Chasing Fairness Under Distribution Shift: A Model Weight Perturbation Approach

Zhimeng Jiang[1]*, **Xiaotian Han**[1]*, **Hongye Jin**[1], **Guanchu Wang**[2], **Rui Chen**[3], **Na Zou**[1], **Xia Hu**[2]
[1]Texas A&M University, [2]Rice University, [3]Samsung Electronics America

Abstract

Fairness in machine learning has attracted increasing attention in recent years. The fairness methods improving algorithmic fairness for in-distribution data may not perform well under distribution shifts. In this paper, we first theoretically demonstrate the inherent connection between distribution shift, data perturbation, and model weight perturbation. Subsequently, we analyze the sufficient conditions to guarantee fairness (i.e., low demographic parity) for the target dataset, including fairness for the source dataset, and low prediction difference between the source and target datasets for each sensitive attribute group. Motivated by these sufficient conditions, we propose robust fairness regularization (RFR) by considering the worst case within the model weight perturbation ball for each sensitive attribute group. We evaluate the effectiveness of our proposed RFR algorithm on synthetic and real distribution shifts across various datasets. Experimental results demonstrate that RFR achieves better fairness-accuracy trade-off performance compared with several baselines. The source code is available at `https://github.com/zhimengj0326/RFR_NeurIPS23`.

1 Introduction

Previous research [1–4] has shown that a classifier trained on a specific source distribution will perform worse when testing on a different target distribution, due to distribution shift. Recently, many studies have focused on investigating the impact of distribution shift on machine learning models, where fairness performance degradation is even more significant than that of prediction performance [5]. The sensitivity of fairness over distribution shift challenges machine learning models in high stake applications, such as criminal justice [6], healthcare [7], and job marketing [8]. Thus the transferability of the fairness performance under distribution shift is a crucial consideration for real-world applications.

To achieve the fairness of the model (already achieved fairness on the source dataset) on the target dataset, we first reveal that distribution shift is equivalent to model weight perturbation, and then seek to achieve fairness under distribution shift via model weight perturbation. Specifically, *i)* we reveal the inherent connection between distribution shift, data perturbation, and model weight perturbation. We theoretically demonstrate that any distribution shift can be equivalent to data perturbation and model weight perturbation in terms of loss value. In other words, the effect of distribution shift on model training can be attributed to data or model perturbation. *ii)* Given the established connection between the distribution shift and model weight perturbation, we next tackle fairness under the distribution shift problem. We first investigate demographic parity relation between source and target datasets. Achieving fair prediction (e.g., low demographic parity) in the source dataset is insufficient for fair prediction in the target dataset. More importantly, the average prediction difference between source and target datasets with the same sensitive attribute also matters for achieving fairness in target datasets.

*Equal contribution.

37th Conference on Neural Information Processing Systems (NeurIPS 2023).

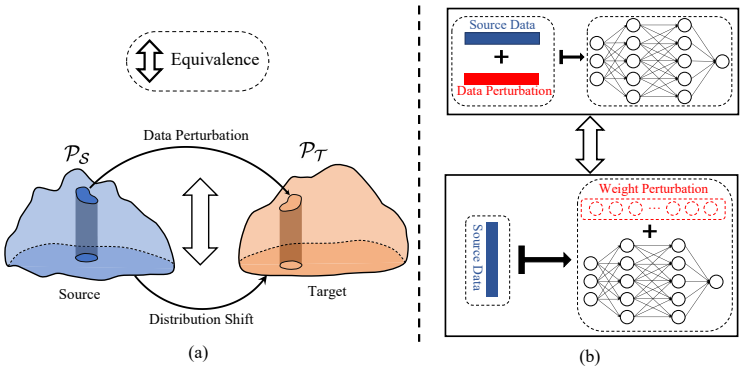

Figure 1: The overview of distribution shift understanding. The left part demonstrates distribution shift can be transformed as data perturbation, while the right part shows that data perturbation and model weight perturbation are equivalent.

Motivated by the established equivalence connection, we propose robust fairness regularization (RFR) to enforce average prediction robustness over model weight. In this way, the well-trained model can tackle the distribution shift problem in terms of demographic parity. Considering the expensive computation for the inner maximization problem in RFR, we accelerate RFR by obtaining the approximate closed-form model weight perturbation using first-order Taylor expansion. In turn, an efficient RFR algorithm, trading the inner maximization problem into two forward and backward propagations for each model update, is proposed to achieve robust fairness under distribution shift. Our contributions are highlighted as follows:

- We theoretically reveal the inherent connection between distribution shift, data perturbation, and model weight perturbation. In other words, distribution shift and model perturbation are equivalent in terms of loss value for any model architecture and loss function.
- We first analyze the sufficient conditions to guarantee fairness transferability under distribution shift. Based on the established connection, we propose RFR explicitly pursuing sufficient conditions for robust fairness by considering the worst case of model perturbation.
- We evaluate the effectiveness of RFR on various real-world datasets with synthetic and real distribution shifts. Experiments results demonstrate that RFR can mostly achieve better fairness-accuracy tradeoff with both synthetic and real distribution shifts.

2 Understanding Distribution Shift

In this section, we first provide the notations used in this paper. And then, we theoretically understand the relations between distribution shift, data perturbation, and model weight perturbation, as shown in Figure 1.

2.1 Notations

We consider source dataset $\mathcal{D}_{\mathcal{S}}$ and target dataset $\mathcal{D}_{\mathcal{T}}$, defined as a probability distribution $\mathcal{P}_{\mathcal{S}}$ and $\mathcal{P}_{\mathcal{T}}$ for samples $S \in \mathcal{S}$ and $T \in \mathcal{T}$, respectively. Each sample defines values for three random variables: features X with arbitrary domain $\mathcal{X}$, binary sensitive attribute $A \in \mathcal{A} = \{0, 1\}$, and label Y arbitrary domain $\mathcal{Y}$, i.e., $\mathcal{S} = \mathcal{T} = \mathcal{X} \times \mathcal{A} \times \mathcal{Y}$. We denote δ as data perturbation on source domain $\mathcal{S}$, where $\delta_X(X)$ and $\delta_Y(Y)$ data perturbation of features and labels. Using $\mathcal{P}(\cdot)$ to denote the space of probability distributions over some domain, we denote the space of distributions over examples as $\mathcal{P}(\mathcal{S})$. We use $||\cdot||_p$ as L_p norm. Let $f_\theta(x)$ be the output of the neural networks parameterized with θ to approximate the true label y. We define $l(f_\theta(x), y)$ as the loss function for neural network training, and the optimal model parameters θ^* training on source data $\mathcal{S}$ is given by $\theta^* = \arg\min_\theta \mathcal{R}_{\mathcal{S}}$, where $\mathcal{R}_{\mathcal{S}} = \mathbb{E}_{(X,Y)\sim\mathcal{P}_{\mathcal{S}}}[l(f_\theta(X), Y)]$. Since the target distribution $\mathcal{P}_{\mathcal{T}}$ maybe be different with source distribution $\mathcal{P}_{\mathcal{S}}$, the well-trained model $f_{\theta^*}(\cdot)$ trained on soure dataset $\mathcal{S}$ typically does not perform well in target dataset $\mathcal{T}$.

2.2 Distribution Shift is Data Perturbation

Deep neural networks are immensely challenged by data quality or dynamic environments [9, 10]. The well-trained deep neural network model may not perform well if existing feature/label noise in training data or distribution shifts among training and target environments. In this subsection, we reveal the inherent equivalence between distribution shift and data perturbation for any neural networks $f_\theta(\cdot)$, source distribution $\mathcal{P}_\mathcal{S}$, and target distribution $\mathcal{P}_\mathcal{T}$ using optimal transport. We first provide the formal definition of optimal transport as follows:

Definition 1 (Optimal Transport [11])**.** *Considering two distributions with probability distribution $P_\mathcal{S}$ and $P_\mathcal{T}$, and cost function moving from s to t as $c(s,t)$, optimal transport between probability distribution $P_\mathcal{S}$ and $P_\mathcal{T}$ is given by*

$$\gamma^*(s,t) = \arg \inf_{\gamma \in \Gamma(P_\mathcal{S}, P_\mathcal{T})} \iint c(s,t)\gamma(s,t)\mathrm{d}s\mathrm{d}t, \tag{1}$$

where the distribution set $\Gamma(P_\mathcal{S}, P_\mathcal{T})$ is the collection of all possible transportation plans and given by $\Gamma(P_\mathcal{S}, P_\mathcal{T}) = \left\{\gamma(s,t) > 0, \int \gamma(s,t)\mathrm{d}t = P_\mathcal{S}(s), \int \gamma(s,t)\mathrm{d}s = P_\mathcal{T}(t)\right\}$. In other words, the distribution set consists of all possible joint distributions with margin distribution $P_\mathcal{S}$ and $P_\mathcal{T}$.

Based on the definition of optimal transport, we demonstrate that distribution shift is equivalent to data perturbation in the following theorem:

Theorem 2.1. *For any two different distributions with probability distribution $P_\mathcal{S}$ and $P_\mathcal{T}$, adding data perturbation δ [2] on source data S can make perturbated source data and target data with the same distribution, where the distribution of data perturbation δ is given by*

$$\mathcal{P}(\delta) = \int_\mathcal{S} \gamma^*(s, s+\delta)\mathrm{d}s. \tag{2}$$

Additionally, for any positive $p > 0$, data perturbation δ with minimal power $\mathbb{E}[||\delta||_p^p]$ is given by Eq.(2) if optimal transport plan $\gamma^*(\cdot,\cdot)$ is calculated based on Eq. (1) with cost function $c(s,t) = ||s-t||_p^p$.

Proof sketch. Given two different distributions, there are many possible perturbations to move one distribution to another. Optimal transport can select the optimal feasible perturbations or transportation plan in terms of specific objectives (e.g., perturbations power). Given the optimal transportation plan, we can derive the distribution of perturbations based on basic probability theory.

Theorem 2.1 demonstrates the equivalence between distribution shift and data perturbation, where data perturbation distribution is dependent on optimal transport between source and target distribution. The intuition is that such distribution shift can be achieved via optimal transportation (i.e., data perturbation). Based on Theorem 2.1, we have the following Corollary 2.2 on the equivalent of neural network behavior for source and target data:

Corollary 2.2. *Given source and target datasets with probability distribution $\mathcal{P}_\mathcal{S}$ and $\mathcal{P}_\mathcal{T}$, there exists data perturbation δ so that the training loss of any neural network $f_\theta(\cdot)$ for target distribution equals that for source distribution with data perturbation δ, i.e.,*

$$\mathbb{E}_{(X,Y)\sim\mathcal{P}_\mathcal{T}}[l(f_\theta(X), Y)] = \mathbb{E}_{\delta_X(X),\delta_Y(Y)}\mathbb{E}_{(X,Y)\sim\mathcal{P}_\mathcal{S}}[l(f_\theta(X+\delta_X(X)), Y+\delta_Y(Y))]. \tag{3}$$

In other words, for the model trained with loss minimization on source data, the deteriorated performance on the target dataset stems from the perturbation of features and labels. The proof sketch is based on Theorem 2.1 since adding a perturbation on the source dataset can be consistent with the target dataset distribution. Therefore, the conclusion can hold for any loss function that is only dependent on the dataset.

