DORO: Distributional and Outlier Robust Optimization

ICML 2021

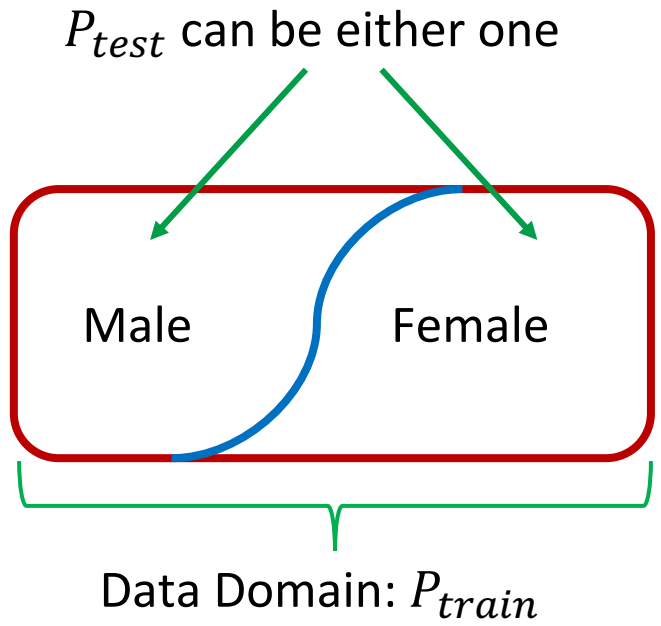
Runtian Zhai*, Chen Dan*,
Zico Kolter, Pradeep Ravikumar

Contents

Distributional Shift

Algorithmic Fairness, Class Imbalance, ...

- P_{train} and P_{test} are different.
- Subpopulation Shift: The data domain contains several subpopulations. P_{test} is the conditional distribution of P_{train} over any subpopulation.
- Eg. In algorithmic fairness, subpopulations are defined by protected features like race and sex.

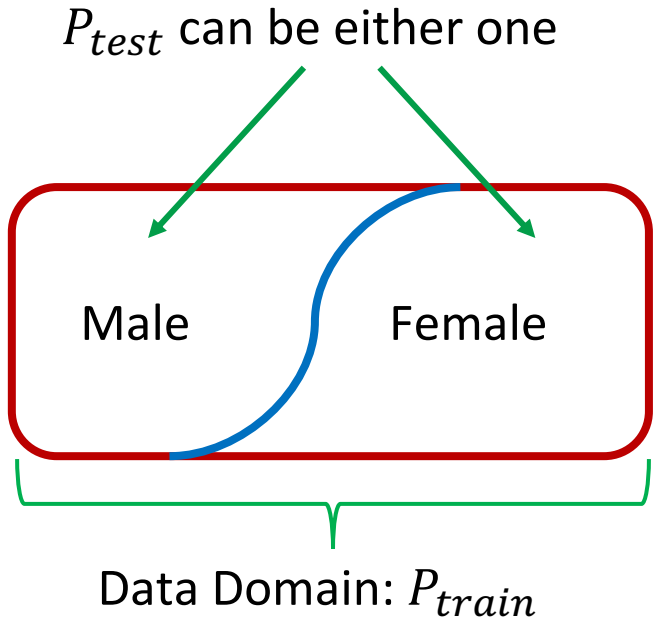


Distributional Shift

- Subpopulations $\mathcal{D}_1, \ldots, \mathcal{D}_K$, loss function ℓ
- **Goal**: Minimize the worst-case risk

$$\mathcal{R}_{\max}(\theta; P_{train}) = \max_{1 \le k \le K} \mathbb{E}[\ell(\theta; Z) | \mathcal{D}_k]$$

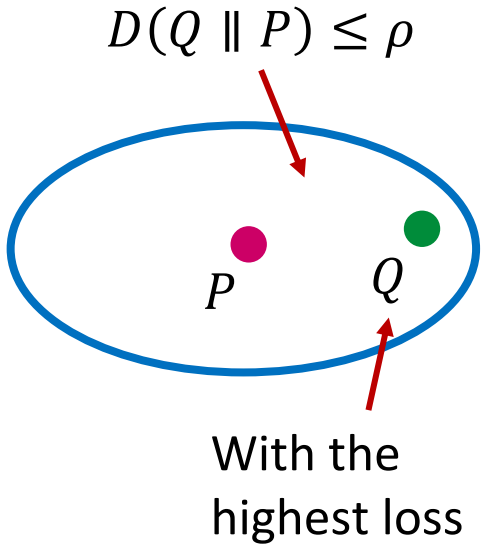
- Domain-oblivious setting: $\mathcal{D}_1, \ldots, \mathcal{D}_K$ and the value K are unknown during training.



