OptiML Group Meeting, 2022-2023 Winter

On Learning Fairness and Accuracy on Multiple Subgroups [Shui et al., 2022]

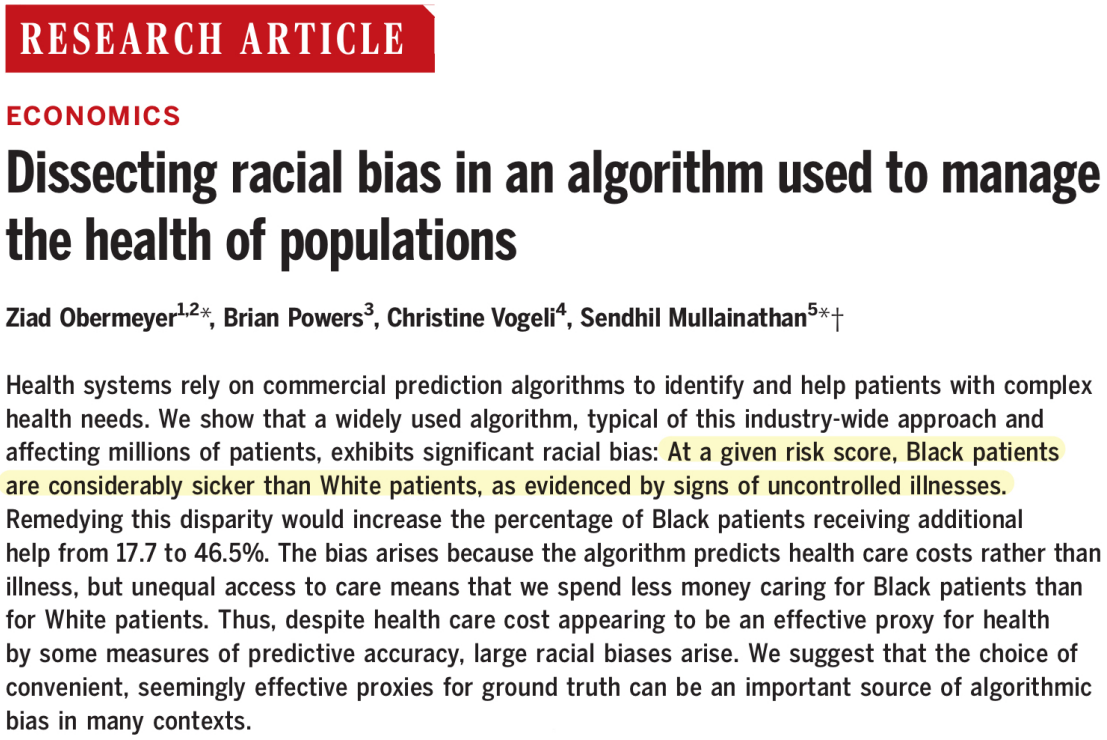
Presenter: Hanseul Cho

January 4th, 2023

Table of Contents

Fair learning

- Algorithmic bias is problematic, especially in socio-technical systems
  - *e.g.*, Medical AI (health risk assessment)—severity of black patients is often underestimated [Obermeyer et al., 2019]



RESEARCH ARTICLE

ECONOMICS

Dissecting racial bias in an algorithm used to manage the health of populations

Ziad Obermeyer[1,2]*, Brian Powers[3], Christine Vogeli[4], Sendhil Mullainathan[5]*†

Health systems rely on commercial prediction algorithms to identify and help patients with complex health needs. We show that a widely used algorithm, typical of this industry-wide approach and affecting millions of patients, exhibits significant racial bias: At a given risk score, Black patients are considerably sicker than White patients, as evidenced by signs of uncontrolled illnesses. Remedying this disparity would increase the percentage of Black patients receiving additional help from 17.7 to 46.5%. The bias arises because the algorithm predicts health care costs rather than illness, but unequal access to care means that we spend less money caring for Black patients than for White patients. Thus, despite health care cost appearing to be an effective proxy for health by some measures of predictive accuracy, large racial biases arise. We suggest that the choice of convenient, seemingly effective proxies for ground truth can be an important source of algorithmic bias in many contexts.

Desiderata in Fair ML

1. **Fair**: no disparities in different demographics
   - No unified definition; several (possibly conflicting) notions of fairness
   - Trivially fair decision can be non-informative; *e.g.*, flipping coin.
2. **Informative**: learning the utility of the data ($\coloneqq$ accuracy)
   - "Fairness-accuracy trade-off": fairness regarded as a constraint

- Can these two be achieved simultaneously?
  - Depending on fairness notions.

