Lookahead Counterfactual Fairness

Anonymous authors
Paper under double-blind review

Abstract

As machine learning (ML) algorithms are used in applications that involve humans, concerns have arisen that these algorithms may be biased against certain social groups. *Counterfactual fairness* (CF) is a fairness notion proposed in Kusner et al. (2017) that measures the unfairness of ML predictions; it requires that the prediction perceived by an individual in the real world has the same marginal distribution as it would be in a counterfactual world, in which the individual belongs to a different group. Although CF ensures fair ML predictions, it fails to consider the downstream effects of ML predictions on individuals. Since humans are strategic and often adapt their behaviors in response to the ML system, predictions that satisfy CF may not lead to a fair future outcome for the individuals. In this paper, we introduce *lookahead counterfactual fairness* (LCF), a fairness notion accounting for the downstream effects of ML models which requires the individual *future status* to be counterfactually fair. We theoretically identify conditions under which LCF can be satisfied and propose an algorithm based on the theorems. We also extend the concept to path-dependent fairness. Experiments on both synthetic and real data validate the proposed method.

1 Introduction

The integration of machine learning (ML) into high-stakes domains (e.g., lending, hiring, college admissions, healthcare) has the potential to enhance traditional human-driven processes. However, it may introduce the risk of perpetuating biases and unfair treatment of protected groups. For instance, the violence risk assessment tool SAVRY has been shown to discriminate against males and foreigners (Tolan et al., 2019); Amazon's previous hiring system exhibited gender bias (Dastin, 2018); the accuracy of a computer-aided clinical diagnostic system varies significantly across patients from different racial groups (Daneshjou et al., 2021). Numerous fairness notions have been proposed in the literature to address unfairness issues, including *unawareness fairness* that prevents the explicit use of demographic attributes in the decision-making process, *parity-based fairness* that equalizes certain statistics (e.g., accuracy, true/false positive rate) across different groups (Hardt et al., 2016b), *preference-based fairness* that ensures a group of individuals, as a whole, regard the results or consequences they receive from the ML system more favorably than those received by another group (Zafar et al., 2017; Do et al., 2022). Unlike these notions that overlook the underlying causal structures among different variables Kusner et al. (2017) introduced the concept of *counterfactual fairness* (CF), which requires that an individual should receive a consistent treatment distribution in a counterfactual world where their sensitive attributes differs. Since then many approaches have been developed to train ML models that satisfy CF (Chiappa, 2019; Zuo et al., 2022; Wu et al., 2019; Xu et al., 2019; Ma et al., 2023).

However, CF is primarily studied in static settings without considering the downstream impacts ML decisions may have on individuals. Because humans in practice often adapt their behaviors in response to the ML system, their future status may be significantly impacted by ML decisions (Miller et al., 2020; Shavit et al., 2020; Hardt et al., 2016a). For example, individuals receiving approvals in loan applications may have more resources and be better equipped to improve their future creditworthiness (Zhang et al., 2020). Content recommended in digital platforms can steer consumer behavior and reshape their preferences (Dean & Morgenstern, 2022; Carroll et al., 2022). As a result, a model that satisfies CF in a static setting without accounting for such downstream effects may lead to unexpected adverse outcomes.

Although the downstream impacts of fair ML have also been studied in prior works (Henzinger et al., 2023a; Ge et al., 2021; Henzinger et al., 2023b; Liu et al., 2018; Zhang et al., 2020), the impact of counterfactually fair decisions remain relatively unexplored. The most related work to this study is (Hu & Zhang, 2022), which considers sequential interactions between individuals and an ML system over time and their goal is to ensure ML *decisions* satisfy path-specific counterfactual fairness constraint throughout the sequential interactions. However, Hu & Zhang (2022) still focuses on the fairness of ML decisions but not the fairness of the individual's actual status. Indeed, it has been well-evidenced that ML decisions satisfying certain fairness constraints during model deployment may reshape the population and unintentionally exacerbate the group disparity (Liu et al., 2018; Zhang et al., 2019; 2020). A prime example is Liu et al. (2018), which studied the lending problem and showed that the lending decisions satisfying statistical parity or equal opportunity fairness (Hardt et al., 2016b) may actually cause harm to disadvantaged groups by lowering their future credit scores, resulting in amplified group disparity. Tang et al. (2023) considered sequential interactions between ML decisions and individuals, where they studied the impact of counterfactual fair predictions on statistical fairness but their goal is still to ensure parity-based fairness at the group level.

In this work, we focus on counterfactual fairness evaluated over individual *future* status (label), which accounts for the downstream effects of ML decisions on individuals. We aim to examine under what conditions and by what algorithms the disparity between individual future status in factual and counterfactual worlds can be mitigated after deploying ML decisions. To this end, we first introduce a new fairness notion called "lookahead counterfactual fairness (LCF)." Unlike the original counterfactual fairness proposed by Kusner et al. (2017) that requires the ML predictions received by individuals to be the same as those in the counterfactual world, LCF takes one step further by enforcing the individual future status (after responding to ML predictions) to be the same.

Given the definition of LCF, we then develop algorithms that learn ML models under LCF. To model the effects of ML decisions on individuals, we focus on scenarios where individuals subject to certain ML decisions adapt their behaviors strategically by increasing their chances of receiving favorable decisions; this can be mathematically formulated as modifying their features toward the direction of the gradient of the decision function (Rosenfeld et al., 2020). We first theoretically identify conditions under which an ML model can satisfy LCF, and then develop an algorithm for training ML models under LCF. We also extend the algorithm and theorems to path-dependent LCF, which only considers unfairness incurred by the causal effect from the sensitive attribute to the outcome along certain paths.

Our contributions can be summarized as follows:

- We propose lookahead counterfactual fairness (LCF), a novel fairness notion that evaluates counterfactual fairness over individual future status (i.e., actual labels after responding to ML systems). Unlike the original CF notion that focuses on current ML predictions, LCF accounts for the subsequent impacts of ML decisions and aims to ensure fairness over individual actual future status. We also extend the definition to path-dependent LCF.

- For scenarios where individuals respond to ML models by changing features toward the direction of the gradient of decision functions, we theoretically identify conditions under which an ML model can satisfy LCF. We further develop an algorithm for training ML models under LCF.

- We conduct extensive experiments on both synthetic and real data to validate the proposed algorithm. Results show that compared to conventional counterfactual fair predictors, our method can improve disparity with respect to the individual actual future status.

2 Problem Formulation

Consider a supervised learning problem with a training dataset consisting of triples (A, X, Y), where $A \in \mathcal{A}$ is a sensitive attribute distinguishing individuals from multiple groups (e.g., race, gender), $X = [X_1, X_2, ..., X_d]^{\mathrm{T}} \in \mathcal{X}$ is a d-dimensional feature vector, and $Y \in \mathcal{Y} \subseteq \mathbb{R}$ is the target variable indicating individual's underlying status (e.g., Y in lending identifies an applicant's ability to repay the loan,

Y in healthcare may represent patients' insulin spike level). The goal is to learn a predictor from training data that can predict Y given inputs A and X. Let $\hat{Y}$ denote the output of the predictor.

We assume (A, X, Y) is associated with a structural causal model (SCM) (Pearl et al., 2000) $\mathcal{M} = (V, U, F)$, where $V = (A, X, Y)$ represents observable variables, U includes unobservable (exogenous) variables that are not caused by any variable in V, and $F = \{f_1, f_2, \ldots, f_{d+2}\}$ is a set of $d + 2$ functions called *structural equations* that determines how each observable variable is constructed. More precisely, we have the following structural equations,

$$\begin{aligned} X_i &= f_i(pa_i, U_{pa_i}),\ \forall i \in \{1, \cdots, d\}, \\ A &= f_A(pa_A, U_{pa_A}), \\ Y &= f_Y(pa_Y, U_{pa_Y}), \end{aligned} \tag{1}$$

where $pa_i \subseteq V$, $pa_A \subseteq V$ and $pa_Y \subseteq V$ are observable variables that are the parents of X_i, A, and Y, respectively. $U_{pa_i} \subseteq U$ are unobservable variables that are the parents of X_i. Similarly, we denote unobservable variables $U_{pa_A} \subseteq U$ and $U_{pa_Y} \subseteq U$ as the parents of A and Y, respectively.

2.1 Background: counterfactuals

If the probability density functions of unobserved variables are known, we can leverage the structural equations in SCM to find the marginal distribution of any observable variable $V_i \in V$ and even study how intervening certain observable variables impacts other variables. Specifically, the **intervention** on variable V_i is equivalent to replacing structural equation $V_i = f_i(pa_i, U_{pa_i})$ with equation $V_i = v$ for some v. Given new structural equation $V_i = v$ and other unchanged structural equations, we can find out how the distribution of other observable variables changes as we change value v.

In addition to understanding the impact of an intervention, SCM can further facilitate **counterfactual inference**, which aims to answer the question "*what would be the value of Y if Z had taken value z in the presence of evidence $O = o$ (both Y and Z are two observable variables)?*" The answer to this question is denoted by $Y_{Z \leftarrow z}(U)$ with U following conditional distribution of $\Pr\{U = u | O = o\}$. Given $U = u$ and structural equations F, the counterfactual value of Y can be computed by replacing the structural equation of Z with $Z = z$ and replacing U with u in the rest of the structural equations. Such counterfactual is typically denoted by $Y_{Z \leftarrow z}(u)$. Given evidence $O = o$, the distribution of counterfactual value $Y_{Z \leftarrow z}(U)$ can be calculated as follows,[1]

$$\Pr\{Y_{Z \leftarrow z}(U) = y | O = o\} = \sum_u \Pr\{Y_{Z \leftarrow z}(u) = y\} \Pr\{U = u | O = o\}. \tag{2}$$

Example 2.1 (**Law school success**)**.** Consider two groups of college students distinguished by gender $A \in \{0, 1\}$ whose first-year average (FYA) in college is denoted by Y. The FYA of each student is causally related to (observable) grade-point average (GPA) before entering college X_G, entrance exam score (LSAT) X_L, and gender A. Suppose there are two unobservable variables $U = (U_A, U_{XY})$, e.g., U_{XY} may be interpreted as the student's knowledge. Consider the following structural equations:

$$\begin{aligned} A &= U_A, & X_G &= b_G + w_G^A A + U_{XY}, \\ X_L &= b_L + w_L^A A + U_{XY}, & Y &= b_F + w_F^A A + U_{XY}, \end{aligned}$$

where $(b_G, w_G^A, b_L, w_L^A, b_F, w_F^A)$ are know parameters of the causal model. Given observation $X_G = 1, A = 0$, the counterfactual value can be calculated with an *abduction-action-prediction* procedure Glymour et al. (2016): (i) *abduction* that finds posterior distribution $\Pr\{U = u | X_G = 1, A = 0\}$. Here, we have $U_{XY} = 1 - b_G$ and $U_A = 0$ with probability 1; (ii) *action* that performs intervention $A = 1$ by replacing structural equations of A; (iii) *prediction* that computes distribution of $Y_{A \leftarrow 1}(U)$ given $X_G = 1, A = 0$ using new structural equations and the posterior. We have:

$$Y_{A \leftarrow 1}(U) = b_f + w_F^A + 1 - b_G \quad \text{with probability 1.}$$

[1] Given structural equations equation 1 and the marginal distribution of U, $\Pr\{U = u, O = o\}$ can be calculated using the Change-of-Variables Technique and the Jacobian factor. As a result, $\Pr\{U = u | O = o\} = \frac{\Pr\{U=u,\ O=o\}}{\Pr\{O=o\}}$ can also be calculated.

