Long-Term Fair Decision Making through Deep Generative Models

Yaowei Hu[1], Yongkai Wu[2], Lu Zhang[1]

[1] University of Arkansas
[2] Clemson University
yaoweihu@uark.edu, yongkaw@clemson.edu, lz006@uark.edu

Abstract

This paper studies long-term fair machine learning which aims to mitigate group disparity over the long term in sequential decision-making systems. To define long-term fairness, we leverage the temporal causal graph and use the 1-Wasserstein distance between the interventional distributions of different demographic groups at a sufficiently large time step as the quantitative metric. Then, we propose a three-phase learning framework where the decision model is trained on high-fidelity data generated by a deep generative model. We formulate the optimization problem as a performative risk minimization and adopt the repeated gradient descent algorithm for learning. The empirical evaluation shows the efficacy of the proposed method using both synthetic and semi-synthetic datasets.

Introduction

Machine learning models are extensively utilized for significant decision-making scenarios, like college admissions (Baker and Hawn 2021), banking loans (Lee and Floridi 2021), job placements (Schumann et al. 2020), and evaluations of recidivism risks (Berk et al. 2021). Fair machine learning, which aims to reduce discrimination and bias in machine-automated decisions, is one of the keys to enabling broad societal acceptance of large-scale deployments of decision-making models. In the past decade, various notions, metrics, and techniques have emerged to address fairness in machine learning. For an overview of fair machine learning studies, readers are directed to recent surveys (Tang, Zhang, and Zhang 2022; Alves et al. 2023).

However, our society is marked by pervasive group disparities. For example, in the context of bank loans, disparities in creditworthiness may be observed among different racial groups. These disparities can arise from systemic or historic factors that influence variables like credit score, employment history, and income. As another example, disparities may be observed in the qualification of students for college admission which could be influenced by factors such as the availability of higher education institutions in rural areas, the economic capacity of families, exposure to college preparatory resources, etc. Currently, the majority of studies in fair machine learning are focused on the problem of

building decision models for fair one-shot decision-making (Mehrabi et al. 2021; Caton and Haas 2020). However, it has been shown that algorithms based on traditional fairness notions, such as demographic parity (DP) and equal opportunity (EO), cannot efficiently mitigate group disparities and could even exacerbate the gap (Liu et al. 2018; Zhang et al. 2020). The challenges lie in the sequential nature of real-world decision-making systems, where each decision may shift the underlying distribution of features or user behavior, which in turn affects the subsequent decisions. For example, a bank loan decision can have an impact on an individual's credit score, income, etc., which may influence his/her future loan application. As a result, long-term fairness has been proposed to focus on the mitigation of group disparities rather than making fair decisions in a single time step, which is more challenging to achieve than traditional fairness notions due to the feedback loops between the decisions and features as well as the data distribution shift during the sequential decision-making process. Please refer to the technical appendix for a literature review.

In political and social science, affirmative action has been implemented as a practice to pursue the goal of group parity (Crosby, Iyer, and Sincharoen 2006). For example, for promoting group diversity, affirmative action has been used in higher education allowing universities to consider race as a factor in admissions. However, affirmative action is controversial and has been criticized on the grounds that race-conscious decisions might lead to reverse discrimination against other groups or stigmatize all members of the target group as unqualified. (Fischer and Massey 2007). Recently, the US Supreme Court ruled that affirmative action policies are unconstitutional and race can no longer be regarded as a factor in admissions to US universities (har 2023).

Given the context, we ask whether we can mitigate group disparity and achieve long-term fairness while limiting the use of the sensitive attribute in decision-making models. Although policies such as expanded outreach or "top X-percent" plans are worth investigating in practice, in this paper, we explore the data-driven prospect of mitigating group disparity only through learning and deploying appropriate decision-making models, taking into account the system dynamics. In other words, we aim to learn a decision model that, once deployed, can lead to group parity by properly reshaping the distribution. To this end, we lever-

age Pearl's structural causal model (SCM) and causal graph (Pearl 2009) for modeling the system dynamics and defining long-term fairness. Formally, we treat longitudinal features and decisions in the sequential decision-making process as temporal variables. Then, we model the system using temporal causal graphs (Pamfil et al. 2020) which are graphical representations that can depict causal relationships between features and decisions over time (see Figure 1 for an example). The deployment of a decision-making model is simulated as soft interventions on the graph so that the influence of feedback can be inferred as the interventional distribution. Thus, group disparity produced by the decision-making model at any time step T can be considered as the causal effect of the sensitive attribute on the features at time T. Meanwhile, sensitive attribute unconsciousness can be formulated as a local requirement that restricts the direct causal effect of the sensitive attribute on the decision at each time step. As a result, we show that our goal can be formulated as the trade-off between two conflicting causal effect objectives.

To strike a balance for the trade-off, we leverage deep generative models to devise a regularized learning problem. Specifically, given the observed time series within the time range $[1, l]$, we train a deep generative model to generate the interventional distribution for the range $[1, T]$ as well as the observational distribution if $T > l$. Then, we integrate the deep generative model and decision-making model into a collaborative training framework so that the predicted data could be used as reliable data for training the decision model. We propose a three-phase training framework. In Phase 1, given the training time series, we first train a base decision model without considering any fairness requirements. In Phase 2, we train a recurrent conditional generative adversarial network (RCGAN) inspired by (Esteban, Hyland, and Rätsch 2017) to fit the training time series in order to generate high-fidelity time series. Finally, in Phase 3, we train a fair decision model on the generated time series within the time range $[1, T]$ by considering both long-term fairness at time T as well as the local fairness requirement at each time step. The optimization problem is treated as performative risk minimization and solved by using the repeated gradient descent algorithm (Perdomo et al. 2020).

