Fairness Transferability Subject to Bounded Distribution Shift

Yatong Chen, Reilly Raab, Jialu Wang, Yang Liu

Presented by Seonghyeon Kim

Formulation

- $(X, Y, G) \in \mathcal{X} \times \mathcal{Y} \times \mathcal{G}$: Random vector of a triplet of output, input and sensitive variable
- $\mathcal{S}, \mathcal{T}$: Source distribution and target distribution
- $\mathcal{P}(\cdot)$: Space of distributions over a given space
- $\Pi := \{\pi : \mathcal{X} \times \mathcal{G} \to \mathcal{P}(\mathcal{Y})\}$: Space of nondeterministic policies
- $\hat{Y}|X, G \sim \pi(X, G)$: Predicted label

Disparity Function

- $\Delta^\star(\pi, \mathcal{T})$: Disparity of a policy π on a distribution $\mathcal{T}$.
- Demographic Parity (DP)

$$\Delta^\star{}_{\mathsf{DP}}(\pi, \mathcal{T}) := \sum_{g,h \in \mathcal{G}} \left| \Pr_{\pi,\mathcal{T}}(\hat{Y}{=}1 \mid G{=}g) - \Pr_{\pi,\mathcal{T}}(\hat{Y}{=}1 \mid G{=}h) \right|$$

- Equal Opportunity (EOp)

$$\Delta^\star{}_{\mathsf{EOp}}(\pi, \mathcal{T}) := \sum_{g,h \in \mathcal{G}} \left| \Pr_{\pi,\mathcal{T}}(\hat{Y}{=}1 \mid G{=}g, Y{=}1) - \Pr_{\pi,\mathcal{T}}(\hat{Y}{=}1 \mid G{=}h, Y{=}1) \right|$$

Vector-Bounded Distribution Shift

- Divergence : $K(P \parallel Q)$ is a *divergence* if and only if K satisfies the followings.
  - $K(P \parallel Q) \geq 0$ for all P and Q
  - $K(P \parallel Q) = 0 \iff P = Q$
- Group-vectorized shift : For given divergences $K_1, \cdots, K_{|\mathcal{G}|}$,

$$\boldsymbol{D}(\mathcal{T} \parallel \mathcal{S}) := \left[K_g\left(\Pr_{\mathcal{T}}(X, Y \mid G{=}g) \parallel \Pr_{\mathcal{S}}(X, Y \mid G{=}g)\right)\right]_{g=1,\cdots,|\mathcal{G}|}.$$

- Vector-bounded distribution shift : For given a vector $\boldsymbol{B} \succeq 0$ (element-wise inequality),

$$\boldsymbol{D}(\mathcal{T} \parallel \mathcal{S}) \preceq \boldsymbol{B}.$$

General Bound

- Supremum disparity function :

$$v(\Delta^\star, \boldsymbol{D}, \pi, \mathcal{S}, \boldsymbol{B}) := \sup_{\boldsymbol{D}(\mathcal{T}\|\mathcal{S}) \preceq \boldsymbol{B}} \Delta^\star(\pi, \mathcal{T})$$

Theorem (Lipshitz Upper Bound)

If there exists an $\boldsymbol{L}$ *such that* $\nabla_{\boldsymbol{b}} v(\Delta^\star, \boldsymbol{D}, \pi, \mathcal{S}, \boldsymbol{b}) \preceq \boldsymbol{L}$*, everywhere along some curve as* $\boldsymbol{b}$ *varies from* 0 *to* $\boldsymbol{B}$*, then for* $\boldsymbol{D}(\mathcal{T} \parallel \mathcal{S}) \preceq \boldsymbol{B}$

$$\Delta^\star(\pi, \mathcal{T}) \leq \Delta^\star(\pi, \mathcal{S}) + \boldsymbol{L} \cdot \boldsymbol{B}.$$