Problem Setting

- Binary classification on a dataset with multiple subgroups
  - Underlying distribution: $\mathcal{D}(X, Y, A)$
  - Input $X \in \mathcal{X}$, label $Y \in \mathcal{Y} = \{0, 1\}$
  - Sensitive attribute $A \in \mathcal{A}$ (scalar, discrete)
  - Predictor: a real-valued function $f : \mathcal{X} \to [0, 1]$
    - ⋆ **Bayesian framework**: we want to learn posterior distribution $\tilde{f} \sim Q$ with a randomized algorithm
    - ⋆ Inference: $f(X) = \mathbb{E}_{\tilde{f} \sim Q}[\tilde{f}(X)]$

Group Fairness

- In real world, sensitive attribute (or subgroup index) $\mathcal{A}$ often have more than two (or even many) elements.
  - Race (black, white, Asian, Latino, ...)
  - Gender (male, female, others/not responded)
  - Religion (Christian, Jewish, Muslim, Buddhist, ...)
  - Nationality
  - Individual customers ...
- Caveat: 'fairness on multiple subgroups' $\neq$ 'multi-group fairness'
  - **Group fairness** ($A \in \mathcal{A}$ is a scalar): considering only a single axis of groups which is not necessarily binary $\rightarrow$ today's setting
  - Multi-group fairness ($A \in \mathcal{A}$ is a vector): multiple axes of dividing groups[1] [Yang et al., 2020, Kang et al., 2021, Alghamdi et al., 2022]

[1] *e.g.*, $(x_1, y_1, \text{male}, \text{black})$, $(x_2, y_2, \text{female}, \text{white})$, $(x_3, y_3, \text{male}, \text{white})$...

Table of Contents

Several definitions of group fairness

- Demographic parity[2] (DP): $\mathbb{E}[f(X)] = \mathbb{E}[f(X)|A]$
  - "Prediction scores must be equal across subgroups"
- Equalized odds (EO or EOD): $\mathbb{E}[f(X)|Y] = \mathbb{E}[f(X)|Y, A]$
  - "Given a label Y, prediction scores must be equal across subgroups"
- **Group Sufficiency**: $\mathbb{E}[Y|f(X)] = \mathbb{E}[Y|f(X), A]$
  - "Given a score $f(X) = \tau$, labels must be equal across subgroups"

Remarks: when A and Y are not completely independent...

- DP/EO based criteria suffers from the fairness-accuracy trade-off [Song et al., 2019, Dutta et al., 2020, Wang et al., 2021].
- Group sufficiency and DP/EO cannot hold simultaneously [Chouldechova, 2017, Pleiss et al., 2017, Barocas et al., 2019].

[2] Also known as 'statistical parity' or 'independence rule'

Fairness violation "gaps"

- Violation of DP $\rightarrow \mathbf{ind}_f := \mathbb{E}_{A,X}[|\mathbb{E}[f(X)] - \mathbb{E}[f(X)|A]|]$
- Violation of EO $\rightarrow \mathbf{sep}_f := \mathbb{E}_{A,X}[|\mathbb{E}[f(X)|Y] - \mathbb{E}[f(X)|Y,A]|]$
- **Group Sufficiency gap**: $\mathbf{Suf}_f := \mathbb{E}_{A,X}[|\mathbb{E}[Y|f(X)] - \mathbb{E}[Y|f(X),A]|]$

Shui et al. [2022, Theorem 4.1]

Group sufficiency gap is upper bounded by

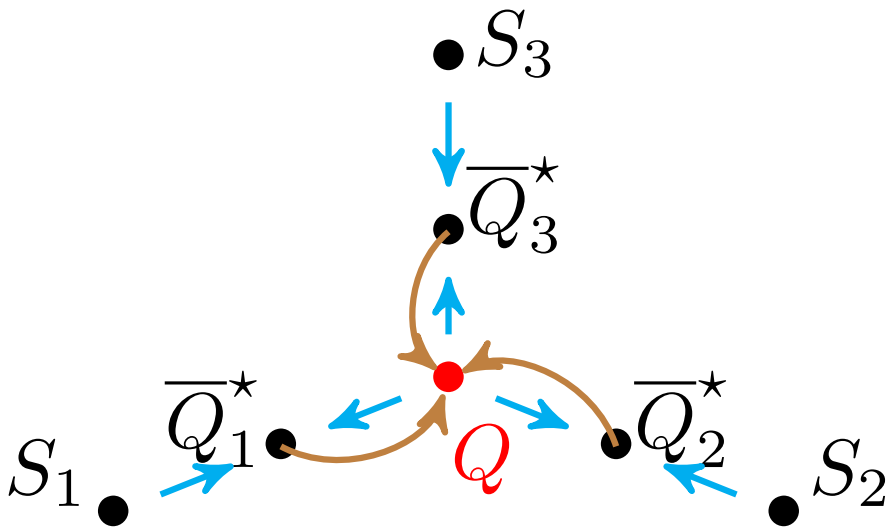
$$\mathbf{Suf}_f \leq 4\mathbb{E}_{A,X}[|f(X) - f_A^{Bayes}(X)|],$$

where $f_A^{Bayes}(X) := \mathbb{E}[Y|X,A]$ is the A-group Bayes predictor.