2.2 Counterfactual Fairness

Counterfactual Fairness (CF) was first proposed by Kusner et al. (2017); it requires that for an individual with $(X = x, A = a)$, the prediction $\hat{Y}$ in the factual world should be the same as that in the counterfactual world in which the individual belongs to a different group. Mathematically, CF is defined as follows: $\forall a, \breve{a} \in \mathcal{A}, X \in \mathcal{X}, y \in \mathcal{Y}$,

$$\Pr\left(\hat{Y}_{A\leftarrow a}(U) = y | X = x, A = a\right) = \Pr\left(\hat{Y}_{A\leftarrow \breve{a}}(U) = y | X = x, A = a\right),$$

While the CF notion has been widely used in the literature, it does not account for the downstream impacts of ML prediction $\hat{Y}$ on individuals in factual and counterfactual worlds. To illustrate the importance of considering such impacts, we provide an example below.

Example 2.2. Consider automatic lending where an ML model is used to decide whether to issue a loan to an applicant based on credit score X and sensitive attribute A. As highlighted in Liu et al. (2018), issuing loans to unqualified people who cannot repay the loan may hurt them by worsening their future credit scores. Assume an applicant in the factual world is qualified for the loan and does not default. But in a counterfactual world where the applicant belongs to another group, he/she is not qualified. Under counterfactually fair predictions, both individuals in the factual and counterfactual worlds should receive the loan with the same probability. Suppose both are issued a loan, then the one in the counterfactual world would have a worse credit score in the future. Thus, it is crucial to consider the downstream effects when learning a fair ML model.

2.3 Characterize downstream effects

Motivated by Example 2.2, this work studies CF in a dynamic setting where the deployed ML decisions may affect individual behavior and change their future features and statuses. Formally, let X' and Y' denote an individual's future feature vector and status, respectively. We use an individual response r to capture the impact of ML prediction $\hat{Y}$ on individuals, as defined below.

Definition 2.1 (Individual response)**.** An individual response $r : \mathcal{U} \times \mathcal{V} \times \mathcal{Y} \mapsto \mathcal{U} \times \mathcal{V}$ is a map from the current exogenous variables $U \in \mathcal{U}$, endogenous variables $V \in \mathcal{V}$, and prediction $\hat{Y} \in \mathcal{Y}$ to the future exogenous variables U' and endogenous variables V'.

One way to tackle the issue in Example 2.2 is to explicitly consider the individual response and impose a fairness constraint on future status Y' instead of the prediction $\hat{Y}$. We call such a fairness notion the *Lookahead Counterfactual Fairness (LCF)* and present it in Section 3.

3 Lookahead Counterfactual Fairness

We consider the fairness over the individual's future outcome Y'. Given structural causal model $\mathcal{M} = (U, V, F)$, individual response r, and data (A, X, Y), we define lookahead counterfactual fairness below.

Definition 3.1. We say an ML model satisfies lookahead counterfactual fairness (LCF) under a response r if the following holds $\forall a, \breve{a} \in \mathcal{A}, X \in \mathcal{X}, y \in \mathcal{Y}$:

$$\Pr\left(Y'_{A\leftarrow a}(U) = y | X = x, A = a\right) = \Pr\left(Y'_{A\leftarrow \breve{a}}(U) = y | X = x, A = a\right), \quad (3)$$

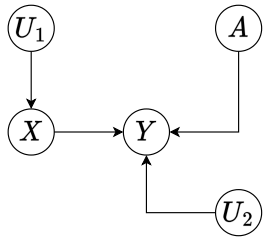


Figure 1: Causal graph in Example 3.1.

LCF implies that the subsequent consequence of ML decisions for a given individual in the factual world should be the same as that in the counterfactual world where the individual belongs to other demographic groups. Note that CF may contradict LCF: even under counterfactually fair predictor, individuals in the factual and counterfactual worlds may end up with very different future statuses. We show this with an example below.

Example 3.1. Consider the causal graph in Figure 1 and the structural functions as follows:

$$\begin{aligned} X &= f_X(U_1) = U_1, & Y &= f_Y(U_2, X, A) = U_2 + X + A, \\ U_1' &= r(U_1, \hat{Y}) = U_1 + \nabla_{U_1}\hat{Y}, & U_2' &= r(U_2, \hat{Y}) = U_2 + \nabla_{U_2}\hat{Y}, \\ X' &= f_X(U_1') = U_1', & Y' &= f_Y(U_2', X', A) = U_2' + X' + A. \end{aligned}$$

Based on Kusner et al. (2017), a predictor that only uses U_1 and U_2 as input is counterfactually fair.[2] Therefore, $\hat{Y} = h(U_1, U_2)$ satisfies CF. Let U_1 and U_2 be uniformly distributed over $[-1, 1]$. Note that the response $r(U_1, \hat{Y})$ and $r(U_2, \hat{Y})$ imply that individuals make efforts to change feature vectors through changing the unobservable variables, which results in higher $\hat{Y}$ in the future. It is easy to see that a CF predictor $h(U_1, U_2) = U_1 + U_2$ minimizes the MSE loss $\mathbb{E}\{(Y - \hat{Y})^2\}$ if $A \in \{-1, 1\}$ and $\Pr\{A = 1\} = 0.5$. However, since $\nabla_{U_1}\hat{Y} = \nabla_{U_2}\hat{Y} = 1$, we have:

$$\begin{aligned} \Pr\left(Y'_{A\leftarrow a}(U) = y | X = x, A = a\right) &= \delta(y - a - x - 2) \\ \Pr\left(Y'_{A\leftarrow \check{a}}(U) = y | X = x, A = a\right) &= \delta(y - \check{a} - x - 2) \end{aligned}$$

where $\delta(y) = 1$ if $y = 0$ and $\delta(y) = 0$ otherwise. It shows that although the decisions in the factual and counterfactual worlds are the same, the future statuses Y' are still different and Definition 3.1 does not hold.

Theorem 3.1 below identifies more general scenarios under which LCF can be violated with a CF predictor.

Theorem 3.1 (**Violation of LCF under a CF predictor**). *Consider a causal model $\mathcal{M} = (U, V, F)$ and individual response r in the following form:*

$$\begin{aligned} U' &= r(U, \hat{Y}) \\ X_i' &= r(X_r, \hat{Y}), \quad X_i \subset V \text{ are the root nodes} \end{aligned}$$

If the response r is a function and the status Y in factual and counterfactual worlds have different distributions, i.e.,

$$\Pr(Y_{A\leftarrow a}(U) = y | X = x, A = a) \neq \Pr(Y_{A\leftarrow \check{a}}(U) = y | X = x, A = a),$$

then imposing a model that satisfies CF will violate LCF, i.e.,

$$\Pr(Y'_{A\leftarrow a}(U) | X = x, A = a) \neq \Pr(Y'_{A\leftarrow \check{a}}(U) = y | X = x, A = a).$$

4 Learning under LCF

This section introduces an algorithm for learning a predictor under LCF. In particular, we focus on a special case with the causal model and the individual response defined below.

Given sets of unobservable variables $U = \{U_1, ..., U_d, U_Y\}$ and observable variables $\{X_1, ..., X_d, A, Y\}$, we consider causal model with the following structural functions:

$$X_i = f_i(U_i, A), \quad Y = f_Y(X_1, ..., X_d, U_Y) \tag{4}$$

where f_i is an invertible function[3], and f_Y is invertible w.r.t. U_Y. After receiving the ML prediction $\hat{Y}$, the individual's future features X' and status Y' change accordingly. Specifically, we consider scenarios where individual unobservable variables U change based on the following

$$\begin{aligned} U_i' &= r_i(U_i, \hat{Y}) = U_i + \eta\nabla_{U_i}\hat{Y}, \quad \forall i \in \{1, ..., d\} \\ U_Y' &= r_Y(U_Y, \hat{Y}) = U_Y + \eta\nabla_{U_Y}\hat{Y} \end{aligned} \tag{5}$$

and the future attributes X_i' and status Y' also change accordingly, i.e.,

$$\begin{aligned} X_i' &= f_i(U_i', A), \\ Y' &= f_Y(X_1', ..., X_d', U_Y') \end{aligned} \tag{6}$$

[2]Note that U_1 and U_2 can be generated for each sample (X, A). See Section 4.1 of (Kusner et al., 2017) for more details.

[3]Several works in causal inference also consider invertible structural function, e.g., *bijective causal models* introduced in Nasr-Esfahany et al. (2023).

The above scenario implies that individuals respond to ML model by strategically moving features toward the **direction that increases their chances of receiving favorable decisions**, step size $\eta > 0$ controls the magnitude of data change and can be interpreted as the effort budget individuals have on changing their data. Note that this type of response has been widely studied in strategic classification literature (Rosenfeld et al., 2020; Hardt et al., 2016a). The above process with $d = 2$ is visualized in Figure 2.

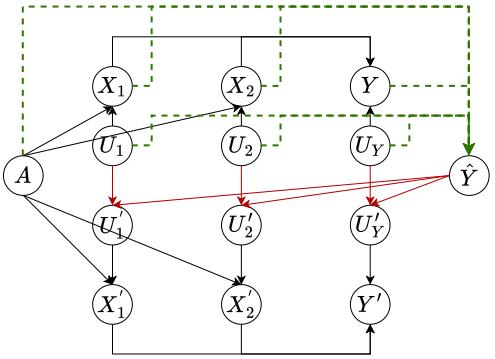


Figure 2: A causal graph and individual responses with two features X_1, X_2. The black arrows represent the connections described in structural functions. The red arrows represent the response process. The green dash arrows are the potential connection to prediction $\hat{Y}$.

Our goal is to train an ML model under LCF constraint. Before presenting our method, we first define the notion of counterfactual random variables.

