Wasserstein Distributionally Robust Optimization Through the Lens of Structural Causal Models and Individual Fairness

Ahmad-Reza Ehyaei [1] Golnoosh Farnadi [2 3] Samira Samadi [1]

[1]Max Planck Institute for Intelligent Systems, Tübingen AI Center [2]Mila - Québec AI Institute [3]McGill University

Introduction

- **Individual Fairness** ensures similar individuals receive similar outcomes. For a classifier h, individual fairness is achieved if:

$$d_{\mathcal{Y}}(h(v), h(w)) \leq L\, d_{\mathcal{V}}(v, w) \quad \forall v, w \in \mathcal{V}$$

where $d_{\mathcal{V}}$ and $d_{\mathcal{Y}}$ are **fair metrics** on feature and label spaces, respectively.

- **Distributionally Robust Optimization** or DRO aims to minimize the gap between in-sample and out-of-sample losses using ambiguity sets based on the **Wasserstein distance**. For distributions $\mathbb{P}, \mathbb{Q}$ with cost function $c(\cdot,\cdot)$, the optimal transport cost is:

$$W_{c,p}(\mathbb{P},\mathbb{Q}) = \min_{\pi \in \mathcal{P}(\mathcal{Z}\times\mathcal{Z})} \left\{ \left(\mathbb{E}_{(z,z')\sim\pi}[c^p(z,z')]\right)^{\frac{1}{p}} : \pi_1 = \mathbb{P}, \pi_2 = \mathbb{Q} \right\}$$

The **Wasserstein ambiguity set** includes all probability measures within a specified distance, defined by the ambiguity radius, $\mathbb{B}_\delta(\mathbb{P}) := \{\mathbb{Q} : W_{c,p}(\mathbb{Q},\mathbb{P}) \leq \delta\}$. If $\ell(z,\theta)$ is the learning loss function the **worst-case loss** quantity is obtained by $\mathcal{R}_\delta(\mathbb{P},\theta) = \sup_{\mathbb{Q}\in\mathbb{B}_\delta(\mathbb{P})} \mathbb{E}_{\mathbb{Q}}[\ell(\mathbf{Z},\theta)]$.

- **Counterfactuals.** In a structural causal model, a **counterfactuals** represents the hypothetical outcome of a variable under an alternative scenario, given a fixed set of observed conditions or interventions on other variables. Counterfactual are derived from **hard** and **soft interventions** in SCMs using *do*-**calculus**:
  - **Hard interventions**: fix features $\mathbf{V}_{\mathcal{I}}$ to constants τ, modifying causal connections for $\mathbf{V}_{\mathcal{I}}$ while keeping other equations intact, represented as $\{\mathbf{V}_i := \tau_i,\ i \in \mathcal{I};\quad \mathbf{V}_i := f_i(\mathbf{V}_{\mathbf{Pa}(i)}, \mathbf{U}_i),\ i \notin \mathcal{I}\}$.
  - **Soft interventions**: adjust equations by adding shifts Δ without altering causal links $\{\mathbf{V}_i := f_i(\mathbf{V}_{\mathbf{Pa}(i)}, \mathbf{U}_i) + \Delta_i\}_{i=1}^n$.

Hard and shift counterfactuals are defined as $\mathsf{CF}(v,\tau)$ and $\mathsf{CF}(v,\Delta)$, respectively, which are central to this analysis. Counterfactuals with modified sensitive attributes, termed **twins**, are essential for individual fairness. Twins are generated by altering a sensitive attribute from a to a', yielding a set $\{\ddot{v}_a = \mathrm{CF}(v,a) : a \in \mathcal{A}\}$ to assess fairness by comparing outcomes across attribute values.

Causally Fair Dissimilarity Function

In the presence of causality and sensitive attributes, we expect that a fair metric should consider two key properties:

- **Zero Dissimilarity for Twin Pairs:** For any $v \in \mathcal{V}$ and $a \in \mathbf{A}$, the dissimilarity $d(v, \ddot{v}_a)$ between an instance and its twins is zero.
- **Guaranteed Similarity for Minor Perturbations:** For every $v \in \mathcal{V}$ and any $\delta > 0$, there exists an ϵ such that for any sufficiently small intervention ($\|\Delta\| \leq \epsilon$) on the non-sensitive attributes ($P_{\mathcal{A}}(\Delta) = 0$), the distance $d(v, \mathsf{CF}(v,\Delta))$ remains less than δ.