To derive the quantitative long-term fairness metric, we consider the interventional distribution of features at time step T and adopt the 1-Wasserstein distance between the interventional distributions of different demographic groups as the measure of the group disparity. We substantiate our metric by demonstrating that a small distance enables reconciliation to occur between DP and EO for any decision models that are unconscious of sensitive attributes, provided the necessary conditions from (Kim, Chen, and Talwalkar 2020) are met, and hence implying systemic equality and group parity.

Our experiments demonstrate that the proposed framework can effectively balance various fairness requirements and utility, while methods based on traditional fairness notions cannot effectively achieve long-term fairness.

Background

Throughout this paper, single variables and their values are denoted by the uppercase letter X and the lowercase letter x.

Corresponding bold letters $\mathbf{X}$ and $\mathbf{x}$ denote the sets of variables and their values respectively. We utilize Pearl's structural causal model (SCM) and causal graph (Pearl 2009) for defining the long-term fairness metric and designing the architecture of the deep generative model. For a gentle introduction to SCM please refer to (Pearl 2010). In this paper, we assume the *Markovian SCM* such that the exogenous variables are mutually independent.

Causal inference in the SCM is facilitated with the interventions (Pearl 2009). The hard intervention forces some variables to take certain constant. The soft intervention, on the other hand, forces some variables to take certain functional relationship in responding to some other variables (Correa and Bareinboim 2020). Symbolicly, the soft intervention that substitutes equation $X = f_X(\mathbf{Pa}_X, \mathbf{U}_X)$ with a new equation $X = g(\mathbf{Z})$ is denoted as $\sigma_{X=g(\mathbf{Z})}$. The distribution of another variable Y after performing the soft intervention is denoted as $P(Y(\sigma_{X=g(\mathbf{Z})}))$.

Formulating Long-term Fairness

Problem Setting

We start by modeling the system dynamics and formulating a long-term fairness metric that permits continuous optimization. To ease the representation, we assume a binary sensitive attribute for indicating different demographic groups denoted by $S \in \mathcal{S} = \{s^+, s^-\}$, as well as a binary decision denoted by $Y \in \mathcal{Y} = \{y^+, y^-\}$. The profile features other than the sensitive feature are denoted by $\mathbf{X} \in \mathcal{X}$. In a sequential decision-making system, if a feature is time-dependent, it means that its value may change from one time step to another. We assume that $\mathbf{X}$ and Y are time-dependent and use the superscript to denote their variants at different time steps, leading to $\mathbf{X}^t$ and Y^t. Many sensitive attributes like gender and race are naturally time-independent. Some sensitive attributes may change over time, but the relative order of individuals in the data does not change, like age. Thus, we treat S as being time-independent in this paper.

Suppose that we have access to a time series $\mathcal{D} = \{(S, \mathbf{X}^t, Y^t)\}_{t=1}^l$. We assume an SCM for describing the data generation mechanism and leverage a temporal causal graph for describing the causal relation among $S, \mathbf{X}^t, Y^t$ in the SCM. We make the *stationarity assumption* such that data distribution may shift over time but the data generation mechanism behind it does not change. Figure 1 gives an example which shows that at each time step the decision Y^t is made based on the value of $\mathbf{X}^t$ and S. Meanwhile, the value of $\mathbf{X}^t$ is affected by the values of $\mathbf{X}^{t-1}$, Y^{t-1} and S. We will use this graph as a running example throughout the remaining of this paper. In practice, the temporal causal graph can be obtained from the domain knowledge or learned from data using structure learning algorithms (e.g., (Vowels, Camgoz, and Bowden 2021; Runge 2020; Pamfil et al. 2020)). Our goal is to learn a decision model $h_\theta : \mathcal{S} \times \mathcal{X} \mapsto \mathcal{Y}$ such that when deployed at every time step, certain fairness requirements can be achieved.

To illustrate our problem setting in a real-world scenario, consider an example of a bank loan system. When people apply for bank loans, their personal information (e.g., race,
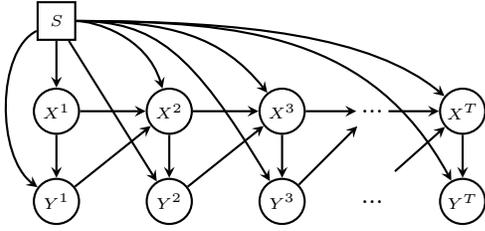
Figure 1: A temporal causal graph for sequential decision making.

job, assets, credit score, etc.) is used by the bank's decision model to decide whether to grant the loans. Except for race, which is a sensitive feature S, other profile features $\mathbf{X}^t$ represent an applicant's qualification at time step t. The bank's decision Y^t can have impacts on the applicants' profile features in $\mathbf{X}^{t+1}$ such as the credit score, which in turn affect the outcomes of their subsequent loans.

Long-term Fairness Metric

To formulate long-term fairness under the decision model deployment, (Hu and Zhang 2022) has proposed to mathematically simulate the model deployment by soft interventions on Y at all time steps. That is, given a decision model h_θ, it replaces the original structural equation of Y in the SCM. We denote the soft intervention by $\sigma_{Y^t=h_\theta(S,\mathbf{X}^t)}$ and abbreviate it as σ_θ. Then, the influence of the model deployment on feature $\mathbf{X}^t$ can be described by its interventional distribution, denoted by $P(\mathbf{X}^t(\sigma_\theta))$. With this formulation, we treat group disparity at time step T as the causal effect of S on $\mathbf{X}^T$, i.e., group parity is achieved when there is no such causal effect and the interventional distribution of $\mathbf{X}^T$ will be equal across demographic groups $\{s^+, s^-\}$.