- Implication: small prediction error and group sufficiency gap can be achieved together
  - Not quite the case of DP/EO [Liu et al., 2019]
- Underlying assumption: "$f_{A=a}^{Bayes}$ are similar $\forall a \in \mathcal{A}$"

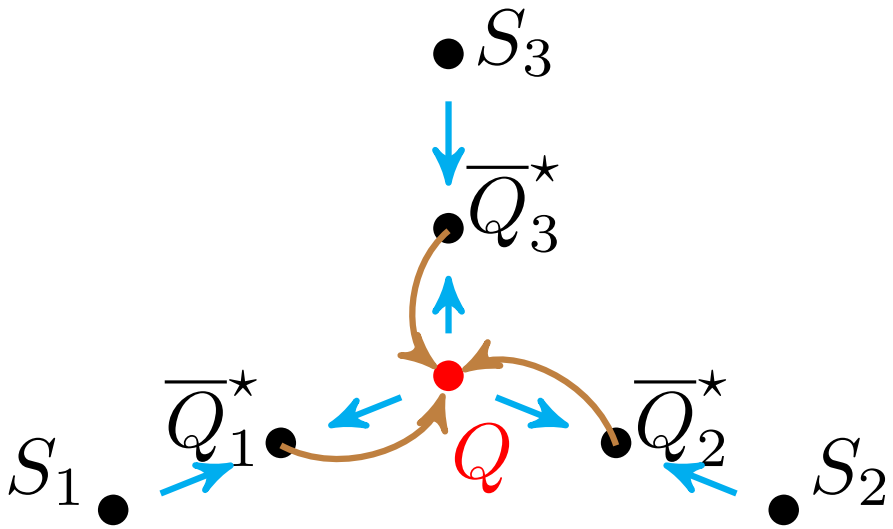
Table of Contents

Illustration of the proposed algorithm



- Subgroups S_a $(a \in \mathcal{A} = \{1, 2, 3\})$
- Learned predictive-distribution $\tilde{f} \sim Q$
- Subgroup-specific predictive-distribution $\overline{Q}_a^\star$ $(a \in \mathcal{A})$
- Lower-level optimization (cyan), Upper-level optimization (brown)
  - Alternating update

Lower-level optimization

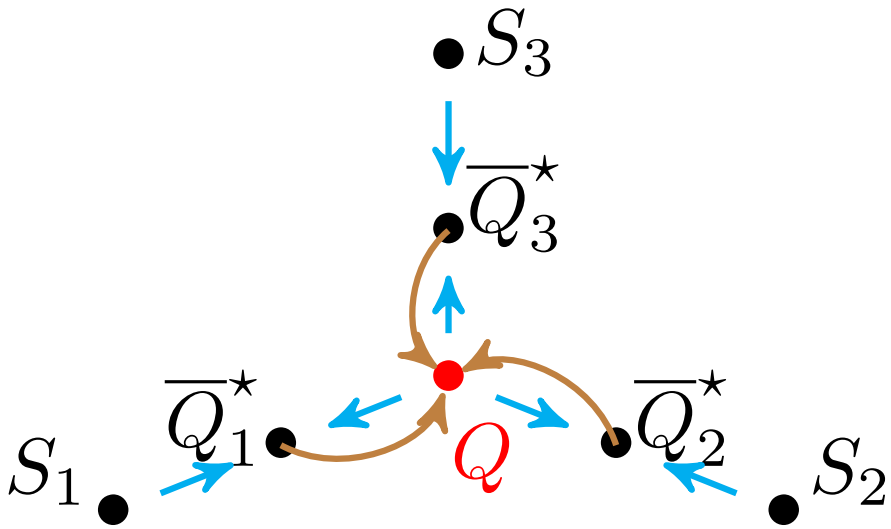


- Learn $\overline{Q}_a^\star$ from S_a's and a **fixed** Q (as a "fair and informative prior"):

$$\overline{Q}_a^\star = \underset{Q_a \in \mathcal{Q}}{\arg\min} \left\{ \mathbb{E}_{\tilde{f}_a \sim Q_a} \hat{\mathcal{L}}_a^{\mathsf{BCE}}(\tilde{f}_a) + \lambda \, \mathrm{KL}(Q_a \| Q) \right\}, \forall a \in \mathcal{A}$$

