Regression under demographic parity constraints via unlabeled post-processing

Gayane Taturyan, Evgenii Chzhen, Mohamed Hebiri

38th Conference on Neural Information Processing Systems, 2024

Regression under DP: The setup

$$(\underbrace{\text{feature}}_{\boldsymbol{X}}, \underbrace{\text{sensitive attribute}}_{S}, \underbrace{\text{label}}_{Y}) \sim \mathbb{P} \text{ on } \mathbb{R}^d \times [K] \times \mathbb{R}$$

A randomized prediction function $\pi : \mathcal{B}(\mathbb{R}) \times \mathbb{R}^d \to [0, 1]$

- For any π define $\hat{Y}_\pi$ $s.t.$ $\mathrm{Law}(\hat{Y}_\pi \mid \boldsymbol{X} = \boldsymbol{x}, S = s) = \pi(\cdot \mid \boldsymbol{x}) \quad \boldsymbol{x} \in \mathbb{R}^d, s \in [K]$

Risk: $\mathcal{R}(\pi) \stackrel{\text{def}}{=} \mathbb{E}[(\hat{Y}_\pi - \eta(\boldsymbol{X}))^2]$

Unfairness: $\mathcal{U}_s(\pi, \hat{y}) \stackrel{\text{def}}{=} |\mathbb{E}\left[\pi(\hat{y} \mid \boldsymbol{X}) \mid S = s\right] - \mathbb{E}\left[\pi(\hat{y} \mid \boldsymbol{X})\right]|$

Optimal fair estimator:

$$\min_\pi\{\mathcal{R}(\pi) : \mathrm{supp}(\pi(\cdot \mid \boldsymbol{x})) = \hat{\mathcal{Y}} \text{ for } \boldsymbol{x} \in \mathbb{R}^d, \mathcal{U}_s(\pi, \hat{y}) \leq \varepsilon_s \text{ for } \hat{y} \in \hat{\mathcal{Y}}, s \in [K]\}$$

Main quantities:

- $\eta(\boldsymbol{x}) \stackrel{\text{def}}{=} \mathbb{E}[Y \mid \boldsymbol{X} = \boldsymbol{x}]$
- $\boldsymbol{p} \stackrel{\text{def}}{=} (p_s)_{s \in [K]}$, with $p_s \stackrel{\text{def}}{=} \mathbb{P}(S = s)$
- $\boldsymbol{\tau}(\boldsymbol{x}) \stackrel{\text{def}}{=} (\tau_s(\boldsymbol{x}))_{s \in [K]}$, with $\tau_s(\boldsymbol{x}) \stackrel{\text{def}}{=} \mathbb{P}(S = s \mid \boldsymbol{X} = \boldsymbol{x})$

Proposed methodology

Assumption

Bounded signal: $|\eta(\boldsymbol{X})| \leq B$ a.s.

Discretization: For every integer $L \geq 0$ and real $B > 0$, a uniform grid

$$\hat{\mathcal{Y}}_L \stackrel{\text{def}}{=} \left\{-B, -\frac{B(L-1)}{L}, \ldots, -\frac{B}{L}, 0, \frac{B}{L} \ldots, \frac{B(L-1)}{L}, B\right\}$$

Entropic regularization: $\mathcal{R}_\beta(\pi) \stackrel{\text{def}}{=} \mathcal{R}(\pi) + \frac{1}{\beta}\mathbb{E}[\Psi(\pi(\cdot \mid \boldsymbol{X}))]$

- $\Psi(\mu) \stackrel{\text{def}}{=} \sum_{\hat{y} \in \operatorname{supp}(\mu)} \mu(\hat{y}) \log(\mu(\hat{y}))$ - negative entropy

Optimal discretized entropic-regularized fair estimator:

$$\min_{\pi}\{\mathcal{R}_\beta(\pi) : \operatorname{supp}(\pi(\cdot \mid \boldsymbol{x})) = \hat{\mathcal{Y}}_L \text{ for } \boldsymbol{x} \in \mathbb{R}^d,$$
$$\mathcal{U}_s(\pi, \hat{y}) \leq \varepsilon_s \text{ for } \hat{y} \in \hat{\mathcal{Y}}_L, s \in [K]\}$$

