APPENDIX: PATH-SPECIFIC CAUSAL FAIR PREDICTION VIA AUXILIARY GRAPH STRUCTURE LEARNING

Anonymous authors
Paper under double-blind review

1 ADJACENCY MATRIX AND FAIRNESS MASK CONSTRUCTION

1.1 FAIRNESS MASK CONSTRUCTION

Our proposed method only requires prior knowledge of what attributes are allowed or not allowed to construct the fairness mask M_F. For example, in loan example, the prior knowledge is that Race R is only allowed to affect Y through income Q. Therefore, in the fairness mask, in the i-th column, only the j-th row is set to be 0, and all others in the i-th column are set as 1, if Race is the i-th variable, and income is the j-th variable. it means that except Race $R \to Q$ income, all other paths including Race $R \to Z$ or $R \to Y$ are all unfair paths.

In the non-root case, the M_F is built in a similar way with the additional prior knowledge regarding the latent variables that the latent variable is not allowed to affect Y through sensitive variables.

1.2 ADJACENCY MATRIX CONSTRUCTION

When the data both have categorical attributes (represented by one-hot encoding or embedding vector) and the numerical attribute (represented by a scalar value), we can use the ℓ_2 norm of the corresponding weight vector as the weight in the adjacency matrix. Specifically, take loan example to illustrate the procedure to construct the adjacency matrix. Suppose A is a categorical attribute, and is represented by a three dimension vector $[A_{(1)}, A_{(2)}, A_{(3)}]$. The reconstruction function is $X_2 = w_{A_1}A_1 + w_{A_2}A_2 + w_{A_3}A_3 + w_{4,2}X_4$. The the $w_{1,2}$ in the adjacency matrix is $w_{1,2} = \sqrt{\frac{w_{A_2}^2 + w_{A_3}^2 + w_{4,2}^2}{3}}$.

2 INITIALIZATION AND OPTIMIZATION

Objective Function. When the sensitive attributes are root nodes, the overall loss function is:

$$\begin{aligned}\mathcal{L} = &||Y - \hat{Y}||_2^2 + \beta||D - \tilde{D}||_2^2 + \gamma_1 \left(tr(e^{W \odot W}) - (d_A + d_X + 1)\right)^2 \\ &+\gamma_2||W||_1 + \alpha||W \odot M_F||_1,\end{aligned} \tag{1}$$

where $\hat{Y}$ is defined in Eqn. (6), and $\tilde{D}$ is the reconstruction of D via the cascade data reconstruction presented in Section 4.1.1.

Initialization. As mentioned previously, the cascade data reconstruction requires acyclic graph. To satisfies this, we can initialize the adjacency matrix by the following two ways: (1) adopt the prior knowledge about the basic acyclic graph; (2) pre-train the parameter by the following objective function:

$$||Y - \hat{Y}||_2^2 + \beta||D - \tilde{D}^{'}||_2^2 + \gamma_1 \left(tr(e^{W \odot W}) - (d_A + d_X + 1)\right)^2, \tag{2}$$

where $\tilde{D}^{'}$ is the data reconstructed by the observed data. Each node V_i in $\tilde{D}^{'}$ is calculated as: $\hat{V}_i^{'} = f_i\left(\Pi(V_i)W[i_\pi, i]\right)$, where $\Pi(V_i)$ is the node V_i's observed value, and $W[i_\pi, i]$ is the same as the one in Eqn. (3). Eqn. (2) replaces the cascade data reconstruction $\tilde{D}$ in Eqn. (1) with regular data reconstruction $\tilde{D}^{'}$, which not strictly requires acyclic graph.

Optimization. We adopt the Adam Kingma & Ba (2014) to optimize both Eqn. (2) and Eqn. (1). Besides, at each iteration of optimizing Eqn. (1), the adjacency matrix W is forced to be acyclic.

3 REPRODUCIBILITY

We are applying for the approval of code releasing from our affiliation. The hyper-parameter settings of the experiment are set as follows:

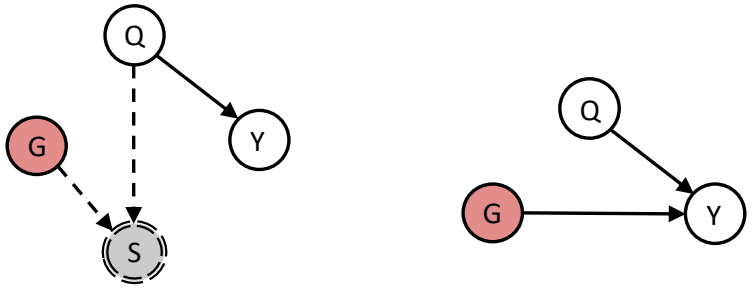
- On the synthetic dataset, α, β are set as 0.5, γ_1 and γ_2 are set as 0.1, and the learning rate is 0.001.
- On adult dataset, the neural network in NN-CFG has two hidden layers and the dimension of each hidden layer is 50. γ_1 is 0.1, γ_2 is 0.01. The hyper-parameter searching range of α and β is $[0, 0.5, 1.0, 1.5, 2.0, 2.5, 3.0, 3.5, 4.0, 4.5, 5.0]$.
- On recommendation dataset, the shared layers have three hidden layers with dimension 32, 16, 8. The dimension of all embeddings is 8. γ_1, γ_2 are set as 0.1. The hyper-parameter searching range of α, β and β_z is $[0, 0.1, 0.5, 1.0, 1.5, 2.0, 2.5, 3.0, 3.5, 4.0, 4.5, 5.0]$.

4 EXPERIMENT ON SYNTHETIC DATASET

We conduct experiments on two synthetic datasets: Unfairness comes from selection bias Bareinboim et al. (2014) and real causality, separately. We experimentally show that when unfairness comes from selection bias, fairness regularization works as the de-bias, and when it comes from real causality, there is a trade-off between utility and fairness.

4.1 DATA GENERATION

Figure 1 shows the causal graph of the above two cases. To better understand these two cases, suppose Figure 1a represents the hiring example, where G is the gender, Q is the candidate's quality, Y is the employment status, and S denotes whether an individual applies for the position. S is affected by gender and quality, which corresponds to the phenomenon that compared to males with the same quality, females prefer to apply to more advanced positions. Therefore, in this dataset, most females are rejected, which causes a spurious correlation between gender and employment status leading to unfairness. In Figure 1b, sensitive attribute G is label node Y's direct cause, which is the source of unfairness. The detailed data generation processes of two cases are in the following.



(a) Case 1: selection bias.

(b) Case 2: real causality.

Figure 1: Causal graphs of synthetic datasets.

Case 1: Unfairness Comes from Selection Bias. We first generate the *underlying set* with 1000 records by: $G \sim N(0, 4)$, $Q \sim N(0, 4)$, $Y \sim N(1.5Q - 1, 1)$. Then we generate the dataset with selection bias, named as *observed set*, by introducing S. For each record $[G_i, Q_i, Y_i]$, if $S_i = 1$, this record is also in the observed set, otherwise not. S_i is sampled from a Bernoulli distribution $Ber(p_i)$, where $p_i = \min(0.99, 7 \times p_N([G_i, Q_i]))$. $p_N([G_i, Q_i])$ is the value of probability density function of the multivariate Gaussian distribution with mean $\mu = [1, 0.5]$, and covariance $\sigma = \begin{bmatrix} 1 & -0.5 \\ -0.5 & 1 \end{bmatrix}$ at point $[G_i, Q_i]$. Totally, there are 439 records in the observed dataset and the covariance between G and Y is -0.434.

Case 2: Unfairness Comes from Real Causality. Based on Figure 1b, the data is generated as: $G \sim N(0, 4)$, $Q \sim N(0, 4)$, $Y \sim N(1.2A + 0.5Q + 2, 1)$. This procedure is repeated 1000 times.

4.2 EXPERIMENT SETTINGS

The objective of this experiment is to reveal the relationship between utility and fairness. We compare our proposed model with the following two base models: (1) Logistic regression using all attributes, denoted as $Y \sim Q, A$; (2) Logistic regression only using attribute Q to predict, denoted as $Y \sim Q$. Among them, $Y \sim Q$ is the ideal fair model in both two cases. Our proposed model is denoted as LR-CGF since the linear logistic regression model is adopted as the causal mechanism function.

Evaluation. On the data of case 1, the observed data is split into training/testing set with a split ratio $0.8/0.2$. All the models are trained on the training set, and evaluated on both the test set split from the observed data and the whole underlying dataset. On the data of case 2, we follow the regular training/testing split with a split ratio $0.8/0.2$. The accuracy is adopted to measure the utility.