Definition 4.1 (Counterfactual random variable)**.** Let x and a be realizations of random variables X and A, and $\check{a} \neq a$. We say $\check{X} := X_{A \leftarrow \check{a}}(U)$ and $\check{Y} := Y_{A \leftarrow \check{a}}(U)$ are the counterfactual random variables associated with (x, a) if U follows the conditional distribution $\Pr\{U|X = x, A = a\}$ as given by the causal model $\mathcal{M}$. The realizations of $\check{X}$, $\check{Y}$ are denoted by $\check{x}$ and $\check{y}$.

The following theorem constructs a predictor g that satisfies LCF, i.e., deploying the predictor g in Theorem 4.1 ensures the future status Y' is counterfactually fair.

Theorem 4.1 (Predictor with perfect LCF)**.** *Consider causal model $\mathcal{M} = (U, V, F)$, where $U = \{U_X, U_Y\}$, $U_X = [U_1, U_2, ..., U_d]^{\mathrm{T}}, V = \{A, X, Y\}, X = [X_1, X_2, ..., X_d]^{\mathrm{T}}$, and the structural equations are given by,*

$$X = \alpha \odot U_X + \beta A, \quad Y = w^{\mathrm{T}} X + \gamma U_Y, \tag{7}$$

where $\alpha = [\alpha_1, \alpha_2, ..., \alpha_d]^{\mathrm{T}}$, $\beta = [\beta_1, \beta_2, ..., \beta_d]^{\mathrm{T}}$, $w = [w_1, w_2, .., w_d]^{\mathrm{T}}$, and $\odot$ denotes the element wise production. Then, the following predictor satisfies LCF,

$$g(\check{Y}, U) = p_1 \check{Y}^2 + p_2 \check{Y} + p_3 + h(U), \tag{8}$$

where $p_1 = \frac{T}{2}$ with $T := \frac{1}{\eta(||w \odot \alpha||_2^2 + \gamma^2)}$, and $\check{Y}$ is the counterfactual random variable associated with Y. Here, p_2 and p_3 and function $h(.)$ are arbitrary and can be trained to improve prediction performance.

The above theorem implies that g should be constructed based on the counterfactual random variable $\check{Y}$ and U. Even though U is unobserved, it can be obtained from the inverse of structural equations. Quantity T in Theorem 4.1 depends on the step size η in individual response, and parameters α, γ, w in structural functions. When $p_1 = \frac{T}{2}$, we can achieve perfect LCF.

It is worth noting that Definition 3.1 can be a very strong constraint and imposing $Y'_{A \leftarrow a}(U)$ and $Y'_{A \leftarrow \check{a}}(U)$ to have the same distribution may degrade the performance of the predictor significantly. To tackle this, we may consider a weaker version of LCF.

Definition 4.2 (Relaxed LCF)**.** We say Relaxed LCF holds if $\forall (a, \check{a}) \in \mathcal{A}^2, a \neq \check{a}, X \in \mathcal{X}, y \in \mathcal{Y}$, we have:

$$\Pr\left(\left\{|Y'_{A \leftarrow a}(U) - Y'_{A \leftarrow \check{a}}(U)| < |Y_{A \leftarrow a}(U) - Y_{A \leftarrow \check{a}}(U)|\right\} \middle| X = x, A = a\right) = 1. \tag{9}$$

Definition 4.2 implies that after individuals respond to ML model, the difference between the future status Y' in factual and counterfactual worlds should be smaller than the difference between original status Y in factual and counterfactual worlds. In other words, it means that the disparity between factual and counterfactual worlds must decrease over time. In Section 6, we empirically show that constraint in equation 9 is weaker than the constraint in equation 3 and can lead to a better prediction performance.

Corollary 4.1 (Relaxed LCF with predictor in equation 8)**.** *Consider the same causal model defined in Theorem 4.1 and the predictor defined in equation 8. Relaxed LCF holds if $p_1 \in [0, T]$.*

Apart from relaxing p_1 in predictor as shown in equation 8, we can also relax the form of the predictor to satisfy Relaxed LCF, as shown in Theorem 4.2.

Theorem 4.2 (Predictor under Relaxed LCF)**.** *Consider the same causal model defined in Theorem 4.1. A predictor $g(\check{Y}, U)$ satisfies Relaxed LCF if g has the following three properties:*

(i) $g(\check{y}, u)$ is strictly convex in $\check{y}$.

(ii) $g(\check{y}, u)$ can be expressed as $g(\check{y}, u) = g_1(\check{y}) + g_2(u)$.

(iii) The derivative of $g(\check{y}, u)$ w.r.t. $\check{y}$ is K-Lipschitz continuous in $\check{y}$ with $K < \frac{2}{\eta(||w \odot \alpha||_2^2 + \gamma^2)}$, i.e.,

$$\left| \frac{\partial g(\check{y}_1, u)}{\partial \check{y}} - \frac{\partial g(\check{y}_2, u)}{\partial \check{y}} \right| \leq K \left| \check{y}_1 - \check{y}_2 \right|.$$

Theorems 4.1 and 4.2 provide insights on designing algorithms to train a predictor with perfect or Relaxed LCF. Specifically, given training data $\mathcal{D} = \{(x^{(i)}, y^{(i)}, a^{(i)})\}_{i=1}^n$, we first estimate the structural equations. Then, we choose a parameterized predictor g that satisfies the conditions in Theorem 4.1 or 4.2. An example is shown in Algorithm 1, which finds an optimal predictor in the form of $g(\check{y}, u) = p_1 \check{y}^2 + p_2 \check{y} + p_3 + h_\theta(u)$ under LCF, where $p_1 = \frac{1}{2\eta(||w \odot \alpha||_2^2 + \gamma^2)}$, θ is the training parameter for function h, and p_2, p_3 are two other training parameters. Under Algorithm 1, we can find the optimal values for p_2, p_3, θ using training data $\mathcal{D}$. If we only want to satisfy Relaxed LCF (Definition 4.2), p_1 can be a training parameter with $0 < p_1 < T$.

Algorithm 1 Training a predictor with perfect LCF

Input: Training data $\mathcal{D} = \{(x^{(i)}, y^{(i)}, a^{(i)})\}_{i=1}^n$, response parameter η.

1: Estimate the structural equations 7 using $\mathcal{D}$ to determine parameters α, β, w, and γ.
2: For each data point $(x^{(i)}, y^{(i)}, a^{(i)})$, draw m samples $\left\{u^{(i)[j]}\right\}_{j=1}^m$ from conditional distribution $\Pr\{U | X = x^{(i)}, A = a^{(i)}\}$ and generate counterfactual $\check{y}^{(i)[j]}$ associated with $u^{(i)[j]}$ based on structural equations 7.
3: Compute $p_1 \leftarrow \frac{1}{2\eta(||w \odot \alpha||_2^2 + \gamma^2)}$.
4: Solve the following optimization problem,

$$\hat{p}_2, \hat{p}_3, \hat{\theta} = \underset{p_2, p_3, \theta}{\arg\min} \frac{1}{mn} \sum_{i=1}^n \sum_{j=1}^m l\left(g\left(\check{y}^{(i)[j]}, u^{(i)[j]}\right), y^{(i)}\right)$$

where

$$g\left(\check{y}^{(i)[j]}, u^{(i)[j]}\right) = p_1 \left(\check{y}^{(i)[j]}\right)^2 + p_2 \check{y}^{(i)[j]} + p_3 + h_\theta(u),$$

θ is a parameter for function h, and l is a loss function.

Output: $\hat{p}_2, \hat{p}_3, \hat{\theta}$

It is worth noting that the results in Theorems 4.1 and 4.2 are for linear causal models. When the causal model is non-linear, it is hard to construct a model satisfying perfect LCF in Definition 3.1. Nonetheless, we can still show that it is possible to satisfy Relaxed LCF (Definition 4.2) for certain non-linear causal models. Theorem 4.3 below focuses on a special case when X is not linearly dependent on A and U_X and it identifies the condition under which Relaxed LCF can be guaranteed.

Theorem 4.3. *Consider causal model $\mathcal{M} = (U, V, F)$, where $U = \{U_X, U_Y\}$, U_X, U_Y are scalar attributes, $V = \{A, X, Y\}$, X is a scalar, and the structural equations are given by,*

$$X = f(\alpha U_X + \beta A), \quad Y = wX + \gamma U_Y,$$

where f is a non-linear function. If f satisfies the following:

- *f is strictly convex;*

- $\forall s_1, s_2, s_1', s_2'$, *if* $|s_1 - s_2| < |s_1' - s_2'|$, *we have* $|f(s_1) - f(s_2)| < |f(s_1') - f(s_2')|$;
- *The derivative of* f *is* K*-Lipschitz continuous with* $K < \frac{2}{|\lambda_1 w \eta \alpha^2|}$, *i.e.,*

$$\left| \frac{\mathrm{d}f(s_1)}{\mathrm{d}s} - \frac{\mathrm{d}f(s_2)}{\mathrm{d}s} \right| \leq K|s_1 - s_2|.$$

Then, the following predictor satisfies Relaxed LCF

$$g(\check{Y}, U) = \lambda_1 \check{Y} + \lambda_2 h(U)$$

where λ_1, λ_2 *are learnable parameters, in which* λ_1 *satisfies* $\lambda_1 w > 0$, *and* h *can be an arbitrary function (e.g., neural network).*

Theorems 4.2 and 4.3 show that designing a predictor under Relaxed LCF highly depends on the form of causal structure and structural equations. To wrap up this section, we would like to identify conditions under which Relaxed LCF holds in a causal graph that X is determined by the product of U_X and A.

Theorem 4.4. *Consider a non-linear causal model* $\mathcal{M} = (U, V, F)$, *where* $U = \{U_X, U_Y\}$, $U_X = [U_1, U_2, ..., U_d]^{\mathrm{T}}, V = \{A, X, Y\}, X = [X_1, X_2, ..., X_d]^{\mathrm{T}}$, $A \in \{a_1, a_2\}$ *is a binary sensitive attribute. Assume that the structural functions are given by,*

$$X = A \cdot (\alpha \odot U_X + \beta), \qquad Y = w^{\mathrm{T}} X + \gamma U_Y \tag{10}$$

where $\alpha = [\alpha_1, \alpha_2, ..., \alpha_d]^{\mathrm{T}}$, $\beta = [\beta_1, \beta_2, ..., \beta_d]^{\mathrm{T}}$, *and* $\odot$ *denotes the element wise production. A predictor* $g(\check{Y})$ *satisfies Relaxed LCF if* g *and the causal model have the following three properties.*

- *The value domain of* A *satisfies* $a_1 a_2 \geq 0$.
- $g(\check{y})$ *is strictly convex.*
- *The derivate of* $g(\check{y})$ *is* K*-Lipschitz continuous with* $K \leq \frac{2}{\eta(a_1 a_2 ||w \odot \alpha||_2^2 + \gamma^2)}$, *i.e.,*

$$\left| \frac{\partial g(\check{y}_1)}{\partial \check{y}} - \frac{\partial g(\check{y}_2)}{\partial \check{y}} \right| < K \left| \check{y}_1 - \check{y}_2 \right|.$$

Although the structural equation associated with Y is still linear in X and U_Y, we emphasize that such a linear assumption has been very common in the literature due to the complex nature of strategic classification Zhang et al. (2022); Liu et al. (2020); Bechavod et al. (2022). For instance, Bechavod et al. (2022) assumed the actual status of individuals is $Y = \beta X$, a linear function of features X. Zhang et al. (2022) assumed that X itself may be non-linear in some underlying traits of the individuals, but the relationship between X and $P(Y = 1|X)$ is still linear. Indeed, due to the individual's strategic response, conducting the theoretical analysis accounting for such responses can be highly challenging. Nonetheless, it is worthwhile extending LCF to non-linear settings and we leave this for future works.

