Long-Term Fair Decision Making through Deep Generative Models (Technical Appendix)

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

Related Work

Fair machine learning in the past decade has been focused on static settings with one-shot decisions being made (Mehrabi et al. 2021; Caton and Haas 2020). In recent years, attention has been paid to dynamic settings where sequential decisions are made over time. Some efforts have been devoted to a compound decision-making process called pipeline (Bower et al. 2017; Dwork and Ilvento 2018). In pipelines, individuals may drop out at any stage, and classification in subsequent stages depends on the remaining cohort of individuals. For instance, hiring is at least a two-stage model: deciding whom to be interviewed from the applicant pool and then deciding whom to be hired from the interview pool. In addition to the pipeline, a more practical and challenging dynamic setting considers that decisions made in the past can reshape the data population and subsequently influence future decisions (Zhang and Liu 2020). In this setting, several studies have demonstrated the inadequacy of static fairness approaches in various application scenarios, including credit lending (Liu et al. 2018), college admission (D'Amour et al. 2020), labor market (Hu and Chen 2018). In (Creager et al. 2020), the authors propose to use causal directed acyclic graphs (DAGs) as a unifying framework to study fairness in dynamical systems but have not reached any approach to achieve long-term fairness. In (Hu et al. 2020), the authors studied fair multiple decision making which also applies SCM and leverages soft interventions to model the deployment of decision models. However, (Hu et al. 2020) is focused on the static fairness of each decision model separately other than the long-term fairness. As a related line of work, some research (e.g., (Jabbari et al. 2017; Zhang et al. 2020; Wen, Bastani, and Topcu 2021; Yu et al. 2022)) studies long-term fairness in the context of reinforcement learning whose setting is different from supervised learning. Another related line of work proposes effort-based fairness measures that balance the effort an individual needs to make to change the decision outcome between two groups (Heidari, Nanda, and Gummadi 2019; Huan et al. 2020; Guldogan et al. 2022). The hypothesis is that effort fairness will encourage rejected individuals to improve their qualifications and prevent the exacerbation of the gap between different groups in the long run. For example, in (Heidari, Nanda, and Gummadi 2019), the authors propose a framework for characterizing the long-term impact of decision making algorithms on reshaping the distribution and leverage social models to simulate how individuals may respond to the decisions. The most relevant work to this paper is (Hu and Zhang 2022). It studied long-term fair decision making and formulated long-term fairness from the causal perspective. However, (Hu and Zhang 2022) requires true causal structure equations for training. In addition, it cannot achieve fairness at a time step that is beyond the training data. These limitations greatly reduce its practical significance.

Proof of Proposition 1

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.

Proof. For simplicity, in this proof we drop the superscript T and the notation of the soft intervention for $\mathbf{X}^T(\sigma_\theta)$. According to the definition of DP, we have

$$\mathrm{DP}(f) = |\mathbb{E}[f(\mathbf{X})|S = s^+] - \mathbb{E}(f(\mathbf{X})|S = s^-)|.$$

Due to the Kantorovich–Rubinstein duality (Villani 2021), it is straightforward that

$$\begin{aligned}\mathrm{DP}(f) &\leq \sup_{\|f\| \leq l_f} \left[\mathbb{E}_{\mathbf{x} \sim P(\mathbf{x}|s^+)}[f(\mathbf{x})] - \mathbb{E}_{\mathbf{x} \sim P(\mathbf{x}|s^-)}[f(\mathbf{x})]\right] \\ &= l_f \cdot W(P(\mathbf{X}|S = s^+), P(\mathbf{X}|S = s^-)) = l_f \cdot d.\end{aligned}$$

On the other hand, we have

$$\mathrm{EO}(f) = |\mathbb{E}[f(\mathbf{X})|Y{=}1, S{=}s^+] - \mathbb{E}(f(\mathbf{X})|Y{=}1, S{=}s^-)|.$$