Closed form expression of the solution

Lemma

Let $\boldsymbol{t}(\boldsymbol{x}) \stackrel{\text{def}}{=} 1 - \frac{\boldsymbol{\tau}(\boldsymbol{x})}{\boldsymbol{p}}$, $r_\ell(\boldsymbol{x}) \stackrel{\text{def}}{=} \left(\eta(\boldsymbol{x}) - \frac{\ell B}{L}\right)^2$, and $\boldsymbol{\lambda}_\ell = (\lambda_{\ell s})_s$, $\boldsymbol{\nu}_\ell = (\nu_{\ell s})_s$. For $L \in \mathbb{N}$ and $\beta > 0$ optimal discretized entropic-regularized fair estimator is given by

$$\pi_{\boldsymbol{\Lambda}^\star, \mathbf{V}^\star}(\ell \mid \boldsymbol{x}) \stackrel{\text{def}}{=} \sigma_\ell \left(\beta \left(\langle \boldsymbol{\lambda}^\star_{\ell'} - \boldsymbol{\nu}^\star_{\ell'}, \boldsymbol{t}(\boldsymbol{x})\rangle - r_{\ell'}(\boldsymbol{x})\right)_{\ell' \in [\![L]\!]}\right) \text{ for } \ell \in [\![L]\!],$$

where $\boldsymbol{\Lambda}^\star = (\lambda^\star_{\ell s})_{\ell,s}$ and $\mathbf{V}^\star = (\nu^\star_{\ell s})_{\ell,s}$ matrices are solutions to

$$\min_{\boldsymbol{\Lambda}, \mathbf{V} \geq 0} \left\{ F(\boldsymbol{\Lambda}, \mathbf{V}) \stackrel{\text{def}}{=} \mathbb{E}\left[\mathrm{LSE}_\beta\left(\left(\langle \boldsymbol{\lambda}_\ell - \boldsymbol{\nu}_\ell, \boldsymbol{t}(\boldsymbol{X})\rangle - r_\ell(\boldsymbol{X})\right)_{\ell \in [\![L]\!]}\right)\right] + \sum_{\ell \in [\![L]\!]} \langle \boldsymbol{\lambda}_\ell + \boldsymbol{\nu}_\ell, \boldsymbol{\varepsilon}\rangle \right\} .$$

F is convex and its gradient is $(\beta\sigma^2)$-Lipschitz, where $\sigma^2 = 2 \sum_{s \in [K]} \frac{1 - p_s}{p_s}$.

Main observation: Gradient of F is crucial

Parametric family: For any $\mathbf{\Lambda}, \mathbf{V} \geq 0$

$$\pi_{\mathbf{\Lambda},\mathbf{V}}(\ell \mid \boldsymbol{x}) \stackrel{\text{def}}{=} \sigma_\ell \left(\beta\left(\langle \boldsymbol{\lambda}_{\ell'} - \boldsymbol{\nu}_{\ell'}, \boldsymbol{t}(\boldsymbol{x})\rangle - r_{\ell'}(\boldsymbol{x})\right)_{\ell' \in [\![L]\!]}\right) \text{ for } \ell \in [\![L]\!]$$

Gradient mapping: For $\alpha > 0$,

$$\boldsymbol{G}_\alpha(\mathbf{\Lambda}, \mathbf{V}) \stackrel{\text{def}}{=} \frac{(\mathbf{\Lambda}, \mathbf{V}) - ((\mathbf{\Lambda}, \mathbf{V}) - \alpha \nabla F(\mathbf{\Lambda}, \mathbf{V}))_+}{\alpha}$$

Lemma

Let $\sigma^2 \stackrel{\text{def}}{=} 2\sum_{s\in[K]} \frac{1-p_s}{p_s}$, $L \in \mathbb{N}$, $\mathbf{\Lambda}, \mathbf{V} \geq 0$, then for any $\alpha > 0, \beta > 0$,

- $\displaystyle\sum_{\ell\in[\![L]\!] s\in[K]} \left(\mathcal{U}_s(\pi_{\mathbf{\Lambda},\mathbf{V}}, \ell) - \varepsilon_s\right)_+^2 \leq \|\boldsymbol{G}_\alpha(\mathbf{\Lambda}, \mathbf{V})\|^2$
- $\mathcal{R}(\pi_{\mathbf{\Lambda},\mathbf{V}}) \leq \mathcal{R}(\pi_{\mathbf{\Lambda}^\star,\mathbf{V}^\star}) + \left(\|(\mathbf{\Lambda}, \mathbf{V})\| + \alpha\left\{\sigma + \|\boldsymbol{\varepsilon}\| \sqrt{2|\hat{\mathcal{Y}}_L|}\right\}\right) \|\boldsymbol{G}_\alpha(\mathbf{\Lambda}, \mathbf{V})\| + \frac{\log|\hat{\mathcal{Y}}_L|}{\beta}$

