Long Term Fairness for Minority Groups via Performative DRO

Liam Peet-Pare, Nidhi Hegde, Alona Fyshe

Department of Computing Science, University of Alberta Contact: peetpare@ualberta.ca

Performative Prediction

Performative prediction attempts to formalize the notion of a model affecting the distribution on which it is making predictions in a type of feedback loop.

Definition

(Performative Risk) The performative risk of a model is:

$$PR(\theta) = \mathbb{E}_{Z\sim\mathcal{D}(\theta)}[\ell(Z;\theta)].$$

Definition

(Performative Stability) A model, $f_{\theta_{PS}}$, is performatively stable if the following relationship holds:

$$\theta_{PS} = \arg\min_{\theta\in\Theta} \mathbb{E}_{Z\sim\mathcal{D}(\theta_{PS})}[\ell(Z;\theta)].$$

Definition

(Repeated Risk Minimization) Repeated risk minimization (RRM) refers to the procedure where, starting from an initial model f_{θ_0}, we perform the following sequence of updates for every $t \geq 0$:

$$\theta_{t+1} = G(\theta_t) := \arg\min_{\theta\in\Theta} \mathbb{E}_{Z\sim\mathcal{D}(\theta_t)}[\ell(Z;\theta)].$$

Distributionally Robust Optimization

The *de facto* objective used in most supervised learning settings is ERM:

$$\hat{h} = \arg\min_{h\in\mathcal{H}} \frac{1}{n} \sum_{i=1}^{n} \ell(h(x_i), y_i),$$

Instead, we can use DRO:

$$\underset{\theta\in\Theta}{\text{minimize}} \left\{ \sup_{Q\ll P_0} \{\mathbb{E}_Q[\ell(\theta;X)] : Q \in \mathcal{U}_f(P_0)\} \right\}.$$

DRO optimizes for the worst-case expected loss in an uncertainty set around the empirical distribution.

Main Takeaway

We build toward realistic, long-term fairness for minority groups, without requiring access to demographic information, by extending *performative prediction* to a *distributionally robust objective* in order to address key limitations of formal fairness criteria:

- They apply only to static supervised learning problems.
- They rely on access to demographic information.
- They ignore intersectionality.

Figure: Paper on arXiv

Theoretical Contributions

We extend several definitions related to smoothness to the distributionally robust objective to prove a convergence result for repeated distributionally robust optmimization.

RDRO is a Contraction Mapping

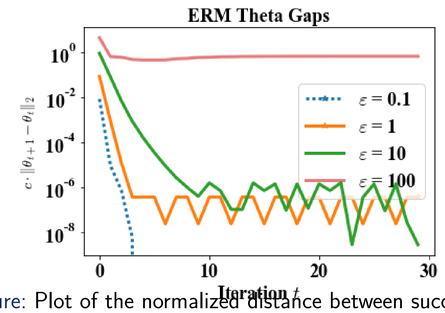

Figure: Plot of the normalized distance between successive values of θ for ERM.

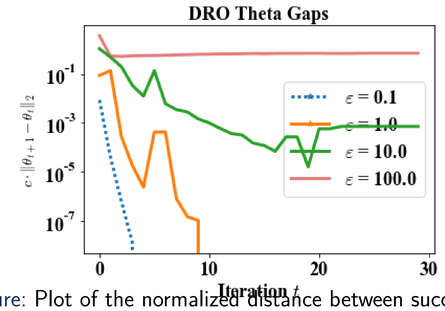

Figure: Plot of the normalized distance between successive values of θ for DRO.

Fair Fixed Points

ERM Performative Accuracy

Group	$\epsilon = 0.01$	$\epsilon = 0.25$	$\epsilon = 0.5$
A	0.893	0.896	0.898
B	0.540	0.540	0.540
All Data	0.834	0.837	0.838

Table: Accuracy by Group for ERM after 30 iterations.

DRO Performative Accuracy

Group	$\epsilon = 0.01$	$\epsilon = 0.25$	$\epsilon = 0.5$
A	0.687	0.710	0.738
B	0.670	0.660	0.660
All Data	0.684	0.701	0.725

Table: Accuracy by Group for DRO after 30 iterations.