Since we assume that Y is given by g and g satisfies the equal base rate condition, we have that $P(Y|\mathbf{X}, S) = g(\mathbf{X})$

and $P(Y|S) = P(Y)$. It then follows that

$$\begin{aligned}&\mathbb{E}[f(\mathbf{X})|y,s] = \int_{\mathbf{x}} f(\mathbf{x})P(\mathbf{x}|y,s)d\mathbf{x} \\ &= \int_{\mathbf{x}} f(\mathbf{x})P(\mathbf{x}|s)\frac{P(y|\mathbf{x},s)}{P(y|s)}d\mathbf{x} = \int_{\mathbf{x}} f(\mathbf{x})P(\mathbf{x}|s)\frac{g(\mathbf{x})}{P(y)}d\mathbf{x} \\ &= \frac{1}{P(y)}\mathbb{E}_{\mathbf{x}\sim P(\mathbf{x}|s)}[f(\mathbf{x})g(\mathbf{x})].\end{aligned}$$

In addition, define $m(x) = f(x)g(x)$ and denote its Lipschitz constant as l_m. It is easy to show that $l_m \leq l_f \cdot \sup_{\mathbf{X}} |f(\mathbf{X})| + l_g \cdot \sup_{\mathbf{X}} |g(\mathbf{X})|$. Since $h(\mathbf{X}) \leq 1$ and $g(\mathbf{X}) \leq 1$, we have $l_m \leq l_f + l_g$. As a result, we have

$$\begin{aligned}EO(f) &\leq \frac{l_f + l_g}{P(y)} W(P(\mathbf{X}|S = s^+), P(\mathbf{X}|S = s^-)) \\ &= \frac{l_f + l_g}{P(y)} \cdot d.\end{aligned}$$

□

Implementations Details and Hyperparameters

Experiments are performed on the computer with Intel Core i7-9700K CPU and NVIDIA GeForce GTX 1180 GPU. Except for **LRLF**, other baselines and our framework are used multi-layer fully-connected networks, i.e., **MLP**, as the classifiers. The details of the model architectures and hyperparameters used in our framework on two datasets are given in Tables 1 and 2. For a fair comparison, we adopt the same network structure and parameter settings for our decision model h_θ. Both datasets are split into train/validation/test sets with the ratio 70/10/20. The models are trained on the train sets and the hyperparameters are chosen on the validation sets. The reported results are calculated on the test sets.

Data Generation

Synthetic Dataset. We generate the synthetic time series dataset based on the causal time series graph shown in Figure 1 in the main paper. 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. We first train a neural network classifier h_{θ^*} on $S, \mathbf{X}^1, Y^1$ and treat it as the ground-truth model. For each time step t, classifier h_{θ^*} takes as inputs S and $\mathbf{X}^t$ and outputs a probability distribution over Y^t. We then sample Y^t from the distribution as shown below:

$$P(Y^t) = h_{\theta^*}(S, \mathbf{X}^t) \qquad Y^t \sim \text{Bernoulli}(P(Y^t)) \tag{1}$$

After that, we update the value of $\mathbf{X}^t$ to obtain $\mathbf{X}^{t+1}$ based on the value of Y^t. We treat Y^t as the ground-truth of loan repayment ($Y^t = 1$) and default ($Y^t = 0$). An individual with $Y^t = 1$ should have a larger probability to be predicted as 1 in the next time step, and vice versa. Therefore, we update the value of $\mathbf{X}^t$ according to the value of Y^t as well as the gradient of a loss function between the predicted probability and label 1, as given below:

$$\mathbf{X}^{t+1} = \mathbf{X}^t - \epsilon \cdot (2Y^t - 1) \cdot \frac{\partial \mathcal{L}(h_{\theta^*}(S, \mathbf{X}^t), \mathbf{1})}{\partial \mathbf{X}^t} \tag{2}$$

where the parameter ϵ controls the magnitude of changes in $\mathbf{X}^t$. As a result, $\mathbf{X}^{t+1}$ will be predicted closer to label 1 if $Y^t = 1$, and will be predicted further from label 1 if $Y^t = 0$. Following above generation rules, we generate a 10-step synthetic time series dataset with 10000 instances and $\mathbf{X}^t$ is 6 dimensional vector. We refer to this dataset SimLoan.