DRO: Distributionally Robust Optimization

- For a given distribution P, DRO minimizes a model's expected risk over the worst-case distribution Q in a ball w.r.t. some divergence function D around P.
- Expected DRO Risk of model θ over P:

$$\mathcal{R}_{D,\rho}(\theta; P) = \sup_{Q \ll P} \{\mathbb{E}_Q[\ell(\theta; Z)] : D(Q \parallel P) \leq \rho\}$$



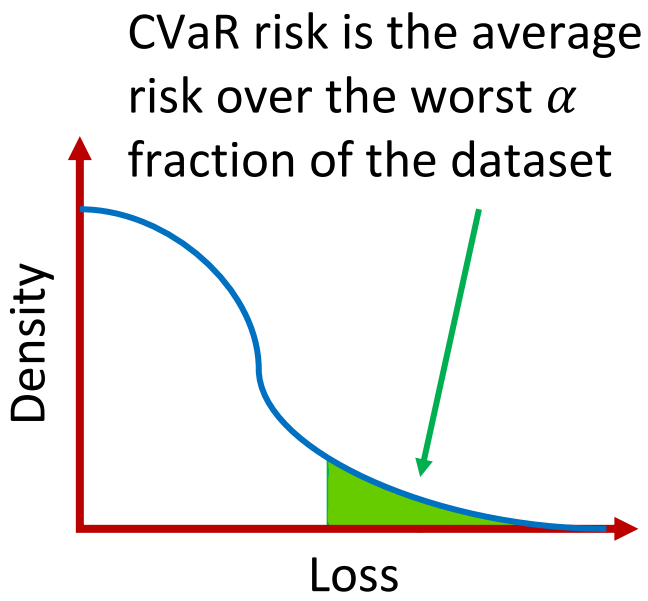
$Q \ll P$: *For any event* A, $P(A) = 0 \Rightarrow Q(A) = 0$

CVaR: Conditional Value at Risk

- $D(Q \parallel P) = \sup \log \frac{dQ}{dP} \qquad \rho = -\log \alpha$

where $0 < \alpha < 1$

- $\mathrm{CVaR}_\alpha(\theta; P) = \sup_{Q \ll P} \left\{ \mathbb{E}_Q[\ell(\theta; Z)] : \frac{dQ}{dP} \leq \frac{1}{\alpha} \right\}$
- If $P(\mathcal{D}_k) \geq \alpha$ for all k, then $\mathcal{R}_{\max}(\theta; P) \leq \mathrm{CVaR}_\alpha(\theta; P)$



χ^2-DRO

- $D_{\chi^2}(Q \parallel P) = \frac{1}{2}\int \left(\frac{dQ}{dP} - 1\right)^2 dP \qquad \rho = \frac{1}{2}\left(\frac{1}{\alpha} - 1\right)^2$
- $\mathcal{R}_{\max}(\theta; P) \leq \mathrm{CVaR}_{\alpha}(\theta; P) \leq \mathcal{R}_{D_{\chi^2},\rho}(\theta; P)$