2.3 Data Perturbation Equals Model Weight Perturbation

Although we understand that the distribution shift can be attributed to the data perturbation of features and labels in source data, it is still unclear how to tackle the distribution shift issue. A natural solution

[2] Data perturbation δ includes the perturbation for features $\delta_X(X)$ and labels $\delta_Y(Y)$.

is to adopt adversarial training to force the well-trained model to be robust over data perturbation. However, it is complicated to generate data perturbation on features and labels simultaneously, and many well-developed adversarial training methods are mainly designed for adversarial feature perturbation. Fortunately, we show that model weight perturbation is equivalent to data perturbation by Theorem 2.3.

Theorem 2.3. *Considering the source dataset with distribution $\mathcal{P}_{\mathcal{S}}$, suppose the source dataset is perturbed with data perturbation δ, and the neural network is given by $f_\theta(\cdot)$, for general case, there exists model weight perturbation $\Delta\theta$ so that the training loss on perturbed source dataset is the same with that for model weight perturbation $\Delta\theta$ on source distribution:*

$$\mathbb{E}_{\delta_X(X),\delta_Y(Y)}\mathbb{E}_{(X,Y)\sim\mathcal{P}_{\mathcal{S}}}[l(f_\theta(X+\delta_X(X)),Y+\delta_Y(Y))] = \mathbb{E}_{(X,Y)\sim\mathcal{P}_{\mathcal{S}}}[l(f_{\theta+\Delta\theta}(X),Y)]. \quad (4)$$

Proof sketch. The loss under data perturbation and weight perturbation can be analyzed using first-order Tayler expansion. The critical step is to find the condition of the equivalence for the first order term of data perturbation and weight perturbation. Fortunately, the existence of such conditions can be easily proved using linear algebra.

Theorem 2.3 demonstrates that the training loss on perturbed data distribution (but fixed model weight) equals the training loss on perturbed model weight (but original data distribution). In other words, the training loss fluctuation from data perturbation can equal model weight perturbation. Furthermore, we conclude that chasing a robust model over data perturbation can be achieved via model weight perturbation, i.e., finding a "flattened" local minimum in terms of the target objective is sufficient for robustness.

3 Methodology

In this section, we first analyze the sufficient condition to achieve robust fairness over distribution shift in terms of demographic parity. Based on the analysis, we propose a simple yet effective robust fairness regularization via explicit adopting group model weight perturbation. Note that the proposed robust fairness regularization involves a computation-expensive maximization problem, we further accelerate model training via first-order Taylor expansion.

3.1 Robust Fairness Analysis

We consider binary sensitive attribute $A \in \{0,1\}$ and demographic parity as fairness metric, i.e., the average prediction gap of model $f_\theta(\cdot)$ for different sensitive attribute groups in target dataset $\Delta DP_{\mathcal{T}} = |\mathbb{E}_{\mathcal{T}_0}[f_\theta(\mathbf{x})] - \mathbb{E}_{\mathcal{T}_1}[f_\theta(\mathbf{x})]|$, where $\mathcal{T}_0$ and $\mathcal{T}_1$ represent the target datasets for sensitive attribute $A=0$ and $A=1$, respectively. However, only source dataset $\mathcal{S}$ is available for neural network training. In other words, even though demographic parity on source dataset $\Delta DP_{\mathcal{S}} = |\mathbb{E}_{\mathcal{S}_0}[f_\theta(\mathbf{x})] - \mathbb{E}_{\mathcal{S}_1}[f_\theta(\mathbf{x})]|$ is low, demographic parity on target dataset $\Delta DP_{\mathcal{T}}$ may not be guaranteed to be low due to distribution shift.

To investigate robust fairness over distribution shift, we try to reveal the connection between demographic parity for source and target datasets. We bound the demographic parity difference for source and target datasets as follows:

$$\begin{aligned} DP_{\mathcal{T}} &\overset{(a)}{\leq} DP_{\mathcal{S}} + \Big||\mathbb{E}_{\mathcal{T}_0}[f_\theta(\mathbf{x})] - \mathbb{E}_{\mathcal{T}_1}[f_\theta(\mathbf{x})]| - |\mathbb{E}_{\mathcal{S}_0}[f_\theta(\mathbf{x})] - \mathbb{E}_{\mathcal{S}_1}[f_\theta(\mathbf{x})]|\Big| \\ &\overset{(b)}{\leq} DP_{\mathcal{S}} + |\mathbb{E}_{\mathcal{S}_0}[f_\theta(\mathbf{x})] - \mathbb{E}_{\mathcal{T}_0}[f_\theta(\mathbf{x})]| + |\mathbb{E}_{\mathcal{S}_1}[f_\theta(\mathbf{x})] - \mathbb{E}_{\mathcal{T}_1}[f_\theta(\mathbf{x})]|, \end{aligned} \quad (5)$$

where inequality (a) and (b) hold due to $a-b \leq |a-b|$ and $\big||a-b|-|a'-b'|\big| \leq |a-a'|+|b-b'|$, respectively, for any a, a', b, b'. In other words, in order to minimize demographic parity for target dataset $DP_{\mathcal{T}}$, the objective of $DP_{\mathcal{S}}$ minimization is insufficient. The minimization of prediction difference for source and target datasets given sensitive attribute groups $A=0$ and $A=1$, defined as $\Delta_0 = |\mathbb{E}_{\mathcal{S}_0}[f_\theta(\mathbf{x})] - \mathbb{E}_{\mathcal{T}_0}[f_\theta(\mathbf{x})]|$ and $\Delta_1 = |\mathbb{E}_{\mathcal{S}_1}[f_\theta(\mathbf{x})] - \mathbb{E}_{\mathcal{T}_1}[f_\theta(\mathbf{x})]|$, are also beneficial to achieve robust fairness over distribution shift in terms of demographic parity.

The bound in Eq. (5) is tight when both condition (a) $DP_{\mathcal{S}} \leq DP_{\mathcal{T}}$ and (b) maximum and minimum of value set $\min\big\{\mathbb{E}_{\mathcal{S}_0}[f_\theta(\mathbf{x})], \mathbb{E}_{\mathcal{T}_0}[f_\theta(\mathbf{x})]\big\}, \min\big\{\mathbb{E}_{\mathcal{S}_1}[f_\theta(\mathbf{x})], \mathbb{E}_{\mathcal{T}_1}[f_\theta(\mathbf{x})]\big\}$ both are from source or target distribution. Even though conditions (a) and (b) may not hold for the neural network model, we

would like to that our goal is not to obtain a tight bound for demographic parity on target distribution. Instead, we aim to find sufficient conditions to guarantee low demographic parity and such low demographic parity can be achieved in model training without any target distribution information. The proposed upper bound Eq. (5) actually reveals sufficient conditions, which can be achieved by our proposed RFR algorithm.

3.2 Robust Fairness Regularization

Motivated by Section 3.1 and distribution shift understanding in Section 2, we develop a robust fairness regularization to achieve robust fairness over distribution shift. Section 3.1 demonstrates that fair prediction on the target dataset requires fair prediction on the source dataset and low prediction difference between the source and target dataset for each sensitive attribute, i.e., low $\Delta_0 = |\mathbb{E}_{\mathcal{S}_0}[f_\theta(\mathbf{x})] - \mathbb{E}_{\mathcal{T}_0}[f_\theta(\mathbf{x})]|$ and $\Delta_1 = |\mathbb{E}_{\mathcal{S}_1}[f_\theta(\mathbf{x})] - \mathbb{E}_{\mathcal{T}_1}[f_\theta(\mathbf{x})]|$. Based on Theorem 2.3, there exists ϵ_0 so that the following equality holds:

$$\Delta_0 = |\mathbb{E}_{\mathcal{S}_0}[f_\theta(\mathbf{x})] - \mathbb{E}_{\epsilon_0}\mathbb{E}_{\mathcal{S}_0}[f_{\theta+\epsilon_0}(\mathbf{x})]|. \tag{6}$$

Note that the distribution shift is unknown, we consider the worst case for model weight perturbation ϵ_0 within L_p-norm perturbation ball with radius ρ as follows:

$$\Delta_0 = |\mathbb{E}_{\mathcal{S}_0}[f_\theta(\mathbf{x})] - \mathbb{E}_{\epsilon_0}\mathbb{E}_{\mathcal{S}_0}[f_{\theta+\epsilon_0}(\mathbf{x})]| \leq \max_{\|\epsilon_0\|_p \leq \rho} |\mathbb{E}_{\mathcal{S}_0}[f_{\theta+\epsilon_0}(\mathbf{x})] - \mathbb{E}_{\mathcal{S}_0}[f_\theta(\mathbf{x})]|, \tag{7}$$

where $\|\cdot\|_p$ represents L_p norm, ρ and p are hyperparameters. Note that the feasible region of model weight perturbation ϵ_0 is symmetric, and the neural network prediction is locally linear around parameter θ, we have $\mathbb{E}_{\mathcal{S}_0}[f_{\theta+\epsilon_0}(\mathbf{x})] - \mathbb{E}_{\mathcal{S}_0}[f_\theta(\mathbf{x})] \approx -\Big(\mathbb{E}_{\mathcal{S}_0}[f_{\theta-\epsilon_0}(\mathbf{x})] - \mathbb{E}_{\mathcal{S}_0}[f_\theta(\mathbf{x})]\Big)$ due to the local linearity.In other words, the absolute operation can be removed if we consider the maximation problem in a symmetric feasible region since there are always non-negative value for the pair perturbation ϵ_0 and $-\epsilon_0$. Therefore, we can further bound Δ_0 as

$$\begin{aligned} \Delta_0 &\leq \max_{\|\epsilon_0\|_p \leq \rho} |\mathbb{E}_{\mathcal{S}_0}[f_{\theta+\epsilon_0}(\mathbf{x})] - \mathbb{E}_{\mathcal{S}_0}[f_\theta(\mathbf{x})]| \\ &\approx \max_{\|\epsilon_0\|_p \leq \rho} \mathbb{E}_{\mathcal{S}_0}[f_{\theta+\epsilon_0}(\mathbf{x})] - \mathbb{E}_{\mathcal{S}_0}[f_\theta(\mathbf{x})] \stackrel{\triangle}{=} \mathcal{L}_{RFR,\mathcal{S}_0}, \end{aligned} \tag{8}$$

Similarly, we can bound the prediction difference for source and target distribution with sensitive group $A = 1$ as follows:

$$\Delta_1 \leq \max_{\|\epsilon_1\|_p \leq \rho} \mathbb{E}_{\mathcal{S}_1}[f_{\theta+\epsilon_1}(\mathbf{x})] - \mathbb{E}_{\mathcal{S}_1}[f_\theta(\mathbf{x})] \stackrel{\triangle}{=} \mathcal{L}_{RFR,\mathcal{S}_1}. \tag{9}$$

Therefore, demographic parity for source and target distribution relation is given by $DP_{\mathcal{T}} \leq DP_{\mathcal{S}} + \mathcal{L}_{RFR,\mathcal{S}_0} + \mathcal{L}_{RFR,\mathcal{S}_1}$, we propose *robust fairness regualarization* (RFR) to achieve robust fairness as

$$\mathcal{L}_{RFR} = \mathcal{L}_{RFR,\mathcal{S}_0} + \mathcal{L}_{RFR,\mathcal{S}_1}. \tag{10}$$

It is worth noting that our proposed RFR is agnostic to the training loss function and model architectures. Additionally, we follow [12] to accelerate model training using sharpness-aware minimization via trading maximization problem can be simplified as two forward and two backward propagations. More details on training acceleration are in Appendix E.