It is worth noting that, the causal effect of S on $\mathbf{X}^T$ is transmitted through multiple causal paths. These paths can be categorized into two sets: those that intersect with decision nodes on the graph (e.g., $S \to \mathbf{X}^1 \to Y^1 \to \mathbf{X}^2 \to \cdots$), and those that bypass these nodes (e.g., $S \to \mathbf{X}^1 \to \mathbf{X}^2 \to \cdots$). The causal effect transmitted through the former set is influenced by updates to the decision model, whereas the effect via the latter remains unchanged. Hence, eliminating the causal effect of S on $\mathbf{X}^T$ via updating the decision model means learning the decision model h_θ such that the causal effects transmitted through the two sets of paths are cancelled out. Consequently, we can formulate the problem of mitigating group disparity as a learning problem.

We note that the causal effect transmitted through h_θ will be limited by the requirement of sensitive attribute unconsciousness that restricts the direct causal effect of S on h_θ at each time step. The formulation of the metric of sensitive attribute unconsciousness will be detailed in the next section. As a result, long-term fairness may not always be achievable only through updating the decision model. In this paper, we aim to develop and conduct the best practices for mitigating group disparity using learning algorithms. The theoretical analysis of the feasibility of long-term fairness will be a valuable direction for future research.

Next, we present the quantitative metric of long-term fairness. We use the 1-Wasserstein distance to measure the difference between the two interventional distributions, defined as follows.

Definition 1. *Given a sequential decision making system, a decision model $h_\theta : \mathcal{S} \times \mathcal{X} \mapsto \mathcal{Y}$, and a time step T, the metric for measuring the long-term fairness produced by deploying h_θ is given by*

$$J_1^T(\theta) \triangleq W(P(\mathbf{X}^T(\sigma_\theta)|S=s^+), P(\mathbf{X}^T(\sigma_\theta)|S=s^-)), \tag{1}$$

where W is the 1-Wasserstein distance and σ_θ is the soft intervention.

We substantiate our metric and the choice of the 1-Wasserstein distance with the following proposition.

Proposition 1. *Let d be the 1-Wasserstein distance given in Definition 1. For any sensitive attribute-unconscious decision model $f : \mathcal{X} \mapsto \mathcal{A}$ that is Lipschitz continuous, its DP is bounded by $l_f \cdot d$ where l_f is the Lipschitz constant of f. If we assume that the true label Y is given by a decision model $g : \mathcal{X} \mapsto \mathcal{A}$ that is Lipschitz continuous and satisfies the equal base rate condition, then the EO of f is bounded by $(l_f + l_g)/P(y) \cdot d$ where l_g is the Lipschitz constant of g.*

Please refer to the technical appendix for the proof. Proposition 1 implies that when the 1-Wasserstein distance is minimized, both DP and EO are mitigated at time T for any sensitive attribute-unconscious decision model that is Lipschitz continuous. As shown in (Kim, Chen, and Talwalkar 2020), the equal base rate condition is the necessary condition for DP and EO to be compatible. Proposition 1 means that minimizing the 1-Wasserstein distance transforms the necessary condition into both a necessary and sufficient condition. This demonstrates a reconciliation between the previously incompatible fairness notions of DP and EO, thereby suggesting group parity and system equity.

In the experiments, to reduce computational complexity, we use the Sinkhorn distance (Cuturi 2013) to approximate the 1-Wasserstein distance. Note that although Proposition 1 applies for sensitive attribute-unconscious decision models, we do allow the decision model h_θ to take the sensitive feature as input. Meanwhile, we impose local fairness constraints to restrict the direct causal effect of S on the decision, as shall be shown in the next section.

Learning Framework for Long-term Fairness

In this section, we introduce the learning framework for achieving long-term fairness.

Problem Formulation

We start by formulating the objective function for the decision model h_θ. In addition to long-term fairness outlined in Definition 1, additional factors must also be taken into account during optimization. The first factor is the utility of the decision model to ensure good prediction performance. We adopt the traditional definitions of the loss function given as follows. In practice, loss functions such as cross-entropy loss can be used to penalize inaccurate predictions.

Definition 2. *Given a time series* $\mathcal{D} = \{(S, \mathbf{X}^t, Y^t)\}_{t=1}^l$, *the loss for the decision model* h_θ *is given by*

$$J_2(\theta) \triangleq \frac{1}{l} \sum_{t=1}^l \mathbb{E}[\mathcal{L}(h_\theta(S, \mathbf{X}^t), Y^t)], \qquad (2)$$

where $\mathcal{L}$ *is any loss function.*

The second factor is the local fairness constraint for ensuring sensitive attribute-unconsciousness as mentioned above. In this paper, we adopt direct discrimination proposed in (Zhang, Wu, and Wu 2016) as the metric for formulating the local fairness constraint, as defined below.

Definition 3. *The local fairness constraint for each time step* $t \in [1, T]$ *is given by direct discrimination at* t, *i.e.,*

$$\begin{aligned} J_3^t(\theta) &\triangleq |\mathbb{E}[h_\theta(S = s^+, \mathbf{X}^t(\sigma_\theta))|S = s^-] - \\ &\quad \mathbb{E}[h_\theta(S = s^-, \mathbf{X}^t(\sigma_\theta))|S = s^-]|. \end{aligned} \qquad (3)$$

We note the trade-off between the three factors. The local fairness and utility trade-off has been studied in previous works (Bakker et al. 2019; Corbett-Davies et al. 2017). What we are more interested is the trade-off between long-term and local fairness. As discussed, mitigating group disparity requires a non-zero causal effect to be transmitted through decision nodes so that the causal effects transmitted through the two sets of paths can be canceled out. However, sensitive attribute unconsciousness requires eliminating the direct causal effect to be transmitted through decision nodes. As a concrete example, making loan decisions exactly according to the credit score is reasonable in terms of local fairness, but might not help in narrowing the gap in credit scores between advantaged and disadvantaged groups. On the other hand, reducing the gap by favoring the disadvantaged group can raise the potential issue of reverse discrimination, which may violate the local fairness constraint.

