Providing Fair Recourse over Plausible Groups

Jayanth Yetukuri[1], Ian Hardy[1], Yevgeniy Vorobeychik[2], Berk Ustun[3], Yang Liu[1]

[1]Computer Science and Engineering, University of California, Santa Cruz, CA, USA
[2]Computer Science and Engineering, Washington University in St. Louis, MO, USA
[3]Halıcıoğlu Data Science Institute, University of California, San Diego, CA, US
jyetukur@ucsc.edu, ihardy@ucsc.edu, eug.vorobey@gmail.com, berk@ucsd.edu, yangliu@ucsc.edu,

Abstract

Machine learning models now automate decisions in applications where we may wish to provide recourse to adversely affected individuals. In practice, existing methods to provide recourse return actions that fail to account for latent characteristics that are not captured in the model (e.g., age, sex, marital status). In this paper, we study how the cost and feasibility of recourse can change across these *latent groups*. We introduce a notion of group-level plausibility to identify groups of individuals with a shared set of latent characteristics. We develop a general-purpose clustering procedure to identify groups from samples. Further, we propose a constrained optimization approach to learn models that equalize the cost of recourse over latent groups. We evaluate our approach through an empirical study on simulated and real-world datasets, showing that it can produce models that have better performance in terms of overall costs and feasibility at a group level.

Introduction

Machine learning models now automate decisions that affect individuals – be it to provide a loan (Siddiqi 2012), a job interview (Ajunwa et al. 2016), or a public service (Chouldechova et al. 2018). Models in such settings should provide *recourse* (Ustun, Spangher, and Liu 2019) – i.e., actions that let individuals overturn their decisions through changes in feature spaces.

Existing methods for recourse provision may output actions that exhibit biases across groups in a target population. Such biases may affect the difficulty or feasibility of recourse. For example, research (Espinosa et al. 2019) suggests that race has a profound correlation with the level of education a person has access to. In the context of a lending model, this relationship would imply that actions that are identical may have diverging "actionability" across protected racial groups. In practice, they may arise due to historical biases within the training data (see e.g., Khosla et al. 2012) or due to the underlying model(see e.g., DeBrusk 2018; Mehrabi et al. 2021).

Some existing literature seeks to address these issues through interventions at the group level. For example, Von Kügelgen et al. (2022) considers an individual's hidden feature(s) in recourse generation, using group-level information to provide subsidies. Likewise, Madras et al. (2019) identify *hidden confounders*, which are unobserved factors that alter the cost and feasibility of recourse at an individual level.

Gupta et al. (2019) argues that negatively impacted individuals from different groups should have equal chances of obtaining recourse, seeking to equalize the distance from the decision boundary across groups.

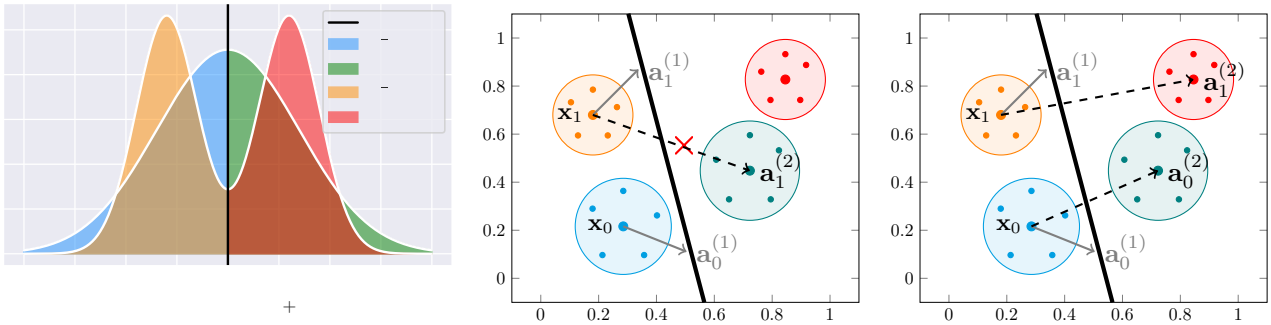
This study aims to consider the actionability at the group level instead of relying on a universal cost function. Consider an individual who applies for a loan and gets denied; we answer:

> "*What actions can I take to be part of the approved subgroup of people with my socioeconomic background?*"

The difference between the notion of *group-level fair actionability* and *fair recourse* is demonstrated using Figure 1 (a). Here, feature distribution for `working hours` follows a high variance unimodal distribution for group A_0, whereas we notice bimodal distribution for group A_1, implying that higher plausibility regime (of recourses) for group A_0 is closer to the decision boundary compared to A_1. Additionally, Figure 1 (b) shows the decision boundary using a scatter plot. Low density of individuals near the decision boundary for A_1, makes the recourse $\mathbf{a}_1^{(1)}$ predominantly undesirable in comparison with $\mathbf{a}_0^{(1)}$ for A_0. Alternatively, $\mathbf{a}_0^{(2)}$ and $\mathbf{a}_1^{(2)}$ from Figure 1 (c) shows post action features which fall within the corresponding high-density regions.

Group-level recourse plausibility of a post-action feature is defined as its believability or realizability with respect to the distribution of the group-specific approved sub-population. Given the spatial proximity nature (Gustafson and Parker 1994) of plausibility, we observe that: "plausibility of post-action features is proportional to the *density* of the resulting region and *similarity* with the resulting region of approved profiles."

This study leverages the *group-level approved sub-population* signals to understand actionability and thereby train a fair actionable model. Here, a group can be any immutable categorical feature in your dataset. We argue that a recourse $\mathbf{a}_0$ for an individual $\mathbf{x}_0 \in \mathcal{H}^-$ has higher chances of actionability if $\mathbf{x}_0 + \mathbf{a}_0 \in \mathcal{H}^+$, where $\mathcal{H}^+$ is the distribution of the approved group to which $\mathbf{x}_0$ belongs.



(a) Group Feature Distributions (b) Implausible $\mathbf{a}_1^{(2)}$ (c) Plausible $\mathbf{a}_0^{(2)}$ and $\mathbf{a}_1^{(2)}$

Figure 1: Toy example scenario demonstrating the existence of group-level actionability unfairness. In these figures, *orange* and *red* represent the negatively (A_1^-) and positively (A_1^+) affected sub-populations of the disadvantaged group (A_1), respectively. *Blue* and *Green* represents the negatively (A_0^-) and positively (A_0^+) affected sub-populations of the advantaged group (A_0). Consider a hypothetical situation where the average cost of recourse $\mathbf{a}_0^{(1)}$ and $\mathbf{a}_1^{(1)}$ for similar individuals $\mathbf{x}_0$ and $\mathbf{x}_1$ from A_0 and A_1, respectively, is identical. Such recourse can be commonly followed by A_0 but not necessarily by A_1.

Motivating Scenarios

We describe two real-world scenarios for motivation for **loan approval**. Applicant A belongs to the *old* group, whereas Applicant B belongs to the *young* group, and both of them have approached a bank for a loan. Both the individuals' loan applications were denied by the bank and were suggested a similar recourse.

Applicant A: Single Parent. The recourse provided by the bank suggests increasing their working hours from 32 per week to 40 per week. Considering that they belong to the sub-population of *denied single parent*, the recourse may not be actionable, as they may not have the flexibility of increasing working hours per week. They are more likely to consider taking a second *remote job* instead. Hence, recourse actions that align with those of other single parents help improve the actionability and benefit such disadvantaged groups.

