Causal Adversarial Perturbations for Individual Fairness and Robustness in Heterogeneous Data Spaces

Ahmad-Reza Ehyaei [1], Kiarash Mohammadi [2 3], Amir-Hossein Karimi [1], Samira Samadi [1], Golnoosh Farnadi [2 3 4]

[1] Max Planck Institute for Intelligent Systems Tübingen, Tübingen, Germany
[2] Université de Montréal, Montréal, Canada
[3] Mila - Québec AI Institute, Montréal, Canada
[4] McGill University, Montréal, Canada
ahmadreza.ehyaei@uni-tuebingen.de, kiarash.mohammadi@mila.quebec, amir@tue.mpg.de, ssamadi@tuebingen.mpg.de, farnadig@mila.quebec

Abstract

As responsible AI gains importance in machine learning algorithms, properties such as fairness, adversarial robustness, and causality have received considerable attention in recent years. However, despite their individual significance, there remains a critical gap in simultaneously exploring and integrating these properties. In this paper, we propose a novel approach that examines the relationship between individual fairness, adversarial robustness, and structural causal models in heterogeneous data spaces, particularly when dealing with discrete sensitive attributes. We use causal structural models and sensitive attributes to create a fair metric and apply it to measure semantic similarity among individuals. By introducing a novel causal adversarial perturbation and applying adversarial training, we create a new regularizer that combines individual fairness, causality, and robustness in the classifier. Our method is evaluated on both real-world and synthetic datasets, demonstrating its effectiveness in achieving an accurate classifier that simultaneously exhibits fairness, adversarial robustness, and causal awareness.

1 Introduction

In the ever-evolving landscape of machine learning, responsible AI has emerged as a pivotal focal point. Attributes such as fairness, adversarial robustness, and causality have taken center stage, each carrying its own weight in shaping ethical and socially reliable AI systems. Yet, the prevailing discourse often falls short of comprehensively addressing these dimensions in a unified manner, leaving a gap in our understanding of how they intersect and influence each other.

Notably, within the realm of fairness, the scientific community has proposed various notions of fairness, broadly categorized as group fairness, examining model's performance across different demographic groups, and individual fairness, assessing model's performance on different individuals (Pessach and Shmueli 2022; Mehrabi et al. 2021). While group fairness can guarantee similar classification performance on different demographic groups, it does not always guarantee individual fairness, i.e., that similarly qualified individuals, receive similar outcomes (Binns 2020).

Various formulations of individual fairness have been proposed in the literature, including Lipschitz (Dwork et al. 2012) and ϵ-δ (John, Vijaykeerthy, and Saha 2020). These formulations presume existence of a metric on the individuals that capture their (qualification) similarity. Such a similarity metric, by definition, is assumed to capture relevant features in the individuals that are important for the classification outcome, and to ignore features that should be irrelevant. For instance, in a hiring scenario, the similarity metric between the individuals could consider work experience and academic degree but should not take into account sensitive attributes. Due to this inherent fairness property in the definition of similarity metric, such metric is often referred to as a *fair metric*. Various similarity functions have been proposed as fair metrics, including weighted ℓ_p norms, Mahalanobis distance, and feature embedding (Benussi et al. 2022).

In the domain of responsible AI, the study of causality is paramount, as the problems addressed often manipulate systems where inter-variable relations are governed by cause-and-effect mechanisms. In fact, many such sensitive attributes such as socio-economic status broadly affect the opportunities presented to individuals, which fair AI aims to rectify. Despite the introduction of causality as a critical lens in fairness literature (Kusner et al. 2017), the aforementioned definitions of fair metrics, and the studied domains therein, have struggled to fully encompass the notion of robustness. While causal reasoning offers a foundation for addressing fairness, the inherent challenges of adversarial perturbations and their potential influence on fairness have remained largely unexplored.

In response to this gap, the initial step in our study is to propose a framework that defines a fair metric based on the functional structure of the underlying structural causal model. We propose a mathematical approach for protecting sensitive attributes by employing the concept of a pseudometric. Our proposed methodology enables the development of a fair metric that effectively mitigates bias across different levels of sensitive features in heterogeneous data spaces. Using our proposed fair metric we establish a causal adversarial perturbation (CAP) set to identify similar individuals. Subsequently, we analyze the characteristics of the CAP and its relationship with counterfactual fairness and adversarial robustness. Finally, we define a novel causal individual fairness notion based on the fair metric, which we refer to as **CAPI** fairness.

After formulating **CAPI** fairness, the next step is to train a classifier that guarantees this notion. This objective can be accomplished by applying bias mitigation methods during the in-processing stage. We ground our theoretical contributions in practicality by demonstrating the implementation of **CAPI** fairness within different classifiers and datasets. We initially examine the underlying cause of unfairness by defining the concept of *unfair area*. We compute the unfair area for a linear model and design a *post-processing* approach to obtain counterfactual fairness. Subsequently, to attain **CAPI** fairness which is a stronger notion, we employ adversarial learning techniques (Madry et al. 2017) and present the first in-processing approach of **CAPI** fairness regularizer. To the best of our knowledge, this work is the first work that simultaneously addresses adversarial robustness, individual fairness, and causal structures in training a machine learning model. Our contributions are as follows:

- **Causal Fair Metric (§ 3.1).** Our primary contribution involves the establishment of a semi-latent space for the formulation of a fair metric. The introduction of this semi-latent space is essential to counteract the inherent bias embedded in the structural causal model. Achieving fairness necessitates the assurance that all potential interventions related to varying levels of sensitive attributes are considered. Based on this concept, we develop a fair metric that not only demonstrates effectiveness across diverse sensitive attributes but also incorporates the intricate aspects of the causal framework.
- **Causal Adversarial Perturbation (§ 3.2)** Building upon the foundation laid by our proposed causal fair metric, we introduce the concept of the causal adversarial perturbation. By leveraging the insights gained from our fair metric, causal adversarial perturbation emerges as a mechanism capable of capturing the similarity set in the presence of causal models.
- **CAPI Fairness (§ 3.3)** Our third contribution entails the introduction of a novel fairness notion **CAPI** fairness. This concept emerges as a pivotal bridge that seamlessly connects individual fairness, adversarial robustness, and the underpinnings of causal structures. Furthermore, we establish a theoretical foundation for **CAPI** fairness, demonstrating its connections with counterfactual fairness and adversarial robustness.
- **Unfair Area (§ 4.1)** We further advance the discourse by defining the notion of the *unfair area*, grounded within the context of **CAPI** fairness, and precisely explain this concept within the framework of a linear structural causal model and a classifier with a post-processing approach.
- **CAPI Fairness Classifier (§ 4.2)** Our fifth contribution is the introduction of a pioneering in-processing adversarial learning method named **CAPIFY**. This method stands as the first of its kind to address **CAPI** fairness—simultaneously embodying individual fairness, adversarial robustness, and an awareness of causal dynamics.
- **Evaluation (§ 5)** We validate the efficacy of our approach through extensive evaluations on both real-world and synthetic datasets. These evaluations demonstrate the effectiveness of our proposed framework to simultaneously embody individual fairness, adversarial robustness, and causal awareness.