In this paper, rather than delving into a theoretical analysis of the trade-off between long-term and local fairness, we utilize the learning algorithm to achieve a balance between them. We formulate a regularized learning problem, leading to the problem formulation below.

Problem Formulation 1. *The problem of long-term fair decision-making is to solve the optimization problem:*

$$\min_\theta \mathcal{L}(\theta) = \lambda_1 J_1^T(\theta) + \lambda_2 J_2(\theta) + \frac{\lambda_3}{T} \sum_{t=1}^T J_3^t(\theta), \qquad (4)$$

where λ_1, λ_2 *and* λ_3 *are weight parameters.*

Overview of Learning Framework

There are two main challenges in solving Problem Formulation 1. First, as shown in Eqs. (1) and (3), the computation of $J_1^T(\theta)$ and $J_3^t(\theta)$ are based on the interventional variants of features $\mathbf{X}^t(\sigma_\theta)$ whose values in turn depend on the model parameter θ. Second, if $T > l$, then $J_1^T(\theta)$ and $\sum_{t=1}^T J_3^t(\theta)$ will be computed on time steps that are beyond the range of the training data. Thus, Problem Formulation 1 cannot be solved by traditional machine learning algorithms.
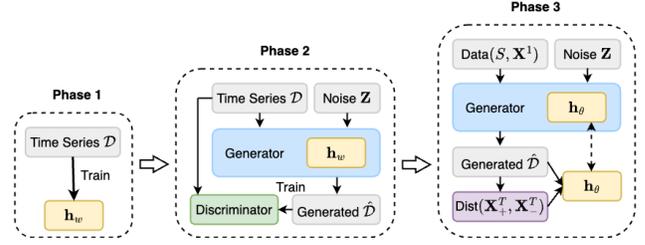

Figure 2: The overview of the proposed framework. Solid arrows represent input, and the dashed arrow represents parameter sharing. For Phase 3 only one generator is shown.

We propose a novel three-phase framework based on causal inference techniques and deep generative networks. The core idea is to use a deep generative network to simulate an SCM for generating both observational and interventional distributions. Our method rests on two theoretical foundations: 1) (Kocaoglu et al. 2018) shows that if the structure of a generative network is arranged to reflect the causal structure, then it can be trained with the observational data such that it will agree with the same SCM in terms of any identifiable interventional distributions; 2) (Kocaoglu et al. 2019) shows that the interventional distribution produced by any soft intervention is identifiable in a Markovian SCM.

Inspired by these prior results, we design the architecture of the deep generative network following the causal structure. We illustrate our framework using the example shown in Figure 1. In this example, the causal structure at each time step can be mathematically described by two structural equations of the SCM:

$$Y^t = f_Y\left(S, \mathbf{X}^t, U_Y\right), \qquad (5)$$

$$\mathbf{X}^t = f_\mathbf{X}\left(S, \mathbf{X}^{t-1}, Y^{t-1}, U_\mathbf{X}\right), \qquad (6)$$

where $U_Y, U_\mathbf{X}$ are exogenous variables. Note that due to the stationarity assumption, f_Y and $f_\mathbf{X}$ are time-independent. According to the principle of independent mechanisms (Peters, Janzing, and Schölkopf 2017), these two structural equations can be learned independently without affecting each other. Motivated by this principle, we propose a three-phase framework. In Phase 1, we train a classifier h_ω on the training time series $\mathcal{D}$ to approximate f_Y and make decisions for each time step. In Phase 2, we train a recurrent conditional generative adversarial network (RCGAN) (Esteban, Hyland, and Rätsch 2017) on the same training time series $\mathcal{D}$ using adversarial training. The generator of the RCGAN uses h_ω obtained in Phase 1 for generating decisions. Finally, in Phase 3, we replace h_ω with the decision model h_θ and train it on the data generated from the RCGAN using the objective function Eq. (4). The overview of the framework is shown in Figure 2 and the pseudo-code is shown in Algorithm 1. Next, we describe the details of each phase.

Phase 1: Train a Decision Classifier

The objective of this phase is to learn a classifier h_ω from the training time series to approximate the mechanism f_Y in Eq. (5) for making decisions, i.e., $\hat{Y}^t = h_\omega(S, \mathbf{X}^t)$. In this

---

Algorithm 1: DeepLF

Input : Dataset $\mathcal{D} = \{(S, \mathbf{X}^t, Y^t)\}_{t=1}^l$, time-lagged causal graph $\mathcal{G}$, parameters λ_1, λ_2 and λ_3
Output: The fair model h_θ

1 Train a classifier h_ω by minimizing Eq. (8) on $\mathcal{D}$;
2 **repeat**
3 $\quad$ Update the discriminator D_ϕ according to Eq. (13);
4 $\quad$ Update the generator $G_{\psi,\omega}$ with the classifier h_ω as one of its components according to Eq. (14);
5 **until** convergence;
6 $i \leftarrow 0$;
7 Initialize h_{θ_0} according to h_ω;
8 **repeat**
9 $\quad$ Generate time series using generator $G_{\psi,\omega}$;
10 $\quad$ Compute $\nabla_\theta \mathcal{L}_l(\theta)$ according to Eq. (5) using the generated data;
11 $\quad$ $\theta_{i+1} \leftarrow \theta_i - \eta_i \nabla_\theta \mathcal{L}_l(\theta_i), i \leftarrow i + 1$;
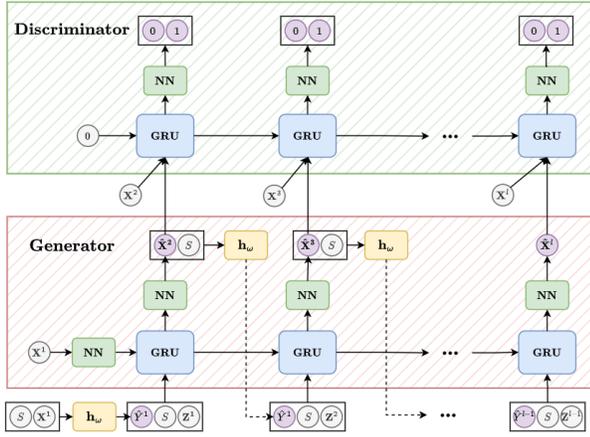12 **until** convergence;
13 **return** h_{θ_i};