5 Path-dependent LCF

An extension of counterfactual fairness called path-dependent fairness has been introduced in Kusner et al. (2017). In this section, we also want to introduce an extension of LCF called path-dependent LCF. We will also modify Algorithm 1 to satisfy path-dependent LCF.

We start by introducing the notion of path-dependent counterfactuals. In a causal model associated with a causal graph $\mathcal{G}$, we denote $\mathcal{P}_{\mathcal{G}_A}$ as a set of unfair paths from sensitive attribute A to Y. We define $X_{\mathcal{P}^c_{\mathcal{G}_A}}$ as the set of features that are not present in any of the unfair paths. Under observation $X = x, A = a$, we call $Y_{A \leftarrow \check{a}, X_{\mathcal{P}^c_{\mathcal{G}_A}} \leftarrow x_{\mathcal{P}^c_{\mathcal{G}_A}}}(U)$ path-dependent counterfactual random variable for Y, and its distribution can be calculated as follows:

$$\Pr\{Y_{A \leftarrow \check{a}, X_{\mathcal{P}^c_{\mathcal{G}_A}} \leftarrow x_{\mathcal{P}^c_{\mathcal{G}_A}}}(U) = y | X = x, A = a\} = \sum_u \Pr\{Y_{A \leftarrow \check{a}, X_{\mathcal{P}^c_{\mathcal{G}_A}} \leftarrow x_{\mathcal{P}^c_{\mathcal{G}_A}}}(u) = y\} \Pr\{U = u | X = x, A = a\}.$$

For simplicity, we use $\check{Y}_{PD}$ and $\check{y}_{PD}$ to represent a path-dependent counterfactual and the corresponding realization. That is, $\check{Y}_{PD} = Y_{A\leftarrow\check{a}, X_{\mathcal{P}^c_{\mathcal{G}_A}} \leftarrow x_{\mathcal{P}^c_{\mathcal{G}_A}}}(U)$ where U follows $\Pr\{U|X=x, A=a\}$. We consider the same kind of causal model described in Section 4, the future attributes X' and outcome Y' are determined by equation 5 and equation 6. We formally define the path-dependent LCF in the following definition.

Definition 5.1. We say an ML model satisfies path-dependent lookahead counterfactual fairness w.r.t. the unfair path set $\mathcal{P}_{\mathcal{G}_A}$ if the following holds $\forall a, \check{a} \in \mathcal{A}, X \in \mathcal{X}, y \in \mathcal{Y}$:

$$\Pr\left(\hat{Y}'_{A\leftarrow a, X_{\mathcal{P}^c_{\mathcal{G}_A}} \leftarrow x_{\mathcal{P}^c_{\mathcal{G}_A}}}(U) = y \middle| X=x, A=a\right) = \Pr\left(\hat{Y}'_{A\leftarrow \check{a}, X_{\mathcal{P}^c_{\mathcal{G}_A}} \leftarrow x_{\mathcal{P}^c_{\mathcal{G}_A}}}(U) = y \middle| X=x, A=a\right).$$

Then we have the following theorem.

Theorem 5.1. *Consider a causal model and structural equations defined in Theorem 4.1. If we denote the features on unfair path as $X_{\mathcal{P}_{\mathcal{G}_A}}$ and remaining features as $X_{\mathcal{P}^c_{\mathcal{G}_A}}$, we can re-write structural equations as*

$$\begin{aligned} X_{\mathcal{P}_{\mathcal{G}_A}} &= \alpha_{\mathcal{P}_{\mathcal{G}_A}} \odot U_{X_{\mathcal{P}_{\mathcal{G}_A}}} + \beta_{\mathcal{P}_{\mathcal{G}_A}} A, \\ X_{\mathcal{P}^c_{\mathcal{G}_A}} &= \alpha_{\mathcal{P}^c_{\mathcal{G}_A}} \odot U_{X_{\mathcal{P}^c_{\mathcal{G}_A}}} + \beta_{\mathcal{P}^c_{\mathcal{G}_A}} A, \\ Y &= w^{\mathrm{T}}_{\mathcal{P}_{\mathcal{G}_A}} X_{\mathcal{P}_{\mathcal{G}_A}} + w^{\mathrm{T}}_{\mathcal{P}^c_{\mathcal{G}_A}} X_{\mathcal{P}^c_{\mathcal{G}_A}} + \gamma U_Y \end{aligned}$$

Then, the following predictor satisfies path-dependent LCF,

$$g(\check{Y}_{PD}, U) = p_1 \check{Y}^2_{PD} + p_2 \check{Y}_{PD} + p_3 + h(U),$$

where $p_1 = \frac{T}{2}$ with

$$T := \frac{1}{\eta(||w_{\mathcal{P}_{\mathcal{G}_A}} \odot \alpha_{\mathcal{P}_{\mathcal{G}_A}}||^2_2 + ||w_{\mathcal{P}^c_{\mathcal{G}_A}} \odot \alpha_{\mathcal{P}^c_{\mathcal{G}_A}}||^2_2 + \gamma^2)},$$

p_2 and p_3 are learnable parameters to improve prediction performance and h is an arbitary function.

6 Experiment

We conduct experiments on both synthetic and real data to validate the proposed method.

6.1 Synthetic Data

We generate the synthetic data based on the causal model described in Theorem 4.1, where we set $d = 10$ and generated 1000 data points. We assume U_X and U_Y follow the uniform distribution over $[0, 1]$ and the sensitive attribute $A \in \{0, 1\}$ is a Bernoulli random variable with $\Pr\{A = 0\} = 0.5$. Then, we generate X and Y using the structural functions described in Theorem 4.1.[4] Based on the causal model, the conditional distribution of U_X and U_Y given $X = x, A = a$ are as follows,

$$U_X|X=x, A=a \ \sim \ \delta\Big(\frac{x-\beta a}{\alpha}\Big) \quad U_Y|X=x, A=a \ \sim \ \text{Uniform}(0,1) \tag{11}$$

Baselines. We used two baselines for comparison: (i) **Unfair predictor (UF)** is a linear model without fairness constraint imposed. It takes feature X as input and predicts Y. (ii) **Counterfactual fair predictor (CF)** only takes the unobservable variables U as the input and was proposed by Kusner et al. (2017).

Implementation Details. To find a predictor satisfying Definition 3.1, we train a predictor in the form of Eq. 8. In our experiment, $h(u)$ is a linear function. To train $g(\check{y}, u)$, we follows Algorithm 1 with $m = 100$. We split the dataset into the training/validation/test set at 60%/20%/20% ratio randomly and repeat the experiment 5 times. We use the validation set to find the optimal number of training epochs and the learning rate. Based on our observation, Adam optimization with a learning rate equal to 10^{-3} and 2000 epochs gives us the best performance.

[4]The exact values for parameters α, β, w and γ can be found in the Appendix B.

Metrics. We use three metrics to evaluate the methods. To evaluate the performance, we use the mean squared error (MSE). Given a dataset $\{x^{(i)}, a^{(i)}, y^{(i)}\}_{i=1}^{n}$, for each $x^{(i)}$ and $a^{(i)}$, we generate $m = 100$ values of $u^{(i)[j]}$ from the posterior distribution. MSE can be estimated as follows,[5]

$$\text{MSE} = \frac{1}{mn} \sum_{i=1}^{n} \sum_{j=1}^{m} \left\| y^{(i)} - \hat{y}^{(i)[j]} \right\|^2, \tag{12}$$

where $\hat{y}^{(i)[j]}$ is the prediction for data $(x^{(i)}, a^{(i)}, u^{(i)[j]})$. Note that for the UF baseline, the prediction does not depend on $u^{(i)[j]}$. Therefore, $\hat{y}^{(i)[j]}$ does not change by j for the UF predictor. To evaluate fairness, we define a metric called average future causal effect (AFCE),

$$\text{AFCE} = \frac{1}{mn} \sum_{i=1}^{n} \sum_{j=1}^{m} \left| y'^{(i)[j]} - \check{y}'^{(i)[j]} \right|$$

It is the average difference between the factual and counterfactual future outcomes. To compare $|Y - \check{Y}|$ with $|Y' - \check{Y}'|$ under different algorithms, we use the unfairness improvement ratio (UIR) defined below. The larger UIR implies a higher improvement in disparity.

$$\text{UIR} = \left(1 - \frac{\sum_{i=1}^{n} \sum_{j=1}^{m} |y'^{(i)[j]} - \check{y}'^{(i)[j]}|}{\sum_{i=1}^{n} \sum_{j=1}^{m} |y^{(i)[j]} - \check{y}^{(i)[j]}|} \right) \times 100\%.$$

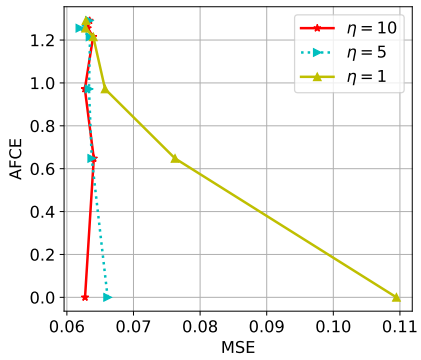
Table 1: Results on Synthetic Data: comparison with two baselines, unfair predictor (UF) and counterfactual fair predictor (CF), in terms of accuracy (MSE) and lookahead counterfactual fairness (AFCE, UIR).