Related Work. Several studies have explored individual fairness by utilizing adversarial robustness techniques. Doherty et al. (2023) investigated the association between adversarial robustness and ϵ-δ individual fairness in Bayesian neural network inference. They considered a specified similarity metric and ensured that the network's output falls within a specified tolerance. Benussi et al. (2022) introduce a method for certifying the ϵ-δ individual fairness formulation in feed-forward neural networks. They define adversarial perturbation using d_{fair} and incorporate an adversarial regularizer in the training loss to achieve a balance between model accuracy and IF. Xu et al. (2021) highlight that adversarial training may lead to notable discrepancies in both performance and robustness concerning group-level fairness. To address this issue, they propose a framework called fair robust learning that aims to enhance a model's robustness while ensuring fairness. Yeom and Fredrikson (2020) employed randomized smoothing techniques to ensure individual fairness in accordance with a specified weighted ℓ_p metric. Several methods tackle individual fairness using Wasserstein distance and distributionally robust optimization (Yurochkin, Bower, and Sun 2019; Yurochkin and Sun 2020; Vargo et al. 2021; Jiang et al. 2020b,a). These approaches employ projected gradient descent and optimal transport with Wasserstein distance to optimize a model with perturbations that substantially modify the sensitive information within a specified distribution. Ruoss et al. (2020) introduced a mixed-integer linear programming approach to develop data representations that exhibit IF. These representations are designed to capture similarities among individuals by generating latent representations that remain unaffected by specific transformations of the input data.

Numerous prior studies (Grari, Lamprier, and Detyniecki 2023; Jung et al. 2019; Kim, Reingold, and Rothblum 2018; John, Vijaykeerthy, and Saha 2020; Adragna et al. 2020; Petersen et al. 2021) have explored the connections among fairness, robustness, and causal structures individually or in pairs. However, to our knowledge, no previous research has explicitly examined the simultaneous interplay of all these properties.

2 Preliminaries

Notation. In this study, random variables are indicated by boldface letters ($\mathbf{V}$), while regular lowercase letters (v) represent assignments or instances. Matrices are denoted by bold uppercase letters, such as $\mathbf{F}$, with $[\mathbf{F}]_i$ referring to the i-th column vector of $\mathbf{F}$ and $[\mathbf{F}]_{i,j}$ representing the entry at row i and column j of $\mathbf{F}$. The feature space $\mathcal{V}$ is constructed using n random variables, denoted as $\mathbf{V} = (\mathbf{V}_1, \dots, \mathbf{V}_n)$.

Structural Causal Model (SCM). We make the assumption that feature variables $\mathcal{V}$ are generated by a SCM (Pearl 2009) denoted as $\mathcal{M}$, as described by a tuple $\langle \mathcal{G}, \mathbf{V}, \mathbf{U}, \mathbf{F}, \mathbf{P}_{\mathcal{U}} \rangle$. Here, $\mathcal{G}$ represents a known directed acyclic graph (DAG), $\mathbf{V} = \{\mathbf{V}_i\}_{i=1}^n$ denotes a set of observed (indigenous) random variables, $\mathbf{U} = \{\mathbf{U}_i\}_{i=1}^n$ represents a set of noise (exogenous) random variables is assumed to be independent, and $\mathbb{F}$ is the set of **structural equations**, defined as $\mathbb{F} = \{\mathbf{V}_i := f_i(\mathbf{V}_{\mathbf{Pa}(i)}, \mathbf{U}_i)\}_{i=1}^n$. These equations describe the causal relationship between each endogenous variable $\mathbf{V}_i$, its direct causes $\mathbf{V}_{\mathbf{Pa}(i)}$, and an exogenous variable $\mathbf{U}_i$ using deterministic functions f_i. Additionally, $\mathbb{P}_{\mathbf{U}}$ represents the probability distribution over the exogenous variables. The structural equations $\mathbb{F}$ establish a mapping $\mathbf{F} : \mathcal{U} \rightarrow \mathcal{V}$ from exogenous to endogenous variables, along with an inverse image $\mathbf{F}^{-1} : \mathcal{V} \rightarrow \mathcal{U}$ that satisfies the property $\mathbf{F}\left(\mathbf{F}^{-1}(v)\right) = v$ for all $v \in \mathcal{V}$.

The latent variable distribution entails a unique distribution $\mathbb{P}(\mathbf{V}) = \prod_{i=1}^{n} \mathbb{P}(\mathbf{V}_i \mid \mathbf{V}_{\mathbf{Pa}(i)})$ over the variables $\mathbf{V}$ (Peters, Janzing, and Schölkopf 2017). The marginal probability distribution of $\mathbb{P}_{\mathbf{V}}$ with respect to the feature $\mathbf{V}_i$ is denoted as $\mathbb{P}_{\mathbf{V}_i}$.

Additive Noise Model (ANM). In order to infer the unique causal structure $\mathcal{G}$ from observational data $\mathcal{V}$, it is necessary to impose additional assumptions on the underlying SCM. One of the causally identifiable classes within SCMs is additive noise models (Hoyer et al. 2009), which posit that the assignments follow the form:

$$\begin{aligned} \mathbb{F} = \{\mathbf{V}_i := f_i(\mathbf{V}_{\mathbf{Pa}(i)}) + \mathbf{U}_i\}_{i=1}^{n} \quad \implies \\ \mathbf{U} = \mathbf{V} - f(\mathbf{V}) \implies \mathbf{V} = (I - f)^{-1}(\mathbf{U}) \end{aligned} \tag{1}$$

where $\mathbf{U}_i$ is an independent known distribution. As observed in Eq. 1, obtaining $\mathbf{U}$ from $\mathbf{V}$ is straightforward, where I represents the identity function ($I(v) = v$). Henceforth, we denote the inverse of $(I - f)^{-1}$ as F. A specific class of ANMs is represented by **linear** SCMs, where the functions f_i are assumed to be linear.

Counterfactuals. SCMs are employed to examine the effects of interventions, which entail external manipulations to modify the data generation process (Peters et al., 2017). Two primary types of interventions exist, hard interventions and soft interventions Interventions facilitate the examination of counterfactual statements v^{CF} for a given instance v under hypothetical interventions on a variable. The counterfactual maps for hard interventions are denoted as $v_{\theta}^{\text{CF}} := \mathbf{CF}(v, do(\mathbf{V}_{\mathcal{I}} := \theta)) = \mathbf{F}^{\theta}(\mathbf{F}^{-1}(v))$ where $\mathbf{F}^{\theta}$ is a simplified notation for $\mathbf{F}^{do(\mathbf{V}_{\mathcal{I}} := \theta)}$.

Sensitive Attribute. A sensitive attribute, such as race, is an ethically or legally significant characteristic used in decision-making processes like hiring, lending, or criminal justice to determine fair treatment or outcomes for individuals or groups. Let $\mathbf{S} \in \{\mathbf{V}_1, \ldots, \mathbf{V}_n\}$ be a sensitive attribute that has finite levels $\mathcal{S} = \{s_1, \ldots, s_k\}$. For each instance v of $\mathcal{V}$, the set of **counterfactual twins** w.r.t protected variable $\mathbf{S}$ is defined as $\ddot{\mathbb{V}} = \{\ddot{v}_s = \mathbf{CF}(v, do(\mathbf{S} := s)) : s \in \mathcal{S}\}$.

