DORO: Distributional and Outlier Robust Optimization

Runtian Zhai* Chen Dan* J. Zico Kolter Pradeep Ravikumar

Carnegie Mellon University

TL;DR

DRO, a popular method for tasks with subpopulation shift, is sensitive to outliers. DORO fixes this issue.

Background: Subpopulation Shift

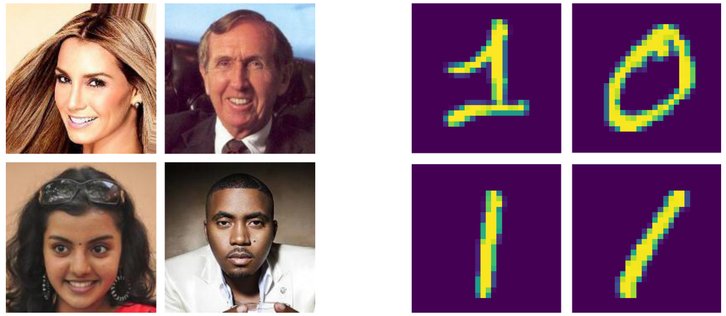
The data domain is divided into several subpopulations (subdomains) $\mathcal{D}_1, \cdots, \mathcal{D}_K$ and we are required to train a model with high performance over each domain.

The problem is often referred to as *subpopulation shift* since the underlying data distribution P is not the same as the test distribution $P_{\text{test}} = P(z|\mathcal{D}_k)$ for some $k = 1, \cdots, K$.

Let the expected risk of model θ over P be $\mathcal{R}(\theta; P)$. Instead of the expected risk, our goal is to minimize the **worst-case risk:** $\mathcal{R}_{\max}(\theta; P) = \max_{k=1,\cdots,K} \mathcal{R}(\theta; P(z|D_k))$.

Domain-Oblivious Setting

We assume that the subdomains $\mathcal{D}_1, \cdots, \mathcal{D}_K$ and the number of subdomains K are unknown during training.



(a) Algorithmic Fairness

(b) Class Imbalance

Figure 1. Two applications of subpopulation shift.

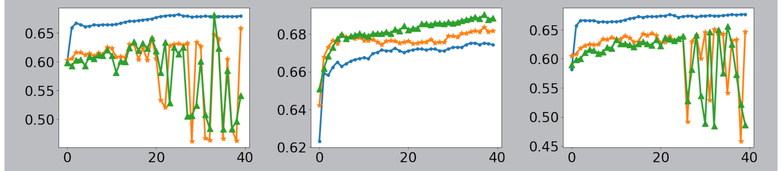
Prior Approach: Distributional Robust Optimization (DRO)

Idea: Construct an *uncertainty set* U containing all possible P_{test}, and minimize the expected risk over the worst distribution in this set (upper bound of the worst-case risk).

$$\mathcal{R}_{DRO}(\theta; P) = \sup_{P' \in U} \mathcal{R}(\theta; P')$$

Issue: DRO is Sensitive to Outliers

Outliers make DRO's performance poor and unstable.



(a) Original (b) No Outliers (c) Noise Added

Figure 2. Worst-case accuracy on the COMPAS dataset. Blue: Standard training (ERM). Green and Orange: DRO methods. X-axis: Epochs. Y-axis: Accuracy.

Proposed Method: DORO

Huber's Model: $P_{\text{train}} = (1-\epsilon)P + \epsilon\tilde{P}$, where ϵ is the noise level and $\tilde{P}$ is an arbitrary outlier distribution.

Equivalently, $P \in \{Q : \exists \tilde{Q} \text{ s.t. } P_{\text{train}} = (1-\epsilon)Q + \epsilon\tilde{Q}\}$.

Idea: Train on the *best* Q in this set, i.e. minimizing

$$\mathcal{R}_{DORO}(\theta; P_{\text{train}}) = \inf_{Q:\exists \tilde{Q} \text{ s.t. } P_{\text{train}}=(1-\epsilon)Q+\epsilon\tilde{Q}} \mathcal{R}_{DRO}(\theta; Q)$$

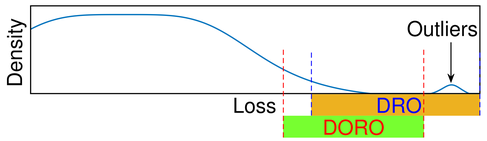


Figure 3. DORO avoids overfitting to outliers.

Theoretical and Empirical Results

Theoretical Results

Given that the loss function has bounded moments,

1. Let $\hat{\theta}$ be the minimizer of $\mathcal{R}_{DORO}(\theta; P_{\text{train}})$. Then $\mathcal{R}_{DRO}(\hat{\theta}; P)$ is close to $\inf_\theta \mathcal{R}_{DRO}(\theta; P)$.
2. $\mathcal{R}_{\max}(\theta; P)$ is upper bounded by $\max\{3\mathcal{R}_{DORO}(\theta; P_{\text{train}}), C\}$ for some constant C.

Empirical Results

DORO improves the average and worst-case accuracy of DRO, and makes the accuracy across epochs more stable.

Method	Average	Worst-case
ERM	95.01 ± 0.38	53.94 ± 2.02
CVaR-DRO	82.83 ± 1.33	66.44 ± 2.34
CVaR-DORO	92.91 ± 0.48	72.17 ± 3.14

Table 1. Accuracy over the CelebA dataset (%)

Method	Average	Worst-case
ERM	0.73 ± 0.06	8.59 ± 0.90
CVaR-DRO	11.53 ± 1.72	21.47 ± 0.71
CVaR-DORO	4.03 ± 1.57	16.84 ± 0.91

Table 2. Standard deviation of accuracy across epochs (%)

ICML 2021: Thirty-eighth International Conference on Machine Learning

rzhai@cmu.edu