Method	MSE	AFCE	UIR
UF	0.036 ± 0.003	1.296 ± 0.000	0% ± 0
CF	0.520 ± 0.045	1.296 ± 0.000	0% ± 0
Ours ($p_1 = T/2$)	0.064 ± 0.001	0.000 ± 0.0016	100% ± 0

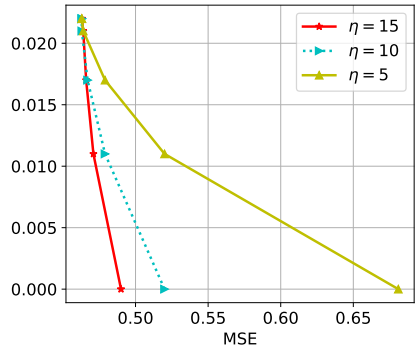
Results. Table 1 illustrates the results when we set $\eta = 10$ and $p_1 = \frac{T}{2}$. The results show that our method can achieve perfect LCF with $p_1 = \frac{T}{2}$. Note that in our experiment, the range of Y is $[0, 3.73]$, and our method and UF can achieve similar MSE. Moreover, our method achieves better performance than the CF method because $\check{Y}$ includes useful predictive information and using it in our predictor can improve performance and decrease the disparity simultaneously. Because both CF and UF do not take into account future outcome Y', $|Y' - \check{Y}'|$ is similar to $|Y - \check{Y}|$, leading to UIR = 0. Based on Corollary 4.1, the value of p_1 can impact the strength of fairness. We examine the tradeoff between accuracy and fairness by changing the value of p_1 from $\frac{T}{512}$ to $\frac{T}{2}$ under different η. Figure 3a shows the MSE as a function of AFCE. The results show that when $\eta = 1$ we can easily control the accuracy-fairness trade-off in our algorithm by adjusting p_1. When η becomes large, we can get a high LCF improvement while maintaining a low MSE. To show how our method impacts a specific individual, we choose the first data point in our test dataset and plot the distribution of factual future status Y' and counterfactual future status $\check{Y}'$ for this specific data point under different methods. Figure 4 illustrates such distributions. It can be seen in the most left plot that there is an obvious gap between factual Y and counterfactual $\check{Y}$. Both UF and CF can not decrease this gap for future outcome Y'. However, with our method, we can observe that the distributions of Y' and $\check{Y}'$ become closer to each other. When $p_1 = \frac{T}{2}$ (the most right plot in Figure 4), the two distributions become the same in the factual and counterfactual worlds.

6.2 Real Data: The Law School Success Dataset

[5]Check Section 4.1 of Kusner et al. (2017) for details on why equation 12 is an empirical estimate of MSE.



(a) Accuracy-fairness trade-off of predictor Eq. 8 on synthetic data: we vary p_1 from $\frac{T}{512}$ to $\frac{T}{2}$ under different η. When $p_1 = \frac{T}{2}$, we attain perfect LCF.



(b) Accuracy-fairness trade-off on the law school dataset: we vary p_1 from $\frac{T}{512}$ to $\frac{T}{2}$ under different η. When $p_1 = \frac{T}{2}$, we attain perfect LCF.

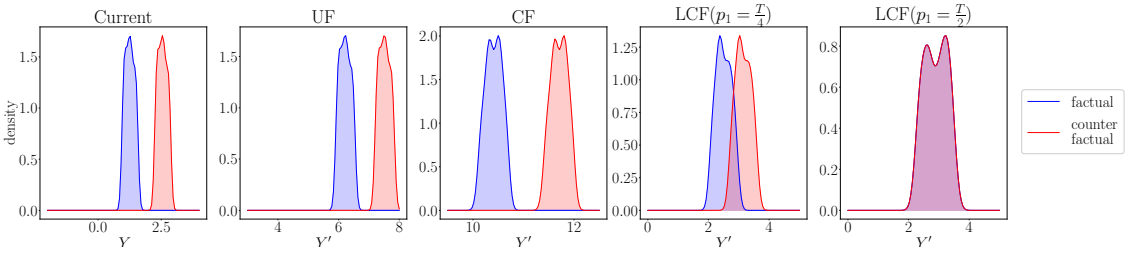


Figure 4: Density plot for Y' and $\check{Y}'$ in synthetic data. For a chosen data point, we sampled a batch of U under the conditional distribution of it and plot the distribution of Y' and $\check{Y}'$.

We further measure the performance of our proposed method using the Law School Admission Dataset Wightman (1998). In this experiment, the objective is to forecast the first-year average grades (FYA) of students in law school using their undergraduate GPA and LSAT scores.

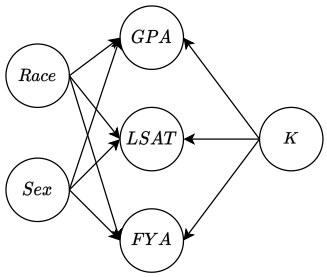


Figure 5: Causal model for the Law School Dataset.

Dataset. The dataset consists of 21,791 records. Each record is characterized by 4 attributes: Sex (S), Race (R), UGPA (G), LSAT (L), and FYA (F). Both Sex and Race are categorical in nature. The Sex attribute can be either male or female, while Race can be Amerindian, Asian, Black, Hispanic, Mexican, Puerto Rican, White, or other. The UGPA is a continuous variable ranging from 0 to 4. LSAT is an integer-based attribute with a range of $[0, 60]$. FYA, which is the target variable for prediction, is a real number ranging from -4 to 4 (it has been normalized). In this study, we consider S as the sensitive attribute, while R, G, and L are treated as features.

Causal Model. We adopt the causal model as presented in Kusner et al. (2017), which can be visualized in Figure 5. In this causal graph, K represents an unobserved variable, which can be interpreted as *knowledge*. Thus, the model suggests that students' grades (UGPA, LSAT, FYA) are influenced by their sex, race, and underlying knowledge. We assume that the prior distribution for K follows a normal distribution, denoted as $\mathcal{N}(0, 1)$. We adopt the same structural equations as Kusner et al. (2017):

$$\begin{aligned} G &= \mathcal{N}(w_G^K K + w_G^R R + w_G^S S + b_G, \sigma_G), \\ L &= \text{Poisson}\left(\exp\left\{w_L^K K + w_L^R R + w_L^S S + b_L\right\}\right), \\ F &= \mathcal{N}(w_F^K K + w_F^R R + w_F^S S, 1). \end{aligned}$$

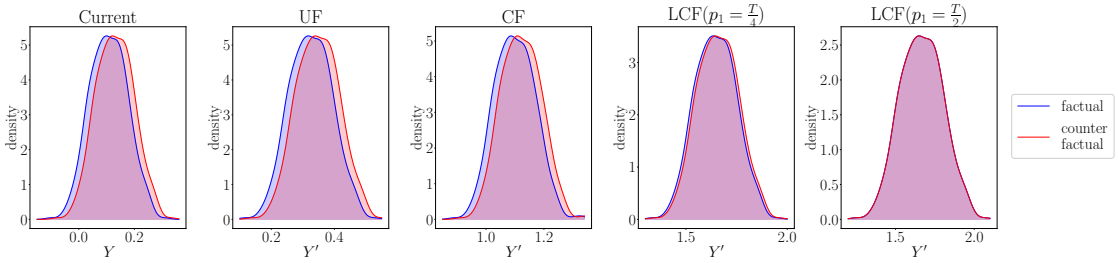


Figure 6: Density plot for F' and $\check{F}'$ in law school data. For a chosen data point, we sampled K from the conditional distribution of K and plot the distribution of F' and $\check{F}'$.

Implementation Details. Note that race is an immutable characteristic. Therefore, we assume that the individuals only adjust their knowledge K in response to the prediction model $\hat{Y}$. That is $K' = K + \eta \nabla_K \hat{Y}$. In contrast to synthetic data, the parameters of structural equations are unknown, and we have to use the training dataset to estimate them. Following the approach of Kusner et al. (2017), we assume that G and F adhere to Gaussian distributions centered at $w_G^K K + w_G^R R + w_G^S S + b_G$ and $w_F^K K + w_F^R R + w_F^S S$, respectively. Note that L is an integer, and it follows a Poisson distribution with the parameter $\exp\{w_L^K K + w_L^R R + w_L^S S + b_L\}$. Using the Markov chain Monte Carlo (MCMC) method Geyer (1992), we can estimate the parameters and the conditional distribution of K given (R, S, G, L). For each given data, we sampled $m = 500$ different k's from this conditional distribution. We partitioned the data into training, validation, and test sets with 60%/20%/20% ratio.

Table 2: Results on Law School Data: comparison with two baselines, unfair predictor (UF) and counterfactual fair predictor (CF), in terms of accuracy (MSE) and lookahead counterfactual fairness (AFCE, UIR).

Method	MSE	AFCE	UIR
UF	0.393 ± 0.046	0.026 ± 0.003	$0\% \pm 0$
CF	0.496 ± 0.051	0.026 ± 0.003	$0\% \pm 0$
Ours ($p_1 = T/4$)	0.493 ± 0.049	0.013 ± 0.002	$50\% \pm 0$
Ours ($p_1 = T/2$)	0.529 ± 0.049	0.000 ± 0.000	$100\% \pm 0$

Results. Table 2 illustrates the results with $\eta = 10$ and $p_1 = \frac{T}{4}$ and $p_1 = \frac{T}{2}$ where $T = 1/(w_K^F)^2$. The results show that our method achieves a similar MSE as the CF predictor. However, it can improve AFCE significantly compared to the baselines. Figure 6 shows the distribution of Y and Y' for the first data point in the test set in the factual and counterfactual worlds. Under the UF and CF predictor, the disparity between factual and factual Y' remains similar to the disparity between factual and counterfactual Y. On the other hand, the disparity between factual and counterfactual Y' under our algorithms gets better for both $p_1 = T/2$ and $p_1 = T/4$. Figure 3b demonstrates that for the law school dataset, the trade-off between MSE and AFCE can be adjusted by changing hyperparameter p_1. Figure 6 show the factual and counterfactual distributions in real data experiment. It can be seen that our method is the only way that can decrease the gap between Y' and $\check{Y}'$ in an obvious way.

7 Conclusion

This work studied the impact of ML decisions on individuals' future status using a counterfactual inference framework. We observed that imposing the CF predictor may not decrease the group disparity in individuals' future status. We thus introduced the lookahead counterfactual fairness (LCF) notion, which takes into account the downstream effects of ML models and requires the individual future status to be counterfactually fair. We proposed a method to train an ML model under LCF and evaluated the method through empirical studies on synthetic and real data.

Impact Statement

This paper advances the field of fair machine learning by studying the downstream effects of counterfactually fair (CF) predictors. It highlights the risks of using CF predictors without considering the downstream effects on individuals. Although this paper proposed LCF, a new fairness notion that requires individual future status (after responding to ML models) to be counterfactual fair, we emphasize that LCF is not always the appropriate notion to consider to pursue fairness. When using the proposed methods in the paper, we need to first check the causal structure carefully and ensure all the conditions are met.

