
CORES^2 [5]	91.14±0.46	83.67±1.29	77.68±2.24	66.47±0.45	58.99±1.49	38.55±3.25
CAL	**92.01±0.75**	**84.96±1.25**	**79.82±2.56**	**69.11±0.46**	**63.17±1.40**	**43.58±3.30**

Relevant Works

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

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

[3] X. Xia, et al. "Are anchor points really indispensable in label-noise learning?" *NeurIPS'19*.

[4] Y. Liu & H. Guo. "Peer loss functions: Learning from noisy labels without knowing noise." *ICML'20*.

[5] H. Cheng, et al. "Learning with instance-dependent label noise: A sample sieve approach." *ICLR'21*.

Related other works from our lab

- CE → f-divergence: When optimizing f-divergence is robust with label noise, *ICLR'21*
- Estimate transition matrix with clusterability: Clusterability as an Alternative to Anchor Points When Learning with Noisy Labels, *ICML'21*

Acknowledgement: Supported in part by National Science Foundation (NSF) under grant IIS-2007951, and in part by Australian Research Council Projects, i.e., DE-190101473.