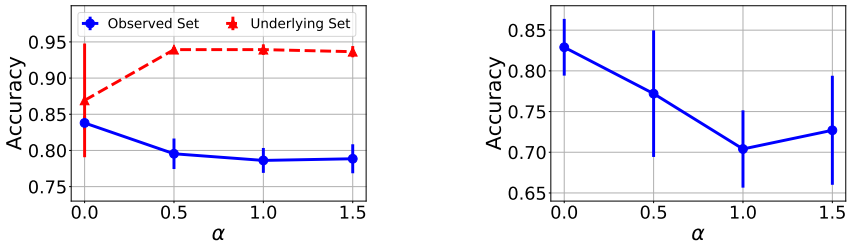
4.3 RESULTS ANALYSIS

Table 1 reports the results with 5-fold cross-validation. From the table, it is observed that in both two cases, the proposed method LR-CGF has a similar performance with the ideal fair model, which confirms the validity of our proposed model. Furthermore, on the underlying set, the fair models $Y \sim Q$ and LR-CGF have superior performance. The reason is that the spurious correlation between G and Y doesn't hold any more on the underlying set and these two fair models successfully remove such spurious correlation. In case 2, the unfairness comes from real causality, thus the fairness and accuracy are actually a trade-off: imposing the fairness constraint would reduce the model utility.

Method	Case 1		Case 2
	Observed Set	Underlying Set	Test Set
$Y \sim Q, G$	$\mathbf{0.841 \pm 0.022}$	0.913 ± 0.004	$\mathbf{0.859 \pm 0.011}$
$Y \sim Q$ (Ideal)	0.800 ± 0.017	0.938 ± 0.002	0.703 ± 0.020
LR-CGF	0.795 ± 0.021	$\mathbf{0.939 \pm 0.007}$	0.727 ± 0.067

Table 1: Results on synthetic dataset.

To further explore the relationship between fairness and utility, we control the strength of fairness regularization by setting different values to hyper-parameter α while others are fixed. The higher the α is, the stronger the fairness regularization strength poses. Figure 2a shows the results on case 1. On the underlying dataset, applying the fairness regularization ($\alpha > 0$) can greatly improve the accuracy compared with no fairness regularization ($\alpha = 0$). Differently, the result shown in Figure 2b clearly indicates the trade-off between fairness and accuracy.



(a) Case 1: selection bias.

(b) Case 2: real causality.

Figure 2: The effect of fairness regularization.

5 ADULT DATASET

5.1 EXPERIMENT SETTINGS

Dataset. The statistics of female and male in the dataset is shown in Table 2. From the table, it is obvious that classic models trained on raw data are highly like to predict the females as low income

compared to males. If banks use those models to predict the loan applicants' income, it is unfair to females. Figure 3 shows the causal graph to the baselines and the arrows marked as red are unfair edges, where G is gender, A is age, M is married, E is higher_edu, O is managerial Occupation, J is gov_jobs, H is high_hours, C is native_country, Y is high_income.

Gender	No. of Low Income	No. of High Income
Female	12909	1660
Male	20797	8503

Table 2: Statistics of Adult Dataset.

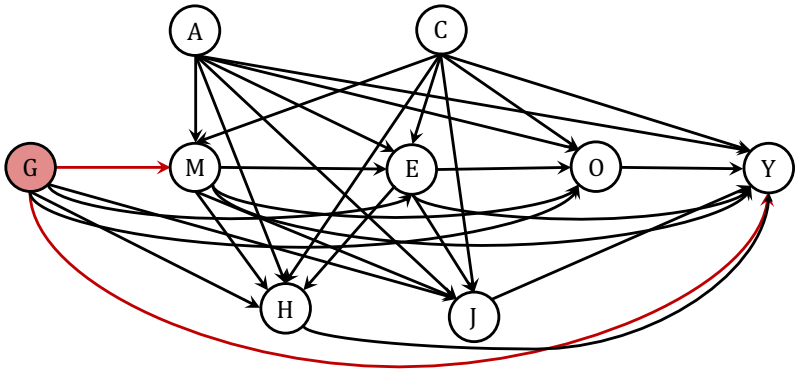


Figure 3: The causal graph of Adult Dataset.

Data Pre-processing for PSE-DR. To run the code of baseline PSE-DR[1] provided by the authors in Zhang et al. (2017), the Adult datasets requires additional stratification step to reduce the number of categories of each variable. The procedure of each variable is: higher_edu: higher_edu: $\lfloor \text{higher\_edu}/10 \rfloor$; high_hours: $\lfloor \text{high\_hours}/20 \rfloor$; managerial_occ: $\lfloor \text{managerial\_occ}/5 \rfloor$; gov_jobs: $\lfloor \text{gov\_jobs}/5 \rfloor$; age: $\lfloor \text{age}/20 \rfloor$; native_country: $\lfloor \text{native\_country}/5 \rfloor$, married: $\lfloor \text{married}/3 \rfloor$, where $\lfloor x \rfloor$ denotes the floor of the scalar x, which is the largest integer i, such that $i <= x$.