Fairness. In fairness, a sensitive attribute defines a protected group, ensuring that machine learning models or algorithms do not disadvantage them. Researchers have proposed different notions of fairness, such as group fairness and individual fairness (IF) (Tang, Zhang, and Zhang 2022; Le Quy et al. 2022; Mehrabi et al. 2021).

Individual-level fairness, introduced by Dwork et al. (2012), ensures that individuals who exhibit similarity according to predefined metrics are treated similarly with regard to outcomes. Various mathematical formulations have been proposed, including the Lipschitz Mapping-based formulation (Dwork et al. 2012) and the ϵ-δ formulation (John, Vijaykeerthy, and Saha 2020). The classifier h satisfies the L-Lipschitz IF condition when:

$$d_{\mathcal{Y}}(h(v), h(w)) \leq L \, d_{\mathcal{V}}(v, w) \quad \forall v, w \in \mathcal{V} \tag{2}$$

where $d_{\mathcal{X}}$ and $d_{\mathcal{Y}}$ represent metrics on the input and output spaces respectively, and $L \in \mathbb{R}_+$.

Counterfactual fairness, introduced by Kusner et al. (2017), is another notion of individual-level fairness that deems a decision fair for an individual if it maintains consistency in both the real and a counterfactual scenario. Formally, it can be expressed as:

$$\mathbb{E}_{\mathbb{P}_{\mathbf{V}}}[\max_{s \in \mathbf{S}} d_{\mathcal{Y}}(h(\mathbf{V}), h(\ddot{\mathbf{V}}_s))] \leq \epsilon \tag{3}$$

Adversarially Robust Learning. Adversarially robust learning aims to create algorithms and models that can withstand adversarial attacks, which involve purposeful perturbations or modifications to input data to induce misclassification or misleading predictions (Goodfellow, Shlens, and Szegedy 2014; Madry et al. 2017). In this framework, models are trained considering the most challenging perturbations of the data rather than the original data itself:

$$\min_{\psi} \mathbb{E}_{(v,y) \sim \mathcal{P}_{\mathcal{D}}}[\max_{\delta \in B_{\Delta}(v)} \ell(h_{\psi}(v + \delta), y)] \tag{4}$$

where, $B_{\Delta}(v)$ is the set of perturbations for the instance v, $\mathcal{P}_{\mathcal{D}}$ is observation distribution, ℓ is the classification loss function, and ψ are the weights of the classifier.

3 Causal Fair Metric

Achieving individual fairness necessitates the formulation of a fair metric, which, in pursuit of this goal, gives rise to two primary challenges. Firstly, the presence of diverse feature types within the SCM, such as categorical or continuous attributes, introduces complexities stemming from its heterogeneous nature. Secondly, inherent biases may be encoded within the SCM, thereby necessitating that our classifier comprehends the full spectrum of hypothetical interventions applied to instances relative to the levels of sensitive attributes. These twin focal points constitute the primary focus of the ensuing chapter.

3.1 Fair Metric

When dealing with independent features, constructing similarity functions based on their attributes and aggregating them through a product metric is relatively straightforward. However, in the context of a causal structure, the integration of causality into metric formulation becomes pivotal. To tackle this, instances undergo a transformation into an independent space where a metric is established. This established metric is subsequently employed in defining a similarity function within the original feature space via the push-forward metric technique.

In the presence of SCM and a sensitive attribute, the similarity function d should be robust to twins and slight perturbations of non-sensitive features. This means that d should not significantly change after a hard intervention ($do(\mathbf{S} := s)$) with respect to the levels of S, or after an additive intervention on continuous features. In ANMs, a hard intervention removes the causal structure of $\mathbf{S}$ and is equivalent to setting f_i to zero and fixing $\mathbf{U}_i := s$. Moreover, additive intervention is equivalent to adding δ to $\mathbf{U}_i$ while keeping f_i unchanged. Consequently, the latent space changes during the hard intervention, replacing the sensitive latent variable $\mathbf{U}_i$ with $\mathbf{S}$ following the distribution $\mathbb{P}_{\mathbf{S}}$. This motivates the definition of a *semi-latent space*.

Definition 1 (Semi-latent Space) *Consider SCM $\mathcal{M}$ with sensitive features indexed by I. We define the semi-latent space $\mathcal{Q}$ as a combination of observed sensitive features $\mathbf{V}_i$ with distribution $\mathbb{P}_{\mathbf{V}_i}$ where $i \in I$, and latent variables $\mathbf{U}_j$ for other features with distribution $\mathbb{P}_{\mathbf{U}_j}$.*

Let $v = (v_1, v_2, \dots, v_n)$ be an instance in the observed space and $u = (u_1, u_2, \dots, u_n) = F^{-1}(v)$ be the corresponding instance in the latent space. The mapping $T : \mathcal{V} \to \mathcal{Q}$ transforms v to the semi-latent space $q = (q_1, q_2, \dots, q_n) = T(v)$, where q_i is defined as follows:

$$q_i := \begin{cases} v_i & i \in I \\ u_i & i \notin I \end{cases} \tag{5}$$

The inverse function $v = T^{-1}(q)$ is determined as follows:

$$v_i := \begin{cases} q_i & i \in I \\ f_i(v_{\mathbf{pa}(i)}) + q_i & i \notin I \end{cases} \tag{6}$$

The identity $v = T^{-1}(T(v))$ holds straightforwardly.

The semi-latent space allows us to describe the counterfactual of instance v w.r.t. hard action $do(\mathbf{V}_I{:=}\theta)$:

$$\mathbf{CF}(v, do(\mathbf{V}_I{:=}\theta)) = T^{-1}(T(v) \odot_I \theta) \tag{7}$$

Here, $v \odot_I \theta$ represents a masking operator that modifies the values of I entries in vector v by replacing θ.

In the semi-latent space, a causal structure-independent similarity function can be readily established. Let $(\mathcal{U}_i, d_{\mathcal{U}_i})$ denote the metric space for the latent space corresponding to $\mathbf{V}_i$. For sensitive variables $\mathbf{S}_i$, $(\mathcal{S}_i, d_{\mathcal{S}_i})$ is considered a pseudometric or metric space. Thus, the semi-latent space $(\mathcal{Q}, d_{\mathcal{Q}})$ has a metric obtained as the product of metrics. To establish a fair metric, incorporating sensitive features into the similarity function is crucial. We adopt the approach by Ehyaei et al. (2023), treating the protected feature as a pseudometric.

Definition 2 (Pseudometric Protected (Ehyaei et al. 2023)) *In SCM $\mathcal{M}$, suppose the sensitive feature $\mathbf{S}$ endowed with a pseudometric space $(\mathcal{S}, d_{\mathcal{S}})$. $\mathbf{S}$ is partially protected if there are two levels with zero distance:*

$$\exists s, s' \in \mathcal{S} \quad s.t. \quad d_{\mathcal{S}}(s, s') = 0 \quad \wedge \quad s \neq s' \tag{8}$$

If for all $s, s' \in \mathcal{S}$ we have $d_{\mathcal{S}}(s, s') = 0$, then $\mathbf{S}$ is called protected feature.

By employing the pseudometric for sensitive attributes within the semi-latent space metric, a fair metric can be established in the feature space using the push-forward metric:

$$d_{fair}(v, w) = d_{\mathcal{Q}}(T(v), T(w)) \tag{9}$$

fair metric enables us to define small perturbations of factual values to identify similar instances.