Worst-case CVaR χ^2-DRO

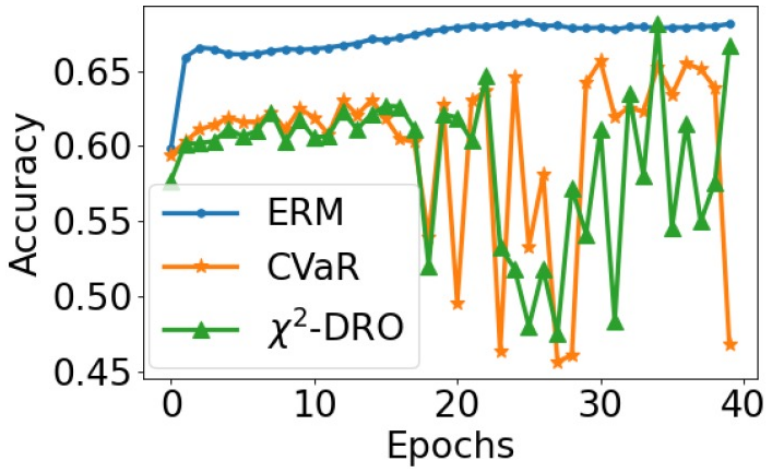
Contents

Experiment Settings

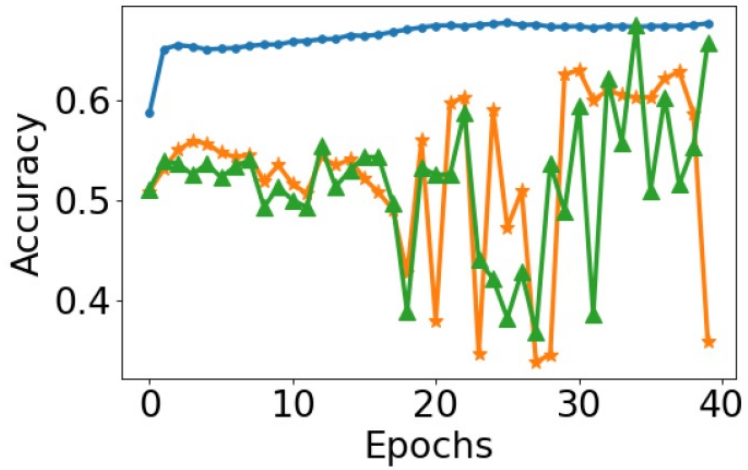
- COMPAS: Recidivism prediction dataset.
- Subpopulations: White, Others, Male, Female.
- Compare CVaR and χ^2-DRO with ERM (Empirical Risk Minimization).

Note that the subpopulations can overlap with each other

Average/Worst-case Accuracy on the Original Dataset



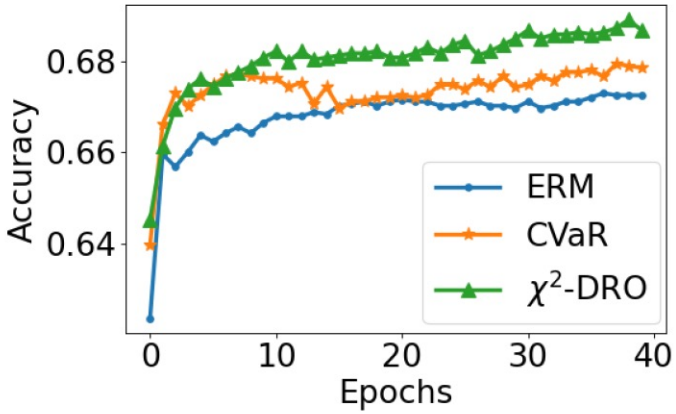
(a) Average (Original)



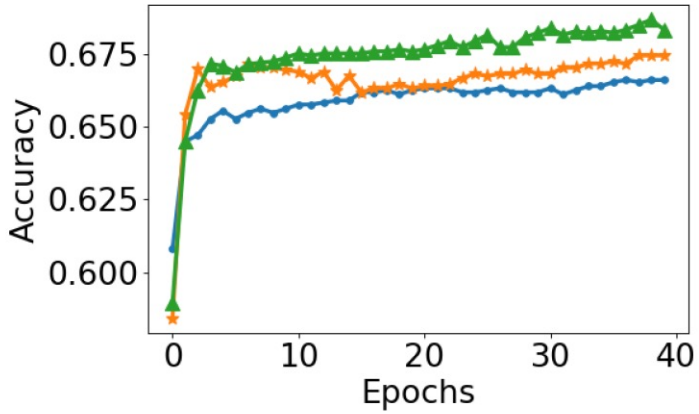
(b) Worst (Original)

Removing Outliers

- Construct a “clean” dataset by removing possible outliers.
- For 5 rounds, train a model with ERM and remove 200 samples with the highest losses. (1000 removed in total)



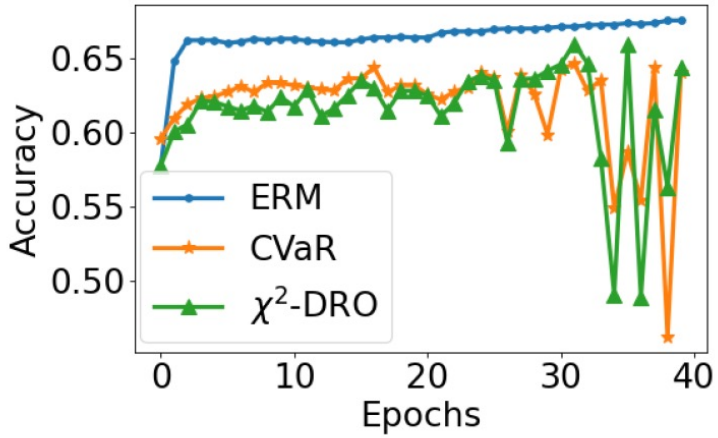
(e) Average (Outliers removed)