3.3 The Proposed Method

In this subsection, we introduce how to use the proposed RFR to achieve robust fairness, i.e., the fair model (low $\Delta DP_{\mathcal{S}}$) trained on the source dataset will also be fair on the target dataset (low $\Delta DP_{\mathcal{T}}$). Considering binary sensitive attribute $A \in \{0, 1\}$ and binary classification problem $Y \in \{0, 1\}$, the classification loss is denoted as

$$\mathcal{L}_{CLF} = \mathbb{E}_{\mathcal{S}}[-Y f_\theta(X) - (1 - Y)\big(1 - f_\theta(X)\big)]. \tag{11}$$

To achieve fairness, we consider demographic parity as fairness regularization, i.e.,

$$\mathcal{L}_{DP} = |\mathbb{E}_{\mathcal{S}_0}[f_\theta(\mathbf{x})] - \mathbb{E}_{\mathcal{S}_1}[f_\theta(\mathbf{x})]|, \tag{12}$$

Table 1: Performance Comparison with Baselines on Synthetic Dataset. (α, β) control distribution shift intensity, and $(0, 1)$ represents no distribution shift. The best/second-best results are highlighted in **boldface**/underlined, respectively.

(α, β)	Methods	Adult			ACS-I			ACS-E		
		Acc (%) ↑	Δ_{DP} (%) ↓	Δ_{EO} (%) ↓	Acc (%) ↑	Δ_{DP} (%) ↓	Δ_{EO} (%) ↓	Acc (%) ↑	Δ_{DP} (%) ↓	Δ_{EO} (%) ↓
(1.0, 2.0)	MLP	82.09±0.05	15.11±0.04	14.33±0.05	77.95±0.52	3.51±0.59	3.77±0.55	80.95±0.10	1.10±0.06	1.43±0.06
	REG	80.60±0.05	3.79±0.06	3.27±0.08	77.77±0.09	2.28±0.32	2.59±0.23	80.44±0.07	0.86±0.09	1.05±0.10
	ADV	78.80±0.68	0.83±0.26	0.79±0.14	75.72±0.63	1.96±0.38	2.00±0.35	79.39±0.15	1.09±0.26	0.95±0.26
	FCR	79.06±0.09	9.98±0.06	9.47±0.07	76.99±0.47	2.94±0.34	2.95±0.28	79.74±0.11	0.97±0.21	1.00±0.22
	RFR	78.84±0.09	**0.44**±0.05	**0.12**±0.06	74.15±0.81	**1.84**±0.27	**1.60**±0.33	80.08±0.08	**0.71**±0.10	**0.06**±0.11
(1.5, 3.0)	MLP	82.05±0.05	15.16±0.09	14.33±0.09	77.85±0.25	3.73±0.53	3.70±0.56	80.42±0.10	1.14±0.07	1.10±0.07
	REG	80.64±0.08	3.74±0.11	3.23±0.10	77.87±0.18	2.25±0.28	2.37±0.27	80.21±0.13	**0.72**±0.04	0.75±0.03
	ADV	78.71±0.41	1.07±0.87	0.87±0.96	75.79±0.68	2.22±0.53	2.44±0.48	79.58±0.13	1.07±0.19	1.26±0.18
	FCR	79.05±0.12	10.01±0.07	9.51±0.06	77.06±0.68	3.39±0.33	3.10±0.36	79.59±0.26	1.17±0.24	1.08±0.23
	RFR	78.91±0.03	**0.46**±0.10	**0.16**±0.09	74.19±0.58	**1.82**±0.29	**2.17**±0.32	80.47±0.03	**0.72**±0.04	**0.71**±0.05
(3.0, 6.0)	MLP	82.07±0.05	15.23±0.14	14.45±0.15	77.89±0.45	3.35±0.36	3.47±0.41	80.30±0.04	1.17±0.04	1.13±0.04
	REG	80.62±0.07	3.72±0.05	3.21±0.04	78.19±0.12	**1.60**±0.48	**1.84**±0.44	80.36±0.09	**0.70**±0.09	0.68±0.11
	ADV	78.97±0.49	**1.28**±0.74	**1.09**±0.50	75.71±0.68	2.28±0.39	2.24±0.41	79.66±0.16	1.34±0.14	1.16±0.13
	FCR	79.03±0.13	10.00±0.05	9.50±0.05	76.71±0.39	2.97±0.34	3.28±0.31	79.89±0.22	1.06±0.14	1.14±0.18
	RFR	80.15±0.07	1.75±0.15	1.30±0.14	74.22±0.56	1.80±0.26	1.89±0.24	80.28±0.12	0.74±0.04	**0.51**±0.04

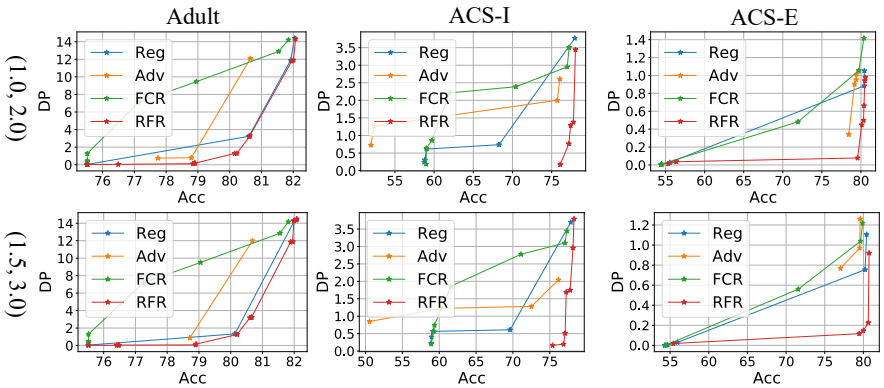

Figure 2: The fairness (DP) and prediction (Acc) trade-off performance on three datasets with different synthetic distribution shifts. The units for x- and y-axis are percentages (%).

Additionally, we also adopt $\mathcal{L}_{RFR}$ as robust fairness regularization. The overall objective is given as follows:

$$\mathcal{L}_{all} = \mathcal{L}_{CLF} + \lambda \cdot \big(\mathcal{L}_{DP} + \mathcal{L}_{RFR}\big), \tag{13}$$

where λ is a hyperparameter to balance the model performance and fairness. $\mathcal{L}_{CLF}$ is the loss function for the downstream task, $\mathcal{L}_{DP}$ is to ensure the demographic parity on the source dataset, and $\mathcal{L}_{RFR}$ is to ensure the transferability of fairness from the source dataset to the target dataset.

In the proposed RFR algorithm, we directly apply approximation to accelerate model training. Note that the optimal model weight perturbation is dependent on the current model weight θ, two forward and backward propagations are required to calculate the final gradient.

4 Experiments

In this section, we conduct experiments to evaluate the effectiveness of our proposed RFR, aiming to answer the three research questions. **Q1**: How effective is RFR to achieve fairness under distribution shift for synthetic distribution shift? **Q2**: How effective is RFR for real spatial and temporal distribution shift? **Q3**: How sensitive is RFR to the key hyperparameter λ?

4.1 Experimental Setting

In this subsection, we present the experimental setting, including datasets, evaluation metrics, baseline methods, distribution shift generation, and implementation details.

Datasets. We adopt the following datasets in our experiments. **UCI Adult** [13] dataset contains information about $45,222$ individuals with 15 attributes from the 1994 US Census. We consider

gender as the sensitive attribute and the task is to predict whether the income of the person is higher than $\$50k$ or not. **ACS-I**ncome [14] is extracted from the American Community Survey (ACS) Public Use Microdata Sample (PUMS) with $3,236,107$ samples. We choose gender as the sensitive attribute. Similar to the task in UCI Adult, the task is to predict whether the individual income is above $\$50k$. **ACS-E**mployment [14] also derives from ACS PUMS, and we also use gender as the sensitive attribute. The task is to predict whether an individual is employed.

Evaluation Metrics. We use accuracy to evaluate the prediction performance for the downstream task. For fairness metrics, we adopt two common-used quantitative group fairness metrics to measure the prediction bias [15, 16], i.e., *demographic parity* $\Delta_{DP} = |\mathbb{P}(\hat{Y} = 1|A = 0) - \mathbb{P}(\hat{Y} = 1|A = 1)|$ and *equal opportunity* $\Delta_{EO} = |\mathbb{P}(\hat{Y} = 1|A = 0, Y = 1) - \mathbb{P}(\hat{Y} = 1|A = 1, Y = 1)|$, where A represents sensitive attribute, Y and $\hat{Y}$ represent the ground-truth label and predicted label, respectively.

Baselines. In our experiments, we consider vanilla multi-layer perceptron (MLP) and two widely adopted in-processing debiasing methods, including fairness regularization (REG), adversarial debiasing (ADV), and fair consistency regularization (FCR). **MLP** directly uses vanilla 3-layer MLP with 50 hidden unit and ReLU activation function [17] to minimize cross-entropy loss with source dataset. In the experiments, we adopt the same model architecture for all other methods (i.e., REG and ADV). **REG** adds a fairness-related metric as a regularization term in the objective function to mitigate the prediction bias [18, 19]. Specifically, we directly adopt demographic parity as a regularization term, and the objective function is $\mathcal{L}_{CLF} + \lambda\mathcal{L}_{DP}$, where $\mathcal{L}_{CLF}$ and $\mathcal{L}_{DP}$ are defined in Eqs. (11) and (12). **ADV** [20] employs a two-player game to mitigate bias, in which a classification network is trained to predict labels based on input features, and an adversarial network takes the output of the classification network as input and aims to identify which sensitive attribute group the sample belongs to. **FCR** [21] aims to minimize and balance consistency loss across groups.

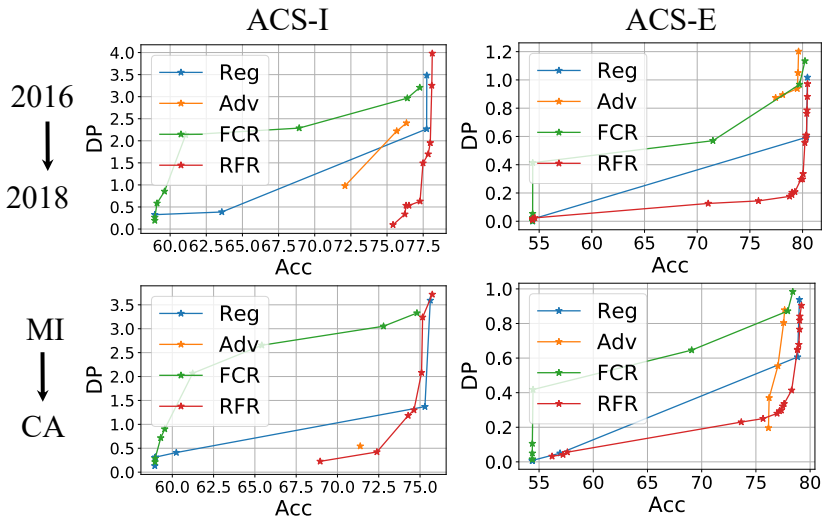

Figure 3: DP and Acc trade-off performance on three real-world datasets with temporal (Top) and spatial (Bottom) distribution shift. The trade-off curve close to the right bottom corner means better trade-off performance. The units for x- and y-axis are percentages (%).