Semi-Synthetic Dataset. We also generage semi-synthetic data by leveraging the real-world Taiwan dataset 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 choose SEX as the sensitive feature S and BILL_ATM1 - BILL_ATM6 as the profile features in $\mathbf{X}$. We refer to this dataset Taiwan.

Table 1: The architectures of CLF and h_θ and hyperparameters for both datasets

Layer	Inputs	Output Dim	
		SimLoan	Taiwan
X		6	6
S		1	1
FC_1	[X, S]	32	16
FC_2	FC_1	64	32
FC_3	FC_2	1	1
Optimizer	Adam		
Learning rate	0.001		
Batch size	512		
λ_u		1.0	1.0
λ_s		2.1	0.2
λ_l		128.4	40.0

References

Bower, A.; Kitchen, S. N.; Niss, L.; Strauss, M. J.; Vargas, A.; and Venkatasubramanian, S. 2017. Fair pipelines. *arXiv preprint arXiv:1707.00391*.

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

Creager, E.; Madras, D.; Pitassi, T.; and Zemel, R. 2020. Causal modeling for fairness in dynamical systems. In *International Conference on Machine Learning*, 2185–2195. PMLR.

Table 2: The architecture of RCGAN and hyperparameters for both datasets

Layer	Inputs	Output Dim	
		SimLoan	Taiwan
X/Z		6	6
S/Y		1	1
Generator			
GRU_1	[Z, S, Y]	64	64
GRU_2	GRU_1	64	64
FC_1	GRU_2	6	6
Penalty			
MMD	[X, FC1]	1	1
Discriminator			
GRU_1	FC_1	64	64
GRU_2	GRU1	64	64
FC_1	GUR_2	1	1
Opimizer	Adam		
Learning rate	0.001		
Batch size	512		
γ	100		

D'Amour, A.; Srinivasan, H.; Atwood, J.; Baljekar, P.; Sculley, D.; and Halpern, Y. 2020. Fairness is not static: deeper understanding of long term fairness via simulation studies. In *Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency*, 525–534.

Dwork, C.; and Ilvento, C. 2018. Fairness under composition. *arXiv preprint arXiv:1806.06122*.

Guldogan, O.; Zeng, Y.; Sohn, J.-y.; Pedarsani, R.; and Lee, K. 2022. Equal Improvability: A New Fairness Notion Considering the Long-term Impact. *arXiv preprint arXiv:2210.06732*.

Heidari, H.; Nanda, V.; and Gummadi, K. 2019. On the Long-term Impact of Algorithmic Decision Policies: Effort Unfairness and Feature Segregation through Social Learning. In *International Conference on Machine Learning*, 2692–2701. PMLR.

Hu, L.; and Chen, Y. 2018. A short-term intervention for long-term fairness in the labor market. In *Proceedings of the 2018 World Wide Web Conference*, 1389–1398.

Hu, Y.; Wu, Y.; Zhang, L.; and Wu, X. 2020. Fair Multiple Decision Making Through Soft Interventions. *Advances in Neural Information Processing Systems*, 33.

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.

Huan, W.; Wu, Y.; Zhang, L.; and Wu, X. 2020. Fairness through equality of effort. In *Companion Proceedings of the Web Conference 2020*, 743–751.

Jabbari, S.; Joseph, M.; Kearns, M.; Morgenstern, J.; and Roth, A. 2017. Fairness in reinforcement learning. In *International Conference on Machine Learning*, 1617–1626. PMLR.

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.

Villani, C. 2021. *Topics in optimal transportation*, volume 58. American Mathematical Soc.

Wen, M.; Bastani, O.; and Topcu, U. 2021. Algorithms for Fairness in Sequential Decision Making. In *International Conference on Artificial Intelligence and Statistics*, 1144–1152. PMLR.

Yu, E. Y.; Qin, Z.; Lee, M. K.; and Gao, S. 2022. Policy Optimization with Advantage Regularization for Long-Term Fairness in Decision Systems. In *Advances in Neural Information Processing Systems*.

Zhang, X.; and Liu, M. 2020. Fairness in learning-based sequential decision algorithms: A survey. *arXiv preprint arXiv:2001.04861*.

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.