- Aims to minimize the upper bound of group-wise generalization error [Shui et al., 2022, Theorem 5.1]

Upper-level optimization



- Update Q from **fixed** $\overline{Q}_a^\star$'s:

$$\min_{Q \in \mathcal{Q}} \frac{1}{|\mathcal{A}|} \sum_{a \in \mathcal{A}} \mathrm{KL}\left(\overline{Q}_a^\star \middle\| Q\right)$$

- Aims to control the upper bound of group sufficiency gap [Shui et al., 2022, Corollary 5.1]

Table of Contents

Upper-level: Upper bound of group sufficiency gap

Shui et al. [2022, Corollary 5.1]

The group sufficiency gap $\mathbf{Suf}_f$ in randomized algorithm w.r.t. learned predictive distribution Q is upper bounded by:

$$\mathbf{Suf}_f \leq \frac{2\sqrt{2}}{|\mathcal{A}|}\left[\sum_{a\in\mathcal{A}} \underbrace{\sqrt{KL(Q_a^\star \| Q)}}_{\text{Optimization error}} + \underbrace{\sqrt{KL(Q_a^\star \| \mathcal{D}(Y|X,A=a))}}_{\text{Approximation error (irreducible)}}\right]$$

where

- $Q_a^\star = \arg\min_{Q_a\in\mathcal{Q}} \mathbb{E}_{\tilde{f}_a\sim Q_a}\left[\mathcal{L}_a^{\mathsf{BCE}}(\tilde{f}_a)\right]$
- $\mathcal{Q}$: family of feasible distributions
- $\mathcal{D}(Y|X,A=a)$: conditional distribution of Y given $(X,A=a)$ (Recall: its expectation is exactly the $f_{A=a}^{Bayes}(X)$)

Lower-level: Upper bound of generalization error

Shui et al. [2022, Theorem 5.1]

Suppose:

- Datasets $\{S_a\}_{a\in\mathcal{A}}$ with $S_a = \{(x_i^a, y_i^a)\}_{i=1}^m \overset{\text{i.i.d.}}{\sim} \mathcal{D}(x, y|A=a)$;
- The BCE loss is upper bounded by L (on $\mathcal{Q}$);
- $Q_a \in \mathcal{Q}$ is any learned distribution from dataset S_a; arbitrary $Q \in \mathcal{Q}$.

Then with high probability $\geq 1-\delta$ with $\forall\delta \in (0,1)$, we have:

$$\frac{1}{|\mathcal{A}|}\sum_{a\in\mathcal{A}} \underbrace{\mathbb{E}_{\tilde{f}_a\sim Q_a}\left[\mathcal{L}_a^{\text{BCE}}(\tilde{f}_a)\right]}_{\text{Generalization error } (A=a)}$$

$$\leq \frac{1}{|\mathcal{A}|}\sum_{a\in\mathcal{A}} \underbrace{\mathbb{E}_{\tilde{f}_a\sim Q_a}\left[\hat{\mathcal{L}}_a^{\text{BCE}}(\tilde{f}_a)\right]}_{\text{Empirical risk } (A=a)} + \frac{L}{\sqrt{|\mathcal{A}|m}}\sum_{a\in\mathcal{A}} \underbrace{\sqrt{KL(Q_a\|Q)}}_{\Rightarrow \text{ Regularizer}} + L\underbrace{\sqrt{\frac{\log(1/\delta)}{|\mathcal{A}|m}}}_{\to 0 \text{ as } |\mathcal{A}|\to\infty}$$

Table of Contents

FAMS: Fair and Informative Learning for Multiple Subgroups[3]

Algorithm 1 Fair and Informative Learning for Multiple Subgroups (FAMS)

Input: Parameters w.r.t. distribution Q:$(\boldsymbol{\theta}, \boldsymbol{\sigma}^2)$, datasets $\{S_a\}$, $a \in \mathcal{A}$.
for Sampling a subset of $\{S_a\}$, where $a \in \mathcal{A}' \subseteq \mathcal{A}$ **do**
    ### Solving the lower-level ###
    Fix Q, optimizing the loss w.r.t. $Q_a = \mathcal{N}(\boldsymbol{\theta}_a, \boldsymbol{\sigma}_a^2)$ through SGD for each $a \in \mathcal{A}'$

$$\mathbb{E}_{\tilde{f}_{\mathbf{w}_a} \sim Q_a} \hat{\mathcal{L}}_a^{\mathrm{BCE}}(\tilde{f}_{\mathbf{w}_a}) + \lambda \mathrm{KL}(Q_a \| Q)$$