Synthetic and Real Distribution Shift. We adopt synthetic and real distribution shifts. For synthetic distribution shift, we follow work [22, 23] to generate distribution shift via biased sampling. Specifically, we adopt applying principal component analysis (PCA) [24] to retrieve the first principal component $\mathcal{C}$ from input attributes. Subsequently, we estimate the mean $\mu(\mathcal{C})$ and standard deviation $\sigma(\mathcal{C})$, and set Gaussian distribution $\mathcal{N}_{\mathcal{S}}(\mu(\mathcal{C}), \sigma(\mathcal{C}))$ to randomly sampling of target dataset. As for source dataset, we choose different parameters α and β to generate distribution shift using another Gaussian distribution $\mathcal{N}_{\mathcal{T}}(\mu(\mathcal{C}) + \alpha, \frac{\sigma(\mathcal{C})}{\beta})$ for randomly sampling. The source and target datasets are constructed by sampling without replacement. For real distribution shift, we adopt the sampling

and pre-processing approaches following *Folktables* [14] to generate spatial and temporal distribution shift via data partition based on different US states and year from 2014 to 2018.

Implementation Details. We run the experiments 5 times and report the average performance for each method. We adopt Adam optimizer with 10^{-5} learning rate and 0.01 weight decay for all models. For baseline ADV, we alternatively train classification and adversarial networks with 70 and 30 epochs, respectively. The hyperparameters for ADV are set as $\{0.0, 1.0, 10.0, 100.0, 500.0\}$. For adding regularization, we adopt the hyperparameters set $\{0.0, 0.5, 1.0, 10.0, 30.0, 50.0\}$.

4.2 Experimental Results on Synthetic Distribution Shift

In this experiment, we evaluate the effectiveness of our proposed Robust Fairness Regularization (RFR) method with synthetic distribution shifts via biased sampling with different parameters (α, β). We compare the performance of RFR with several other baseline methods, including standard training, adversarial training, and fairness regularization methods. The results of performance value are presented in Table 1, and the results of fairness-accuracy tradeoff are presented in Figure 2. We have the following observations:

- The results in Table 1 demonstrate that RFR consistently outperforms the baselines in terms of fairness for small distribution shifts, achieving a better balance between fairness and accuracy. For example, RFR achieves an improvement of 47.0% on metric ΔDP and 84.8% on metric ΔEO compared to the second-best method in the Adult dataset with $(1.0, 2.0)$-synthetic distribution shift. Furthermore, we observed that the outperforms of RFR compared with baselines decrease as the distribution shift intensity increases. The reason is that the approximation RFR involved Taylor expansion over model perturbation and is effective with mild distribution shifts.
- The results in Figure 1 show that RFR achieved a better fairness-accuracy tradeoff compared to the baseline methods for mild distribution shift, We observed that our proposed method achieved a better Pareto frontier compared to the existing methods, and the bias can be mitigated with tiny accuracy drop.

The experimental results show that our method can effectively address the fairness problem under mild distribution shift while maintaining high accuracy, outperforming the existing state-of-the-art baseline methods.

4.3 Experimental Results on Real Distribution Shift

In this experiment, we evaluate the performance of RFR on multiple real-world datasets with real distribution shifts. We use ACS dataset in the experiment. The distribution of this dataset varies across different time periods or geographic locations, which also causes fairness performance degradation under the distribution shift. The results are presented in Table 2, and the results of the fairness-accuracy tradeoff are presented in Figure 3. The results show that our method consistently outperforms the baselines across all datasets, achieving a better balance between fairness and accuracy. Specifically, we have the following observations:

- Table 2 demonstrates that RFR consistently outperforms the baselines across all datasets in terms of fairness-accuracy tradeoff. This suggests that our method is effective in achieving robust fairness under real temporal and spatial distribution shifts. For example, RFR achieved an improvement of 34.9% on metric ΔDP and 34.4% on metric ΔEO compared to the second-best method in ACS-I dataset with temporal distribution shift.
- Figure 3 shows that RFR can achieve better fairness-accuracy tradeoff than baselines, i.e., RFR is particularly effective in addressing the fairness problem on the ACS dataset with source/target data varying across different time periods or geographic locations. This is important for many practical applications, where fairness is a critical requirement, such as in credit scoring, and loan approval.
- The variance for spatial shift on ACS-I is higher than that of temporal shift on all methods, which indicates neural networks easily converge to different local minima for spatial distribution shift. The proposed methods can achieve better fairness results for most cases and the tradeoff results clearly demonstrate the effectiveness.

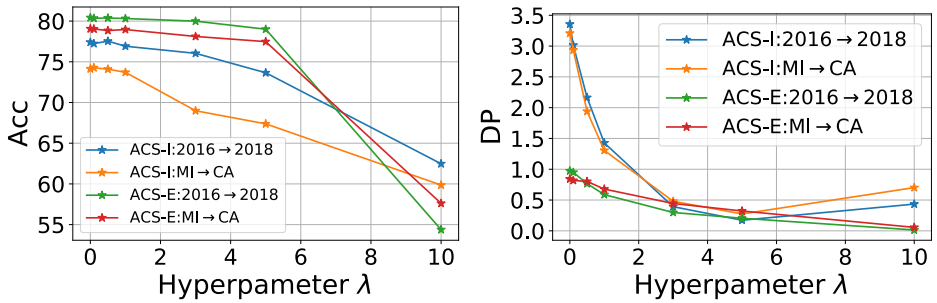

Figure 4: The hyperparameters study for λ. The **Left** subfigure and **Right** subfigure show the results of accuracy and fairness performance, respectively.

Overall, the experimental results on temporal and spatial distribution shifts further support the effectiveness of our proposed method RFR and its potential for practical applications in diverse settings.

Table 2: Performance comparison with baselines on real temporal (the year 2016 to the year 2018) and spatial (Michigan State to California State) distribution shift. The best and second-best results are highlighted with **hold** and underline, respectively.

Real	Methods	ACS-I			ACS-E		
		Acc (%) ↑	Δ_{DP} (%) ↓	Δ_{EO} (%) ↓	Acc (%) ↑	Δ_{DP} (%) ↓	Δ_{EO} (%) ↓
2016 → 2018	MLP	77.75±0.44	3.26±0.38	3.48±0.41	80.46±0.05	1.07±0.10	1.02±0.10
	REG	77.74±0.62	2.09±0.21	2.27±0.24	80.37±0.12	0.77±0.08	0.74±0.08
	ADV	75.94±0.40	2.41±0.49	2.53±0.55	79.62±0.14	1.17±0.14	1.10±0.14
	FCR	76.40±0.45	2.81±0.30	2.96±0.30	79.59±0.38	0.95±0.42	0.91±0.34
	RFR	77.49±0.32	**1.36**±0.17	**1.49**±0.17	80.36±0.05	**0.61**±0.11	**0.58**±0.10
MI → CA	MLP	75.62±0.80	5.22±0.86	3.60±0.34	79.02±0.20	0.73±0.07	0.94±0.05
	REG	75.52±0.78	2.88±0.44	2.17±0.22	75.34±1.11	**0.42**±0.09	**0.61**±0.11
	ADV	73.38±1.07	**1.04**±0.58	**0.54**±0.38	77.56±0.41	0.61±0.18	0.80±0.13
	FCR	74.28±0.35	5.06±0.62	3.67±0.51	77.96±0.22	0.44±0.14	0.67±0.38
	RFR	74.63±0.45	1.35±0.39	1.30±0.24	78.84±0.21	0.44±0.09	0.65±0.07

4.4 Hyperparameter Study

In this experiment, we investigate the sensitivity of the hyperparameter λ in Equation $\mathcal{L}_{all} = \mathcal{L}_{CLF} + \lambda(\mathcal{L}_{DP} + \mathcal{L}_{RFR})$ for spatial and temporal distribution shift across different datasets. Specifically, we tune the hyperparameter as $\lambda = \{0.0, 0.1, 0.5, 1.0, 3.0, 5.0, 10.0\}$. From the results in Figure 4, it is seen that the accuracy and demographic parity are both sensitive to hyperparameter λ, which implies the capability of accuracy-fairness control. With the increase of λ, accuracy decreases while demographic parity also decreases. When λ is smaller than 5, accuracy drops slowly while demographic parity drops faster. Such observation represents that an appropriate hyperparameter can mitigate prediction bias while preserving comparable prediction performance.

5 Related Work

In this section, we present two lines of related work, including fairness in machine learning and distribution shift.

Fairness in Machine Learning. Fairness [25**?** –31] is a legal requirement for machine learning models for various high-stake real-world predictions, such as healthcare [7, 32, 33], education [34–36], and job market [37, 38]. Achieving fairness, either from a data or model perspective [39–42], in machine learning is a challenging problem. As such, there has been an increasing interest in both the industrial and research community to develop algorithms to mitigate these issues and ensure that machine learning models make fair and unbiased decisions. Extensive efforts led to the development of various techniques and metrics for fairness and proposed various definitions of fairness, such as group fairness [43–47, 40], individual fairness [48–53], and counterfactual fairness [54–56]. In this paper, we focus on group fairness, and the widely used methods to achieve group fairness are fair regularization and adversarial debias method. [19, 18] proposed to add a fairness regularization term

to the objective function to achieve group fairness, and [20] proposes to jointly train a classification network and an adversarial network to mitigate the bias for different demographic groups to achieve group fairness. Overall, ensuring fairness in machine learning is a critical and ongoing research area that will continue to be an important focus for the development of responsible machine learning.

Distribution Shift. Previous work [1–4, 57] reveals that a classifier trained on a source distribution will perform worse on a given target distribution because of the distribution shift, and recently extensive works have explored the influence of distribution shift in model prediction. Moreover, distribution shift can significantly affect the fairness performance of machine learning models. The sensitivity of fairness to the distribution shift is notorious for the legal requirement. The degraded performance of fair models under a distribution shift would trigger new bias and discrimination issues. There are some works [23, 58, 21, 59, 60] that solve fairness under various distribution shifts. For example, [23] explore the fairness under covariate shift, where the inputs change with the in-distribution label. [60] proposes Shifty algorithms to hold fairness guarantees when the dataset in the deployment environment is out-of-distribution of the training datasets (distribution shift). Our work is different from those prior works from model weight perturbation perspective.