References

Yahav Bechavod, Chara Podimata, Steven Wu, and Juba Ziani. Information discrepancy in strategic learning. In *International Conference on Machine Learning*, pp. 1691–1715, 2022.

Micah D Carroll, Anca Dragan, Stuart Russell, and Dylan Hadfield-Menell. Estimating and penalizing induced preference shifts in recommender systems. In *International Conference on Machine Learning*, pp. 2686–2708. PMLR, 2022.

Silvia Chiappa. Path-specific counterfactual fairness. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 33, pp. 7801–7808, 2019.

Roxana Daneshjou, Kailas Vodrahalli, Weixin Liang, Roberto A Novoa, Melissa Jenkins, Veronica Rotemberg, Justin Ko, Susan M Swetter, Elizabeth E Bailey, Olivier Gevaert, et al. Disparities in dermatology ai: Assessments using diverse clinical images. *arXiv preprint arXiv:2111.08006*, 2021.

Jeffrey Dastin. Amazon scraps secret ai recruiting tool that showed bias against women. `http://reut.rs/2MXzkly`, 2018.

Sarah Dean and Jamie Morgenstern. Preference dynamics under personalized recommendations. In *Proceedings of the 23rd ACM Conference on Economics and Computation*, pp. 795–816, 2022.

Virginie Do, Sam Corbett-Davies, Jamal Atif, and Nicolas Usunier. Online certification of preference-based fairness for personalized recommender systems. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 36, pp. 6532–6540, 2022.

Yingqiang Ge, Shuchang Liu, Ruoyuan Gao, Yikun Xian, Yunqi Li, Xiangyu Zhao, Changhua Pei, Fei Sun, Junfeng Ge, Wenwu Ou, et al. Towards long-term fairness in recommendation. In *Proceedings of the 14th ACM international conference on web search and data mining*, pp. 445–453, 2021.

Charles J Geyer. Practical markov chain monte carlo. *Statistical science*, pp. 473–483, 1992.

Madelyn Glymour, Judea Pearl, and Nicholas P Jewell. *Causal inference in statistics: A primer*. John Wiley & Sons, 2016.

Moritz Hardt, Nimrod Megiddo, Christos Papadimitriou, and Mary Wootters. Strategic classification. In *Proceedings of the 2016 ACM conference on innovations in theoretical computer science*, pp. 111–122, 2016a.

Moritz Hardt, Eric Price, and Nati Srebro. Equality of opportunity in supervised learning. *Advances in neural information processing systems*, 29:3315–3323, 2016b.

Thomas Henzinger, Mahyar Karimi, Konstantin Kueffner, and Kaushik Mallik. Runtime monitoring of dynamic fairness properties. In *Proceedings of the 2023 ACM Conference on Fairness, Accountability, and Transparency*, pp. 604–614, 2023a.

Thomas A Henzinger, Mahyar Karimi, Konstantin Kueffner, and Kaushik Mallik. Monitoring algorithmic fairness. *arXiv preprint arXiv:2305.15979*, 2023b.

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

Matt J Kusner, Joshua Loftus, Chris Russell, and Ricardo Silva. Counterfactual fairness. *Advances in neural information processing systems*, 30, 2017.

Lydia T Liu, Sarah Dean, Esther Rolf, Max Simchowitz, and Moritz Hardt. Delayed impact of fair machine learning. In *International Conference on Machine Learning*, pp. 3150–3158. PMLR, 2018.

Lydia T Liu, Ashia Wilson, Nika Haghtalab, Adam Tauman Kalai, Christian Borgs, and Jennifer Chayes. The disparate equilibria of algorithmic decision making when individuals invest rationally. In *Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency*, pp. 381–391, 2020.

Jing Ma, Ruocheng Guo, Aidong Zhang, and Jundong Li. Learning for counterfactual fairness from observational data. In *Proceedings of the 29th ACM SIGKDD Conference on Knowledge Discovery and Data Mining*, pp. 1620–1630, 2023.

John Miller, Smitha Milli, and Moritz Hardt. Strategic classification is causal modeling in disguise. In *International Conference on Machine Learning*, pp. 6917–6926. PMLR, 2020.

Arash Nasr-Esfahany, Mohammad Alizadeh, and Devavrat Shah. Counterfactual identifiability of bijective causal models. *arXiv preprint arXiv:2302.02228*, 2023.

Judea Pearl et al. Models, reasoning and inference. *Cambridge, UK: CambridgeUniversityPress*, 19(2), 2000.

Nir Rosenfeld, Anna Hilgard, Sai Srivatsa Ravindranath, and David C Parkes. From predictions to decisions: Using lookahead regularization. *Advances in Neural Information Processing Systems*, 33:4115–4126, 2020.

Yonadav Shavit, Benjamin Edelman, and Brian Axelrod. Causal strategic linear regression. In *International Conference on Machine Learning*, pp. 8676–8686. PMLR, 2020.

Zeyu Tang, Yatong Chen, Yang Liu, and Kun Zhang. Tier balancing: Towards dynamic fairness over underlying causal factors. *arXiv preprint arXiv:2301.08987*, 2023.

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*, pp. 83–92, 2019.

Linda F Wightman. Lsac national longitudinal bar passage study. lsac research report series. 1998.

Yongkai Wu, Lu Zhang, and Xintao Wu. Counterfactual fairness: Unidentification, bound and algorithm. In *Proceedings of the twenty-eighth international joint conference on Artificial Intelligence*, 2019.

Depeng Xu, Shuhan Yuan, and Xintao Wu. Achieving differential privacy and fairness in logistic regression. In *Companion Proceedings of The 2019 World Wide Web Conference*, pp. 594–599, 2019.

Muhammad Bilal Zafar, Isabel Valera, Manuel Gomez Rodriguez, and Krishna P Gummadi. Fairness beyond disparate treatment & disparate impact: Learning classification without disparate mistreatment. In *Proceedings of the 26th international conference on world wide web*, pp. 1171–1180, 2017.

Xueru Zhang, Mohammadmahdi Khaliligarekani, Cem Tekin, et al. Group retention when using machine learning in sequential decision making: the interplay between user dynamics and fairness. *Advances in Neural Information Processing Systems*, 32:15269–15278, 2019.

Xueru Zhang, Ruibo Tu, Yang Liu, Mingyan Liu, Hedvig Kjellstrom, Kun Zhang, and Cheng Zhang. How do fair decisions fare in long-term qualification? *Advances in Neural Information Processing Systems*, 33: 18457–18469, 2020.

Xueru Zhang, Mohammad Mahdi Khalili, Kun Jin, Parinaz Naghizadeh, and Mingyan Liu. Fairness interventions as (Dis)Incentives for strategic manipulation. In *Proceedings of the 39th International Conference on Machine Learning*, pp. 26239–26264, 2022.

Aoqi Zuo, Susan Wei, Tongliang Liu, Bo Han, Kun Zhang, and Mingming Gong. Counterfactual fairness with partially known causal graph. *Advances in Neural Information Processing Systems*, 35:1238–1252, 2022.

A Proofs

A.1 Proof of Theorem 4.1 and Theorem 4.2

Proof. For any given x, a, we can find the conditional distribution $U_X|X = x, A = a$ and $U_Y|X = x, A = a$ based on causal model $\mathcal{M}$. Consider sample $u = [u_X, u_Y]$ drawn from this conditional distribution. For this sample, we have,

$$\check{x} = \alpha \odot u_X + \beta \check{a}$$

$$\check{y} = w^{\mathrm{T}} \check{x} + \gamma u_Y$$

So, the gradient of $g(\check{y}, u_X, u_Y)$ w.r.t. u_X, u_Y are

$$\nabla_{u_X} g = \frac{\partial g(\check{y}, u_X, u_Y)}{\partial u_X} + \frac{\partial g(\check{y}, u_X, u_Y)}{\partial \check{y}} \odot w \odot \alpha \tag{13}$$

$$\nabla_{u_Y} g = \frac{\partial g(\check{y}, u_X, u_Y)}{\partial u_Y} + \frac{\partial g(\check{y}, u_X, u_Y)}{\partial \check{y}} \gamma. \tag{14}$$

Then, y' can be calculated using response r as follows,

$$y' = y + \eta w^{\mathrm{T}} \left(\alpha \odot \frac{\partial g(\check{y}, u_X, u_Y)}{\partial u_X} \right) + \eta \, ||w \odot \alpha| \, |_2^2 \frac{\partial g(\check{y}, u_X, u_Y)}{\partial \check{y}} + \eta\gamma \frac{\partial g(\check{y}, u_X, u_Y)}{\partial u_Y} + \eta\gamma^2 \frac{\partial g(\check{y}, u_X, u_Y)}{\partial \check{y}}. \tag{15}$$

Similarly, we can calculate counterfactual value $\check{y}'$ as follows,

$$\check{y}' = \check{y} + \eta w^{\mathrm{T}} \left(\alpha \odot \frac{\partial g(y, u_X, u_Y)}{\partial u_X} \right) + \eta \, ||w \odot \alpha| \, |_2^2 \frac{\partial g(y, u_X, u_Y)}{\partial y} + \eta\gamma \frac{\partial g(y, u_X, u_Y)}{\partial u_Y} + \eta\gamma^2 \frac{\partial g(y, u_X, u_Y)}{\partial y} \tag{16}$$

Note that the following hold for g,

$$\frac{\partial g(\check{y}, u_X, u_Y)}{\partial u_X} = \frac{\partial g(y, u_X, u_Y)}{\partial u_X} \tag{17}$$

$$\frac{\partial g(\check{y}, u_X, u_Y)}{\partial u_Y} = \frac{\partial g(y, u_X, u_Y)}{\partial u_Y} \tag{18}$$

Thus,

$$|\check{y}' - y'| = \left| \check{y} - y + \eta \left(||w \odot \alpha||_2^2 + \gamma^2 \right) \left(\frac{\partial g(y, u_X, u_Y)}{\partial y} - \frac{\partial g(\check{y}, u_X, u_Y)}{\partial \check{y}} \right) \right| \tag{19}$$

Given above equation, now we can prove Theorem 4.1 and Corollary 4.2,

- For g in Theorem 4.1, we have,

$$g(\check{y}, u_X, u_Y) = p_1 \check{y}^2 + p_2 \check{y} + p_3 + h(u) \tag{20}$$

$$\frac{\partial g(\check{y}, u_X, u_Y)}{\partial \check{y}} = 2 p_1 \check{y}. \tag{21}$$

  Equations 19 and 21 together imply that,

$$|y' - \check{y}'| = |y - \check{y} + \check{y} - y| = 0 \tag{22}$$

  Since, for any realization of u, the above equation holds, we can conclude that the following holds,

$$\Pr(\hat{Y}_{A \leftarrow a}(U) = y | X = x, A = a) = \Pr(\hat{Y}_{A \leftarrow \check{a}}(U) = y | X = x, A = a) \tag{23}$$