This function is called **causally fair dissimilarity function (CFDF)**. In the case, that $\mathcal{M}$ is an ANM, there exists the bijection map $g : \mathcal{U} \to \mathcal{X}$ from exogenous to endogenous space such that $x = g(u)$. If $P_{\mathcal{X}}(u)$ the projection of vector u to the non-sensitive part $\mathcal{U}_{\mathcal{X}}$ then CFDF d can be represented as a dissimilarity function $d_{\mathcal{X}}$ dependent solely on the non-sensitive components $\mathcal{U}_{\mathcal{X}}$ i.e.,

$$d(v,v') = d_{\mathcal{X}}(P_{\mathcal{X}}(g(v)), P_{\mathcal{X}}(g(v'))).$$

We have below **Assumption 1.** for the feature space and CFDF:

1. The CFDF is defined as $d(v,v') = \left\|P_{\mathcal{X}}(g(v)) - P_{\mathcal{X}}(g(v'))\right\|$, where $\|.\|$ is a some norm.
2. Cost function over $\mathcal{Z}$ has form $c((v,y),(v',y')) = d(v,v') + \infty \cdot |y - y'|$.
3. The ambiguity set is defined as: $\mathbb{B}_\delta(\mathbb{P}) = \{\mathbb{Q} \in \mathcal{P}(\mathcal{V}) : W_{c,p}(\mathbb{P},\mathbb{Q}) \leq \delta\}$, for $p \in [1,\infty)$.

Causally Fair Distributionally Robust Optimization

Theorem 1 (Causally Fair Strong Duality). Let $\mathbb{P}$ be probability distribution and $\psi : \mathcal{V} \to \mathbb{R}$ be upper semi-continuous and L_1-integrable function, then following duality holds:

$$\sup_{\mathbb{Q}\in\mathbb{B}_\delta(\mathbb{P})} \left\{ \mathop{\mathbb{E}}_{v\sim\mathbb{Q}}[\psi(v)] \right\} = \inf_{\lambda\geq 0} \left\{ \lambda\delta^p + \mathop{\mathbb{E}}_{v\sim\mathbb{P}} \left[\sup_{a\in\mathcal{A}} \psi_\lambda(\ddot{v}_a) \right] \right\},$$

where $\psi_\lambda(v)$ is defined as

$$\psi_\lambda(v) := \sup_{\Delta\in\mathcal{X}} \left\{ \psi(\mathsf{CF}_0(v,\Delta)) - \lambda^p d(v, \mathsf{CF}_0(v,\Delta)) \right\},$$

and CF_0 is counterfactual regarding parent-free SCM $\mathcal{M}_0$.

Theorem 2 (Higher Order Linear Loss). Given Assumptions, let $\mathcal{M}$ be a linear SCM and the loss function $\ell(z,\theta)^p$, where $\ell(z,\theta)$ is of the form $h(y - \langle\theta, v,\rangle)$ or $h(y \cdot \langle\theta, v,\rangle)$ for functions $h(t)$ such as $|t|$, $\max(0,t)$, $|t-\tau|$, or $\max(0, t-\tau)$ for some $\tau \geq 0$, and $p \in [1,\infty)$. Then the DRO problem can be reduced to:

$$\mathcal{R}_\delta(\mathbb{P}_N,\theta) = \begin{cases} \left(\mathcal{R}_\delta^{cf}(\mathbb{P}_N,\theta)^{\frac{1}{p}} + \delta \left\|P_{\mathcal{X}}(M^T\theta)\right\|_*\right)^p, & \text{diam}(\mathcal{A}) < \infty \\ \left(\mathcal{R}(\mathbb{P}_N,\theta)^{\frac{1}{p}} + \delta \left\|P_{\mathcal{X}}(M^T\theta)\right\|_*\right)^p, \quad \text{s.t.} \quad P_{\mathcal{A}}(M^T\theta) = 0; & \text{diam}(\mathcal{A}) = \infty \end{cases}$$

where M is the corresponding matrix for the linear map g^{-1} and

$$\mathcal{R}_\delta^{cf}(\mathbb{P},\theta) = \mathop{\mathbb{E}}_{v\sim\mathbb{P}} \left[\sup_{a\in\mathcal{A}} \ell(\ddot{v}_a, y, \theta) \right]$$

is a counterfactual loss function.

Theorem 3 (Nonlinear Loss). Let $p = 1$, $\mathcal{M}$ be linear SCM with matrix M corresponding to map g^{-1}, and $\ell(z,\theta)$ be a loss function of the form $h(y - \langle\theta, v\rangle)$ for regression and $h(y \cdot \langle\theta, v\rangle)$ for classification, where h has the following two properties:

1. h is Lipschitz on $\mathbb{R}$ with L_h constant, i.e., $|h(t_2) - h(t_1)| \leq L_h|t_2 - t_1|, \quad \forall t_1, t_2 \in \mathbb{R}$.
2. There exists sequence of $\{t_k\}_{k=1}^\infty$ goes to ∞ such that for each $t_0 \in \mathbb{R}$ we have $lim_{k\to\infty} \frac{|h(t_0 + t_k) - h(t_0)|}{|t_k|} = L_h$.