    Obtaining the solution $\overline{Q}_a^\star$, $a \in \mathcal{A}'$.
    ### Solving the upper-level ###
    Fix $\overline{Q}_a^\star$ with $a \in \mathcal{A}'$, optimizing the loss w.r.t. Q through SGD: $\frac{1}{|\mathcal{A}'|} \sum_a \mathrm{KL}(\overline{Q}_a^\star \| Q)$
    Obtaining updated parameter $(\boldsymbol{\theta}, \boldsymbol{\sigma}^2)$ in Q
end for
Return: Parameter of distribution Q: $(\boldsymbol{\theta}, \boldsymbol{\sigma}^2)$

- Choose Q and Q_a from the family of isotropic Gaussian distributions
  - Computationally efficient: Closed-form & differentiable KL divergences:
  - Sample $\boldsymbol{w} \sim Q = \mathcal{N}(\boldsymbol{\theta}, \boldsymbol{\sigma}^2)$ and $\boldsymbol{w}_a \sim Q_a = \mathcal{N}(\boldsymbol{\theta}_a, \boldsymbol{\sigma}_a^2)$ for $\tilde{f}_{\boldsymbol{w}}$ & $\tilde{f}_{\boldsymbol{w}_a}$
  - Thus, the algorithm learns $(\boldsymbol{\theta}, \boldsymbol{\sigma}^2)$ and $(\boldsymbol{\theta}_a, \boldsymbol{\sigma}_a^2)$ $(\forall a \in \mathcal{A})$

[3] Perhaps the authors intended to call the algorithm 'FILMS'? I *guess* the old name FAMS stands for something like 'Fair and Accurate learning for Multiple Subgroups'

FAMS: Fair and Informative Learning for Multiple Subgroups

Algorithm 1 Fair and Informative Learning for Multiple Subgroups (FAMS)

Input: Parameters w.r.t. distribution Q:$(\boldsymbol{\theta}, \boldsymbol{\sigma}^2)$, datasets $\{S_a\}$, $a \in \mathcal{A}$.
for Sampling a subset of $\{S_a\}$, where $a \in \mathcal{A}' \subseteq \mathcal{A}$ **do**
    ### Solving the lower-level ###
    Fix Q, optimizing the loss w.r.t. $Q_a = \mathcal{N}(\boldsymbol{\theta}_a, \boldsymbol{\sigma}_a^2)$ through SGD for each $a \in \mathcal{A}'$

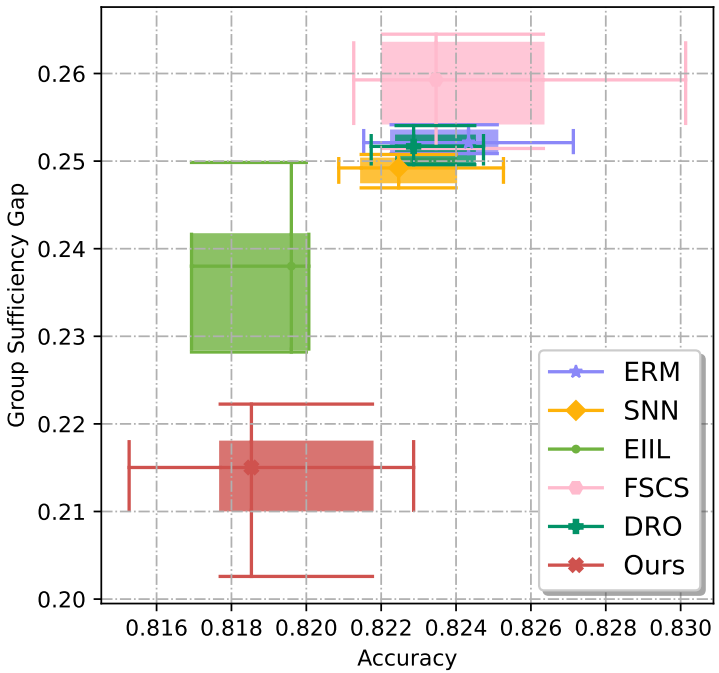
$$\mathbb{E}_{\tilde{f}_{\mathbf{w}_a} \sim Q_a} \hat{\mathcal{L}}_a^{\text{BCE}}(\tilde{f}_{\mathbf{w}_a}) + \lambda \text{KL}(Q_a \| Q)$$