Figure 3: The architecture of the RCGAN.

phase, we train the classifier by maximizing the accuracy without considering any fairness requirement. Since f_Y does not change with time, we aggregate the losses at all time steps in the loss function. Specifically, if the cross-entropy loss is used, then the loss function for training h_ω on time series $\mathcal{D}$ is given by

$$\min_\omega \mathcal{L}_c(\omega) = -\frac{1}{l} \sum_{t=1}^l \mathbb{E}\left[Y^t \log h_\omega(S, X^t)\right]. \quad (7)$$

Phase 2: Train an RCGAN

We then train an RCGAN to simulate the system dynamics and generate values of features $\mathbf{X}$ by taking the predictions of h_ω as the input. Our purpose is to generate both observational and intervention distributions of $\mathbf{X}$, so it is important to design the architecture of the RCGAN following the causal structure.

The architecture of the RCGAN is illustrated in Figure 3, which consists of one generator and one discriminator.

We use the gated recurrent unit (GRU) (Cho et al. 2014) as the core structure of the generator and discriminator. For the generator, it takes the sensitive feature S, a set of noise vectors $\mathbf{Z}$, as well as the features at the first time step $\mathbf{X}^1$ as the input. The hidden state is then initialized by a non-linear transformation (e.g., a multi-layer perceptron) of $\mathbf{X}^1$ as:

$$\mathbf{h}^1 = \text{MLP}\left(\mathbf{X}^1\right). \quad (8)$$

Then, for each time step t, we concatenate the noise vector $\mathbf{Z}^{t-1}$ with the conditional information $\hat{Y}^{t-1}$ and S as the input to each GRU. After the calculation of GRU, we again use a non-linear transformation to convert the hidden states $\mathbf{h}^t$ to predicted $\hat{\mathbf{X}}^{t+1}$. These three steps are shown by the three equations below.

$$\mathbf{I}^{t-1} \leftarrow \left[\hat{Y}^{t-1}, S, \mathbf{Z}^{t-1}\right], \quad (9)$$

$$\mathbf{h}^t = \text{GRU}\left(\mathbf{I}^{t-1}, \mathbf{h}^{t-1}\right), \quad (10)$$

$$\hat{\mathbf{X}}^{t+1} = \text{MLP}\left(\mathbf{h}^t\right). \quad (11)$$

We also use GRUs for the discriminator to distinguish between generated data and real data. For the generated data, the discriminator attempts to predict label 0 for each time step, and vice versa, for the real data, the discriminator attempts to predict label 1 for each time step.

Finally, for training the generator and discriminator, in addition to the objective of the original GAN for minimizing the likelihood of generated data given by the discriminator, the maximum mean discrepancy (MMD) (Gretton et al. 2006) between original data and generated data is also explicitly minimized. The MMD brings two distributions together by comparing their statistics. As a result, the loss functions of the discriminator and the generator are shown below, which are optimized alternatively.

$$\max_\phi \mathcal{L}_d(\phi) = \mathbb{E}_\mathbf{X}[\log(D_\phi(\mathbf{X}))] + \\ \mathbb{E}_\mathbf{Z}[\log(1 - D_\phi(G_{\psi,\omega}(S, \mathbf{Z}, \mathbf{X}^1)))], \quad (12)$$

$$\min_\psi \mathcal{L}_g(\psi) = \mathbb{E}_\mathbf{Z}[\log(1 - D_\phi(G_{\psi,\omega}(S, \mathbf{Z}, \mathbf{X}^1)))] + \\ \gamma \text{MMD}(\mathbf{X}, G_{\psi,\omega}(S, \mathbf{Z}, \mathbf{X}^1)), \quad (13)$$

where D_ϕ represents the discriminator, $G_{\psi,\omega}$ represents the generator that uses h_ω as the classifier, and γ controls the strength of the regularization.

Phase 3: Train a Long-term Fair Decision Model

In the last phase, we train a decision model h_θ on the data generated by the RCGAN using the objective function Eq. (4). We use the generator obtained in Phase 2 as well as a variant of this generator where h_ω is replaced with h_θ. The former is for generating the observational distribution and the latter is for generating interventional distribution. Specifically, we first directly apply the generator $G_{\psi,\omega}$ obtained in Phase 2 to generate data for time steps from 1 to l which are used to compute $J_2(\theta)$ in Eq. (2). Then, we perform soft intervention σ_θ by replacing h_ω with h_θ to obtain a variant generator $G_{\psi,\theta}$ which is used to generate data for

time steps from 1 to T for computing $J_1^T(\theta)$ in Eq. (1) and $J_3^t(\theta)$ in Eq. (3). In other words, we use $G_{\psi,\theta}$ to generate the interventional data $\mathbf{X}^t(\sigma_\theta)$. Finally, the loss $\mathcal{L}(\theta)$ is computed and h_θ is updated accordingly. Note that the RCGAN trained in Phase 2 will not be updated in this phase.

