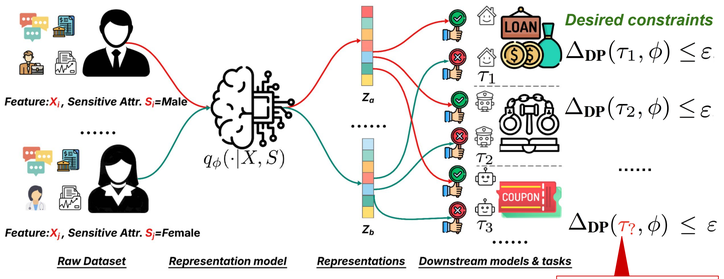
Fair Representation Learning with *Controllable High Confidence Guarantees* via Adversarial Inference (FRG)

Yuhong Luo — Rutgers University
Austin Hoag — Sony AI
Xintong Wang — Rutgers University
Philip S. Thomas — UMass Amherst
Przemyslaw A. Grabowicz — UMass Amherst; University College Dublin

NEURAL INFORMATION PROCESSING SYSTEMS

Fair Representation Learning (*FRL*)



* A metric of evaluating how unfair a downstream model is:

$$\Delta_{\text{DP}}(\tau,\phi) := |\Pr(\hat{Y}=1|S=1) - \Pr(\hat{Y}=1|S=0)|$$

Given a **dataset** D, an **algorithm** a outputs the **parameter** ϕ for a **representation model**. A **representations** Z is used for prediction by an *arbitrary* **downstream model** τ in any downstream task.

A representation model is considered "ε-fair" if the unfairness on *every* downstream task and model is upper-bounded by ε.

Problem Formulation

- As ML algorithms are *probabilistic*, any useful model can hardly guarantee the above with *certainty for unseen datasets*.
- Thus, we aim for **probabilistic guarantees**.

Desired constraints

$$\underbrace{\Delta_{\text{DP}}(\tau_1,\phi) \leq \varepsilon \quad \Delta_{\text{DP}}(\tau_2,\phi) \leq \varepsilon \quad \ldots\ldots \quad \Delta_{\text{DP}}(\tau_?,\phi) \leq \varepsilon}_{\text{With a } \textbf{high probability} \text{ - at least } 1-\delta}$$

for an unknown task

Given $\varepsilon \in [0,1], \delta \in (0,1)$ and a *dataset* D a *representation learning algorithm* a is said to provide a $1-\delta$ confidence ε-fairness guarantee if and only if $\Pr(g_\varepsilon(a(D)) \leq 0) \geq 1-\delta$, where $g_\varepsilon(\phi) := \sup_\tau \Delta_{\text{DP}}(\tau,\phi) - \varepsilon$. ***(Worst-case unfairness among all downstream models to be bounded with a high probability.)***

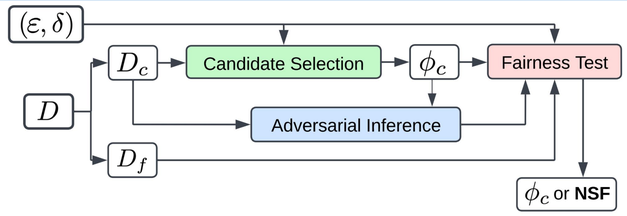
Adversarial Inference

Definition 5.1 (Optimal adversary). Given a representation model q_ϕ, a downstream model τ^*_{adv} is an optimal adversary if and only if $\tau^*_{\text{adv}} \in \arg\max_\tau \Delta_{\text{DP}}(\tau,\phi)$.

Theorem 5.2. *Suppose* $S, \hat{Y} \in \{0,1\}$ *are Bernoulli random variables. We have* $\Delta_{\text{DP}}(\tau,\phi) = \frac{|\text{Cov}(\hat{Y},S)|}{\text{Var}(S)}$.

- The optimal adversary can be *approximated* via **SGD** on $\max_\tau |\text{Cov}(\hat{Y},S)|$

Methodology: *FRG*



* ***NSF*** stands for **N**o **S**olution **F**ound. This can be due to small ε, δ and/or size of D.

1. Split the data into ***disjoint*** sets D_c and D_f.
2. The ***Candidate selection*** uses D_c to *optimize and propose* ϕ_c.
3. Train an ***adversarial model*** to *predict sensitive attributes* using representations of D_c generated by model q_{ϕ_c}.
4. The ***fairness test*** uses predictions of the adversarial model on D_f to evaluate whether $g_\varepsilon(\phi_c) \leq 0$ with *sufficient confidence.*
5. Return ϕ_c if the ***fairness test*** is passed, and ***NSF*** otherwise.

Fairness Test

Theorem 5.3. *Suppose* ***fairness test*** *have access to an* ***optimal adversary*** *then* ***FRG*** *provides a* $1-\delta$ confidence ε-fair guarantee.