3.2 Causal Adversarial Perturbation

Adversarial perturbation involves the manipulation of input data to evaluate the resilience of machine learning models. The introduction of a fair metric contributes to the definition of adversarial perturbation in alignment with causal relationships.

Definition 3 (Causal Adversarial Perturbation) *Let $\mathcal{M}$ be an SCM with sensitive attributes, and d_{fair} be its fair metric. The CAP for instance v is defined as:*

$$B^{\mathbf{CAP}}_{\Delta}(v) = \{w \in \mathcal{V} : d_{fair}(v, w) \leq \Delta\} \tag{10}$$

where $\Delta \in \mathbb{R}_{\geq 0}$. CAP can be seen as transforming the unit ball in the semi-latent space using the inverse mapping function T^{-1}:

$$B^{\mathcal{Q}}_{\Delta}(q) = \{p \in \mathcal{Q} : d_{\mathcal{Q}}(q, p) \leq \Delta\}. \tag{11}$$

then $B^{\mathbf{CAP}}_{\Delta}(v) = T^{-1}(B^{\mathcal{Q}}_{\Delta}(T(v)))$.

Remark 1 *When all features are continuous or all sensitive features don't have parents, CAP simplifies interpretation. In these cases, the semi-latent space coincides with the latent space, and CAP is achieved by transforming the unit ball in the latent space using the mapping function F. Specifically, $B^{\mathbf{CAP}}_{\Delta}(v) = F(B^{\mathcal{U}}_{\Delta}(F^{-1}(v)))$, where $B^{\mathcal{U}}_{\Delta}$ represents a closed ball with radius Δ in the latent space.*

Building upon Remark 1, we seek a concise geometric interpretation of CAP by perturbing only the continuous feature of the SCM. Let $q = (z, x) \in \mathcal{Q}$ with x as the continuous part and z as the categorical part of features. We define $B^{\mathcal{Q}+}_{\Delta}$ as the unit ball with a radius of Δ, specifically designed for the continuous part:

$$B^{\mathcal{Q}+}_{\Delta}(q) = \{q' = (z', x') \in \mathcal{Q} : z' = z \ \wedge \ d_{\mathcal{X}}(x', x) \leq \Delta\}$$

Without loss of generality, assuming a norm on the continuous part, we define the closed unit disk as $D^{\mathcal{X}}_{\Delta} = \{\delta : \|\delta\| \leq \Delta \quad \wedge \quad \delta_{|\mathcal{Z}} = 0\}$ where $\mathcal{Z}$ is categorical part of feature space. Thus, in this scenario, $B^{\mathcal{Q}+}_{\Delta}$ is derived from:

$$B^{\mathcal{Q}+}_{\Delta}(q) = \{q + \delta : \delta_{|\mathcal{X}} \in D^{\mathcal{X}}_{\Delta}\} \tag{12}$$

By defining $B^{+}_{\Delta}(v) = T^{-1}(B^{\mathcal{Q}+}_{\Delta}(T(v)))$, the CAP can be decomposed into B^{+}_{Δ}, as stated in the following proposition.

Proposition 1 *Let $B^{\mathbf{CAP}}_{\Delta}(v)$ represent the CAP around instance $v = (z, x)$ with radius Δ, and let $\Theta_{\Delta} = \{\theta \in \mathcal{Z} : (\theta, .) \in B^{\mathcal{Q}}_{\Delta}(T(v))\}$ denote the set of categorical levels within the perturbation ball. The counterfactual perturbation can be expressed as:*

$$B^{\mathbf{CAP}}_{\Delta}(v) = \bigcup_{\theta \in \Theta_{\Delta}} B^{+}_{\Delta_\theta}(\boldsymbol{CF}(v, \theta)) \tag{13}$$

where Δ_θ represents the value of the continuous part of Δ. For instance, in the case of using the L_2 product metric, $\Delta_\theta = \sqrt{\Delta^2 - d_{\mathcal{Z}}(\theta, s)^2}$.

The decomposition of perturbation allows analyzing the shape of CAP for a sensitive attribute, especially for small Δ values. This aspect is elaborated upon in the subsequent corollary.

Corollary 1 *If $\mathbf{S}$ is a protected feature and other categorical variables in $\mathcal{M}$ are not partially protected, there exists a Δ_0 such that for all $\Delta \leq \Delta_0$:*

$$B^{\mathbf{CAP}}_{\Delta}(v) = \bigcup_{s \in \mathcal{S}} B^{+}_{\Delta}(\ddot{v}_s) \tag{14}$$

Consequently, for all $v, w \in \ddot{\mathbb{V}}$, we have $B^{\mathbf{CAP}}_{\Delta}(v) = B^{\mathbf{CAP}}_{\Delta}(w)$.

The CAP definition considers causal similarity in relation to counterfactuals. The subsequent lemma shows that a CAP with a diameter 0 represents the set of twins.

Corollary 2 *If $\mathbf{S}$ is a protected feature and other categorical variables in $\mathcal{M}$ are not partially protected, the counterfactual twins correspond to the zero-radius CAP:*

$$\ddot{\mathbb{V}} = B^{\mathbf{CAP}}_{0}(v) := \lim_{\Delta \to 0} B^{\mathbf{CAP}}_{\Delta}(v)$$

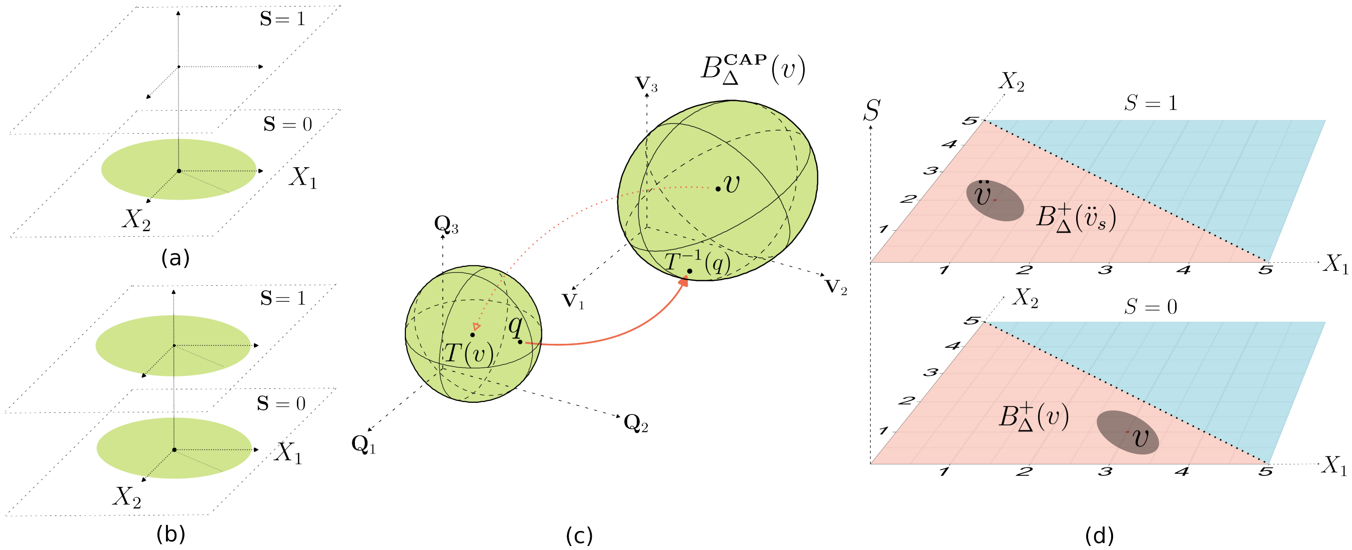

Figure 1: The difference in unit ball shape between considering the sensitive attribute as a Euclidean metric (a) and as a trivial pseudometric (b). The geometric interpretation of CAP is mapping a closed unit ball in semi-latent space (c). Causal adversarial perturbation is the union of continuous perturbations around each twin (d).