Evaluation Metric: Path-specific Effects (PSE). As suggested in Nabi & Shpitser (2018), we formulate the path-specific effects (PSE) as the form of nested counterfactuals Shpitser (2013). Intuitively, the variables along the pathways of interest are set as the value if the treatment variable is set to the treated value. Along other pathways, they are set as if the treatment variable is set to the control value, which turning off the effect passed along those pathways. Under this scheme, the PSE of effect of Gender (denoted as G) along the paths $G \to M \to \cdots \to Y$ and $G \to Y$ in Figure 3 is :

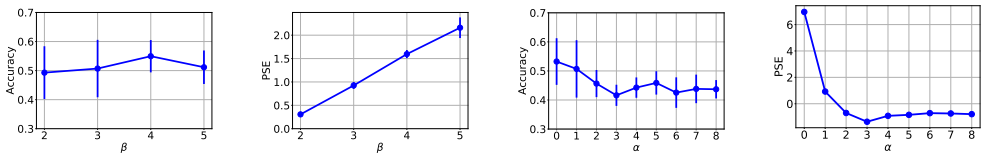
$$\begin{aligned}\mathbb{E}\,[Y\,(G=1, M(G=1), E(G=0, M(G=1)), O(G=0, E(G=0, M(G=1))),\\ H(G=0, M(G=1)), J(M(G=1), E(G=0, M(G=1))))] - \mathbb{E}[Y(G=0)],\end{aligned} \tag{3}$$

where $M(G=1)$ denotes the variable M is its parent G had been set as 1; $E(G=0, M(G=1)$ denotes the variable E if its parent G have been set as 0, and its another parent M have been set as the value if M's parent G had been set as $G=1$. Other nested counterfactuals in Eqn. (3) can be expressed similarly.

5.2 Results of LR-CGF

To explore the relationship between the reconstruction and fairness regularization, we fix one part's hyper-parameter and tunes the other. Figure 4 reports the results of LR-CGF, and similar trends to NN-CGF can be observed.

[1] https://www.yongkaiwu.com/publication/zhang-2017-causal/zhang-2017-causal.zip



(a) AUC ($\Uparrow$) v.s. β. (b) PSE ($\Rightarrow 0$) v.s. β. (c) AUC ($\Uparrow$) v.s. α. (d) PSE ($\Rightarrow 0$) v.s. α.

Figure 4: Effects of reconstruction and fairness regularization.

5.3 Graph Structure Learning Results

Figure 5 shows the adjacency matrix learned by NN-CGF on Adult dataset, which reflects the model graph. The cell located in the i-th row and j-th column denotes the weight associated with $V_i \rightarrow V_j$. If the weight is equal to 0, then the edge $V_i \rightarrow V_j$ does not exist in the causal graph. From the figure, it can be observed that in the model graph, the edge $G \rightarrow M$ and $G \rightarrow Y$ are reduced.

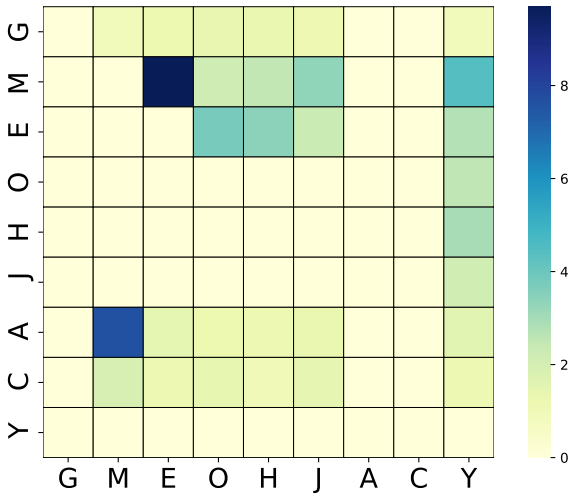


Figure 5: The adjacency matrix of NN-CGF on Adult Dataset.

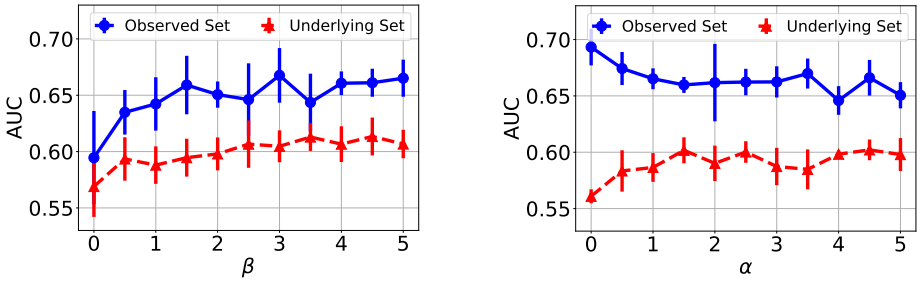
5.4 Results on the Underlying Set.