6 Conclusion

This paper aims the solve the fairness problem under the distribution shifts from the model weight perturbation perspective. We first establish a theoretical connection between distribution shift, data perturbation, and model weight perturbation, which allowed us to conclude that distribution shift and model perturbation are equivalent. We then propose a sufficient condition for ensuring fairness transference under distribution shift. To explicitly chase such sufficient conditions, we introduce Robust Fairness Regularization (RFR) method based on the established understanding, to achieve robust fairness. Our experiments on both synthetic and real distribution shifts demonstrate the effectiveness of RFR in achieving a better fairness-accuracy tradeoff compared to existing baselines. We believe that our understanding of distribution shift is valuable and intriguing to the development of robust machine learning models, and the proposed RFR approach can be of great practical value to build fair and robust machine learning models in real-world applications.

7 Acknowledgement

The authors thank the anonymous reviewers for their helpful comments. The work is in part supported by NSF grants NSF IIS-2224843, IIS-1939716, and IIS-1900990. The views and conclusions contained in this paper are those of the authors and should not be interpreted as representing any funding agencies.

References

[1] Dan Hendrycks and Thomas Dietterich. Benchmarking neural network robustness to common corruptions and perturbations. *arXiv preprint arXiv:1903.12261*, 2019.

[2] Dan Hendrycks, Steven Basart, Norman Mu, Saurav Kadavath, Frank Wang, Evan Dorundo, Rahul Desai, Tyler Zhu, Samyak Parajuli, Mike Guo, et al. The many faces of robustness: A critical analysis of out-of-distribution generalization. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 8340–8349, 2021.

[3] Rohan Taori, Achal Dave, Vaishaal Shankar, Nicholas Carlini, Benjamin Recht, and Ludwig Schmidt. Measuring robustness to natural distribution shifts in image classification. *Advances in Neural Information Processing Systems*, 33:18583–18599, 2020.

[4] Steffen Schneider, Evgenia Rusak, Luisa Eck, Oliver Bringmann, Wieland Brendel, and Matthias Bethge. Improving robustness against common corruptions by covariate shift adaptation. *Advances in Neural Information Processing Systems*, 33:11539–11551, 2020.

[5] L Elisa Celis, Anay Mehrotra, and Nisheeth Vishnoi. Fair classification with adversarial perturbations. *Advances in Neural Information Processing Systems*, 34:8158–8171, 2021.

[6] Songül Tolan, Marius Miron, Emilia Gómez, and Carlos Castillo. Why machine learning may lead to unfairness: Evidence from risk assessment for juvenile justice in catalonia. In *Proceedings of the Seventeenth International Conference on Artificial Intelligence and Law*, pages 83–92, 2019.

[7] Muhammad Aurangzeb Ahmad, Arpit Patel, Carly Eckert, Vikas Kumar, and Ankur Teredesai. Fairness in machine learning for healthcare. In *Proceedings of the 26th ACM SIGKDD*, pages 3529–3530, 2020.

[8] Jeff Johnson, Donald M Truxillo, Berrin Erdogan, Talya N Bauer, and Leslie Hammer. Perceptions of overall fairness: are effects on job performance moderated by leader-member exchange? *Human Performance*, 22(5), 2009.

[9] Ainhize Barrainkua, Paula Gordaliza, Jose A Lozano, and Novi Quadrianto. A survey on preserving fairness guarantees in changing environments. *arXiv preprint arXiv:2211.07530*, 2022.

[10] Olivia Wiles, Sven Gowal, Florian Stimberg, Sylvestre-Alvise Rebuffi, Ira Ktena, Krishnamurthy Dj Dvijotham, and Ali Taylan Cemgil. A fine-grained analysis on distribution shift. In *International Conference on Learning Representations*, 2022.

[11] Cédric Villani et al. *Optimal transport: old and new*, volume 338. Springer, 2009.

[12] Pierre Foret, Ariel Kleiner, Hossein Mobahi, and Behnam Neyshabur. Sharpness-aware minimization for efficiently improving generalization. *arXiv preprint arXiv:2010.01412*, 2020.

[13] Dheeru Dua and Casey Graff. UCI machine learning repository, 2017.

[14] Frances Ding, Moritz Hardt, John Miller, and Ludwig Schmidt. Retiring adult: New datasets for fair machine learning. *NeurIPS*, 2021.

[15] Christos Louizos, Kevin Swersky, Yujia Li, Max Welling, and Richard Zemel. The variational fair autoencoder. *arXiv preprint arXiv:1511.00830*, 2015.

[16] Alex Beutel, Jilin Chen, Zhe Zhao, and Ed H Chi. Data decisions and theoretical implications when adversarially learning fair representations. *arXiv preprint arXiv:1707.00075*, 2017.

[17] Vinod Nair and Geoffrey E Hinton. Rectified linear units improve restricted boltzmann machines. In *ICML*, 2010.

[18] Toshihiro Kamishima, Shotaro Akaho, Hideki Asoh, and Jun Sakuma. Fairness-aware classifier with prejudice remover regularizer. In *Joint European conference on machine learning and knowledge discovery in databases*. Springer, 2012.

[19] Ching-Yao Chuang and Youssef Mroueh. Fair mixup: Fairness via interpolation. In *ICLR*, 2020.

[20] Gilles Louppe, Michael Kagan, and Kyle Cranmer. Learning to pivot with adversarial networks. *NeurIPS*, 30, 2017.

[21] Bang An, Zora Che, Mucong Ding, and Furong Huang. Transferring fairness under distribution shifts via fair consistency regularization. *arXiv preprint arXiv:2206.12796*, 2022.

[22] Arthur Gretton, Alex Smola, Jiayuan Huang, Marcel Schmittfull, Karsten Borgwardt, and Bernhard Schölkopf. Covariate shift by kernel mean matching. *Dataset shift in machine learning*, 3(4):5, 2009.

[23] Ashkan Rezaei, Anqi Liu, Omid Memarrast, and Brian D Ziebart. Robust fairness under covariate shift. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 35, pages 9419–9427, 2021.

[24] Hervé Abdi and Lynne J Williams. Principal component analysis. *Wiley interdisciplinary reviews: computational statistics*, 2(4):433–459, 2010.

[25] Ninareh Mehrabi, Fred Morstatter, Nripsuta Saxena, Kristina Lerman, and Aram Galstyan. A survey on bias and fairness in machine learning. *ACM Computing Surveys (CSUR)*, 54(6):1–35, 2021.

[26] Dana Pessach and Erez Shmueli. A review on fairness in machine learning. *ACM Computing Surveys (CSUR)*, 55(3):1–44, 2022.

[27] Michael Madaio, Lisa Egede, Hariharan Subramonyam, Jennifer Wortman Vaughan, and Hanna Wallach. Assessing the fairness of ai systems: Ai practitioners' processes, challenges, and needs for support. *Proceedings of the ACM on Human-Computer Interaction*, 6(CSCW1):1–26, 2022.

[28] Dexun Li and Pradeep Varakantham. Efficient resource allocation with fairness constraints in restless multi-armed bandits. In James Cussens and Kun Zhang, editors, *Uncertainty in Artificial Intelligence, Proceedings of the Thirty-Eighth Conference on Uncertainty in Artificial Intelligence, UAI 2022, 1-5 August 2022, Eindhoven, The Netherlands*, volume 180 of *Proceedings of Machine Learning Research*, pages 1158–1167. PMLR, 2022.

[29] Kirtan Padh, Diego Antognini, Emma Lejal Glaude, Boi Faltings, and Claudiu Musat. Addressing fairness in classification with a model-agnostic multi-objective algorithm. In Cassio P. de Campos, Marloes H. Maathuis, and Erik Quaeghebeur, editors, *Proceedings of the Thirty-Seventh Conference on Uncertainty in Artificial Intelligence, UAI 2021, Virtual Event, 27-30 July 2021*, volume 161 of *Proceedings of Machine Learning Research*, pages 600–609. AUAI Press, 2021.

[30] Zhimeng Jiang, Xiaotian Han, Chao Fan, Zirui Liu, Na Zou, Ali Mostafavi, and Xia Hu. Fmp: Toward fair graph message passing against topology bias. *arXiv preprint arXiv:2202.04187*, 2022.

[31] Xiaotian Han, Zhimeng Jiang, Hongye Jin, Zirui Liu, Na Zou, Qifan Wang, and Xia Hu. Retiring ΔDP: New distribution-level metrics for demographic parity. *Transactions on Machine Learning Research*, 2023.

[32] Margrét Vilborg Bjarnadóttir and David Anderson. Machine learning in healthcare: Fairness, issues, and challenges. In *Pushing the Boundaries: Frontiers in Impactful OR/OM Research*, pages 64–83. INFORMS, 2020.

[33] Thomas Grote and Geoff Keeling. Enabling fairness in healthcare through machine learning. *Ethics and Information Technology*, 24(3):39, 2022.

[34] Steinar Bøyum. Fairness in education–a normative analysis of oecd policy documents. *Journal of Education Policy*, 29(6):856–870, 2014.

[35] Paolo Brunori, Vito Peragine, and Laura Serlenga. Fairness in education: The italian university before and after the reform. *Economics of Education Review*, 31(5):764–777, 2012.

[36] René F Kizilcec and Hansol Lee. Algorithmic fairness in education. In *The ethics of artificial intelligence in education*, pages 174–202. Routledge, 2022.

[37] Lily Hu and Yiling Chen. A short-term intervention for long-term fairness in the labor market. In *Proceedings of the 2018 World Wide Web Conference*, pages 1389–1398, 2018.

[38] G Stoney Alder and Joseph Gilbert. Achieving ethics and fairness in hiring: Going beyond the law. *Journal of Business Ethics*, 68:449–464, 2006.

[39] Daochen Zha, Zaid Pervaiz Bhat, Kwei-Herng Lai, Fan Yang, Zhimeng Jiang, Shaochen Zhong, and Xia Hu. Data-centric artificial intelligence: A survey. *arXiv preprint arXiv:2303.10158*, 2023.

[40] Qizhang Feng, Zhimeng Jiang, Ruiquan Li, Yicheng Wang, Na Zou, Jiang Bian, and Xia Hu. Fair graph distillation. 2023.

[41] Chia-Yuan Chang, Yu-Neng Chuang, Zhimeng Jiang, Kwei-Herng Lai, Anxiao Jiang, and Na Zou. Coda: Temporal domain generalization via concept drift simulator. *arXiv preprint arXiv:2310.01508*, 2023.

[42] Chia-Yuan Chang, Yu-Neng Chuang, Kwei-Herng Lai, Xiaotian Han, Xia Hu, and Na Zou. Towards assumption-free bias mitigation. *arXiv preprint arXiv:2307.04105*, 2023.

[43] Moritz Hardt, Eric Price, and Nati Srebro. Equality of opportunity in supervised learning. *NeurIPS*, 29, 2016.

[44] Sahil Verma and Julia Rubin. Fairness definitions explained. In *2018 ieee/acm international workshop on software fairness (fairware)*, pages 1–7. IEEE, 2018.

[45] Peizhao Li, Yifei Wang, Han Zhao, Pengyu Hong, and Hongfu Liu. On dyadic fairness: Exploring and mitigating bias in graph connections. In *International Conference on Learning Representations*, 2020.