3.3 CAPI Fairness

This section presents our innovative concept of causal individual fairness, denoted as **CAPI** fairness. Within the Lipschitz formulation of IF, we introduce the metric d_{fair} as a measure in the feature space:

$$d_{\mathcal{Y}}(h(v), h(w)) \leq d_{fair}(v, w)$$

Here, $d_{\mathcal{Y}}$ represents the metric applied in the outcome space. By incorporating a fair metric as a similarity function, individual fairness now encompasses both the causal structure and the sensitive protected feature.

Proposition 2 *CAPI Fairness implies both Counterfactual Fairness and Adversarial Robustness:*

CAPI Fairness ⇒ Counterfactual Fairness
CAPI Fairness ⇒ Adversarial Robustness

However, the inverse statements are not necessarily true.

4 Fair Classifier

In this section, we will initially explore the origins of unfairness concerning IF in the context of an SCM. Following that, we will introduce IF classifiers based on CAPI fairness.

4.1 Unfair Area

To analyze the bottlenecks in designing fair classifiers, we should understand the origins of unfairness. We begin by defining unfair areas for CAPI fairness, inspired by Ehyaei et al. (2023).

Definition 4 (Unfair Area) *Let $\mathcal{M}$ denote an SCM, Δ diameter of CAP, and h be a binary classifier operating on $\mathbf{V}$. The unfair area includes instances where the CAPI fairness property is not met:*

$$A_\Delta^{\neq} := \{v \in \mathcal{V} : \exists v' \in B_\Delta^{CAP}(v) \quad s.t. \quad h(v) \neq h(v')\}$$

To understand the shape of the unfair area, we aim to determine $A_\Delta^{\neq}$ assuming linear SCMs and classifiers (see Fig. 2).

Proposition 3 *Consider a linear SCM with a binary linear classifier $h(v) = sign(w^T \cdot v - b)$, where $w \in \mathbb{R}^n$. Assume $\mathcal{M}$ has one binary sensitive attribute $S \in \{0, 1\}$ and other features X are continuous. Without loss of generality, let V_1 represent the sensitive attribute. The unfair area $A_C^{\neq}$ for counterfactual fairness is delineated as follows:*

$$\{v = (s, x) \in \mathcal{V} : \quad \text{sign}((s - (1 - s)) * h(v)) \geq 0 \quad \wedge$$
$$dist(x, L) \leq \frac{|w^T \cdot [F]_1|}{\|w\|_{p^*}}\}$$

The unfair area $A_\Delta^{\neq}$ is defined as the band parallel to the classifier boundary L:

$$A_\Delta^{\neq} = \{v = (s, x) \in \mathcal{V} : dist(x, A_C^{\neq}) \leq \frac{\Delta\|w_{-1}^T \times F_{-1}\|_{p^*}}{\|w\|_{p^*}}\}$$

Here, L denotes the decision boundary of the classifier, while w_{-1} and F_{-1} represent the continuous components of w and F, respectively.

According to Prop. 3, a straightforward condition can be derived for ensuring counterfactual fairness.

Corollary 3 *Considering the condition in Prop. 3, achieving a counterfactually fair classifier for $\mathcal{M}$ is impossible unless F and w satisfy the equation $w^T \cdot [F]_1 = 0$. This implies that the classifier h relies solely on a subset of variables that are non-descendants of* **S** *in $\mathcal{M}$.*

In assessing CAPI unfairness, a meaningful indicator is the probability associated with the unfair area in the trained classifier.

Definition 5 (Unfair Area Indicator (UAI)) *Let $\mathcal{M}$ be the SCM with parameters denoted by $\langle \mathcal{G}, \mathbf{V}, \mathbf{U}, \mathbb{F}, \mathbb{P}_{\mathbf{U}} \rangle$, and $\hat{h}$*

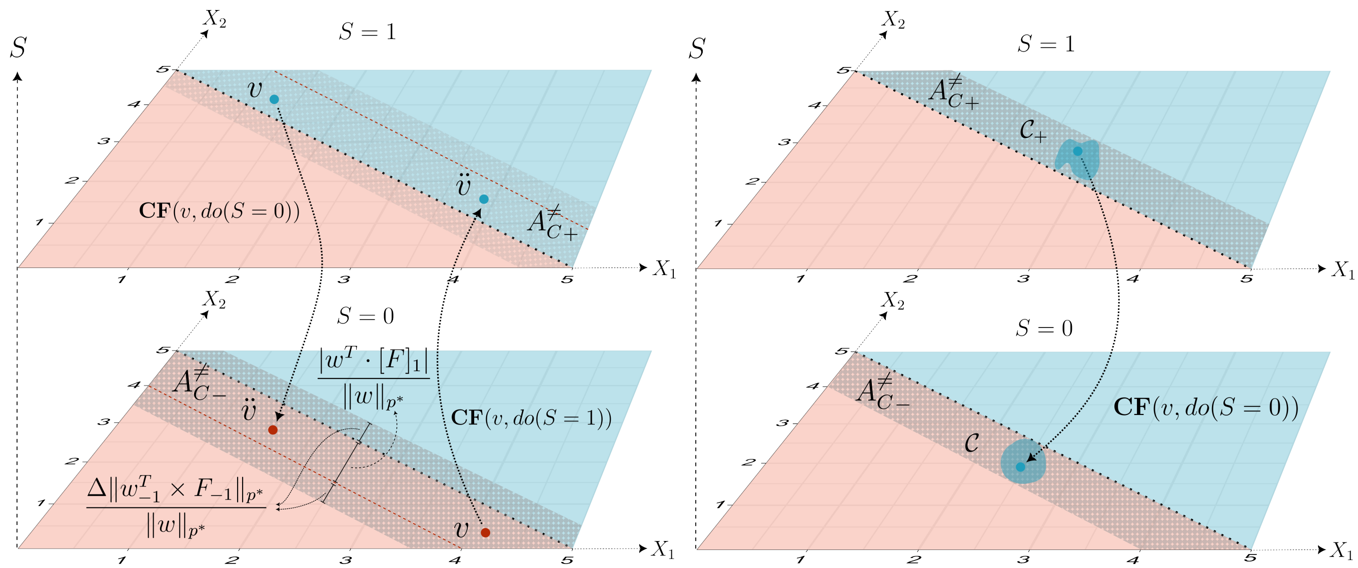

Figure 2: The unfair area for linear SCM and classifier consists of two parts: counterfactual and adversarial robustness (left) The counterfactual fairness mitigation idea is based on the property that the twin of the twin is equal to the instance v (right).

be the trained binary classifier. The probability $\mathbb{P}_{\mathcal{V}}(A^{\neq}_{\Delta})$, referred to as the Unfair Area Indicator, quantifies the likelihood of the CAPI unfairness for $\hat{h}$.