It is important to note that when we use the RCGAN to generate data samples for computing $\mathcal{L}(\theta)$, those data samples are affected by h_θ as well, due to the fact that h_θ is trained on the interventional distribution after performing soft intervention σ_θ. This optimization problem is different from the traditional empirical risk minimization and is called the performative risk minimization (Perdomo et al. 2020). In our work, we adopt the repeated gradient descent algorithm (RGD) (Perdomo et al. 2020) which is an iterative training approach to address this problem. In the training process of Phase 3, we first initialize h_θ according to h_ω. Then, in each iteration, we use the current version of h_θ for generating data and computing the empirical loss, and h_θ is updated based on the empirical gradient $\nabla_\theta \mathcal{L}(\theta)$. After that, we replace h_θ with its updated version and conduct another iteration of training. This process is repeated until the parameters of h_θ converge.

Experiments

In this section, we conduct empirical evaluations of our method[1]. We refer to our method as deep long-term fair decision making (**DeepLF**). We use a multi-layer perceptron for both h_ω and h_θ in our method.

Datasets

Many commonly used datasets in fair machine learning (Le Quy et al. 2022) are not for dynamic fairness research. In (Ding et al. 2021), the authors construct a dataset that spans multiple years and allows researchers to study temporal shifts in the distribution level. However, our study requires the longitudinal data that track each instance over time, other than multiple datasets with temporal distribution shifts. Thus, following (Hu and Zhang 2022), we generate synthetic and semi-synthetic time series datasets as follows.
Synthetic Dataset. We generate the synthetic time series dataset based on the temporal causal graph shown in Figure 1. Each sample at each time step in the time series includes a sensitive feature S, profile features $\mathbf{X}^t$ and a decision Y^t. The samples at the initial time step $\mathbf{X}^1, Y^1$ are generated by calling the data generation function (i.e., make_classification) of scikit-learn package. Then, we cluster the generated samples into two groups and assign S to each sample according to the cluster it belongs to. To generate the data samples in the remaining time steps, we design a procedure by simulating the bank loan system in the real world. Please refer to the technical appendix for the detailed generation rules, following which we generate a 10-step synthetic time series dataset with 10000 instances where $\mathbf{X}^t$ is a 6-dimensional vector. We refer to this dataset SimLoan.
Semi-Synthetic Dataset. We also generate semi-synthetic data by leveraging the real-world Taiwan dataset (Yeh and

Lien 2009) as the initial data at $t = 1$. A ground-truth classifier and similar generation rules of change are used to generate subsequent decisions $Y^1, ..., Y^l$ and profile features $\mathbf{X}^2, ..., \mathbf{X}^l$. There are 10000 instances in the initial data and they are randomly and equally sampled from groups by S and Y for balance. Like the SimLoan dataset, this dataset is also made up of 10 steps. We refer to this dataset Taiwan.

Experimental Setting

Baselines. A multi-layer perceptron with the same number of layers as h_ω and h_θ that is trained on the training time series $\mathcal{D}$ without any fairness constraints is used as the first baseline, denoted as (**MLP**). Two common static fairness constraints, i.e., demographic parity and equal opportunity, are applied to the **MLP** model as fairness constraints respectively, referred to as **MLP-DP** and **MLP-EO**. We also implement the method proposed in (Hu and Zhang 2022) that formulates long-term fairness as path-specific effects and trains the model using repeated risk minimization, referred to as **LRLF**. It requires the true causal structure equations for training, so we provide it with the true data generation rules. Implementing details are included in the appendix.
Evaluation. To evaluate the performance of models after deployment, the RCGAN trained in Phase 2 and the decision models that we evaluate are used together to generate interventional data on which the local and long-term fairness are computed. The long-term fairness is measured by Eq. (1) computed on the evaluated decision model; the local fairness is measured by direct discrimination Eq. (3) at each time step; and the accuracy of predictions is evaluated based on the ground-truth classifier h_ω at each time step.

Results

To evaluate the performance of our algorithm and baselines, we conduct experiments with two settings on both SimLoan and Taiwan datasets for 5 times and calculate their mean and std. In the first setting, the time step T for achieving long-term fairness is set to 10. We train the models on the 10-step training data (i.e., within the time range $[1, 10]$) and evaluate the models on the 10-step generated datasets with $\mathbf{X}^1$ as input (i.e., also within the time range $[1, 10]$). The results of accuracy and unfairness of all algorithms on the two datasets are shown in Figures 4 (a) and (c). As can be seen, both the local and long-term fairness produced by **DeepLF** are comparable with or better than those of **LRLF**, and they markedly outperform other baselines. For **LRLF**, although it also produces relatively small local and long-term unfairness, it requires true causal structure equations for training which may not be available in practice. For the other three baselines, there is no clear decreasing trend in both local and long-term unfairness, although a relatively higher level of accuracy is achieved. The results demonstrate that our method strikes an effective balance between long-term fairness, local fairness, and utility, and it requires only the information from the historical data.

To illustrate the difference in the qualification distribution produced by different methods, we adopt T-SNE to visualize the distribution of $\mathbf{X}^{10}$ produced by **MLP** and **DeepLF**, as