Restricted Distribution Shift

- Covariate shift :

$$\Pr_{\mathcal{T}}(Y \mid X, G) = \Pr_{\mathcal{S}}(Y \mid X, G)$$

- Label shift :

$$\Pr_{\mathcal{T}}(X \mid Y, G) = \Pr_{\mathcal{S}}(X \mid Y, G)$$

Covariate Shift - Demographic Parity

- Covariate shift : $\Pr_{\mathcal{T}}(Y \mid X, G) = \Pr_{\mathcal{S}}(Y \mid X, G)$
- Density ratio : $\omega_g(\mathcal{T}, \mathcal{S}, x) := \frac{p_{\mathcal{T}}(X{=}x \mid G{=}g)}{p_{\mathcal{S}}(X{=}x \mid G{=}g)}$
- Note that $\omega_g(\mathcal{T}, \mathcal{S}, x)$ is extremely large, when $p_{\mathcal{T}}(X{=}x \mid G{=}g)$ is sufficiently large and $p_{\mathcal{S}}(X{=}x \mid G{=}g) \approx 0$
- $\beta_g := \Pr_{\pi,\mathcal{S}}(\hat{Y}{=}1 \mid G{=}g)$

Theorem (Bound for DP under Covariate Shift)

For demographic parity between two groups under covariate shift,

$$\Delta^{\star}{}_{DP}(\pi, \mathcal{T}) \leq \Delta^{\star}{}_{DP}(\pi, \mathcal{S}) + \sum_{g} \left(\beta_g(1-\beta_g) \cdot \operatorname*{Var}_{\mathcal{S}}[\omega_g(\mathcal{T}, \mathcal{S}, x)]\right)^{1/2}$$

Covariate Shift - Equal Opportunity

- Covariate shift : $\Pr_{\mathcal{T}}(Y \mid X, G) = \Pr_{\mathcal{S}}(Y \mid X, G)$
- $\beta_g^+(\pi, \mathcal{T}) := \Pr_{\pi,\mathcal{T}}(\hat{Y}{=}1 \mid Y{=}1, G{=}g)$
- $\Delta^\star{}_{\mathsf{EOp}}(\pi, \mathcal{T}) = \sum_{g,g' \in \mathcal{G}} \left| \beta_g^+(\pi, \mathcal{T}) - \beta_{g'}^+(\pi, \mathcal{T}) \right|$

Theorem (Bound for EOp under Covariate Shift)

Subject to covariate shift and any given $\boldsymbol{D}, \boldsymbol{B}$, assume extremal values for β_g^+, i.e.,

$$\forall g, \quad \left(D_g(\mathcal{T} \parallel \mathcal{S}) < B_g\right) \implies \left(l_g \leq \beta_g^+(\pi, \mathcal{T}) \leq u_g\right)$$

it follows that

$$v(\Delta^\star{}_{EOp}, \boldsymbol{D}, \pi, \mathcal{S}, \boldsymbol{B}) \leq \max_{\boldsymbol{r} \in \prod_{g=\mathbf{1}}^{|\mathcal{G}|} \{l_g, u_g\}} \sum_{h,h' \in \mathcal{G}} \left| r_h - r_{h'} \right|.$$

Label Shift - Demographic Parity

- Label shift : $\Pr_{\mathcal{T}}(X \mid Y, G) = \Pr_{\mathcal{S}}(X \mid Y, G)$
- Note that $\Pr_{\pi,\mathcal{T}}(\hat{Y} \mid Y, G) = \Pr_{\pi,\mathcal{S}}(\hat{Y} \mid Y, G)$, so EOp is invariant.
- $\beta_g^+ := \Pr_\pi(\hat{Y}{=}1 \mid Y{=}1, G{=}g)$; $\quad \beta_g^- := \Pr_\pi(\hat{Y}{=}1 \mid Y{=}0, G{=}g)$
- $Q_g(\mathcal{T}) := \Pr_{\mathcal{T}}(Y = 1 \mid G = g)$

Theorem (Bound for DP under Label Shift)

For DP under the bounded label-shift assumption $\forall g, |Q_g(\mathcal{S}) - Q_g(\mathcal{T})| \leq B_g$,

$$\Delta^\star_{DP}(\pi, \mathcal{T}) \leq \Delta^\star_{DP}(\pi, \mathcal{S}) + (|\mathcal{G}| - 1) \sum_g B_g \left| \beta_g^+ - \beta_g^- \right|$$