Taking into account the concept of unfair areas and the inherent property that a twin's twin is an identity function, we present a post-processing technique involving label-flipping. This method aims to mitigate counterfactual fairness issues.

Proposition 4 (Counterfactual Unfairness Mitigation) *Let $A^{\neq}_{C+}$ and $A^{\neq}_{C-}$ represent the positive and negative regions of counterfactual unfairness $A^{\neq}_{C}$, respectively (see Fig. 2). Assuming $\mathcal{C} \subset A^{\neq}_{C-}$, the unfair area mitigation method involves flipping the labels of instances in $\mathcal{C}$ to positive. By changing labels, the reduction in unfairness area is given by:*

$$\mathbb{P}_{\mathcal{V}}(A^{\neq}_{C}) - \mathbb{P}_{\mathcal{V}}(\mathcal{C}) - \mathbb{P}_{\mathcal{V}}(\mathcal{C}_{+}) \tag{15}$$

Here, $\mathcal{C}_+$ is a subset of $A^{\neq}_{\Delta+}$, representing the points in $A^{\neq}_{\Delta+}$ whose corresponding twins belong to the set $\mathcal{C}$. If we set $\mathcal{C} = A^{\neq}_{C-}$, complete mitigation of counterfactual fairness can be achieved.

Remark 2 *The label-flipping direction ($+$ to $-$) does not inherently impact counterfactual unfairness mitigation. However, fairness considerations often involve a preferred direction. In such cases, flipping the sign of the unfair region in relation to this preferred direction can be employed to promote fairness.*

Label flipping alone is insufficient to remove CAPI unfairness. Therefore, in the next section, we introduce an additional in-processing method to mitigate unfairness.

4.2 Causal Adversarial Learning

Fair adversarial learning aims to achieve high accuracy in predicting the target variable while ensuring fairness regarding sensitive attributes. This involves formulating a min-max optimization problem, where the model simultaneously minimizes the classification error and maximizes the adversarial loss. In previous chapters, the concept of CAP was discussed. Now, we formulate the objective function for Causal Adversarial Learning (CAL). Let $\mathcal{D} = \{(v_i, y_i)\}_{i=1}^{n}$ represent the set of observations. The objective function to be minimized over the classifier space in CAL is as follows:

$$\min_{\psi} \mathbb{E}_{(v,y)\sim\mathcal{P}_{\mathcal{D}}}[\max_{w\in B^{\mathbf{CAP}}_{\Delta}(v)} \ell(h_{\psi}(w), y)] \tag{16}$$

The optimization objective in Eq. 16 promotes the proximity of values for h within the neighborhood $B^{\mathbf{CAP}}_{\Delta}(v)$ to $h(v)$. According to Lem. 1 in appendix, we can establish the inequality $f(v+\delta) \leq f(v) + |\delta^T \nabla_v f(v)| + \gamma(\Delta, v)$. By setting $f(v+\delta) = \ell(h(T^{-1}(T(v)+\delta), y)$, we can utilize Cor. 1 to represent the expression within the expectation of Eq. 16 as follows:

$$\begin{aligned}
&\max_{w\in B^{\mathbf{CAP}}_{\Delta}(v)} \ell(h(w), y) = \max_{s\in\mathcal{S}} \max_{w\in B^{+}_{\Delta}(\ddot{v}_s)} \ell(h(w), y) = \\
&\max_{s\in\mathcal{S}} \max_{\delta\in D^{\mathcal{X}}_{\Delta}} \ell(h(T^{-1}(T(\ddot{v}_s)+\delta), y) \leq \\
&\max_{s\in\mathcal{S}} \max_{\delta\in D^{\mathcal{X}}_{\Delta}} \ell(h(T^{-1}(T(\ddot{v}_s)), y) + |\delta^T \nabla^{\mathcal{X}}_{\ddot{v}_s} f(\ddot{v}_s)| + \\
&\gamma(\Delta, \ddot{v}_s) \leq \ell(h(v), y) + \max_{s\in\mathcal{S}} \ell(h(\ddot{v}_s), y) + \\
&\max_{s\in\mathcal{S}} \max_{\delta\in D^{\mathcal{X}}_{\Delta}} |\delta^T \nabla^{\mathcal{X}}_{\ddot{v}_s} f(\ddot{v}_s)| + \gamma(\Delta, \ddot{v}_s) = \\
&\ell(h(v), y) + \underbrace{\max_{s\in\mathcal{S}} \ell(h(\ddot{v}_s), y)}_{II} + \\
&\underbrace{\max_{s\in\mathcal{S}} \left(\|\nabla^{\mathcal{X}}_{\ddot{v}_s} f(\ddot{v}_s)\|_* + \gamma(\Delta, \ddot{v}_s)\right)}_{II}
\end{aligned}$$

The symbol $\nabla^{\mathcal{X}}$ denotes the gradient operator for continuous features. The validity of the final equation can be es-