Similar to the underlying set of the synthetic dataset, we construct an underlying set by randomly select the same number of individuals in four groups shown in Table 2. Finally, in the underlying set, the numbers of low-income female, high-income female, low-income male and high-income male are all 1000. Figure 6 shows the effect of reconstruction part and the fairness on the model utility of NN-CGF. For better comparison, we also plot the results on the test set split from the original observed dataset, marked as blue line. In Figure 6a, the reconstruction part improves the model utility in both two test sets. In Figure 6b, it is worth mentioning that in the underlying set, the fairness regularization no longer sacrifices the model utility, instead, it improves the utility on the underlying dataset. The reason is that the correlation between gender and income doesn't hold anymore in the underlying set, and the fairness regularization successfully corrects the bias due to such correlation.

6 Experiment in Recommendation Dataset

6.1 Experiment Settings

We adopt the Gini Index and Popularity Rate (PR) are also adopted to measure the fairness. Given the item impression list $\mathcal{K} = [k_1, k_2, \cdots, k_{\mathcal{I}}]$, where k_i represents the number of exposures of the



(a) AUC (⇑) v.s. β. (b) AUC (⇑) v.s. α.

Figure 6: Results on underlying set.

i-th item, the Gini Index is defined as: $\text{Gini Index}(\mathcal{K}) = \frac{1}{2|\mathcal{I}|^2\bar{k}} \sum_{i=1}^{\mathcal{I}} \sum_{j=1}^{\mathcal{I}} |k_i - k_j|$, where $\mathcal{I}$ is the number of total items, $\bar{k}$ is the mean of item impression list $\mathcal{K}$. Gini Index measures the statistical dispersion of the item exposure. Popularity rate is the ratio of popular items among the total items recommended to the users, and is defined as: $\text{PR}(\mathcal{K}) = \sum_{i=1}^{\mathcal{I}} P_i k_i / \sum_{i=1}^{\mathcal{I}} k_i$, where P_i is i-th item's value of item popularity, which is binary. For HR and NDCG, the higher the value is, the better the performance is. For Gini Index and PR, the lower the value is, the fairer the model is.

6.2 Neural Network Structure of CGF in Recommendation Dataset

Figure 7 shows the neural network structure of MLP-CGF, where Z is the concatenation of user embedding Z_u, rating related item embedding Z_l, popularity related item embedding Z_s and popularity P. The weight W_l and W_s in the first layer control the flow of information into rating prediction and popularity prediction, separately. After the first layer, the ratings and popularity tasks shared several common layers, followed by their specific prediction layers.

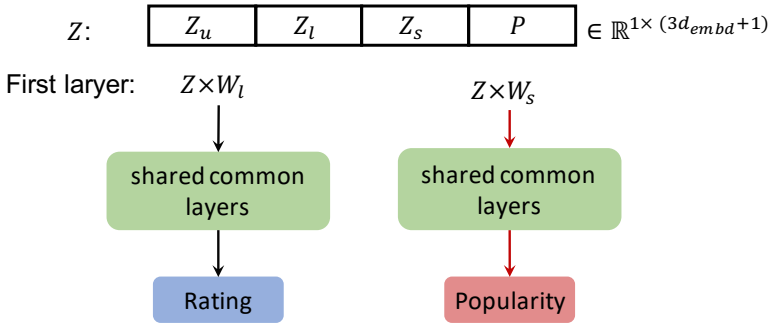


Figure 7: Neural network structure of MLP-CGF.

The weighting matrices W_l and W_s control the information flow from Z to rating and popularity prediction, respectively. The details of W_l and W_s are shown in Figure 8. The dimension of W_l and W_s are both $(3d_{embd} + 1) \times d_{share}$, where d_{embd} is the embedding size, d_{share} is the dimension of the first layer in shared common layers. Each of the weights contains four parts that are Z_u related weights, Z_l related weights, Z_s related weights and P related weights. the P related weights W_s^l in W_s is zero matrix because in popularity prediction, ground-truth popularity value should be the input. Notice that Z_l and X_u should not affect item popularity, we also minimize the norm of W_s^u and W_s^l. Since the paths $Z_s \to Y$ and $P \to Y$ are unfair as shown in Figure 9, the fairness regularization is: $L_F = \alpha \left(||W_l^s||_1 + ||W_l^P||_1 \right)$.