    Obtaining the solution $\overline{Q}_a^\star$, $a \in \mathcal{A}'$.
    ### Solving the upper-level ###
    Fix $\overline{Q}_a^\star$ with $a \in \mathcal{A}'$, optimizing the loss w.r.t. Q through SGD: $\frac{1}{|\mathcal{A}'|} \sum_a \text{KL}(\overline{Q}_a^\star \| Q)$
    Obtaining updated parameter $(\boldsymbol{\theta}, \boldsymbol{\sigma}^2)$ in Q
end for
Return: Parameter of distribution Q: $(\boldsymbol{\theta}, \boldsymbol{\sigma}^2)$

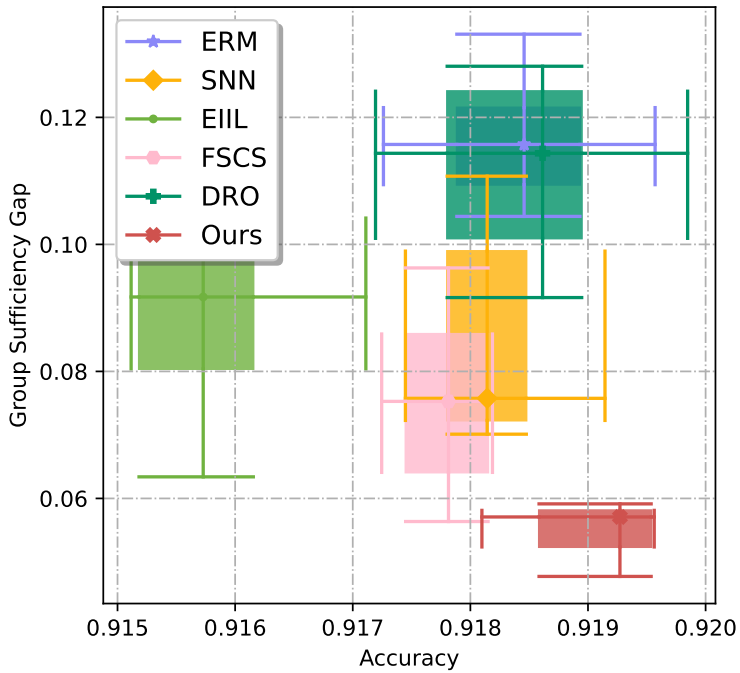
- Alternating update rule btw lower-/upper-level optimization
  - *e.g.*, run lower-loop for 10-15 iter $\Rightarrow$ upper-loop for 1-5 iter
  - If there are too many subgroups, randomly sample $\mathcal{A}' \subsetneq \mathcal{A}$ per epoch
- Inference: $f(x) \approx \frac{1}{N} \sum_{i=1}^{N} \tilde{f}_{\boldsymbol{w}^{(i)}}(x)$ with $\boldsymbol{w}^{(i)} \overset{\text{i.i.d.}}{\sim} Q$ (Monte-Carlo)

Table of Contents

Accuracy v.s. Group sufficiency gap



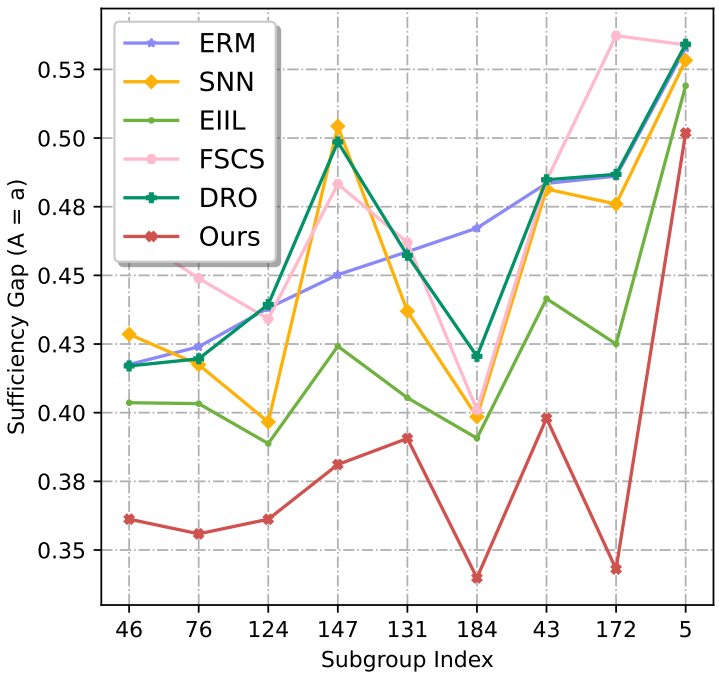
(a) Amazon Reviews dataset
$\mathcal{A}$ = (all individual users)