- For g in Theorem 4.2, since $g(\check{y}, u_x, u_y)$ is strictly convex in $\check{y}$, we have,

$$(\check{y} - y)\left(\frac{\partial g(y, u_X, u_Y)}{\partial y} - \frac{\partial g(\check{y}, u_X, u_Y)}{\partial y}\right) < 0 \tag{24}$$

Note that derivative of $g(\check{y}, u_x, u_y)$ with respect to $\check{y}$ is K-Lipschitz continuous in $\check{y}$,

$$\left|\frac{\partial g(y, u_X, u_Y)}{\partial y} - \frac{\partial g(\check{y}, u_X, u_Y)}{\partial \check{y}}\right| < \frac{2|y - \check{y}|}{\eta(||w \odot \alpha||_2^2 + \gamma^2)} \tag{25}$$

we proved that

$$|y' - \check{y}'| < |y - \check{y}| \tag{26}$$

So we have

$$\Pr(\{|Y'_{A\leftarrow a}(U) - Y'_{A\leftarrow \check{a}}(U)| < |Y_{A\leftarrow a}(U) - Y_{A\leftarrow \check{a}}(U)|\}|X = x, A = a) = 1 \tag{27}$$

□

A.2 Theorem 4.2 for non-binary A

Let $\{a\} \cup \{\check{a}^{[1]}, \check{a}^{[2]}, ..., \check{a}^{[m]}\}$ be a set of all possible values for A. Let $\check{Y}^{[j]}$ be the counterfactual random variable associated with $\check{a}^{[j]}$ given observation $X = x$ and $A = a$. Then, $g\left(\frac{\check{Y}^{[1]} + ... \check{Y}^{[m]}}{m}, U\right)$ satisfies LCF, where g defined in Theorem 4.2.

Proof. For any given x, a, we assume the set of counterfactual a is $\{\check{a}^{[1]}, \check{a}^{[2]}, ..., \check{a}^{m}\}$. Consider sample $u = [u_X, u_Y]$ drawn from the condition distribution of $U_X|X = x, A = a$ and $U_Y|X = x, A = a$, with a predictor $g\left(\frac{\check{y}^{[1]} + \cdots \check{y}^{[m]}}{m}, u\right)$, use the same way in A.1, we can get

$$\left|\check{y}'^{[j]} - y'\right| = \left|\check{y}^{[j]} - y + \eta(||w \odot \alpha||^2 + \gamma^2)\left(\frac{\partial g(\check{y}^{[1]} + \cdots \check{y}^{[m]}, u)}{\partial \check{y}^{[1]} + \cdots \check{y}^{[m]}} - \frac{\partial g(y + \check{y}^{[1]} + \cdots + \check{y}^{[j-1]} + \check{y}^{[j+1]} \cdots \check{y}^{[m]}, u)}{\partial y + \check{y}^{[1]} + \cdots \check{y}^{[j-1]} + \check{y}^{[j+1]} \cdots \check{y}^{[m]}}\right)\right| \tag{28}$$

When $y > \check{y}^{[j]}$, we have

$$\check{y}^{[1]} + \cdots \check{y}^{[m]} < y + \check{y}^{[1]} + \cdots \check{y}^{[j-1]} + \check{y}^{[j+1]} \cdots \check{y}^{[m]} \tag{29}$$

and when $y < \check{y}^{[j]}$,

$$\check{y}^{[1]} + \cdots \check{y}^{[m]} > y + \check{y}^{[1]} + \cdots \check{y}^{[j-1]} + \check{y}^{[j+1]} \cdots \check{y}^{[m]} \tag{30}$$

Because g is strictly convex and Lipschitz continuous, we have

$$|\check{y}'^{[j]} - y'| < |\check{y}^{[j]} - y| \tag{31}$$

So we proved that, for any $j \in \{1, 2, ..., m\}$

$$\Pr(\{|Y'_{A\leftarrow a}(U) - Y'_{A\leftarrow \check{a}^{[j]}}(U)| < |Y_{A\leftarrow a}(U) - Y_{A\leftarrow \check{a}^{[j]}}(U)|\}|X = x, A = a) = 1 \tag{32}$$

□

A.3 Proof of Theorem 4.3

Proof. For any given x, a, we can find the conditional distribution $U_X|X = x, A = a$ and $U_Y|X = x, A = a$ based on causal model $\mathcal{M}$. Consider a sample $u = [u_X, u_Y]$ drawn from this conditional distribution. For this sample, we have

$$\check{x} = f(\alpha u_X + \beta a) \tag{33}$$

$$\check{y} = w\check{x} + \gamma u_Y \tag{34}$$

So, the gradient of $g(\check{y}, u_X, u_Y)$ w.r.t. u_X, u_Y are

$$\nabla_{u_X} g = \lambda_1 w\alpha f'(\alpha u_X + \beta\check{a}) + \lambda_2 \frac{\partial g(u_X, u_Y)}{\partial u_X} \tag{35}$$

$$\nabla_{u_Y} g = \lambda_1 \gamma + \lambda_2 \frac{\partial g(u_X, u_Y)}{\partial u_Y} \tag{36}$$

Then y' can be calculated using response r as follows,

$$\begin{aligned} y' =& wf\left(\alpha u_X + \lambda_1 w\eta\alpha^2 f'(\alpha u_X + \beta\check{a}) + \lambda_2\alpha\frac{\partial g(u_X, u_Y)}{\partial u_X} + \beta a\right) \\ &+ \gamma\left(u_Y + \lambda_1\eta\gamma + \lambda_2\eta\frac{\partial g(u_X, u_Y)}{\partial u_Y}\right) \end{aligned} \tag{37}$$

Similarly, we can calculate counterfactual value $\check{y}'$ as follows,

$$\begin{aligned} \check{y}' =& wf\left(\alpha u_X + \lambda_1 w\eta\alpha^2 f'(\alpha u_X + \beta a) + \lambda_2\alpha\frac{\partial g(u_X, u_Y)}{\partial u_X} + \beta\check{a}\right) \\ &+ \gamma\left(u_Y + \lambda_1\eta\gamma + \lambda_2\eta\frac{\partial g(u_X, u_Y)}{\partial u_Y}\right) \end{aligned} \tag{38}$$

So,

$$|y' - \check{y}'| = |w| \cdot \left| f\left(\alpha u_X + \lambda_1 w\eta\alpha^2 f'(\alpha u_X + \beta\check{a}) + \lambda_2\alpha\frac{\partial g(u_X, u_Y)}{\partial u_X} + \beta a\right) - f\left(\alpha u_X + \lambda_1 w\eta\alpha^2 f'(\alpha u_X + \beta a) + \lambda_2\alpha\frac{\partial g(u_X, u_Y)}{\partial u_X} + \beta\check{a}\right)\right| \tag{39}$$

Since f is strictly convex,

$$[(\alpha u_X + \beta a) - (\alpha u_X - \beta\check{a})] \cdot [f'(\alpha u_X + \beta a) - f'(\alpha u_X - \beta\check{a})] < 0 \tag{40}$$

And because f' is Lipschitz continuous,

$$\begin{aligned} &\left|\left(\alpha u_X + \lambda_1 w\eta\alpha^2 f'(\alpha u_X + \beta\check{a}) + \lambda_2\alpha\frac{\partial g(u_X, u_Y)}{\partial u_X} + \beta a\right) - \left(\alpha u_X + \lambda_1 w\eta\alpha^2 f'(\alpha u_X + \beta a) + \lambda_2\alpha\frac{\partial g(u_X, u_Y)}{\partial u_X} + \beta\check{a}\right)\right| \\ &< |(\alpha u_X + \beta a) - (\alpha u_X - \beta\check{a})| \end{aligned} \tag{41}$$

From the second property of f, we know that

$$|y' - \check{y}'| < |wf(\alpha u_X + \beta a) - wf(\alpha u_X - \beta\check{a})| \tag{42}$$

which is exactly $|y - \check{y}|$. □

A.4 Proof of Theorem 4.4

Proof. From the causal functions defined in Theorem 4.4, given any x, a, we can find the conditional distribution $U_X|X = x, A = a$ and $U_Y|X = x, A = a$. Similar to the proof of Theorem 4.2, we have

$$\check{x} = \check{a}(\alpha \odot u_X + \beta) \tag{43}$$

$$\check{y} = w^{\mathrm{T}}\check{x} + \gamma u_Y \tag{44}$$

So, the gradient of $g(\check{y})$ w.r.t u_X, u_Y are

$$\nabla_{u_X} g = \frac{\partial g(\check{y})}{\partial \check{y}} \check{a} w \odot \alpha \tag{45}$$

$$\nabla_{u_Y} g = \frac{\partial g(\check{y})}{\partial \check{y}} \gamma \tag{46}$$

Then, y' can be calculated using the response r as follows,

$$y' = y + \eta \left(a\check{a}||w \odot \alpha||^2 + \gamma^2\right) \frac{\partial g(\check{y})}{\partial \check{y}} \tag{47}$$

In the counterfactual world,

$$\check{y}' = \check{y} + \eta(\check{a}a||w \odot \alpha||^2 + \gamma^2)\frac{\partial g(y)}{\partial y} \tag{48}$$

So,

$$|y' - \check{y}'| = \left| y - \check{y} + \eta(a\check{a}||w \odot \alpha||^2 + \gamma^2)(\frac{\partial g(\check{y})}{\partial \check{y}} - \frac{\partial g(y)}{\partial y}) \right| \tag{49}$$

Because A is a binary attributes, we have

$$a\check{a} = a_1 a_2 \tag{50}$$

From the property of g, we have

$$(y - \check{y})(\frac{\partial g(\check{y})}{\partial \check{y}} - \frac{\partial g(y)}{\partial y}) < 0 \tag{51}$$

Note that the derivate of $g(\check{y})$ is K-Lipschitz continuous,

$$\left| \frac{\partial g(\check{y})}{\partial \check{y}} - \frac{\partial g(y)}{\partial y} \right| < \frac{2|\check{y} - y|}{\eta(a\check{a}||w \odot \alpha||_2^2 + \gamma^2)} \tag{52}$$

which is to say, for every u sampled from the conditional distribution, $|\check{y}' - y'| < |\check{y} - y|$. So we proved

$$\Pr(\{|Y'_{A\leftarrow a}(U) - Y'_{A\leftarrow \check{a}}(U)| < |Y_{A\leftarrow a}(U) - Y_{A\leftarrow \check{a}}(U)|\}|X = x, A = a) = 1 \tag{53}$$

