Learning from Noisy Labels via Conditional Distributionally Robust Optimization

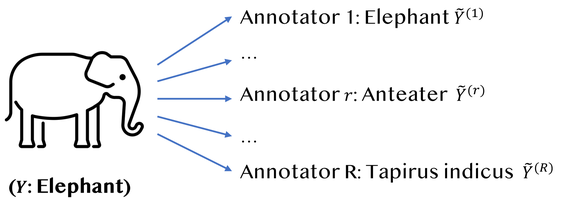
Hui Guo, Grace Y. Yi , Boyu Wang University of Western Ontario

hguo288@uwo.ca
@hguo288uwo
13102210978
2369670759
Contact

GitHub Arxiv WeChat

Problem

- **Learning with noisy labels:** the true label is unobserved; instead, noisy labels are collected
- **Crowdsourcing:** each data item is labeled by multiple annotators with diverse expertise or skills



Data (X) **Crowdsourced Noisy Labels ($\widetilde{Y}$)**

- **Noisy training data:** $\mathcal{D} = \{\mathbf{X}_i, \widetilde{\mathbf{Y}}_i\}_{i=1}^n$
  - ➢ $\mathbf{X}_i$: input data
  - ➢ Y_j: unobserved true label
  - ➢ $\widetilde{\mathbf{Y}}_i$: noisy labels provided by R annotators for $\mathbf{X}_i$
- **Goal:** train a classifier ψ using the noisy dataset $\mathcal{D}$ to accurately predict the true label for future instances

Motivation

- **Existing methods:**
  - ➢ Approximate the posterior distribution of the true label Y, given the observed data $\mathbf{X}$ and $\widetilde{\mathbf{Y}}$, denoted $P_{\mathrm{y}|\mathbf{x},\tilde{\mathbf{y}}}$
  - ➢ Directly apply $P_{\mathrm{y}|\mathbf{x},\tilde{\mathbf{y}}}$ to either infer the true labels or to weight the loss functions, **without considering potential mis-specification of the associated model**
  - ➢ **Our goal:** mitigate the effects of model misspecifications
- **Conditional distributionally robust risk optimization problem (CDRP):**

$$\inf_{\psi\in\Psi} \mathsf{R}_\epsilon(\psi; P_{\mathrm{y}|\mathbf{x},\tilde{\mathbf{y}}}), \quad \text{(robust/worst-case risk)}$$

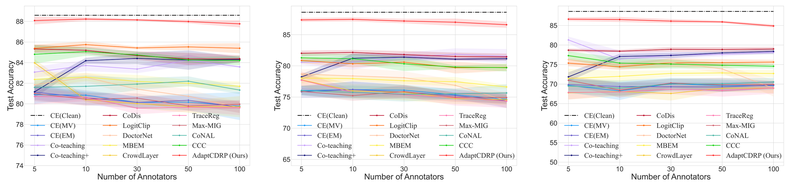
$$\text{with } \mathsf{R}_\epsilon(\psi; P_{\mathrm{y}|\mathbf{x},\tilde{\mathbf{y}}}) \triangleq \mathbb{E}_{\mathbf{x},\tilde{\mathbf{y}}}\left[\sup_{Q_{\mathrm{y}|\mathbf{x},\tilde{\mathbf{y}}}\in \Gamma_\epsilon(P_{\mathrm{y}|\mathbf{x},\tilde{\mathbf{y}}})} \mathbb{E}_{Q_{\mathrm{y}|\mathbf{x},\tilde{\mathbf{y}}}}\left\{\ell(\psi(\mathbf{X}), \mathrm{Y})\right\}\right] \quad \text{(loss function)}$$

ambiguity set centered around the reference probability distribution $P_{\mathrm{y}|\mathbf{x},\tilde{\mathbf{y}}}$:
$\Gamma_\epsilon(P_{\mathrm{y}|\mathbf{x},\tilde{\mathbf{y}}}) = \left\{Q_{\mathrm{y}|\mathbf{x},\tilde{\mathbf{y}}} \in \mathcal{P}(\mathcal{Y}) : d(Q_{\mathrm{y}|\mathbf{x},\tilde{\mathbf{y}}}, P_{\mathrm{y}|\mathbf{x},\tilde{\mathbf{y}}}) \le \epsilon\right\}$ (d: p-Wasserstein distance)

- ➢ **Challenge 1:** construct a reliable reference distribution without access to true labels
- ➢ **Challenge 2:** solve the minimax optimization problem

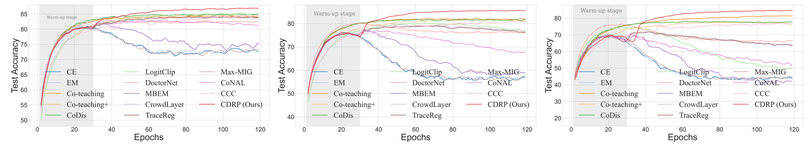
Empirical Results

Average test accuracy on CIFAR-10 with varying numbers of annotators:



(a) IDN-LOW (b) IDN-MID (c) IDN-HIGH

Training dynamics on CIFAR-10 ($R = 5$):



(d) IDN-LOW (e) IDN-MID (f) IDN-HIGH

Accuracy of robust pseudo-labels with varying numbers of annotators:

(g) IDN-LOW (h) IDN-LOW (i) IDN-LOW

Table 1: Average test accuracies on CIFAR-10 and real datasets with different transition matrix estimation methods.