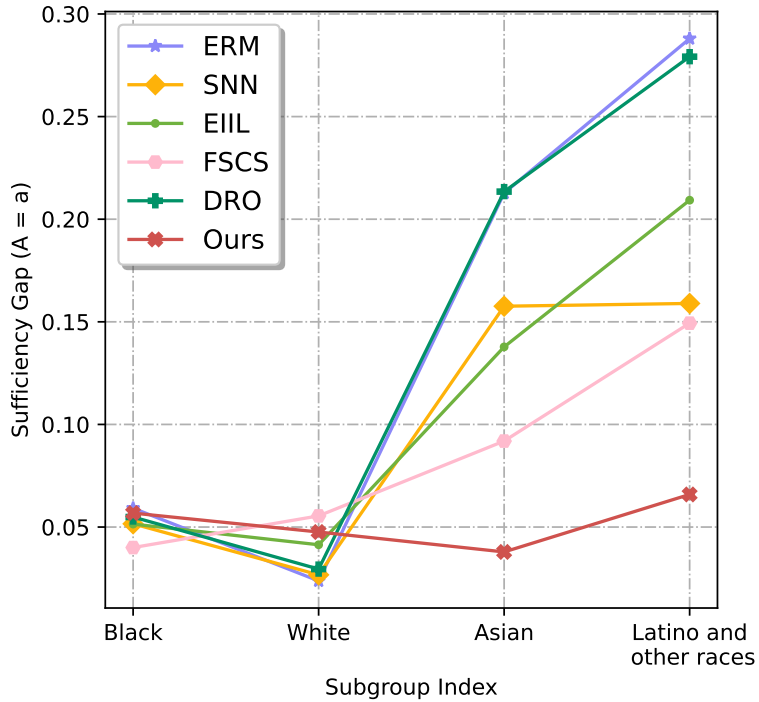
(b) Toxic Comments dataset
$\mathcal{A} = \{\text{black,white,Asian,Latino\&others}\}$

- Experiment details:
  - Shared embedding with DistilBERT [Sanh et al., 2019]
  - $\tilde{f}_{\boldsymbol{w}}$ and $\tilde{f}_{\boldsymbol{w}_a}$ as 4-layer FCNN
  - Baselines: ERM, SNN(stochastic neural net), EIIL, FSCS (both target $\mathbf{Suf}_f$), & DRO (encourages identical losses)

Group sufficiency gap on each subgroup



(c) Amazon Reviews dataset
$\mathcal{A}$ = (all individual users)



(d) Toxic Comments dataset
$\mathcal{A} = \{\text{black,white,Asian,Latino\&others}\}$

- Same experiment details
- Group sufficiency gap *for subgroup* a: $\mathbb{E}_X[|\mathbb{E}[Y|f(X)] - \mathbb{E}[Y|f(X), A = a]|]$
  - Compare with $\mathbf{Suf}_f := \mathbb{E}_{A,X}[|\mathbb{E}[Y|f(X)] - \mathbb{E}[Y|f(X), A]|]$
- Summary: "Lower group sufficiency gap, comparable accuracy."

Table of Contents

Discussion points

- Takeaway: a provable bi-level optimization framework for...
  - mitigating the group sufficiency bias, and
  - preserving the utility of data.
- Similar and useful idea appears in Probabilistic/Bayesian MAML [Finn et al., 2018, Yoon et al., 2018, Chen and Chen, 2022], implicit MAML [Rajeswaran et al., 2019], `Ditto`(fair & robust FL) [Li et al., 2021], and more.
- Why $\overline{Q}_a^\star$'s are regarded as constant in upper-level optimization?
  - In bilevel optimization literature, the lower-level solution is not regarded as just a constant. For instance,

$$\min_{x,y^*} f(x,y^*) \quad \text{s.t.} \quad y^* \in \arg\min_y g(x,y).$$

  - Likewise, $\overline{Q}_a^\star$ depends on Q.
  - However, it seems difficult to deal with $KL(Q_a^\star(Q)\|Q)$ as a differentiable function of Q (maybe or not)
  - Maybe the reason of absence of "convergence" analysis?