Post-processing algorithm

- Gradient of F:
  $\nabla_{\square_{\ell s}} F(\mathbf{\Lambda}, \mathbf{V}) = \triangle \mathbb{E}\left[\sigma_\ell\left(\beta\left(\langle \boldsymbol{\lambda}_{\ell'} - \boldsymbol{\nu}_{\ell'}, \boldsymbol{t}(\boldsymbol{X})\rangle - r_{\ell'}(\boldsymbol{X})\right)_{\ell'=-L}^{L}\right) t_s(\boldsymbol{X})\right] + \varepsilon_s$
- Stochastic gradient of F:
  $g_{\square_{\ell s}}(\mathbf{\Lambda}, \mathbf{V}) = \triangle \sigma_\ell\left(\beta\left(\langle \boldsymbol{\lambda}_{\ell'} - \boldsymbol{\nu}_{\ell'}, \boldsymbol{t}(\boldsymbol{X})\rangle - r_{\ell'}(\boldsymbol{X})\right)_{\ell'=-L}^{L}\right) t_s(\boldsymbol{X}) + \varepsilon_s$

  (where $\square \in \{\lambda, \nu\}$ and $\triangle = 1$ if $\square = \lambda$ and $\triangle = -1$ otherwise)

Controlled variance: $\mathbb{E}\|\boldsymbol{g}(\mathbf{\Lambda}, \mathbf{V}) - \nabla F(\mathbf{\Lambda}, \mathbf{V})\|^2 \leq \sigma^2$, where $\sigma^2 = 2\sum_{s\in[K]} \frac{1-p_s}{p_s}$.

The algorithm

- **Input:** $L, T, \beta, \boldsymbol{p}, B, \eta, \boldsymbol{\tau}$
- Build uniform grid $\hat{\mathcal{Y}}_L$ over $[-B, B]$
- Set parameters: $\sigma^2 = 2\sum_{s\in[K]} \frac{1-p_s}{p_s}$, $M = \beta\sigma^2$
- Set $(\mathbf{\Lambda}, \mathbf{V}) \mapsto F(\mathbf{\Lambda}, \mathbf{V})$
- Run a black-box optimizer $\mathcal{A}(F, \sigma^2, M, T)$ to obtain $(\hat{\mathbf{\Lambda}}, \hat{\mathbf{V}})$
- **Return:** $\pi_{(\hat{\mathbf{\Lambda}}, \hat{\mathbf{V}})}(\cdot \mid \cdot)$

Theoretical guarantees

For deterministic prediction:

$$\mathcal{R}^\star \stackrel{\text{def}}{=} \inf_{h:\mathbb{R}^d \to [-B,B]} \left\{ \mathcal{R}(h) \,:\, \sup_{t\in\mathbb{R}} |\mathbb{P}(h(\boldsymbol{X}) \le t \mid S = s) - \mathbb{P}(h(\boldsymbol{X}) \le t)| \le \frac{\varepsilon_s}{2}, \forall s \in [K] \right\}$$

Theorem

With $\boldsymbol{\varepsilon} = (\varepsilon_s)_{s\in[K]} \in [0,1]^K$, $\sigma^2 = 2 \sum_{s\in[K]} \frac{1-p_s}{p_s}$, setting $\beta = \frac{T}{8\log_2(T)}$ and $L = \sqrt{T}$

- $\mathbf{E}^{1/2}\left[\sum_{\ell\in[\![L]\!] s\in[K]} \left(\mathcal{U}_s(\pi_{\hat{\boldsymbol{\Lambda}},\hat{\mathbf{V}}}, \ell) - \varepsilon_s \right)_+^2 \right] \le \tilde{\mathcal{O}}\left(\frac{\sigma}{\sqrt{T}} \left(1 + \frac{\sigma}{\sqrt{T}} \|(\boldsymbol{\Lambda}^\star, \mathbf{V}^\star)\| \right) \right)$
- $\mathcal{E}(\pi_{\hat{\boldsymbol{\Lambda}},\hat{\mathbf{V}}}) \le \tilde{\mathcal{O}}\left(\left(\frac{\sigma}{\sqrt{T}} \mathbf{E}^{1/2}\left[\|(\hat{\boldsymbol{\Lambda}}, \hat{\mathbf{V}})\|^2 \right] + \frac{\|\boldsymbol{\varepsilon}\|}{T^{5/4}} \right) \left(1 + \frac{\sigma}{\sqrt{T}} \|(\boldsymbol{\Lambda}^\star, \mathbf{V}^\star)\| \right) + \frac{B}{\sqrt{T}} \right)$
  where $\mathcal{E}(\pi_{\hat{\boldsymbol{\Lambda}},\hat{\mathbf{V}}}) \stackrel{\text{def}}{=} \mathbb{E}\left[\mathcal{R}(\pi_{\hat{\boldsymbol{\Lambda}},\hat{\mathbf{V}}}) \right] - \mathcal{R}^\star$.

- Extension to unknown η and $\boldsymbol{\tau}$: The guarantees still hold when replacing η and $\boldsymbol{\tau}$ with their estimates $\hat{\eta}$ and $\hat{\boldsymbol{\tau}}$ if we pay additional price for estimation of η and $\boldsymbol{\tau}$.

Thank you!