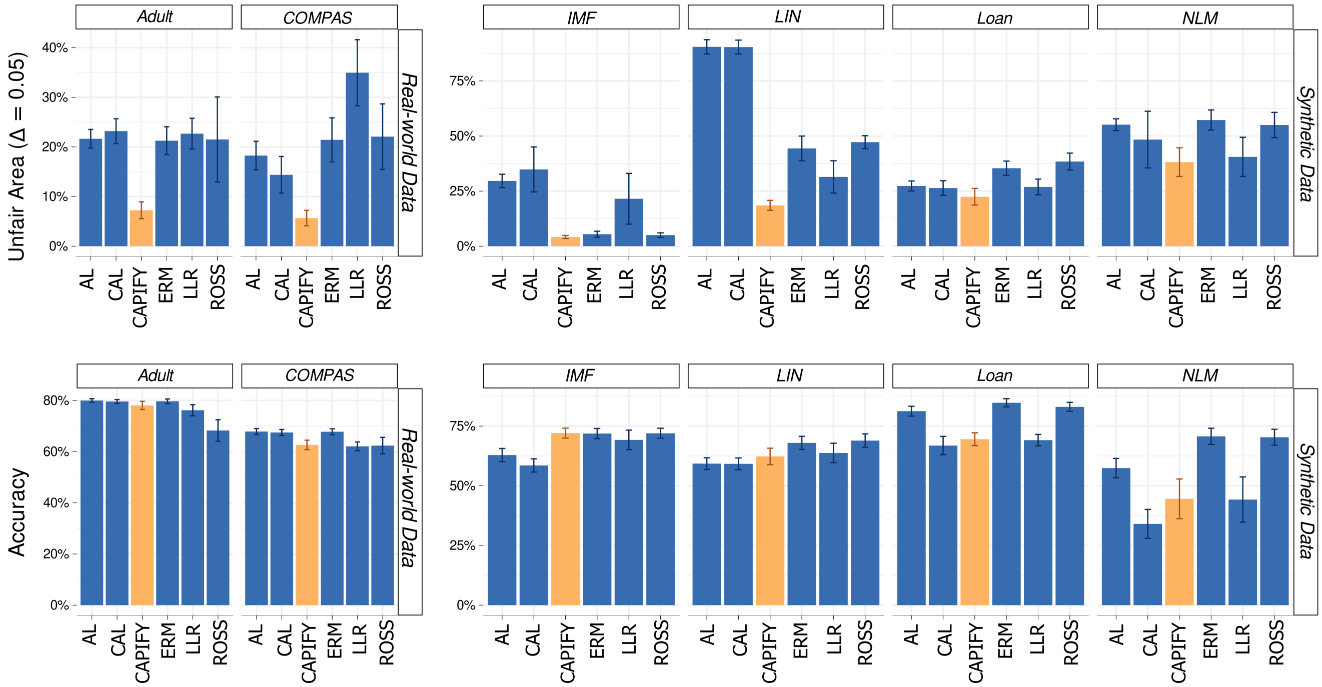

Figure 3: Figure depicts our numerical experiment's results, showcasing diverse trainers and datasets to evaluate CAPIFY performance. The initial bar plot assesses trainer performance through UAI values (favoring lower values) at $\Delta = .05$. The subsequent bar plot contrasts methods based on prediction performance (favoring higher values).

tablished by bounding $|\delta^T \nabla^{\mathcal{X}}_{\ddot{v}_s} f(\ddot{v}_s)|$ using the dual norm $\|\nabla^{\mathcal{X}}_{\ddot{v}_s} f(\ddot{v}_s)\|_*$.

The adversarial loss function, in the above equation, comprises a regular loss function and a regularizer, which can be decomposed into two components. The first component addresses counterfactual fairness by capturing the discrepancy between the instance y and the corresponding twins' classifier label. The second component measures the adversarial robustness of classifier h regarding the continuous features surrounding each twin. Assuming random observations, the evaluation of the robustness property is narrowed down to the instance denoted as v. Hence, the reformulated expression for the regularizer can be stated as follows:

$$\begin{aligned} \mathcal{R}(v) = \quad & \mu_1 * \max_{s \in \mathcal{S}} \ell(h(\ddot{v}_s), y) \quad + \\ & \mu_2 * \gamma(\Delta, v) + \mu_3 * \|\nabla^{\mathcal{X}}_v f(v)\|_* \end{aligned} \tag{17}$$

where the hyperparameters $\mu_i \in \mathbb{R}$ determine the extent of regularization in the model.

5 Numerical Experiments

In this study, we empirically validate the theoretical propositions presented in the paper. We assess the performance of the CAPIFY and CAL training methods in comparison to conventional empirical risk minimization (ERM) and other pertinent techniques, including Adversarial Learning (AL) (Madry et al. 2017), Locally Linear Regularizer (LLR) training (Qin et al. 2019), and Ross method (Ross, Lakkaraju, and Bastani 2021). Our experimentation involves real datasets, specifically Adult (Kohavi and Becker 1996) and COMPAS (Washington 2018), which are pre-processed according to (Dominguez-Olmedo, Karimi, and Schölkopf 2022). Furthermore, we consider three synthetic datasets related to Linear (LIN), Non-linear (NLM), and independent futures (IMF) SCMs, along with the semi-synthetic Loan dataset based on (Karimi et al. 2020).

We utilize a multi-layer perceptron with three hidden layers, each comprising 100 nodes, for the COMPAS, Adult, NLM, and Loan datasets. Logistic regression is employed for the remaining datasets. To evaluate classifier performance, we measure accuracy and Matthews correlation coefficient (MCC). Furthermore, we quantify CAPI fairness using UAI across various Δ values, including 0.05, 0.01, and 0.0. Additionally, we compute UAI for non-sensitive scenarios, employing Δ values of 0.05 and 0.01 to represent the non-robust data percentage. Additional comprehensive details about the computational experiments are available in the appendix.

We performed our experiment using 100 different seeds, and the results are presented in Table 1. Figure 3 illustrate that the CAPIFY method exhibits a lower unfair area (U_Δ) for $\Delta = 0.05$, $\Delta = 0.01$, and $\Delta = 0.0$. However, the CAL method shows unsatisfactory accuracy due to the issues reported previously (Qin et al. 2019). Compared to ERM, CAPIFY shows slightly lower accuracy, a trade-off noted in multiple studies (Pessach and Shmueli 2022). Notably, real-world data indicates a greater reduction in unfairness than in accuracy. Moreover, CAPIFY exhibits robustness and counterfactual fairness attributes (see Tab. 1), making it the favored model when assessing both concepts. For

	Real-World Data						Synthetic Data											
	Adult			COMPAS			IMF			LIN			Loan			NLM		
Trainer	$U_{.05}$	CF	$\mathcal{R}_{.05}$	$U_{.05}$	CF	$\mathcal{R}_{.05}$	$U_{.05}$	CF	$\mathcal{R}_{.05}$	$U_{.05}$	CF	$\mathcal{R}_{.05}$	$U_{.05}$	CF	$\mathcal{R}_{.05}$	$U_{.05}$	CF	$\mathcal{R}_{.05}$
AL	0.22	0.18	**0.04**	0.18	0.14	0.04	0.30	0.28	0.11	0.90	0.90	0.26	0.27	0.27	0.16	0.55	0.53	0.37
CAL	0.23	0.18	0.05	0.14	0.10	0.04	0.35	0.34	0.13	0.90	0.90	0.26	0.26	0.26	0.19	0.48	0.46	0.24
CAPIFY	**0.07**	**0.03**	0.05	**0.06**	**0.01**	0.04	**0.04**	**0.00**	0.04	**0.19**	**0.15**	**0.09**	**0.22**	**0.22**	0.14	**0.38**	**0.36**	0.24
ERM	0.21	0.18	0.04	0.21	0.17	0.04	0.05	0.02	**0.04**	0.44	0.43	0.18	0.35	0.35	0.23	0.57	0.54	0.41
LLR	0.23	0.19	0.04	0.35	0.32	**0.04**	0.22	0.20	0.08	0.31	0.31	0.27	0.27	0.26	**0.12**	0.41	0.39	**0.20**
ROSS	0.22	0.08	0.11	0.22	0.16	0.06	0.05	0.02	0.04	0.47	0.44	0.11	0.38	0.38	0.27	0.55	0.52	0.43

Table 1: The table displays the outcomes of our numerical experiment, wherein different trainers are compared based on their input sets in terms of CAPI fairness metrics ($U_{.05}$, lower values are better), Counterfactual Unfair area (CF, lower values are better), and the non-robust percentage concerning adversarial perturbation with radii 0.05 ($\mathcal{R}_{.05}$, lower values are better). The best-performing techniques for each trainer, dataset, and metric are indicated in bold. The findings highlight that CAPIFY outperforms other trainers in reducing CAPI unfairness. The standard deviation average for CAPIFY is 0.028, whereas for the other methods, it is 0.038.

more results, see appendix

6 Discussion and Future Work

In this study, we introduce a comprehensive method considering individual fairness (IF) and robustness within an underlying causal model. We establish adversarial learning through the use of CAP. Remarkably, our CAP strategy sets itself apart by not requiring assumptions for all categorical features, a departure from the approach by Ehyaei et al. (2023). Our CAP framework exclusively focuses on sensitive features.