□

A.5 Proof of Theorem 5.1

Proof. For any given x, a we can find the consitional distribution $U_X|X = x, A = a$ and $U_Y|X = x, A = a$ based on causal model $\mathcal{M}$. Consider sample $u = [u_X, u_Y]$ drawn from this conditional distribution. For this sample, we have,

$$\check{x}_{\mathcal{P}_{\mathcal{G}_A}} = \alpha_{\mathcal{P}_{\mathcal{G}_A}} \odot u_{X_{\mathcal{P}_{\mathcal{G}_A}}} + \beta_{\mathcal{P}_{\mathcal{G}_A}} \check{a} \tag{54}$$

$$\check{y}_{PD} = w_{\mathcal{P}_{\mathcal{G}_A}}^{\mathrm{T}} \check{x}_{\mathcal{P}_{\mathcal{G}_A}} + w_{\mathcal{P}^c_{\mathcal{G}_A}}^{\mathrm{T}} x_{\mathcal{P}^c_{\mathcal{G}_A}} + \gamma u_Y \tag{55}$$

So, the gradient of $g(\check{y}_{PD}, u_{X_{\mathcal{P}_{\mathcal{G}_A}}}, u_{X_{\mathcal{P}^c_{\mathcal{G}_A}}}, u_Y)$ w.r.t. $u_{X_{\mathcal{P}_{\mathcal{G}_A}}}, u_{X_{\mathcal{P}^c_{\mathcal{G}_A}}}, u_Y$ are

$$\nabla_{u_{X_{\mathcal{P}_{\mathcal{G}_A}}}} g = \frac{\partial g(\check{y}_{PD}, u_{X_{\mathcal{P}_{\mathcal{G}_A}}}, u_{X_{\mathcal{P}^c_{\mathcal{G}_A}}}, u_Y)}{\partial u_{X_{\mathcal{P}_{\mathcal{G}_A}}}} + \frac{\partial g(\check{y}_{PD}, u_{X_{\mathcal{P}_{\mathcal{G}_A}}}, u_{X_{\mathcal{P}^c_{\mathcal{G}_A}}}, u_Y)}{\partial \check{y}} \odot w_{\mathcal{P}_{\mathcal{G}_A}} \odot \alpha_{\mathcal{P}_{\mathcal{G}_A}} \tag{56}$$

$$\nabla_{u_{X_{\mathcal{P}^c_{\mathcal{G}_A}}}} g = \frac{\partial g(\check{y}_{PD}, u_{X_{\mathcal{P}_{\mathcal{G}_A}}}, u_{X_{\mathcal{P}^c_{\mathcal{G}_A}}}, u_Y)}{\partial u_{X_{\mathcal{P}^c_{\mathcal{G}_A}}}} + \frac{\partial g(\check{y}_{PD}, u_{X_{\mathcal{P}_{\mathcal{G}_A}}}, u_{X_{\mathcal{P}^c_{\mathcal{G}_A}}}, u_Y)}{\partial \check{y}} \odot w_{\mathcal{P}^c_{\mathcal{G}_A}} \odot \alpha_{\mathcal{P}^c_{\mathcal{G}_A}} \tag{57}$$

$$\nabla_{u_Y} g = \frac{\partial g(\check{y}, u_{X_{\mathcal{P}_{\mathcal{G}_A}}}, u_{X_{\mathcal{P}^c_{\mathcal{G}_A}}}, u_Y)}{\partial u_Y} + \frac{\partial g(\check{y}, u_{X_{\mathcal{P}_{\mathcal{G}_A}}}, u_{X_{\mathcal{P}^c_{\mathcal{G}_A}}}, u_Y)}{\partial \check{y_{PD}}} \gamma. \tag{58}$$

Then, y' can be calculated using response r as follows,

$$\begin{aligned} y' =& y + \eta w_{\mathcal{P}_{\mathcal{G}_A}}^{\mathrm{T}} \left(\alpha \odot \frac{\partial g(\check{y}_{PD}, u_{X_{\mathcal{P}_{\mathcal{G}_A}}}, u_{X_{\mathcal{P}^c_{\mathcal{G}_A}}}, u_Y)}{\partial u_{X_{\mathcal{P}_{\mathcal{G}_A}}}} \right) + \eta w_{\mathcal{P}^c_{\mathcal{G}_A}}^{\mathrm{T}} \left(\alpha \odot \frac{\partial g(\check{y}_{PD}, u_{X_{\mathcal{P}_{\mathcal{G}_A}}}, u_{X_{\mathcal{P}^c_{\mathcal{G}_A}}}, u_Y)}{\partial u_{X_{\mathcal{P}^c_{\mathcal{G}_A}}}} \right) + \\ & \eta \left\| w_{\mathcal{P}_{\mathcal{G}_A}} \odot \alpha_{\mathcal{P}_{\mathcal{G}_A}} \right\|_2^2 \frac{\partial g(\check{y}_{PD}, u_{X_{\mathcal{P}_{\mathcal{G}_A}}}, u_{X_{\mathcal{P}^c_{\mathcal{G}_A}}}, u_Y)}{\partial \check{y}_{PD}} + \eta \left\| w_{\mathcal{P}^c_{\mathcal{G}_A}} \odot \alpha_{\mathcal{P}^c_{\mathcal{G}_A}} \right\|_2^2 \frac{\partial g(\check{y}_{PD}, u_{X_{\mathcal{P}_{\mathcal{G}_A}}}, u_{X_{\mathcal{P}^c_{\mathcal{G}_A}}}, u_Y)}{\partial \check{y}_{PD}} + \\ & \eta\gamma \frac{\partial g(\check{y}_{PD}, u_{X_{\mathcal{P}_{\mathcal{G}_A}}}, u_Y, u_Y)}{\partial u_Y} + \eta\gamma^2 \frac{\partial g(\check{y}_{PD}, u_{X_{\mathcal{P}_{\mathcal{G}_A}}}, u_Y, u_Y)}{\partial \check{y}_{PD}}. \end{aligned} \tag{59}$$

Similarly, we can calculate path-dependent counterfactual value $\check{y}'_{PD}$ as follows,

$$\begin{aligned} \check{y}'_{PD} =& \check{y}_{PD} + \eta w_{\mathcal{P}_{\mathcal{G}_A}}^{\mathrm{T}} \left(\alpha \odot \frac{\partial g(y, u_{X_{\mathcal{P}_{\mathcal{G}_A}}}, u_{X_{\mathcal{P}^c_{\mathcal{G}_A}}}, u_Y)}{\partial u_{X_{\mathcal{P}_{\mathcal{G}_A}}}} \right) + \eta w_{\mathcal{P}^c_{\mathcal{G}_A}}^{\mathrm{T}} \left(\alpha \odot \frac{\partial g(y, u_{X_{\mathcal{P}_{\mathcal{G}_A}}}, u_{X_{\mathcal{P}^c_{\mathcal{G}_A}}}, u_Y)}{\partial u_{X_{\mathcal{P}^c_{\mathcal{G}_A}}}} \right) + \\ & \eta \left\| w_{\mathcal{P}_{\mathcal{G}_A}} \odot \alpha_{\mathcal{P}_{\mathcal{G}_A}} \right\|_2^2 \frac{\partial g(y, u_{X_{\mathcal{P}_{\mathcal{G}_A}}}, u_{X_{\mathcal{P}^c_{\mathcal{G}_A}}}, u_Y)}{\partial y} + \eta \left\| w_{\mathcal{P}^c_{\mathcal{G}_A}} \odot \alpha_{\mathcal{P}^c_{\mathcal{G}_A}} \right\|_2^2 \frac{\partial g(y, u_{X_{\mathcal{P}_{\mathcal{G}_A}}}, u_{X_{\mathcal{P}^c_{\mathcal{G}_A}}}, u_Y)}{\partial y} + \\ & \eta\gamma \frac{\partial g(y, u_{X_{\mathcal{P}_{\mathcal{G}_A}}}, u_Y, u_Y)}{\partial u_Y} + \eta\gamma^2 \frac{\partial g(y, u_{X_{\mathcal{P}_{\mathcal{G}_A}}}, u_Y, u_Y)}{\partial y}. \end{aligned} \tag{60}$$

Thus,

$$\begin{aligned} & |\check{y}'_{PD} - y'| = \\ & \left| \check{y}_{PD} - y + \eta(||w_{\mathcal{P}_{\mathcal{G}_A}} \odot \alpha_{\mathcal{P}_{\mathcal{G}_A}}||_2^2 + ||w_{\mathcal{P}^c_{\mathcal{G}_A}} \odot \alpha_{\mathcal{P}^c_{\mathcal{G}_A}}||_2^2 + \gamma^2) \left(\frac{\partial g(y, u_{X_{\mathcal{P}_{\mathcal{G}_A}}}, u_{X_{\mathcal{P}^c_{\mathcal{G}_A}}}, u_Y)}{\partial y} - \frac{\partial g(\check{y}_{PD}, u_{X_{\mathcal{P}_{\mathcal{G}_A}}}, u_{X_{\mathcal{P}^c_{\mathcal{G}_A}}}, u_Y)}{\partial \check{y}_{PD}} \right) \right| \end{aligned} \tag{61}$$

We denote $p_1 = \frac{1}{2\eta(||w_{\mathcal{P}_{\mathcal{G}_A}} \odot \alpha_{\mathcal{P}_{\mathcal{G}_A}}||_2^2 + ||w_{\mathcal{P}^c_{\mathcal{G}_A}} \odot \alpha_{\mathcal{P}^c_{\mathcal{G}_A}}||_2^2 + \gamma^2)}$. Since we know the partial gradient of $g(\check{y}_{PD}, u_{X_{\mathcal{P}_{\mathcal{G}_A}}}, u_{X_{\mathcal{P}^c_{\mathcal{G}_A}}}, u_Y)$ is $2p_1\check{y}_{PD}$, we know that $|y' - \check{y}'_{PD}| = 0$. Since for any realization of u, the equation holds, we can conclude that the path-dependent LCF holds. □

B Parameters for Synthetic Data Simulation

When generating the synthetic data, we used α = [0.37454012, 0.95071431, 0.73199394, 0.59865848, 0.15601864, 0.15599452, 0.05808361, 0.86617615, 0.60111501, 0.70807258]$^{\mathrm{T}}$. β =[0.02058449, 0.96990985, 0.83244264, 0.21233911, 0.18182497, 0.18340451, 0.30424224, 0.52475643, 0.43194502, 0.29122914]$^{\mathrm{T}}$. w =[0.61185289, 0.13949386, 0.29214465, 0.36636184, 0.45606998, 0.78517596, 0.19967378, 0.51423444, 0.59241457, 0.04645041]$^{\mathrm{T}}$. $\gamma = 0.60754485$ (These values are generated randomly).