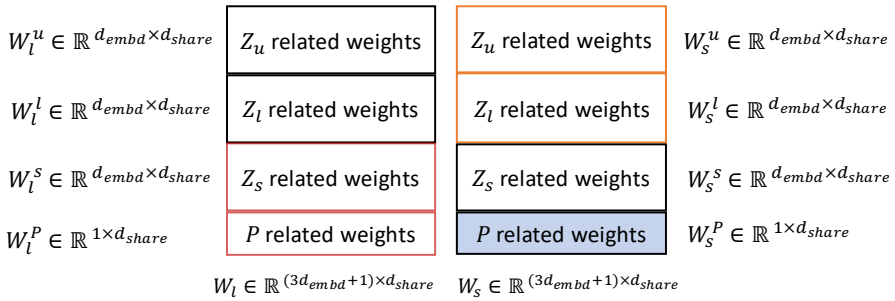


Figure 8: Details of W_l and W_s in the first layer.

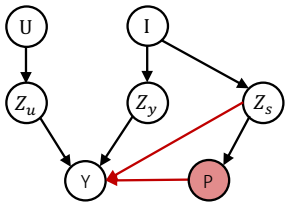


Figure 9: Recommendation: Causal Graph with Effect Diversion.

7 Proof of Theorem 4.1

The theorem shows that the generalization error relates to the reconstruction error (the first term) and fairness regularization ($\mathcal{R}_F$). It explains the benefits of minimizing the reconstruction error and the fairness regularization jointly, which provides the theoretical support for our proposed framework. We also notice that these two terms cannot achieve the minimal value simultaneously, except that the observed graph of the original dataset is exactly the same as the fair graph. When the reconstruction error is small, the model graph would be more likely close to the observed graph, which leads to a high value of the fairness regularization term. Similarly, when the value of the fairness regularization term is small, the model graph would be close to the fair graph, which results in a high value of the reconstruction error. Our proposed method minimizes these two terms targeting a better trade-off between fairness and prediction accuracy. In the following, we give the proof of theorem.

The generalization error of fair classifier h_F on the distribution of observed data $\mathcal{D}^{ob}$ can be written as:

$$
\begin{aligned}
&\epsilon_{h_F}^{\mathcal{D}^{ob}} \\
=&(\epsilon_{h_F}^{\mathcal{D}^{ob}} - \epsilon_{h_F}^{\mathcal{D}^F}) + \epsilon_{h_F}^{\mathcal{D}^F} \\
\leq& \left| \int_{\mathcal{A}\times\mathcal{X}} \ell_{h_F}(a,x) \left[p^{\mathcal{D}^F}(a,x) - p^{\mathcal{D}^{ob}}(a,x) \right] dadx \right| + \epsilon_{h_F}^{\mathcal{D}^F} \\
\leq& \sup_{g\in\mathcal{H}} \left| \int_{\mathcal{A}\times\mathcal{X}} g(a,x) \left[p^{\mathcal{D}^F}(a,x) - p^{\mathcal{D}^{ob}}(a,x) \right] dadx \right| + \epsilon_{h_F}^{\mathcal{D}^F} \\
=& \text{wass}_1 \left(p^{\mathcal{D}^F}, p^{\mathcal{D}^{ob}} \right) + \epsilon_{h_F}^{\mathcal{D}^F},
\end{aligned}
\tag{4}
$$

where $\mathcal{D}^{ob}$ is the distribution of observed data, and $\mathcal{D}^F$ is the distribution of data reconstructed from fair graph.

7.1 THE BOUND OF $\text{WASS}_1\left(p^{\mathcal{D}^F}, p^{\mathcal{D}^{ob}}\right)$

When $p^{\mathcal{D}^{ob}}$ is normal distribution, i.e., $x_i|\text{pa}_{\mathcal{G}}(x_i) \sim \mathcal{N}(f_{\text{pa}_{\mathcal{G}}}(x_i), 1)$, where $\text{pa}_{\mathcal{G}}(x_i)$ is x_i's parent nodes in graph $\mathcal{G}$. wass $\left(p^{\mathcal{D}^F}, p^{\mathcal{D}^{ob}}\right)$ has the following close form Chafai & Malrieu (2010):

$$\begin{aligned}
&\text{wass}_1\left(p^{\mathcal{D}^F}, p^{\mathcal{D}^{ob}}\right) \\
=& |\mu^{ob} - \mu^F| + \sqrt{\textstyle\sum_{i=1}^{d_A+d_X} \left|\sqrt{\lambda_i^{ob}}\upsilon_i^{ob} - \sqrt{\lambda_i^F}\upsilon_i^F\right|^2} \\
=& |\mu^{ob} - \mu^F| \\
&+\sqrt{\textstyle\sum_{i=1}^{d_A+d_X} \left\{\left(\sqrt{\lambda_i^{ob}} - \sqrt{\lambda_i^F}\right)^2 + 2\sqrt{\lambda_i^{ob}\lambda_i^F}(1 - \upsilon_i^{ob}\cdot\upsilon_i^F)\right\}},
\end{aligned} \tag{5}$$