Method	CIFAR-10			Real datasets	
	IDN-LOW	IDN-MID	IDN-HIGH	Animal10N	LabelMe
TraceReg	$80.72_{\pm 0.79}$	$77.71_{\pm 1.36}$	$67.86_{\pm 1.77}$	$80.34_{\pm 0.66}$	$83.10_{\pm 0.15}$
TraceReg+Ours	$87.74_{\pm 0.26}$	$86.76_{\pm 0.07}$	$85.83_{\pm 0.37}$	$83.05_{\pm 0.26}$	$83.80_{\pm 0.44}$
GeoCrowdNet (F)	$84.73_{\pm 0.39}$	$81.44_{\pm 1.00}$	$77.29_{\pm 1.23}$	$81.07_{\pm 0.45}$	$84.59_{\pm 0.19}$
GeoCrowdNet (F) + Ours	$88.06_{\pm 0.33}$	$87.43_{\pm 0.29}$	$86.69_{\pm 0.13}$	$83.12_{\pm 0.42}$	$86.20_{\pm 0.48}$
GeoCrowdNet (W)	$83.82_{\pm 0.53}$	$75.72_{\pm 1.10}$	$64.64_{\pm 2.23}$	$80.19_{\pm 0.33}$	$81.63_{\pm 1.49}$
GeoCrowdNet (W) + Ours	$87.94_{\pm 0.35}$	$87.21_{\pm 0.33}$	$83.48_{\pm 5.69}$	$82.41_{\pm 0.04}$	$83.32_{\pm 0.51}$
BayesianIDNT	$86.46_{\pm 1.07}$	$85.14_{\pm 0.96}$	$82.49_{\pm 2.86}$	$81.22_{\pm 0.59}$	$83.01_{\pm 0.32}$
BayesianIDNT + Ours	$87.66_{\pm 0.85}$	$86.44_{\pm 0.57}$	$84.38_{\pm 0.10}$	$83.80_{\pm 0.44}$	$84.09_{\pm 0.53}$

AdaptCDRP: Theoretical Analysis and Methodology

- **Proposition (Dual problem):** $\mathsf{R}_\epsilon(\psi; P_{\mathrm{y}|\mathbf{x},\tilde{\mathbf{y}}}) = \mathbb{E}_{\mathbf{x},\tilde{\mathbf{y}}}\left\{\inf_{\gamma\ge 0}\left(\gamma\epsilon^p + \mathbb{E}_{P_{\mathrm{y}|\mathbf{x},\tilde{\mathbf{y}}}}\left[\sup_{y'\in\mathcal{Y}}\left\{\ell(\psi(\mathbf{X}), y') - \gamma c^p(y', \mathrm{Y})\right\}\right]\right)\right\}$
- **Relaxed problem:** $\mathfrak{R}_\epsilon(\psi; P_{\mathrm{y}|\mathbf{x},\tilde{\mathbf{y}}}) \triangleq \inf_{\gamma\ge 0}\mathbb{E}_{\mathbf{x},\tilde{\mathbf{y}}}\left(\gamma\epsilon^p + \mathbb{E}_{P_{\mathrm{y}|\mathbf{x},\tilde{\mathbf{y}}}}\left[\sup_{y'\in\mathcal{Y}}\left\{\ell(\psi(\mathbf{X}), y') - \gamma c^p(y', \mathrm{Y})\right\}\right]\right) \Rightarrow$ Empirical version: $\widehat{\mathfrak{R}}_\epsilon(\psi; P_{\mathrm{y}|\mathbf{x},\tilde{\mathbf{y}}})$, or $\widehat{\mathfrak{R}}_\epsilon$

Learning from Noisy Labels via Conditional Distributionally Robust True Label Posterior with an Adaptive Lagrange multiplier (AdaptCDRP)

❖ Warm up classifiers $\psi^{(1)}$ and $\psi^{(2)}$; Approximate noise transition probabilities $\widehat{\tau}_j(\tilde{\mathbf{y}})$ for $P^*(\widetilde{\mathbf{Y}} = \tilde{\mathbf{y}}|\mathrm{Y} = j, \mathbf{x})$

