

# Variance Regularization of Causal Effects for Fair Classification

近原 鷹一 藤野 昭典


(NTTコミュニケーション科学基礎研究所)

- 目的** 差別の定義が因果パスで表されるとき、各人にとって公平な分類器を学習
- 手法** 因果効果の平均/分散を制約する最適化問題に対し、弱凸な目的関数を提案
- 結果** 停留点収束を理論保証、因果効果の平均/分散が0になることを実験で確認

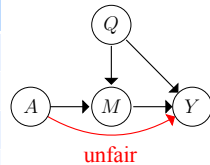
## 1. 問題設定

入力: 訓練データ + 因果グラフ

出力: 分類器



$A$	$Q$	$M$	$Y$
0	1	175	1
1	2	165	0
0	0	156	0
0	0	147	0



$$Y = \hat{h}(A, M, Q)$$

## 3. 提案

上側信頼区間(平均+標準偏差)を制約

$$\begin{aligned} & \mathbb{E}_{\hat{P}_n} [ |g_\theta(\mathbf{X})| ] + \sqrt{\frac{2\rho}{n} \text{Var}_{\hat{P}_n} [ |g_\theta(\mathbf{X})| ]} \\ &= \max_{\mathbf{p} \in \mathcal{P}_{\rho,n}} \mathbb{E}_{\mathbf{p}} [ |g_\theta(\mathbf{X})| ] \end{aligned}$$

cf., [Namkoong+; NeurIPS2017]

## 2. 解きたい問題

$$\begin{aligned} & \underset{\theta}{\text{minimize}} && \mathbb{E}_{\hat{P}_n} [l(h_\theta(\mathbf{X}), Y)] \\ & \text{subject to} && -\delta \leq \mathbb{E}_{\hat{P}_n} [g_\theta(\mathbf{X})] \leq \delta, \text{Var}_{\hat{P}_n} [g_\theta(\mathbf{X})] \leq \zeta \end{aligned}$$

$g_\theta(\mathbf{X})$ : 因果効果

目的関数を弱凸関数として定式化

$$\min_{\theta} \max_{\mathbf{p} \in \mathcal{P}_{\rho,n}} \mathbb{E}_{\hat{P}_n} [l(h_\theta(\mathbf{X}), Y)] + \nu \mathbb{E}_{\mathbf{p}} [ |g_\theta(\mathbf{X})| ]$$



難しさ: 分散制約のせいで非凸かつ非滑らかな問題に

分類器がDNNでも収束保証可能