where $\left\{\lambda_i^{ob}\right\}_{i=1}^{d_A+d_X}$, $\left\{\lambda_i^F\right\}_{i=1}^{d_A+d_X}$ are ordered spectrum of Σ^{ob} and Σ^F, respectively. μ^{ob} and μ^F are the mean vectors of distribution $\mathcal{D}^{ob}$ and $\mathcal{D}^F$. Σ^{ob} and Σ^F are the covariance matrices of distribution $\mathcal{D}^{ob}$ and $\mathcal{D}^F$. $\left\{\upsilon_i^{ob}\right\}_{i=1}^{d_A+d_X}$, $\left\{\upsilon_i^F\right\}_{i=1}^{d_A+d_X}$ are the associated orthonormal basis of eigenvectors. We assume $\upsilon_i^{ob}\cdot\upsilon_i^F > 0$.

First term: If we assume: $\tau = ||\mathcal{L}_{\mathcal{W}_{\mathcal{D}}}(\cdot,\cdot)||^2 < \infty$, where $\mathcal{L}_{\mathcal{W}_{\mathcal{D}}}(\cdot,\cdot)$ is the Euclidean distance function, according to the Hoeffding's Inequality, the following inequality regarding the first term in Eqn (5) holds with probability $1-\delta$,

$$\begin{aligned}
&||\mu^{ob} - \mu^F||^2 \\
=& \textstyle\int_{\mathcal{A}\times\mathcal{X}} \mathcal{L}_{W_D}([a^{ob}, x^{ob}], [a^F, x^F])dadx \\
\leq& ||D - \tilde{D}||_{fro}^2 + \tau\sqrt{\frac{2\log\frac{2}{\delta}}{m}},
\end{aligned} \tag{6}$$

where D is the observed samples, and $\tilde{D}$ is the reconstructed data.

Second term: According to corollary 4.2 and theorem 4.1 in Loukas (2017), with probability $1-\delta$, we have the following:

$$\begin{aligned}
&\textstyle\sum_{i=1}^{d_A+d_X}\left(\sqrt{\lambda_i^{ob}} - \sqrt{\lambda_i^F}\right)^2 \\
=& \textstyle\sum_{i=1}^{d_A+d_X}\left(\sqrt{\lambda_i^{ob}} - \sqrt{\hat{\lambda}_i^{ob}}\right)^2 + \left(\sqrt{\hat{\lambda}_i^{ob}} - \sqrt{\hat{\lambda}_i^F}\right)^2 + \left(\sqrt{\hat{\lambda}_i^F} - \sqrt{\hat{\lambda}_i^F}\right)^2 \\
\leq& \textstyle\sum_{i=1}^{d_A+d_X}\frac{\kappa_i^{ob}+\kappa_i^F}{\sqrt{m\delta}} + \left(\sqrt{\hat{\lambda}_i^{ob}} - \sqrt{\hat{\lambda}_i^F}\right)^2,
\end{aligned} \tag{7}$$

where $(\kappa_i^{ob})^2 = \lambda_i^{ob}(\lambda_i^{ob} + tr(\Sigma^{ob}))$, $(\kappa_i^F)^2 = \lambda_i^F(\lambda_i^F + tr(\Sigma^F))$, $\hat{\lambda}_i^F$ and $\hat{\lambda}_i^{ob}$ are the empirical estimation of λ_i^F and λ_i^{ob}, separately.

$$\begin{aligned}
&2\sqrt{\lambda_i^{ob}\lambda_i^F}(1 - \upsilon_i^{ob}\cdot\upsilon_i^F) \\
\leq& 2\sqrt{(\hat{\lambda}_i^{ob} - \lambda_i^{ob})(\hat{\lambda}_i^F - \lambda_i^F) + \hat{\lambda}_i^F(\hat{\lambda}_i^{ob} - \lambda_i^{ob}) + \hat{\lambda}_i^{ob}(\hat{\lambda}_i^F - \lambda_i^F) + \hat{\lambda}_i^{ob}\hat{\lambda}_i^F} \\
\leq& 2\sqrt{\frac{\kappa_i^{ob}+\kappa_i^F}{m\delta} + \frac{1}{\sqrt{m\delta}}(\kappa_i^{ob}\hat{\lambda}_i^F + \kappa_i^F\hat{\lambda}_i^{ob}) + \hat{\lambda}_i^{ob}\hat{\lambda}_i^F} \\
\leq& 2\left(\frac{\sqrt{\kappa_i^{ob}+\kappa_i^F}}{\sqrt{m\delta}} + \frac{\sqrt{\kappa_i^{ob}\hat{\lambda}_i^F + \kappa_i^F\hat{\lambda}_i^{ob}}}{\sqrt[4]{m\delta}} + \sqrt{\hat{\lambda}_i^{ob}\hat{\lambda}_i^F}\right).
\end{aligned} \tag{8}$$