---

[1]The code and hyperparameter settings are available online: https://github.com/yaoweihu/Generative-Models-for-Fairness.
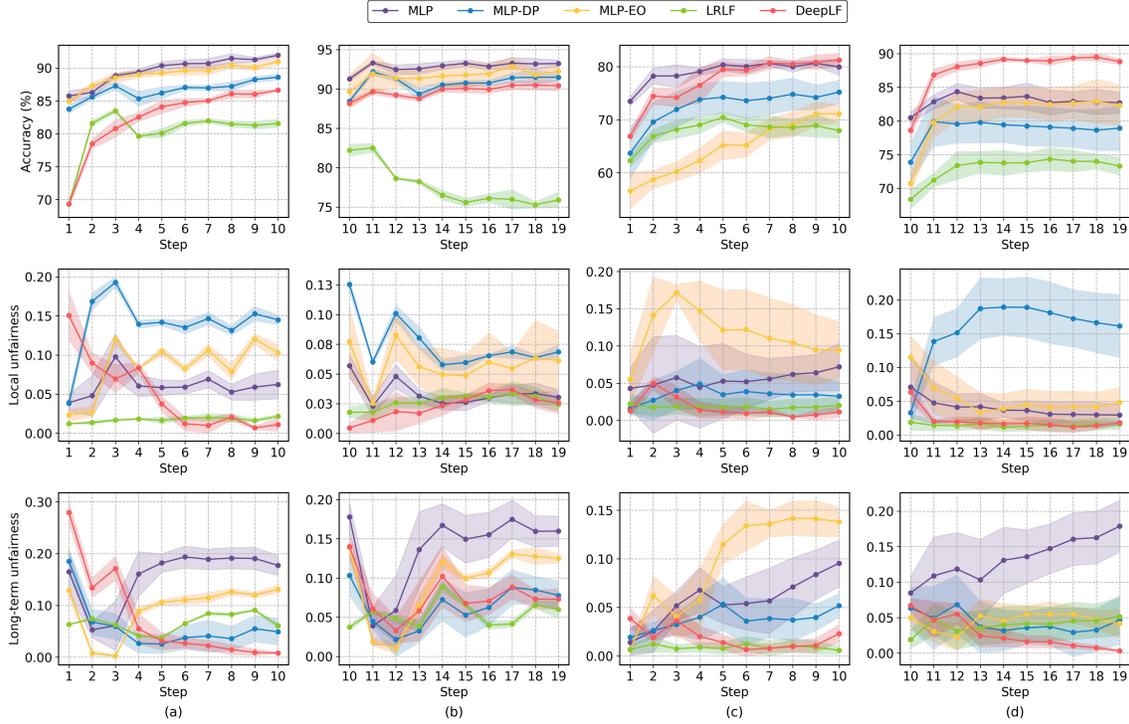
Figure 4: Accuracy ($\uparrow$), local and long-term unfairness ($\downarrow$) of different algorithms on SimLoan ((a) and (b)) and Taiwan ((c) and (d)) datasets. The decision models are trained on generated data within the time range $[1, 10]$. (a) and (c): Results of evaluation on generated data within time range $[1, 10]$. (b) and (d): Results of evaluation on generated data within the time range $[10, 19]$.
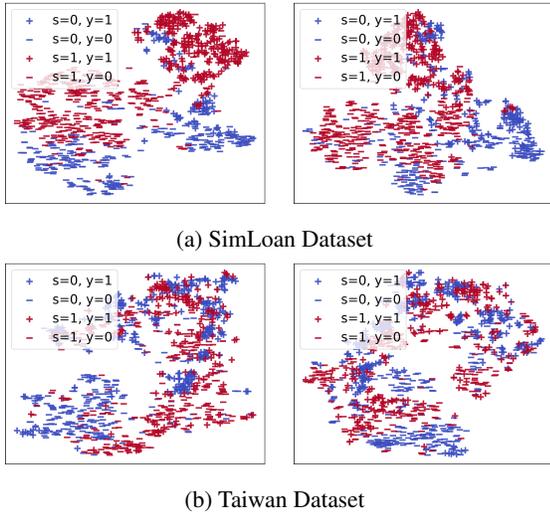

(a) SimLoan Dataset

(b) Taiwan Dataset

Figure 5: T-SNE of generated data distributions at time step $t = 10$ produced by MLP (left) and DeepLF (right).

shown in Figure 5. It can be seen that compared with the distribution obtained by using the **MLP** as the decision model (left figures), the data samples of two groups ($s = 0, 1$) produced by **DeepLF** (right figures) are more evenly mixed together which implies a fairer qualification distribution.

In the second setting, the time step T for achieving long-term fairness is set to 19. We train the decision models on the same training data as in the first setting but evaluate the models on the 10-step generated data with $\mathbf{X}^{10}$ as the input, i.e., the generated data within the time range $[10, 19]$. The difference in the second setting is that we only modify the decision model that will be deployed in the future (i.e., starting from $t = 10$). The results are shown in Figures 4 (c) and (d). In general, we observe similar results to the first setting where our algorithm outperforms all the baseline methods in achieving the best trade-off.

Conclusions

In this paper, we studied the problem of mitigating group disparity and achieving long-term fairness while limiting the use of sensitive attribute in decision-making. We proposed a data-driven method that requires only the information from the historical data. Leveraging the temporal causal graph, we formulated long-term fairness as the 1-Wasserstein distance between the interventional distributions of different demographic groups. Then, we proposed a three-phase learning framework to achieve long-term fairness by training an RC-GAN to predictively generate observational and interventional data and then training a classifier upon the generated data. Experiments on both synthetic and semi-synthetic data show that our method can achieve a more effective balance between long-term fairness, local fairness, and utility compared with methods based on traditional fairness notions.

Acknowledgments

References

2023. Students for Fair Admissions v. Harvard. https://www.supremecourt.gov/opinions/22pdf/20-1199_hgdj.pdf.

Alves, G.; Bernier, F.; Couceiro, M.; Makhlouf, K.; Palamidessi, C.; and Zhioua, S. 2023. Survey on fairness notions and related tensions. *EURO Journal on Decision Processes*, 100033.

Baker, R. S.; and Hawn, A. 2021. Algorithmic bias in education. *International Journal of Artificial Intelligence in Education*, 1–41.

Bakker, M. A.; Noriega-Campero, A.; Tu, D. P.; Sattigeri, P.; Varshney, K. R.; and Pentland, A. 2019. On fairness in budget-constrained decision making. In *Proceedings of the KDD Workshop on Explainable Artificial Intelligence*.

Berk, R.; Heidari, H.; Jabbari, S.; Kearns, M.; and Roth, A. 2021. Fairness in criminal justice risk assessments: The state of the art. *Sociological Methods & Research*, 50(1): 3–44.