In this study we use the discrete sensitive features for simplicity, every theoretical and numerical part are satisfied for continuous sensitive attribute as well. Our approach avoids specific assumptions for defining IF based on the L-Lipschitz formulation. Instead, we can reframe everything using the ϵ-δ formulation. The optimization in Eq. 16 may yield nonlinear decision boundaries, particularly with numerous features. To tackle this, we adopt the locally linear regularizer (LLR) proposed by Qin et al. (2019). LLR is advantageous in deep learning for countering overfitting, enhancing generalization through smoother function learning, and attaining leading computational performance.

An objection to our approach is that, like many fair learning methods, although we address unfairness by introducing a regularizer, there's no assured theoretical guarantee for the resulting classifier to uphold individual fairness. In future research, our goal is to develop a classifier with theoretical foundations that endorse CAPI fairness principles.

Acknowledgments

Funding support for project activities has been partially provided by Canada CIFAR AI Chair, Google scholar award, NSERC Discovery Grants program and Grant 01IS20051 from the German Federal Ministry of Education and Research. A.Ehyaei thanks the Max Planck Intelligent Systems for the support this project. We also express our gratitude to Compute Canada for their support in providing facilities for our evaluations.

References

Adragna, R.; Creager, E.; Madras, D.; and Zemel, R. 2020. Fairness and robustness in invariant learning: A case study in toxicity classification. *arXiv preprint arXiv:2011.06485*.

Benussi, E.; Patane, A.; Wicker, M.; Laurenti, L.; and Kwiatkowska, M. 2022. Individual fairness guarantees for neural networks. *arXiv preprint arXiv:2205.05763*.

Binns, R. 2020. On the apparent conflict between individual and group fairness. In *Proceedings of the 2020 conference on fairness, accountability, and transparency*, 514–524.

Doherty, A.; Wicker, M.; Laurenti, L.; and Patane, A. 2023. Individual Fairness in Bayesian Neural Networks. *arXiv preprint arXiv:2304.10828*.

Dominguez-Olmedo, R.; Karimi, A. H.; and Schölkopf, B. 2022. On the adversarial robustness of causal algorithmic recourse. In *International Conference on Machine Learning*, 5324–5342. PMLR.

Dwork, C.; Hardt, M.; Pitassi, T.; Reingold, O.; and Zemel, R. 2012. Fairness through awareness. In *Proceedings of the 3rd innovations in theoretical computer science conference*, 214–226.

Ehyaei, A.-R.; Karimi, A.-H.; Schölkopf, B.; and Maghsudi, S. 2023. Robustness Implies Fairness in Casual Algorithmic Recourse. *arXiv preprint arXiv:2302.03465*.

Goodfellow, I. J.; Shlens, J.; and Szegedy, C. 2014. Explaining and harnessing adversarial examples. *arXiv preprint arXiv:1412.6572*.

Grari, V.; Lamprier, S.; and Detyniecki, M. 2023. Adversarial learning for counterfactual fairness. *Machine Learning*, 112(3): 741–763.

Hoyer, P. O.; Janzing, D.; Mooij, J. M.; Peters, J.; and Schölkopf, B. 2009. Nonlinear causal discovery with additive noise models. In *Advances in neural information processing systems*, 689–696.

Jiang, R.; Pacchiano, A.; Stepleton, T.; Jiang, H.; and Chiappa, S. 2020a. A Distributionally Robust Approach to Fair Classification. 108: 1–10.

Jiang, R.; Pacchiano, A.; Stepleton, T.; Jiang, H.; and Chiappa, S. 2020b. Wasserstein Fair Classification. In Adams,

R. P.; and Gogate, V., eds., *Proceedings of The 35th Uncertainty in Artificial Intelligence Conference*, volume 115 of *Proceedings of Machine Learning Research*, 862–872.

John, P. G.; Vijaykeerthy, D.; and Saha, D. 2020. Verifying individual fairness in machine learning models. In *Conference on Uncertainty in Artificial Intelligence*, 749–758. PMLR.

Jung, C.; Kearns, M.; Neel, S.; Roth, A.; Stapleton, L.; and Wu, Z. S. 2019. An algorithmic framework for fairness elicitation. *arXiv preprint arXiv:1905.10660*.

Karimi, A.-H.; Von Kügelgen, J.; Schölkopf, B.; and Valera, I. 2020. Algorithmic recourse under imperfect causal knowledge: a probabilistic approach. *Advances in neural information processing systems*, 33: 265–277.

Kim, M.; Reingold, O.; and Rothblum, G. 2018. Fairness through computationally-bounded awareness. *Advances in Neural Information Processing Systems*, 31.

Kohavi, R.; and Becker, B. 1996. Uci adult data set. *UCI Meachine Learning Repository*, 5.

Kusner, M. J.; Loftus, J.; Russell, C.; and Silva, R. 2017. Counterfactual fairness. In *Advances in Neural Information Processing Systems*, 4069–4079.

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*, 12(3): e1452.

Madry, A.; Makelov, A.; Schmidt, L.; Tsipras, D.; and Vladu, A. 2017. Towards deep learning models resistant to adversarial attacks. *arXiv preprint arXiv:1706.06083*.

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.

Pearl, J. 2009. *Causality: Models, Reasoning, and Inference*. Cambridge University Press.

Pessach, D.; and Shmueli, E. 2022. A review on fairness in machine learning. *ACM Computing Surveys (CSUR)*, 55(3): 1–44.

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

Petersen, F.; Mukherjee, D.; Sun, Y.; and Yurochkin, M. 2021. Post-processing for individual fairness. *Advances in Neural Information Processing Systems*, 34: 25944–25955.

Qin, C.; Martens, J.; Gowal, S.; Krishnan, D.; Dvijotham, K.; Fawzi, A.; De, S.; Stanforth, R.; and Kohli, P. 2019. Adversarial robustness through local linearization. *Advances in Neural Information Processing Systems*, 32.

Ross, A.; Lakkaraju, H.; and Bastani, O. 2021. Learning models for actionable recourse. *Advances in Neural Information Processing Systems*, 34: 18734–18746.

Ruoss, A.; Balunovic, M.; Fischer, M.; and Vechev, M. 2020. Learning certified individually fair representations. *Advances in neural information processing systems*, 33: 7584–7596.

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

Vargo, A.; Zhang, F.; Yurochkin, M.; and Sun, Y. 2021. Individually fair gradient boosting. *arXiv preprint arXiv:2103.16785*.

Washington, A. L. 2018. How to argue with an algorithm: Lessons from the COMPAS-ProPublica debate. *Colo. Tech. LJ*, 17: 131.

Xu, H.; Liu, X.; Li, Y.; Jain, A.; and Tang, J. 2021. To be robust or to be fair: Towards fairness in adversarial training. In *International Conference on Machine Learning*, 11492–11501. PMLR.

Yeom, S.; and Fredrikson, M. 2020. Individual fairness revisited: Transferring techniques from adversarial robustness. *arXiv preprint arXiv:2002.07738*.

Yurochkin, M.; Bower, A.; and Sun, Y. 2019. Training individually fair ML models with sensitive subspace robustness. *arXiv preprint arXiv:1907.00020*.

Yurochkin, M.; and Sun, Y. 2020. Sensei: Sensitive set invariance for enforcing individual fairness. *arXiv preprint arXiv:2006.14168*.