Combining Eqn. (5), Eqn. (6), Eqn. (7) and Eqn. (8), the following equality holds with probability $1-\delta$, for $\forall\delta > 0$,

$$\begin{aligned}
&\text{wass}_1\left(p^{\mathcal{D}^F}, p^{\mathcal{D}^{ob}}\right) \\
\leq& \left(||D - \tilde{D}||_{fro}^2 + \tau\sqrt{\frac{2\log\frac{2}{\delta}}{m}} + \frac{C_1}{\sqrt{m\delta}} + \frac{C_2}{\sqrt[4]{m\delta}} + C_3\right)^{\frac{1}{2}},
\end{aligned} \tag{9}$$

where $C_1 = \sum_{i=1}^{d_A+d_X}(\kappa_i^{ob} + \kappa_i^F + 2\sqrt{\kappa_i^{ob} + \kappa_i^F})$, $C_2 = 2\sum_{i=1}^{d_A+d_X}\sqrt{\kappa_i^{ob}\hat{\lambda}_i^F + \kappa_i^F\hat{\lambda}_i^{ob}}$, $C_3 = \sum_{i=1}^{d_A+d_X}\left(\hat{\lambda}_i^{ob} + \hat{\lambda}_i^F\right)$.

7.2 THE BOUND OF $\epsilon_{h_F}^{\mathcal{D}^F}$.

According to the Theorem 1 in Kyono et al. (2020), the expected error of h_F on the modified dataset, with probability $1-\delta, \forall\delta, \gamma > 0$ satisfies the following:

$$\epsilon_{h_F}^{\mathcal{D}^F} \leq 4\hat{\epsilon}_{h_F}^{\mathcal{D}^F} + \frac{1}{m}\left[\mathcal{R}_{dag} + C_4(\mathcal{R}_{l_1} + \mathcal{R}_F) + \log(\frac{8}{\delta})\right] + C_5, \tag{10}$$

where C_4 and C_5 are the same as the C_1 and C_2 defined in Kyono et al. (2020) separately.

7.3 SUMMARY

The generalization error of h_F on the observed dataset satisfies the following inequality with probability $1-\delta, \forall\delta > 0$:

$$\begin{aligned}\epsilon_{h_F}^{\mathcal{D}^{ob}} \leq &\left(||D-\tilde{D}||_{fro}^2 + \tau\sqrt{\tfrac{2\log\frac{2}{\delta}}{m}} + \tfrac{C_1}{\sqrt{m\delta}} + \tfrac{C_2}{\sqrt[4]{m\delta}} + C_3\right)^{\frac{1}{2}} \\ &+4\hat{\epsilon}_{h_F}^{\mathcal{D}^F} + \tfrac{1}{m}\left[\mathcal{R}_{dag} + C_4(\mathcal{R}_{l_1} + \mathcal{R}_F) + \log(\tfrac{8}{\delta})\right] + C_5.\end{aligned} \tag{11}$$

REFERENCES

Elias Bareinboim, Jin Tian, and Judea Pearl. Recovering from selection bias in causal and statistical inference. In *Proc. of AAAI'14*, volume 28, 2014.

Djalil Chafai and Florent Malrieu. On fine properties of mixtures with respect to concentration of measure and sobolev type inequalities. In *Annales de l'IHP Probabilités et statistiques*, volume 46, pp. 72–96, 2010.

Diederik P Kingma and Jimmy Ba. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*, 2014.

Trent Kyono, Yao Zhang, and Mihaela van der Schaar. CASTLE: regularization via auxiliary causal graph discovery. In *Proc. of NeurIPS'20*, 2020.

Andreas Loukas. How close are the eigenvectors of the sample and actual covariance matrices? In Doina Precup and Yee Whye Teh (eds.), *Proc. of ICML'17*, volume 70, pp. 2228–2237, 2017.

Razieh Nabi and Ilya Shpitser. Fair inference on outcomes. In *Proc. of AAAI'18*, pp. 1931–1940. AAAI Press, 2018.

Ilya Shpitser. Counterfactual graphical models for longitudinal mediation analysis with unobserved confounding. *Cognitive science*, 37(6):1011–1035, 2013.

Lu Zhang, Yongkai Wu, and Xintao Wu. A causal framework for discovering and removing direct and indirect discrimination. In *Proc. of IJCAI'17*, pp. 3929–3935, 2017.