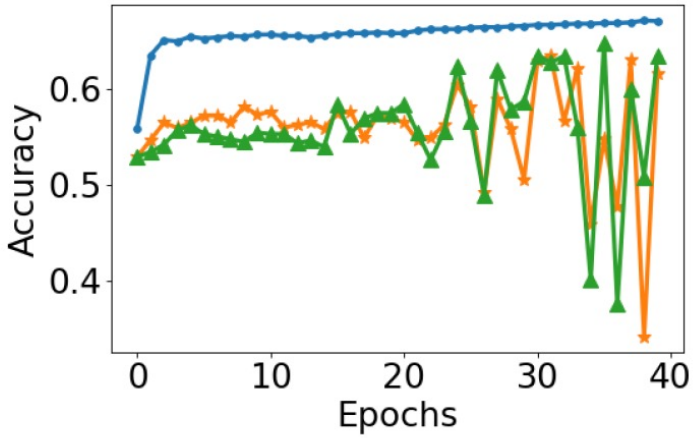
(f) Worst (Outliers removed)

Flipping Labels

- Add outliers by randomly flipping 20% of the labels in the "clean" dataset.



(g) Average (Labels flipped)



(h) Worst (Labels flipped)

Summary of the Results

- DRO is poor and unstable on the original dataset
- Removing outliers mitigates the problem
- Adding outliers exacerbates the problem
- Conclusion: DRO is sensitive to outliers

Open question from Hashimoto et al., 2018: Is it possible to make DRO robust to outliers?

Hashimoto et al. Fairness without demographics in repeated loss minimization. ICML 2018.

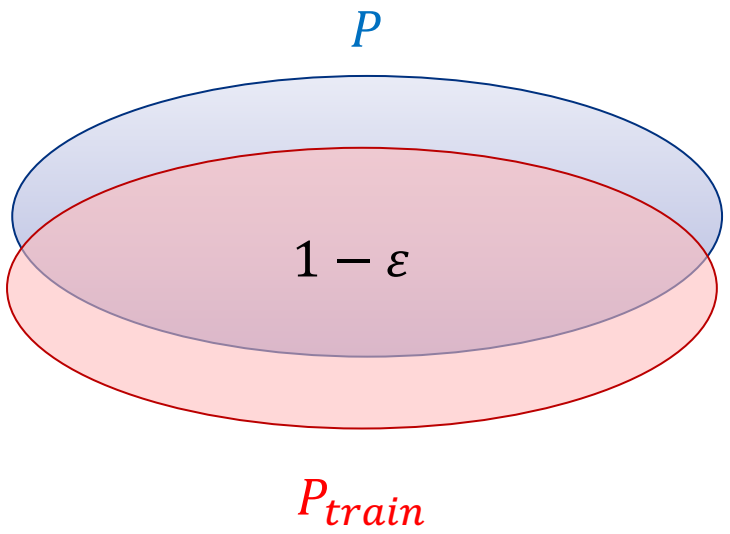
Contents

Huber's Contamination Model

- P: real data distribution
- The observed data distribution is

$$P_{train} = (1 - \varepsilon)P + \varepsilon\tilde{P}$$

where ε is the noise level, and $\tilde{P}$ is an arbitrary distribution.