[46] Gaurush Hiranandani, Jatin Mathur, Harikrishna Narasimhan, and Oluwasanmi Koyejo. Quadratic metric elicitation for fairness and beyond. In James Cussens and Kun Zhang, editors, *Uncertainty in Artificial Intelligence, Proceedings of the Thirty-Eighth Conference on Uncertainty in Artificial Intelligence, UAI 2022, 1-5 August 2022, Eindhoven, The Netherlands*, volume 180 of *Proceedings of Machine Learning Research*, pages 811–821. PMLR, 2022.

[47] Zhimeng Jiang, Xiaotian Han, Chao Fan, Fan Yang, Ali Mostafavi, and Xia Hu. Generalized demographic parity for group fairness. In *International Conference on Learning Representations*, 2022.

[48] Mikhail Yurochkin, Amanda Bower, and Yuekai Sun. Training individually fair ml models with sensitive subspace robustness. *International Conference on Learning Representations*, 2020.

[49] Debarghya Mukherjee, Mikhail Yurochkin, Moulinath Banerjee, and Yuekai Sun. Two simple ways to learn individual fairness metrics from data. In *ICML*, 2020.

[50] Mikhail Yurochkin and Yuekai Sun. Sensei: Sensitive set invariance for enforcing individual fairness. In *International Conference on Learning Representations*, 2021.

[51] Jian Kang, Jingrui He, Ross Maciejewski, and Hanghang Tong. Inform: Individual fairness on graph mining. In *Proceedings of the 26th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pages 379–389, 2020.

[52] Debarghya Mukherjee, Felix Petersen, Mikhail Yurochkin, and Yuekai Sun. Domain adaptation meets individual fairness. and they get along. In Alice H. Oh, Alekh Agarwal, Danielle Belgrave, and Kyunghyun Cho, editors, *Advances in Neural Information Processing Systems*, 2022.

[53] Shantanu Das, Swapnil Dhamal, Ganesh Ghalme, Shweta Jain, and Sujit Gujar. Individual fairness in feature-based pricing for monopoly markets. In James Cussens and Kun Zhang, editors, *Uncertainty in Artificial Intelligence, Proceedings of the Thirty-Eighth Conference on Uncertainty in Artificial Intelligence, UAI 2022, 1-5 August 2022, Eindhoven, The Netherlands*, volume 180 of *Proceedings of Machine Learning Research*, pages 486–495. PMLR, 2022.

[54] Matt J Kusner, Joshua Loftus, Chris Russell, and Ricardo Silva. Counterfactual fairness. *NeurIPS*, 30, 2017.

[55] Chirag Agarwal, Himabindu Lakkaraju, and Marinka Zitnik. Towards a unified framework for fair and stable graph representation learning. In *UAI 2021: Uncertainty in Artificial Intelligence*, 2021.

[56] Aoqi Zuo, Susan Wei, Tongliang Liu, Bo Han, Kun Zhang, and Mingming Gong. Counterfactual fairness with partially known causal graph. In Alice H. Oh, Alekh Agarwal, Danielle Belgrave, and Kyunghyun Cho, editors, *Advances in Neural Information Processing Systems*, 2022.

[57] Chia-Yuan Chang, Yu-Neng Chuang, Guanchu Wang, Mengnan Du, and Zou Na. Dispel: Domain generalization via domain-specific liberating. *arXiv preprint arXiv:2307.07181*, 2023.

[58] Jessica Schrouff, Natalie Harris, Oluwasanmi Koyejo, Ibrahim Alabdulmohsin, Eva Schnider, Krista Opsahl-Ong, Alex Brown, Subhrajit Roy, Diana Mincu, Christina Chen, et al. Maintaining fairness across distribution shift: do we have viable solutions for real-world applications? *arXiv preprint arXiv:2202.01034*, 2022.

[59] Harvineet Singh, Rina Singh, Vishwali Mhasawade, and Rumi Chunara. Fairness violations and mitigation under covariate shift. In *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency*, pages 3–13, 2021.

[60] Stephen Giguere, Blossom Metevier, Bruno Castro da Silva, Yuriy Brun, Philip Thomas, and Scott Niekum. Fairness guarantees under demographic shift. In *International Conference on Learning Representations*, 2022.

[61] Jiongli Zhu, Sainyam Galhotra, Nazanin Sabri, and Babak Salimi. Consistent range approximation for fair predictive modeling. *Proceedings of the VLDB Endowment*, 16(11):2925–2938, 2023.

[62] Jiawei Du, Hanshu Yan, Jiashi Feng, Joey Tianyi Zhou, Liangli Zhen, Rick Siow Mong Goh, and Vincent YF Tan. Efficient sharpness-aware minimization for improved training of neural networks. *International Conference on Learning Representations*, 2022.

[63] Maksym Andriushchenko and Nicolas Flammarion. Towards understanding sharpness-aware minimization. In *International Conference on Machine Learning*, pages 639–668. PMLR, 2022.

[64] Yuzhe Lu, Zhenlin Wang, Runtian Zhai, Soheil Kolouri, Joseph Campbell, and Katia Sycara. Predicting out-of-distribution error with confidence optimal transport. *arXiv preprint arXiv:2302.05018*, 2023.

[65] Yaodong Yu, Zitong Yang, Alexander Wei, Yi Ma, and Jacob Steinhardt. Predicting out-of-distribution error with the projection norm. In *International Conference on Machine Learning*, pages 25721–25746. PMLR, 2022.

A Proof of Theorem 2.1

For any two different distributions with probability distribution $P_{\mathcal{S}}$ and $P_{\mathcal{T}}$, based on optimal transport definition, any transport plan in $\Gamma(P_{\mathcal{S}}, P_{\mathcal{T}})$ can move source distribution $P_{\mathcal{S}}$ to target distribution $P_{\mathcal{T}}$. Define data perturbation $\delta = T - S$ and joint distribution of S and T is given by $\gamma^*(s,t)$. We can obtain the following probability for any u

$$F(u) = \mathbb{P}(\delta \leq u) = \int_{\mathcal{S}} \int_{\mathcal{T}_{s,u}} \gamma^*(s,t) \mathrm{d}t \mathrm{d}s, \tag{14}$$

where $\mathcal{T}_{s,u} = \big\{t : t \leq u + s\big\}$. The probability of data perturbation δ is given by

$$\mathcal{P}(\delta) = \frac{\partial F(u)}{\partial u}\big\|_{u=\delta} = \int_{\mathcal{S}} \gamma^*(s, s+\delta) \mathrm{d}s, \tag{15}$$

Note that, for any positive $p > 0$, the power of data perturbation δ satisfy

$$\mathbb{E}[||\delta||_p^p] = \mathbb{E}[||T - S||_p^p] = \iint ||s - t||_p^p \gamma^*(s,t) \mathrm{d}s \mathrm{d}t, \tag{16}$$

where $||\cdot||_p$ represents L_p norm. The data perturbation δ based on Eq. (1)with square cost function $c(s,t) = ||s-t||_p^p$ has minimal power.

B Proof of Corollary 2.2

Based on Theorem 2.1, it is easy to see $(X + \delta_X(X), Y + \delta_Y(Y)) \sim \mathcal{P}_{\mathcal{T}}$ if $(X, Y) \sim \mathcal{P}_{\mathcal{S}}$, therefore, the following equality holds for any loss function $l(\cdot, \cdot)$:

$$\begin{aligned} & \mathbb{E}_{(X,Y)\sim\mathcal{P}_{\mathcal{T}}}[l(f_\theta(X), Y)] \\ = \; & \mathbb{E}_{\delta_X(X),\delta_Y(Y)} \mathbb{E}_{(X,Y)\sim\mathcal{P}_{\mathcal{S}}}[l(f_\theta(X + \delta_X(X)), Y + \delta_Y(Y))]. \end{aligned}$$

C Proof of Theorem 2.3

We consider the small distribution shift, and smooth training loss and model prediction function, the equivalent data perturbation is also small. We conduct first-order Taylor expansion over loss function $l(f_\theta(X + \delta_X(X)), Y + \delta_Y(Y))$, we have

$$\begin{aligned} & l(f_\theta(X + \delta_X(X)), Y + \delta_Y(Y)) \\ = \; & l(f_\theta(X), Y) + \frac{\partial l}{\partial f}\frac{\partial f}{\partial X}\Big|_{X=\hat{X}} \delta_X(X) + \frac{\partial l}{\partial f}\frac{\partial f}{\partial Y}\Big|_{Y=\hat{Y}} \delta_Y(Y), \end{aligned} \tag{17}$$

where $\hat{X}$ is between X and $X + \delta_X(X)$, and $\hat{Y}$ is between Y and $Y + \delta_Y(Y)$. For simplicity, we ignore $\hat{X}$ and $\hat{Y}$ in the following derivation. Subsequently, we can approximate the training loss on the target dataset as follows:

$$\begin{aligned} & \mathbb{E}_{\delta_X(X),\delta_Y(Y)} \mathbb{E}_{(X,Y)\sim\mathcal{P}_{\mathcal{S}}}[l(f_\theta(X + \delta_X(X)), Y + \delta_Y(Y))] \\ = \; & \mathcal{R}_{\mathcal{S}} + \mathbb{E}_{\delta_X(X),\delta_Y(Y)} \mathbb{E}_{(X,Y)\sim\mathcal{P}_{\mathcal{S}}}[\frac{\partial l}{\partial f}\frac{\partial f}{\partial X}\delta_X(X) + \frac{\partial l}{\partial f}\frac{\partial f}{\partial Y}\delta_Y(Y)] \\ = \; & \mathcal{R}_{\mathcal{S}} + \mathbb{E}_{(X,Y)\sim\mathcal{P}_{\mathcal{S}}}[\frac{\partial l}{\partial f}\frac{\partial f}{\partial X}] \cdot \mathbb{E}_{\delta_X(X)}[\delta_X(X)] \\ & + \mathbb{E}_{(X,Y)\sim\mathcal{P}_{\mathcal{S}}}[\frac{\partial l}{\partial f}\frac{\partial f}{\partial Y}] \cdot \mathbb{E}_{\delta_Y(Y)}[\delta_Y(Y)]. \end{aligned} \tag{18}$$

As for model weight perturbation $\Delta\theta$, according to first-order Taylor expansion, we have

$$l(f_{\theta+\Delta\theta}(X), Y)] = l(f_\theta(X), Y)] + \frac{\partial l}{\partial f}\frac{\partial f}{\partial \theta}\Big|_{\theta=\hat{\theta}} \Delta\theta, \tag{19}$$

where $\hat{\theta}$ is between θ and $\theta + \Delta\theta$. Subsequently, we can approximate the training loss on the source dataset as follows:

$$\begin{aligned} & \mathbb{E}_{(X,Y)\sim\mathcal{P}_{\mathcal{S}}}[l(f_{\theta+\Delta\theta}(X), Y)] \\ = \quad & \mathcal{R}_{\mathcal{S}} + \mathbb{E}_{(X,Y)\sim\mathcal{P}_{\mathcal{S}}}[\frac{\partial l}{\partial f}\frac{\partial f}{\partial \theta}]\Big|_{\theta=\hat{\theta}} \Delta\theta, \end{aligned} \tag{20}$$

Compared Eqs. (18) and (20), for any data perturbation, the model weight perturbation is treated as multivariate but with only one equation. In other words, considering linear equation $\mathbf{A}\mathbf{x} = b$, where $\mathbf{A} \in \mathbb{R}^{1\times n}$ and $b \in \mathbb{R}^{1\times 1}$ are both constant, $\mathbf{x} \in \mathbb{R}^{n\times 1}$ is variables, the goal is to find whether the solution $\mathbf{x}$ exists for linear equation $\mathbf{A}\mathbf{x} = b$. Note that we consider distribution shift problem, i.e., $b \neq 0$, the solution for linear equation $\mathbf{A}\mathbf{x} = \mathbf{b}$ exists when $rank([A|b]) = 1 = rank(A)$, i.e., $\|A\| = \left\| E_{(X,Y)\sim\mathcal{P}_{\mathcal{S}}}[\frac{\partial l}{\partial f}\frac{\partial f}{\partial \theta}]\big|_{\theta=\hat{\theta}} \right\| > 0$. Note that such a non-zero gradient condition is easily satisfied for models that are not well-trained. Therefore, we can always find a model weight perturbation so that the training loss on source dataset with data perturbation δ and model weight perturbation $\Delta\theta$ are the same.