The DRO formula has the below formulation:

$$\mathcal{R}_\delta(\mathbb{P}_N,\theta) = \begin{cases} \mathcal{R}_\delta^{cf}(\mathbb{P}_N,\theta) + \delta L_h \left\|P_{\mathcal{X}}(M^T\theta)\right\|_*, & \text{diam}(\mathcal{A}) < \infty \\ \mathcal{R}(\mathbb{P}_N,\theta) + \delta L_h \left\|P_{\mathcal{X}}(M^T\theta)\right\|_*, \quad \text{s.t.} \quad P_{\mathcal{A}}(M^T\theta) = 0; & \text{diam}(\mathcal{A}) = \infty \end{cases}$$

Theorem 4 (Finite Sample Guarantee). With mild assumptions for $\hat{\theta}_N^{\text{dro}}$ we have:

$$\mathcal{R}_\delta(\mathbb{P}_*, \hat{\theta}_N^{\text{dro}}) - \inf_{\theta\in\Theta} \mathcal{R}_\delta(\mathbb{P}_*,\theta) \leq N^{-1/2} \left[c_0 + c_1\delta^{1-p} + c_2\delta^{1-p}N^{-\eta+1/2} + c_3\sqrt{\log(2/\epsilon)} \right],$$

With probability at least $1-2\epsilon$. With $\mathfrak{C}(\mathcal{L})$ denoting the Dudley entropy integral for the function class $\{\ell(\cdot,\theta) : \theta \in \Theta\}$, the constants c_0, c_1 and c_2 are identified as follows:

$$c_0 := 96\mathfrak{C}(\mathcal{L}),\ c_1 := 96L \cdot \text{diam}(\mathcal{V})^p,\ c_2 := 2pL \cdot \text{diam}(\mathcal{V})^{p-1} \cdot M_d, \text{ and } c_3 := 2\sqrt{2} \times M.$$

Proposition (Approximation by Robust Optimization). Suppose $\mathcal{A}$ is a finite set and let $\{(v^i, y^i)\}_{i=1}^N$ be observational data. Under Assumption 1, assume that for the loss function ℓ there exist constants $L, M \geq 0$ such that

$$|\ell(v,y,\theta) - \ell(v',y,\theta)| < L d^p(v,v') + M \quad \text{for all} \quad v, v' \in \mathcal{V} \text{ and } p \in [1,\infty).$$

For an arbitrary $K \in \mathbb{N}$, consider the adversarial loss within the setting:

$$\tilde{\mathcal{R}}_\delta^{adv}(\mathbb{P}_N) := \sup_{(w^{ik})_{i,k}\in\tilde{B}_\delta} \left\{ \frac{1}{NK} \sum_{i=1}^N \sum_{k=1}^K \sup_{a\in\mathcal{A}} \ell(\ddot{w}_a^{ik}, y_i, \theta) \right\},$$

where the uncertainty set $\tilde{B}_\delta$ is defined as:

$$\tilde{B}_\delta := \left\{ (w^{ik})_{i,k} : \frac{1}{N} \sum_{i=1}^N \sum_{k=1}^K d^p(v^i, w^{ik}) \leq \delta,\ w^{ik} \in \mathcal{V} \right\}.$$

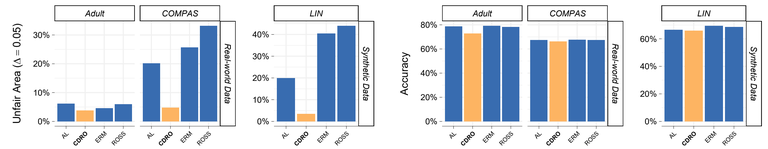
Then, the DRO can be approximated by adversarial optimization as follows:

$$\tilde{\mathcal{R}}_\delta^{adv}(\mathbb{P}_N) \leq \mathcal{R}_\delta(\mathbb{P}_N) \leq \tilde{\mathcal{R}}_\delta^{adv}(\mathbb{P}_N) + \frac{LD + M}{NK},$$

where D is independent of K.

Numerical Experiments

In our numerical studies, we assess the effectiveness of causally fair DRO, referred to as **CDRO**, in mitigating individual unfairness. We compare CDRO's performance against Empirical Risk Minimization (ERM), non-causal Adversarial Learning (AL), and the Ross method. Our experiments use real-world datasets, specifically the Adult and COMPAS datasets, as well as a synthetic dataset based on a linear structural causal model (LIN).



The figure displays the findings from our numerical experiment, assessing the performance of DRO across different models and datasets. (left) Bar plot showing the comparison of models based on the unfair area percentage (lower values are better) for $\Delta = .05$, where $\mathrm{U}_\Delta := \mathbb{P}(\{v \in \mathcal{V} : \exists v' \in \mathcal{V} \text{s.t.} d(v,v') \leq \Delta \wedge h(v) \neq h(v')\})$ is the unfairness area measure. (right) Bar plot comparing methods by prediction accuracy performance (higher values are better).

Main Contributions

- Define a causally fair dissimilarity function, an individual fair metric incorporating causal structures and sensitive attributes, along with its representation form.
- Define a causally fair DRO problem with a causally fair dissimilarity function cost.
- Present the strong duality theorem for causally fair DRO.
- Provide the exact regularizer for linear SCM under mild conditions for the loss function in regression and classification problems.
- Estimate the first-order causally fair DRO regularizer for non-linear SCM.
- Provide the relation between classical robust optimization and causally fair DRO.
- Demonstrate that under unknown SCM assumptions, by estimating the SCM or cost function, we have finite sample guarantees for convergence of empirical DRO problems.

ahmad.ehyaei@tuebingen.mpg.de

The 38th Annual AAAI Conference on Artificial Intelligence (AAAI)

farnadig@mila.quebec