Discussion points

- Isn't it memory-inefficient?
  - All parameters of Q and Q_a's ($a \in \mathcal{A}$) should be stored in memory appropriately: $\dim(w) \times (|\mathcal{A}| + 1)$ parameters.
  - How about learning a policy function $\pi_\phi : \mathcal{A} \to \mathcal{W}$ so that $w_a := \pi_\phi(a)$ is the parameter for Q_a? ($\mathcal{W}$: parameter space)
    - ⋆ Regarding $\mathcal{A}$ as an action space in RL.
    - ⋆ Applying similar bilevel framework, we could learn the models w.r.t. ϕ (for π_ϕ) and θ (for Q)
    - ⋆ Might be useful when the sensitive attribute is ordinal or continuous (*e.g.*, age or height)
- Can a fair binary classification be done by only learning appropriate threshold $\tau \in [0, 1)$?
  - For a given predictor $f : \mathcal{X} \to [0, 1]$, return 0 if $f(X) < \tau_a$; 1 if $f(X) \geq \tau_a$. (Something like 'personalization' in federated learning)
  - Or, similar idea could be implemented with some form of nonlinear mapping from $[0, 1)$ to itself.

References I

W. Alghamdi, H. Hsu, H. Jeong, H. Wang, P. W. Michalak, S. Asoodeh, and F. Calmon. Beyond adult and compas: Fair multi-class prediction via information projection. In *Advances in Neural Information Processing Systems*, 2022.

S. Barocas, M. Hardt, and A. Narayanan. *Fairness and Machine Learning: Limitations and Opportunities*. fairmlbook.org, 2019. `http://www.fairmlbook.org`.

L. Chen and T. Chen. Is bayesian model-agnostic meta learning better than model-agnostic meta learning, provably? In *International Conference on Artificial Intelligence and Statistics*, pages 1733–1774. PMLR, 2022.

A. Chouldechova. Fair prediction with disparate impact: A study of bias in recidivism prediction instruments. *Big Data*, 5(2): 153–163, 2017. doi: 10.1089/big.2016.0047. URL `https://doi.org/10.1089/big.2016.0047`. PMID: 28632438.

S. Dutta, D. Wei, H. Yueksel, P.-Y. Chen, S. Liu, and K. Varshney. Is there a trade-off between fairness and accuracy? a perspective using mismatched hypothesis testing. In *International Conference on Machine Learning*, pages 2803–2813. PMLR, 2020.

C. Finn, K. Xu, and S. Levine. Probabilistic model-agnostic meta-learning. *Advances in neural information processing systems*, 31, 2018.

J. Kang, T. Xie, X. Wu, R. Maciejewski, and H. Tong. Multifair: Multi-group fairness in machine learning. *arXiv preprint arXiv:2105.11069*, 2021.

T. Li, S. Hu, A. Beirami, and V. Smith. Ditto: Fair and robust federated learning through personalization. In *International Conference on Machine Learning*, pages 6357–6368. PMLR, 2021.

L. T. Liu, M. Simchowitz, and M. Hardt. The implicit fairness criterion of unconstrained learning. In *International Conference on Machine Learning*, pages 4051–4060. PMLR, 2019.

Z. Obermeyer, B. Powers, C. Vogeli, and S. Mullainathan. Dissecting racial bias in an algorithm used to manage the health of populations. *Science*, 366(6464):447–453, 2019.

G. Pleiss, M. Raghavan, F. Wu, J. Kleinberg, and K. Q. Weinberger. On fairness and calibration. In *Advances in neural information processing systems*, volume 30, 2017.

A. Rajeswaran, C. Finn, S. M. Kakade, and S. Levine. Meta-learning with implicit gradients. *Advances in neural information processing systems*, 32, 2019.

V. Sanh, L. Debut, J. Chaumond, and T. Wolf. Distilbert, a distilled version of bert: smaller, faster, cheaper and lighter. *ArXiv*, abs/1910.01108, 2019.

References II

C. Shui, G. Xu, Q. Chen, J. Li, C. X. Ling, T. Arbel, B. Wang, and C. Gagné. On learning fairness and accuracy on multiple subgroups. In *Advances in Neural Information Processing Systems*, 2022.

J. Song, P. Kalluri, A. Grover, S. Zhao, and S. Ermon. Learning controllable fair representations. In *The 22nd International Conference on Artificial Intelligence and Statistics*, pages 2164–2173. PMLR, 2019.

Y. Wang, X. Wang, A. Beutel, F. Prost, J. Chen, and E. H. Chi. Understanding and improving fairness-accuracy trade-offs in multi-task learning. In *Proceedings of the 27th ACM SIGKDD Conference on Knowledge Discovery & Data Mining*, pages 1748–1757, 2021.

F. Yang, M. Cisse, and S. Koyejo. Fairness with overlapping groups; a probabilistic perspective. *Advances in neural information processing systems*, 33:4067–4078, 2020.

J. Yoon, T. Kim, O. Dia, S. Kim, Y. Bengio, and S. Ahn. Bayesian model-agnostic meta-learning. *Advances in neural information processing systems*, 31, 2018.