D More details on Eq. (5)

Lemma D.1. *For any scale a_1, a_2, b_1, and b_2, we have* $\big||a_1 - b_1| - |a_2 - b_2|\big| \leq |a_1 - a_2| + |b_1 - b_2|$.

For any scale a_1, a_2, b_1, and b_2, it is easy to check that

$$\begin{aligned} & \big||a_1 - b_1| - |a_2 - b_2|\big| \leq |a_1 - a_2| + |b_1 - b_2| \\ \Longleftrightarrow \quad & -2a_1b_1 - 2a_2b_2 - 2|a_1 - b_1||a_2 - b_2| \\ & \leq -2a_1a_2 - 2b_1b_2 + 2|a_1 - a_2||b_1 - b_2| \\ \Longleftrightarrow \quad & |a_1 - a_2||b_1 - b_2| + |a_1 - b_1||a_2 - b_2| \\ & + a_1b_1 + a_2b_2 - a_1a_2 - b_1b_2 \geq 0, \end{aligned} \tag{21}$$

Notice that

$$\begin{aligned} & -(a_1 - a_2)(b_1 - b_2) + (a_1 - b_1)(a_2 - b_2) \\ & + a_1b_1 + a_2b_2 - a_1a_2 - b_1b_2 = 0, \end{aligned} \tag{22}$$

Eq. (21) holds, and the proof is completed.

E More details on Training Acceleration

Considering that the proposed RFR is computation-expensive due to the inherent maximization problem, it is intractable to adopt RFR during model training. To this end, we develop an efficient and effective approximation to model weight perturbation for the worst case and the gradient of RFR. In this way, the optimizer (e.g., stochastic gradient descent) can be directly adopted for training. Specifically, we first approximate the maximization problem via a first-order Taylor expansion. For $\mathcal{L}_{RFR,\mathcal{S}_0}$, the optimal model weight perturbation is given by

$$\begin{aligned} \epsilon_0^*(\theta) \quad &= \quad \arg\max_{\|\epsilon_0\|_p \leq \rho} \mathbb{E}_{\mathcal{S}_0}[f_{\theta+\epsilon_0}(\mathbf{x})] - \mathbb{E}_{\mathcal{S}_0}[f_\theta(\mathbf{x})] \\ &\approx \quad \arg\max_{\|\epsilon_0\|_p \leq \rho} \frac{\partial \mathbb{E}_{\mathcal{S}_0}[f_\theta(\mathbf{x})]}{\partial \theta}\epsilon_0 \triangleq \arg\max_{\|\epsilon_0\|_p \leq \rho} g_0\epsilon_0, \end{aligned}$$

where $g_0 = \frac{\partial \mathbb{E}_{\mathcal{S}_0}[f_\theta(\mathbf{x})]}{\partial \theta}$ represents the gradient of average prediction for source data with sensitive attribute $A = 0$. Note that the number of model parameters is usually high, the higher order terms computation is time-consuming. For example, the second term involves Hessian matrix with square polynomial complexity. Therefore, we omit high-order terms and only keep one order term to accelerate training. In turn, the optimal model weight perturbation is given by the solution of a classical dual norm problem, i.e.,

$$\epsilon_0^*(\theta) = \rho \cdot \text{sign}(g_0)\frac{|g_0|^{q-1}}{\left(\|g_0\|_q^q\right)^{1/p}}, \tag{23}$$

where $\frac{1}{p}+\frac{1}{q}=1$, sign$(\cdot)$ is element-wise sign function, and $|\cdot|^{q-1}$ denotes element-wise absolute value and power. Considering gradient-based optimizer for model training, the gradient for $\mathcal{L}_{RFR,\mathcal{S}_0}$ is given by

$$\begin{aligned}\nabla_\theta \mathcal{L}_{RFR,\mathcal{S}_0} &\approx \frac{\partial \mathbb{E}_{\mathcal{S}_0}[f_{\theta+\epsilon_0^*(\theta)}(\mathbf{x})]}{\partial \theta} \\ &= \frac{\partial \mathbb{E}_{\mathcal{S}_0}[f_\theta(\mathbf{x})]}{\partial \theta} + \frac{\partial \epsilon_0^*(\theta)}{\partial \theta}\frac{\partial \mathbb{E}_{\mathcal{S}_0}[f_\theta(\mathbf{x})]}{\partial \theta}.\end{aligned} \tag{24}$$

It is seen that the approximation of $\nabla_\theta \mathcal{L}_{RFR,\mathcal{S}_0}$ can be directly calculated via automatic differentiation. However, the calculation of the term $\frac{\partial \epsilon_0^*(\theta)}{\partial \theta}$ implicitly depends on the Hessian of $\mathbb{E}_{\mathcal{S}_0}[f_\theta(\mathbf{x})]$ due to $\epsilon_0^*(\theta)$ is a function of g_0. To further accelerate the computation, we drop the second-order term and the final approximation of $\nabla_\theta \mathcal{L}_{RFR,\mathcal{S}_0}$ is given by

$$\nabla_\theta \mathcal{L}_{RFR,\mathcal{S}_0} \approx \frac{\partial \mathbb{E}_{\mathcal{S}_0}[f_\theta(\mathbf{x})]}{\partial \theta}|_{\theta+\epsilon_0^*(\theta)}, \tag{25}$$

where $\epsilon_0^*(\theta)$ is given by Eq. (23). Similarly, we can obtain the approximation of $\nabla_\theta \mathcal{L}_{RFR,\mathcal{S}_1}$ as follows:

$$\nabla_\theta \mathcal{L}_{RFR,\mathcal{S}_1} \approx \frac{\partial \mathbb{E}_{\mathcal{S}_1}[f_\theta(\mathbf{x})]}{\partial \theta}|_{\theta+\epsilon_1^*(\theta)}, \tag{26}$$

where $g_1 = \frac{\partial \mathbb{E}_{\mathcal{S}_1}[f_\theta(\mathbf{x})]}{\partial \theta}$ and model weight perturbation for group $A=1$ is given by

$$\epsilon_1^*(\theta) = \rho \cdot \text{sign}(g_1)\frac{|g_1|^{q-1}}{\left(\|g_1\|_q^q\right)^{1/p}}. \tag{27}$$

F More Experimental Results

We provide more experimental results to further support the effectiveness of our proposed RFR.

F.1 More Experimental Results on Training Datasets

In this experiment, we evaluate the fairness-accuracy performance, as shown in Figure 5, for source and target datasets on two real-world datasets with temporal and spatial distribution shifts. We observe that RFR achieves a better fairness-accuracy tradeoff on the target dataset compared to the baseline methods for temporal and spatial distribution shifts, while RFR does not always perform best on the source dataset in terms of fairness-accuracy tradeoff.

F.2 More Experimental Results on EO

We also report the acc-EO tradeoff performance for two real-world datasets compared with many baselines in Figure 6. It is seen that similar to acc-DP tradeoff performance, the tradeoff performance of RFR is the best among all baselines. Please notice that there is no revision for REG and ADV methods, i.e., REG and ADV are designed for DP.

F.3 Experimental Results without Distribution Shifts

To further investigate the effect of distribution shifts, we also provide the experimental results without distribution shifts. Specifically, we avoid the data partition and randomly split the data samples across multiple years and multiple states in two real-world datasets. The fairness and accuracy and the corresponding results are shown in Table 3 and Figure 5. We have the following observations:

- Table 3 demonstrate that REG and ADV serve as strong baselines for this setting and our method can only achieve the best results in ACS-I dataset. This suggests that the applicable scope of our method is limited. For example, for the setting without distribution shifts, the conventional methods, such as REG and ADV, may perform better. Such observation further validates the effectiveness of our proposed RFR in tackling the distribution shifts problem.

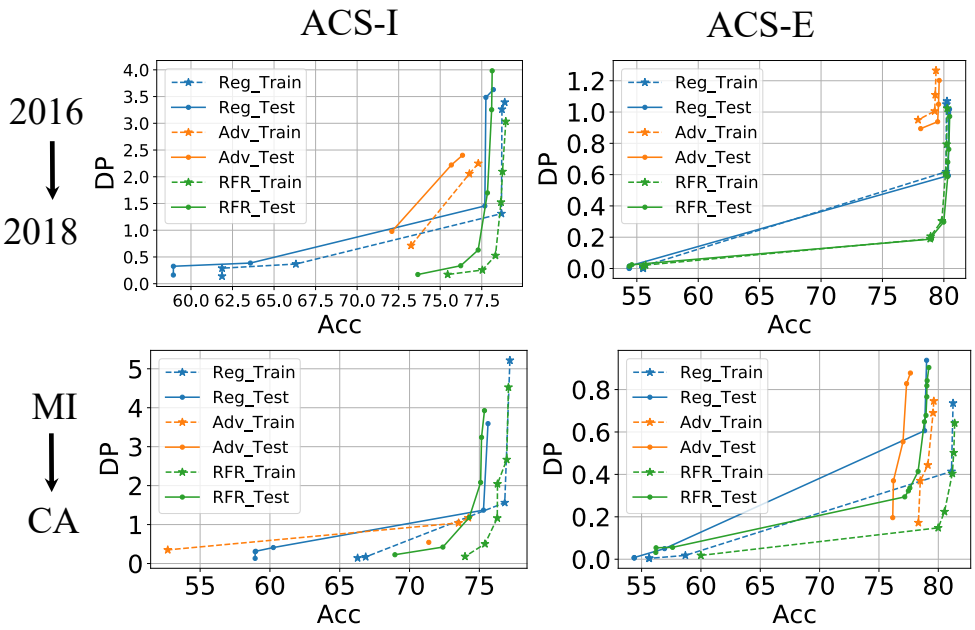

Figure 5: DP and Acc trade-off performance on three real-world datasets with temporal (Top) and spatial (Bottom) distribution shifts for source (train) and target (test) datasets. The trade-off curve close to the right bottom corner means better trade-off performance. The units for x- and y-axis are percentages (%).