Caton, S.; and Haas, C. 2020. Fairness in machine learning: A survey. *arXiv preprint arXiv:2010.04053*.

Cho, K.; Van Merriënboer, B.; Bahdanau, D.; and Bengio, Y. 2014. On the properties of neural machine translation: Encoder-decoder approaches. *arXiv preprint arXiv:1409.1259*.

Corbett-Davies, S.; Pierson, E.; Feller, A.; Goel, S.; and Huq, A. 2017. Algorithmic decision making and the cost of fairness. In *Proceedings of the 23rd acm sigkdd international conference on knowledge discovery and data mining*, 797–806.

Correa, J.; and Bareinboim, E. 2020. A calculus for stochastic interventions: Causal effect identification and surrogate experiments. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 34, 10093–10100.

Crosby, F. J.; Iyer, A.; and Sincharoen, S. 2006. Understanding affirmative action. *Annu. Rev. Psychol.*, 57: 585–611.

Cuturi, M. 2013. Sinkhorn distances: Lightspeed computation of optimal transport. *Advances in neural information processing systems*, 26.

Ding, F.; Hardt, M.; Miller, J.; and Schmidt, L. 2021. Retiring adult: New datasets for fair machine learning. *Advances in Neural Information Processing Systems*, 34: 6478–6490.

Esteban, C.; Hyland, S. L.; and Rätsch, G. 2017. Real-valued (medical) time series generation with recurrent conditional gans. *arXiv preprint arXiv:1706.02633*.

Fischer, M. J.; and Massey, D. S. 2007. The effects of affirmative action in higher education. *Social Science Research*, 36(2): 531–549.

Gretton, A.; Borgwardt, K.; Rasch, M.; Schölkopf, B.; and Smola, A. 2006. A kernel method for the two-sample-problem. *Advances in neural information processing systems*, 19.

Hu, Y.; and Zhang, L. 2022. Achieving long-term fairness in sequential decision making. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 36, 9549–9557.

Kim, J. S.; Chen, J.; and Talwalkar, A. 2020. FACT: A diagnostic for group fairness trade-offs. In *International Conference on Machine Learning*, 5264–5274. PMLR.

Kocaoglu, M.; Jaber, A.; Shanmugam, K.; and Bareinboim, E. 2019. Characterization and learning of causal graphs with latent variables from soft interventions. *Advances in Neural Information Processing Systems*, 32.

Kocaoglu, M.; Snyder, C.; Dimakis, A. G.; and Vishwanath, S. 2018. CausalGAN: Learning Causal Implicit Generative Models with Adversarial Training. In *International Conference on Learning Representations*.

Le Quy, T.; Roy, A.; Iosifidis, V.; Zhang, W.; and Ntoutsi, E. 2022. A survey on datasets for fairness-aware machine learning. *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery*, e1452.

Lee, M. S. A.; and Floridi, L. 2021. Algorithmic fairness in mortgage lending: from absolute conditions to relational trade-offs. *Minds and Machines*, 31(1): 165–191.

Liu, L. T.; Dean, S.; Rolf, E.; Simchowitz, M.; and Hardt, M. 2018. Delayed impact of fair machine learning. In *International Conference on Machine Learning*, 3150–3158. PMLR.

Mehrabi, N.; Morstatter, F.; Saxena, N.; Lerman, K.; and Galstyan, A. 2021. A survey on bias and fairness in machine learning. *ACM Computing Surveys (CSUR)*, 54(6): 1–35.

Pamfil, R.; Sriwattanaworachai, N.; Desai, S.; Pilgerstorfer, P.; Georgatzis, K.; Beaumont, P.; and Aragam, B. 2020. Dynotears: Structure learning from time-series data. In *International Conference on Artificial Intelligence and Statistics*, 1595–1605. PMLR.

Pearl, J. 2009. *Causality*. Cambridge university press.

Pearl, J. 2010. Causal inference. *NIPS Causality: Objectives and Assessment*, 39–58.

Perdomo, J.; Zrnic, T.; Mendler-Dünner, C.; and Hardt, M. 2020. Performative prediction. In *International Conference on Machine Learning*, 7599–7609. PMLR.

Peters, J.; Janzing, D.; and Schölkopf, B. 2017. *Elements of causal inference: foundations and learning algorithms*. The MIT Press.

Runge, J. 2020. Discovering contemporaneous and lagged causal relations in autocorrelated nonlinear time series datasets. In *Conference on Uncertainty in Artificial Intelligence*, 1388–1397. PMLR.

Schumann, C.; Foster, J.; Mattei, N.; and Dickerson, J. 2020. We need fairness and explainability in algorithmic hiring. In *International Conference on Autonomous Agents and Multi-Agent Systems (AAMAS)*.

Tang, Z.; Zhang, J.; and Zhang, K. 2022. What-Is and How-To for Fairness in Machine Learning: A Survey, Reflection, and Perspective. *arXiv preprint arXiv:2206.04101*.

Vowels, M. J.; Camgoz, N. C.; and Bowden, R. 2021. D'ya like dags? a survey on structure learning and causal discovery. *arXiv preprint arXiv:2103.02582*.

Yeh, I.-C.; and Lien, C.-h. 2009. The comparisons of data mining techniques for the predictive accuracy of probability of default of credit card clients. *Expert systems with applications*, 36(2): 2473–2480.

Zhang, L.; Wu, Y.; and Wu, X. 2016. A causal framework for discovering and removing direct and indirect discrimination. *arXiv preprint arXiv:1611.07509*.

Zhang, X.; Tu, R.; Liu, Y.; Liu, M.; Kjellstrom, H.; Zhang, K.; and Zhang, C. 2020. How do fair decisions fare in long-term qualification? *Advances in Neural Information Processing Systems*, 33: 18457–18469.