❖ **Epoch t: Update the classifiers with pseudo-empirical distribution**

- ➢ Update approximated true label posteriors (Bayes' Rule):
  $\widehat{P}_j^{(\iota)}(\mathbf{x}, \tilde{\mathbf{y}}) \triangleq \widehat{P}^{(\iota)}(\mathrm{Y} = j|\widetilde{\mathbf{Y}} = \tilde{\mathbf{y}}, \mathbf{X} = \mathbf{x}) \propto \psi_j^{(\iota)}(\mathbf{x}) \cdot \widehat{\tau}_j(\tilde{\mathbf{y}})$
- ➢ For each instance $\mathbf{x}_i$:
  $\frac{\widehat{P}_{k^\star}^{(\iota)}(\mathbf{x}_i, \tilde{\mathbf{y}}_i)}{\max_{j\ne k^\star}\widehat{P}_j^{(\iota)}(\mathbf{x}, \tilde{\mathbf{y}})} \ge \mathcal{C} \implies \mathrm{y}_i^\star = k^\star$ and collect $(\mathbf{x}_i, \tilde{\mathbf{y}}_i, \mathrm{y}_i^\star)$ into $\mathcal{D}_{t,\iota}^\star$
- ➢ Update the pseudo-empirical distribution $P_{t,\iota}^\star$ based on $\mathcal{D}_{t,\iota}^\star$
- ➢ Update $\psi^{(\iota)}$ by minimizing $\mathfrak{R}_\epsilon(\psi; P_{\mathrm{y}|\mathbf{x},\tilde{\mathbf{y}}})$
  - – reference distribution $P_{t,\backslash\iota}^\star$; Lagrange multiplier $\gamma_{t-1}^{(\iota)}$

Addressing Challenge 1:

- **Find the optimal action $\psi^\star$ for each data point $(\mathbf{x}, \tilde{\mathbf{y}})$**
- **Construct the reference distribution**

Theorem (*Informal*). Let $P_j \triangleq P(Y = j|\mathbf{x}, \tilde{\mathbf{y}})$, and let $P^{(j)}$ denote the j-th largest element of $\{P_k\}_{k=1}^K$. Minimizing $\widehat{\mathfrak{R}}_\epsilon$ yields $\psi^{\star(j)} = \frac{1}{k_0}\mathbf{1}(j \le k_0)$ for some $k_0 \in [K]$, where $\psi^{\star(j)}$ corresponds to the index of $P^{(j)}$. In particular, if the difference between $P^{(1)}$ and $P^{(2)}$ is large enough, then $\psi^{\star(1)} = 1$ and $\psi^{\star(j)} = 0$ for $j = 2, ..., K$. ⇒ **Robust pseudo-label $\mathrm{y}^\star$**

❖ **Epoch t: Update the Lagrange multiplier**

- ➢ Compute $\alpha_i's$ for $i \in [nK]$ and s^*
- ➢ Compute the reference value for γ: $\gamma_{0,t} = \alpha^{(s^*)}$
- ➢ Update γ: $\gamma_t^{(\iota)} = \gamma_{0,t} - \frac{1}{\lambda}\{\epsilon^p - \mathbb{E}_{P_{t,\backslash\iota}^\star} c^p(y', \mathrm{Y})\}$

Addressing Challenge 2: Derive the optimal Lagrange multiplier $\gamma^\star$

Theorem (*Informal*). Consider the loss function of the form $\ell(\psi(\mathbf{x}), \mathrm{y}) = \sum_{j=1}^K \mathbf{1}(\mathrm{y} = j)\mathcal{T}(\psi(\mathbf{x})_j)$. Let $P_{i,j} \triangleq P(Y = j|\mathbf{x}_i, \tilde{\mathbf{y}}_i)$ and $\psi_{i,j} \triangleq \psi(\mathbf{x}_i)_j$ for $i \in [n]$ and $j \in [K]$. An integer $\alpha^* \in [nK + 1]$ can be defined using $\{P_{i,j}\}$. Let $\alpha_{i,j} \triangleq \mathcal{T}(\min_{j\in[K]}\psi_{i,j}) - \mathcal{T}(\psi_{i,j})$, and sort $\{\alpha_{i,j}\}$ as $\alpha^{(1)} \ge ... \ge \alpha^{(nK)}$. Then, $\gamma^\star = \alpha^{(s^*)}$.

Contributions

1. We **formulate learning with noisy labels as a CDRO problem** and **develop its dual form** to tackle the challenge of potential model misspecification in estimating the true label posterior using noisy data.
2. We derive an **analytical solution to the dual problem** for each data point, and propose a novel algorithm that **constructs a robust reference distribution** for this problem.
3. By deriving **the optimal Lagrange multiplier for the empirical robust risk**, we develop an one-step update method for the Lagrange multiplier, allowing for a principled balance between robustness and model fitting.