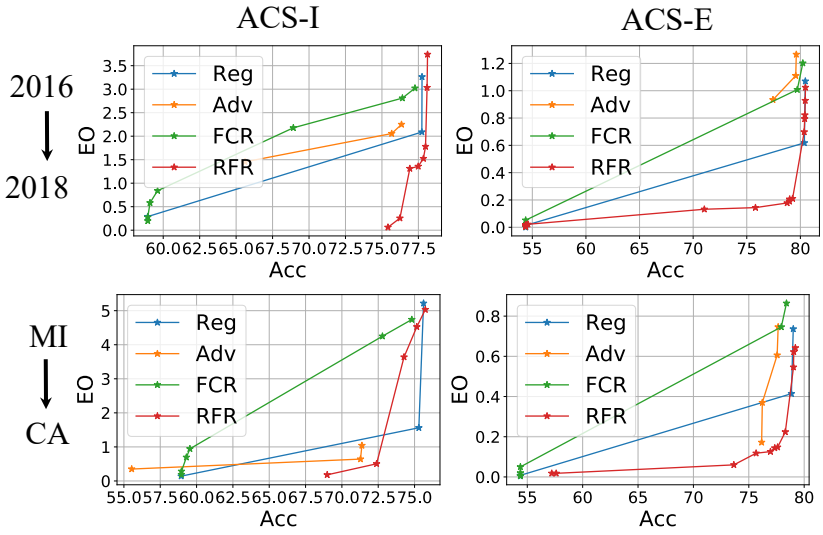

Figure 6: EO and Acc trade-off performance on two real-world datasets without distribution shift.

- Figure 5 shows that REG and ADV are comparable with our proposed RFR in terms of accuracy and fairness tradeoff performance. This observation implies the importance of distribution shift intensity identification, which will serve as important prior knowledge for algorithm or model selection.

F.4 Hyperparameter study on ρ

We conduct the hyperparameter study on ρ to validate the effectiveness of our proposed RFR considering the worst of weight perturbation, as shown in Figure 8. We have the following observations:

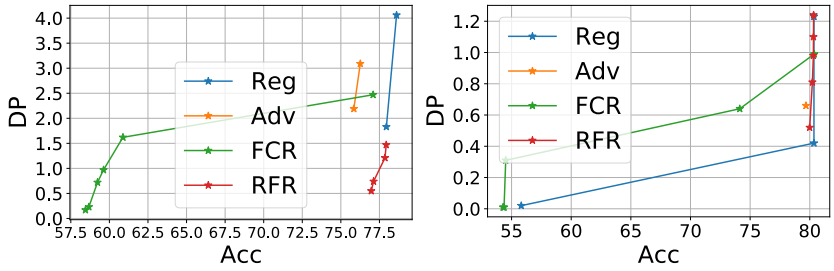

Figure 7: DP and Acc trade-off performance on two real-world datasets without distribution shift. The trade-off curve close to the right bottom corner means better trade-off performance.

Table 3: Performance comparison with baselines without distribution shift. The best and second-best results are highlighted with **hold** and underline, respectively.

Methods	ACS-I			ACS-E		
	Acc (%) ↑	Δ_{DP} (%) ↓	Δ_{EO} (%) ↓	Acc (%) ↑	Δ_{DP} (%) ↓	Δ_{EO} (%) ↓
MLP	78.61±0.42	3.82±0.21	4.06±0.43	80.29±0.08	1.23±0.18	0.89±0.21
REG	77.82±0.51	2.66±0.28	2.61±0.34	80.35±0.17	**1.05**±0.13	1.23±0.16
ADV	75.85±0.62	2.13±0.59	1.90±0.45	79.48±0.24	1.15±0.18	**0.59**±0.12
FCR	75.88±0.81	2.60±0.31	2.94±0.33	79.90±0.15	1.07±0.07	1.18±0.06
RFR	77.57±0.51	**1.88**±0.36	**1.40**±0.36	80.12±0.13	1.09±0.07	0.70±0.07

- The accuracy and DP metrics are both sensitive to hyperparameter ρ, where ACS-I dataset is even more sensitive to ACS-E. In other words, the optimal hyperparameter ρ to achieve robust fairness while preserving accuracy is dependent on the dataset and distribution shifts intensity, which implies that there are opportunities to further improve the tradeoff performance by hyperparameter selection.
- It is not necessarily that a large perturbation ball leads to a better fairness performance. The reasons are two-fold. Firstly, we only consider the worst case of weight perturbation, which is not consistent with real distribution shifts. Secondly, the optimal (worst case) perturbation vector is hard to find in practice. In the implementation of RFR, we use the Taler expansion to approximately find the weight perturbation for training acceleration in Appendix E.

F.5 Ablation study on $\mathcal{L}_{DP}$ for different hyperparameter λ

We conduct the aablation study on $\mathcal{L}_{DP}$ to validate the effectiveness of our proposed RFR as shown in Figure 9. We have the following observations:

- $\mathcal{L}_{DP}$ loss term is very important to achieve robust fairness under distribution shift. Without $\mathcal{L}_{DP}$ loss, the DP cannot be mitigated significantly with a large hyperparameter. Such observation is consistent with Eq. (5).
- ACS-I dataset is more sensitive to hyperparameter λ compared with ACS-E. This suggests the necessity of tuning the hyperparameter carefully for each dataset.

G RFR Algorithms

The pseudo-code of RFR algorithm is given by Algorithm 1. The two forward and two backward propagations happen in Lines 6 and 7.

H More Discussion

H.1 Discussion on Decision Tree Extention

Our method is designed for neural networks and can not be used for decision tree methods (e.g., XGBoost, GBDT). The key reason is that our algorithm (See Appendix G) involves gradient computation

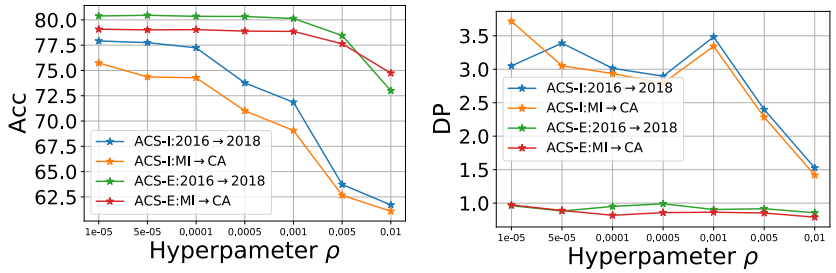

Figure 8: Hyperparameter study on ρ.

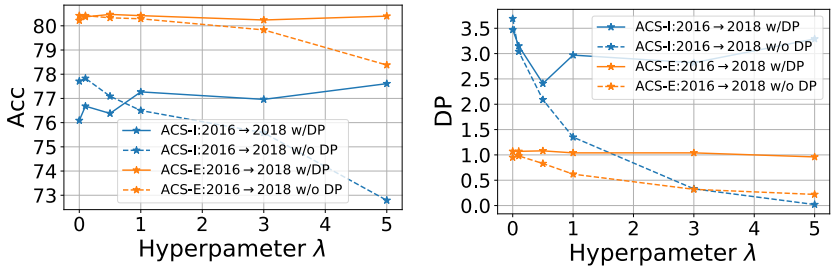

Figure 9: Ablation study on $\mathcal{L}_{DP}$ for different hyperparameter λ.

and weight perturbation to accelerate the bi-level optimization problem-solving. We summarize the details as follows:

- Nature of Parameters: In a neural network, the parameters are continuous values (weights, biases) that are updated using gradients to minimize the loss. In GBDT or XGBoost, the "parameters" are the structure of the trees themselves, including the split points and leaf values. These are not continuous values that can be updated with gradient descent in the traditional sense.
- Training Mechanism: While GBDT and XGBoost use the concept of gradients (specifically, gradient boosting works by fitting new trees to the negative gradient of the loss), this is not the same as computing gradients with respect to weights in a neural network and updating them. Instead, trees are added to the model to correct the errors (residuals) of the current ensemble.
- Non-Differentiability: Decision trees involve making decisions based on hard thresholds, which are inherently non-differentiable operations. This makes them unsuitable for traditional gradient-based optimization.

Therefore, we leave the extension of this work for decision tree-based models in future work.

H.2 Discussion Related to Existing Work

The approach of using the worst-case bound as a regularizer has been explored before for selection bias [61]. The difference is two-fold: (1) The problem setting. Work [61] considers the fairness problem with selection bias, where the selection bias is described with available auxiliary information. Our paper mainly focuses on the fairness problem under distribution shift without any information on target distribution, which can be but is not necessarily caused by selection bias. (2) Methodology. Work [61] mainly focuses on Consistent Range Approximation (CRA) of a fairness query using probability information in data collection. Our paper mainly focuses on DP metric difference in source and target distribution, and then further derives a model perturbation approach to achieve fairness under distribution.

Additionally, Sharpness-Aware Minimization (SAM) [12, 62, 63] aims to encourage the training to converge to a flatter region in which the training losses in the neighborhood around the minimizer are lower. In this paper, we mainly focus on the fairness performance under distribution shift while SAM

Algorithm 1 Robust Fairness Regularization (RFR)

Input: Training (source) dataset $\mathcal{S}=\cup_{i=1}^{N}\{(x_i, y_i, a_i)\}$, hyperparameters λ, ρ, p.
Output: Robust fair model $f_\theta(\cdot)$.
Initialize model weight θ_0, step size η, update step $t=0$.
while not convergence **do**
    Compute gradient for $\nabla_\theta \mathcal{L}_{CLF} + \lambda \nabla_\theta \mathcal{L}_{DP}$.
    Calculate model weight perturbation ϵ_0^* and ϵ_1^* based on Eqs. (23) and (27).
    Approximate gradient of RFR $\nabla_\theta \mathcal{L}_{RFR}$ based on Eqs (24) and (26).
    Update model weights: $\theta_{t+1} = \theta_t - \eta\big(\nabla_\theta \mathcal{L}_{CLF} + \lambda \nabla_\theta \mathcal{L}_{DP}\big) + \lambda \nabla_\theta \mathcal{L}_{RFR}$.
    $t = t + 1$.
end while

improves the generalization performance. Techniquely, the objectives and the considered sample batch of RFR are different from that of SAM.

H.3 Future Work

There are several future directions: (1) The proposed method is specifically developed for demographic parity. As for the robust fairness for other metrics, such as counterfactual fairness, individual fairness, and other group fairness, we leave these extension works in future work. (2) We mainly focus on model weight perturbation to achieve robust fairness. Another possible approach is the input perturbation method, which is complicated since the input perturbation is dependent on input samples. Additionally, input perturbation is highly related to feature type (numerical, categorial, or mixed). (3) Note that it is intractable to select the optimal without accessing target data distribution. One promising future direction is to access more information (e.g., few-shot target samples or input features in the target dataset), the hyperparameter tuning can be done and achieve better performance . For example, if the input features in (unlabeled) target dataset are available, it is tractable to predict out-of-distribution error and then use it for hyperparameter selection [64, 65]. Note that the input feature in the target dataset should be available in the inference stage.

I Broader Impact

Algorithmic Fairness focuses on ensuring fairness and lack of bias in the decisions made by algorithms or machine learning models. Fairness in socio-technical systems goes beyond algorithmic fairness, considering broader societal impacts, ethical considerations, data biases, and interdisciplinary collaboration. It focuses on human-centered design, policy changes, and ongoing assessment to ensure technology aligns with societal values and promotes equitable outcomes.