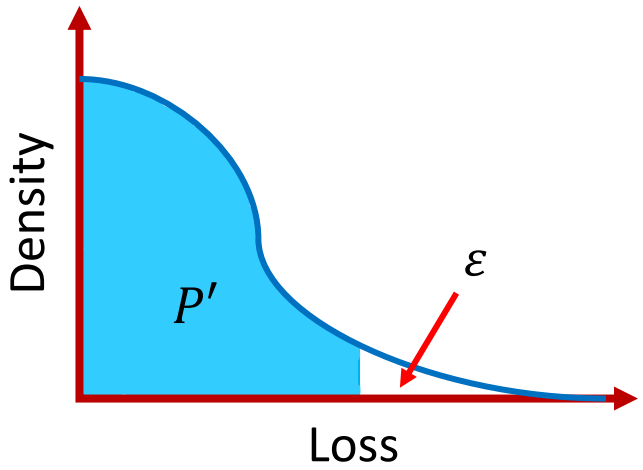


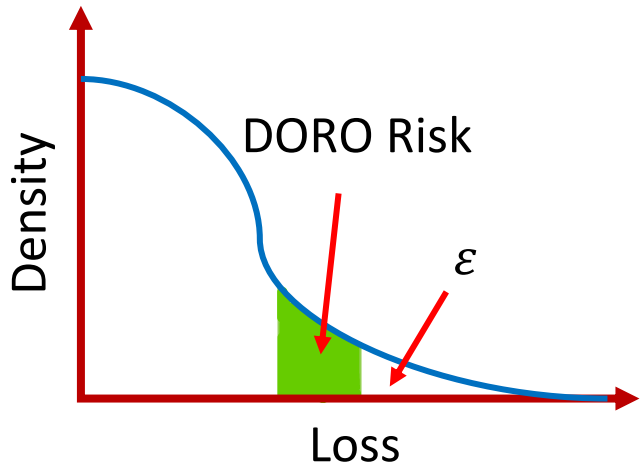
DORO: Distributional and Outlier Robust Optimization

- Minimize the ε-DORO risk:

DRO Risk over the best P'

$$\mathcal{R}_{D,\rho,\varepsilon}(\theta; P_{train}) = \inf_{P'} \left\{ \mathcal{R}_{D,\rho}(\theta; P') : \exists \tilde{P}' \; s.t. \; P_{train} = (1-\varepsilon)P' + \varepsilon \tilde{P}' \right\}$$





Theoretical Results

- Assumption: ℓ has a bounded second moment over P: $\mathbb{E}_P[\ell(\theta; Z)^2] \leq \sigma^2$

1. The minimizer of the DORO risk over P_{train} achieves a DRO risk close to the minimum over P.

2. $$\mathcal{R}_{\max}(\theta; P) \leq \max\left\{3\mathrm{CVaR}_{\alpha,\varepsilon}(\theta; P_{train}), 3\alpha^{-1}\sigma\sqrt{\frac{\varepsilon}{1-\varepsilon}}\right\}$$

Worst-case risk over P

CVaR-DORO risk over P_{train}

Can be replaced by χ^2-DORO as it upper bounds CVaR-DORO

Contents

Experimental Results

- Datasets: COMPAS, CelebA, CivilComments-Wilds.
- DORO improves the worst-case accuracy and the training stability of DRO.

	Method	Average	Worst
CelebA	ERM	95.01	53.94
	CVaR	82.83	66.44
	CVaR-DORO	92.91	72.17

Average/Worst-case Accuracy (%)
(Average over 10 runs)

	Method	Average	Worst
CelebA	ERM	0.73	8.59
	CVaR	11.53	21.47
	CVaR-DORO	4.03	16.84

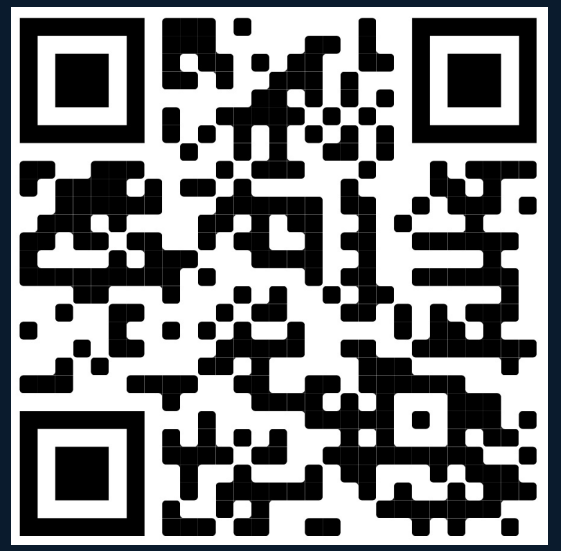
Std. Dev. of Average/Worst-case Accuracy across epochs (%)
(Average over 10 runs)

Open Question: Model Selection

- We use group labels during validation in our experiments, which should be unknown under the domain-oblivious setting.
- Problem: No known model selection method that makes DRO or DORO significantly better than ERM without group labels.

Thanks

Paper



Code