Assuming access to an optimal adversary τ^*_{adv}, given a candidate ϕ_c

1. **Construct** a $1-\delta$ ***confidence upper bound*** on $g_\varepsilon(\phi)$ using D_f.
   a. Define $U_\varepsilon(\phi, D_f)$ as the $1-\delta$ confidence upper bound on $g_\varepsilon(\phi)$ i.e., $\Pr\Big(g_\varepsilon(\phi) \leq U_\varepsilon(\phi, D_f)\Big) \geq 1-\delta.$
   b. Construct $U_\varepsilon(\phi, D_f)$ using *statistical tools* and τ^*_{adv} on D_f.
2. If $U_\varepsilon(\phi_c, D_f) \leq 0$, there is at least confidence $1-\delta$ that $g_\varepsilon(\phi_c) \leq 0$. Therefore, ϕ_c passes the test and is returned.

   Otherwise, return ***NSF*** as there is *not sufficient* confidence that $g_\varepsilon(\phi_c) \leq 0$.

Candidate Selection

FRG's design of *candidate selection* searches for candidates that are "*likely*" to pass the *fairness test*, thus, avoiding ***NSF***.

1. **Predict** whether a *candidate solution* will pass the *fairness test* via ***adversarial training*** and constructing a ***confidence upper bound*** with a similar procedure as ***fairness test*** for. ***(To Avoid NSF.)***
2. **Optimize** for a *candidate solution* with a ***constrained*** *VAE-based* objective using the *KKT conditions*. ***(To keep high utility.)***

Experiment

- **3 Datasets**, each with **2-3 downstream tasks** including one *adversarial task* that predicts the sensitive attribute.
- **6 Baselines**. One baseline ***FARE*** provides provides high-confidence fairness guarantees. But in ***FARE*** the unfairness threshold ε is not controllable by the user but is determined by their framework. The ε guaranteed by ***FARE*** is typically *higher than the desired* ε.

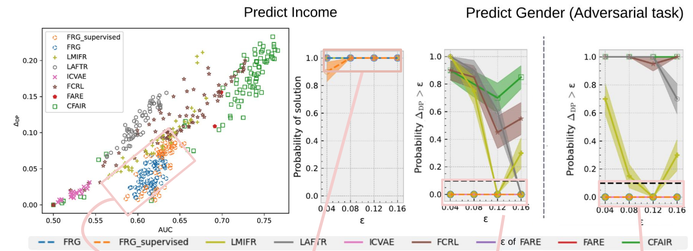


Figure 2: The evaluation on the **Adult** dataset. The target label is *income* and the sensitive attribute is *gender*.

Competitive prediction under similar Δ_{DP}	Likely to return a solution and avoid **NSF**.	Bounded unfairness with a high prob.	Robust to adversarial tasks

1. ***FRG*** can satisfy $\Delta_{\text{DP}}(\tau,\phi) \leq \varepsilon$ (***guaranteed fairness***) for all evaluated downstream models and tasks ***with a high probability.***
2. *FRG* can ***match or outperform baselines*** in terms of ***prediction performance*** (AUC). Baselines either violate the fairness constraints with *non-trivial prob.* or *perform worse* than ***FRG***.
3. *FRG* keeps a ***high probability of returning a solution*** overall.

Additional Results

- The *theoretical* results can be generalized to
  - ***Multi-class*** sensitive attributes.
  - Other *group-fairness* notions including **Equal Opportunity and Equalized Odds**.
- *Empirically*, ***FRG*** works for ***multi-class*** sensitive attributes.
- *Interpretation*:
  - as ***FRG*** does not rely on labels of a downstream task, the fairness guarantees are naturally ***robust*** to certain *distribution shifts*, including ***label shift and concept drift***.

References

Christos Louizos, et al. The variational fair autoencoder. In ICLR, 2016.
Barry Becker and Ronny Kohavi. Adult. UCI Machine Learning Repository, 1996.
Frances Ding, et al. Retiring adult: New datasets for fair machine learning. In NeurIPS, 2021.
Kaggle. Health heritage prize, 2012. URL https://www.kaggle.com/c/hhp.
David Madras, et al. Learning adversarially fair and transferable representations. In ICML, 2018.
Daniel Moyer, et al. Invariant representations without adversarial training. In NeurIPS, 2018.
Jiaming Song, et al. Learning controllable fair representations. International Conference on AISTATS, 2019.
Han Zhao, et al. Conditional learning of fair representations. In ICLR, 2020.
Umang Gupta, et al. Controllable guarantees for fair outcomes via contrastive information estimation. In AAAI, 2021.
Nikola Jovanovic, et al. Fare: Provably fair representation learning with practical certificates. In ICML, 2023.