Applicant B: International Student. Applicant B is an undocumented employee with severe restrictions due to his immigration status, often limiting their flexibility in acting on the recourse provided. He may need more capabilities to act upon several features such as *income, working hours, job sector* etc. Such constraints are further exacerbated if Applicant B is a student. For the holistic benefit of society and improved trust in machine learning systems, the suggested recourses must be unbiased in terms of plausibility metrics. The main contributions of this work include:

1. We introduce a notion of group-level plausibility using latent characteristics related to immutable categorical features.
2. We introduce a fairness notion *group-level plausibility bias* and provide metrics for quantification using a general purpose clustering procedure.
3. We provide evidence of group-level plausibility bias using a real-world dataset dataset to show its detrimental effects on the trustworthiness of a model.
4. We consolidate the traditional performance metrics of recourse generation and compare the proposed fairness metric between naturally trained models and trained with our proposed optimization.

Broader Impacts

This work is primarily designed to mitigate specific failure modes of machine learning models used in consumer-facing applications such as lending, hiring, and the allocation of services. In particular, we seek to study how these models can assign predictions that are difficult or impossible to change across groups that are difficult to identify using features that are not used by the model. Our work studies these biases in responsiveness through the lens of recourse and outlines a general-purpose approach to correct them. In particular, we (re)introduce plausible recourse as an alternative to a low-cost recourse.

Framework

We consider a classification task where a model $f : \mathcal{X} \to \mathcal{Y}$ assigns a binary label $\mathbf{y} \in \{\pm 1\}$ to an individual with features $\mathbf{x} = [x_1, \ldots, x_d] \in \mathcal{X} = \mathcal{X}_1 \times \cdots \times \mathcal{X}_d \subseteq \mathbb{R}^d$. Let $\mathcal{D} = \left\{\left(\mathbf{x}^{(i)}, \mathbf{y}^{(i)}\right)\right\}_{i=1}^n$ be the set of data samples observed from the true underlying distribution.

Let g observing values $g \in \mathcal{G} = \{1, \ldots, K\}$ denote a categorical attribute encodes a protected characteristics.

We define the following subspaces based on the true label $\mathbf{y}$ and predicted label $f(\mathbf{x})$: $\mathcal{D}^- = \{\mathbf{x} \in \mathcal{X} : \mathbf{y} = -1\}$, $\mathcal{D}^+ = \{\mathbf{x} \in \mathcal{X} : \mathbf{y} = +1\}$, $\mathcal{H}^- = \{\mathbf{x} \in \mathcal{X} : f(\mathbf{x}) = -1\}$, and $\mathcal{H}^+ = \{\mathbf{x} \in \mathcal{X} : f(\mathbf{x}) = -1\}$.

Let $v^{(i)} \in \mathcal{D}$ be a labeled example where each $v^{(i)}$ is associated with a group $g \in \mathcal{G}$. Given a *group membership function* $m : \mathbb{R}^d \to \{\pm 1\}$, we define $\mathcal{H}_g^+ = \{v^{(i)} \in \mathcal{D} |\ m\left(v^{(i)}\right) = g, f(\mathbf{x}) = +1\}$. and $\mathcal{H}_g^- = \{v^{(i)} \in \mathcal{D} |\ m\left(v^{(i)}\right) = g, f(\mathbf{x}) = -1\}$.

Recourse. Given an individual with features $\mathbf{x}_0$ such that $f(\mathbf{x}_0) = -1$, we return an action $\mathbf{a}_0$ that achieves recourse by solving an optimization problem of the form:

$$\begin{aligned} \min_{\mathbf{a}_0} \quad & \text{cost}(\mathbf{x}_0, \mathbf{a}_0) \\ \text{s.t.} \quad & f(\mathbf{x}_0 + \mathbf{a}_0) = +1, \\ & \mathbf{a}_0 \in \mathcal{A}(\mathbf{x}_0). \end{aligned} \tag{1}$$

Here, $\text{cost}(\mathbf{x}_0, \mathbf{a}_0) : \mathcal{A}(\mathbf{x}_0) \to \mathbb{R}^+$ is any cost function used to capture the difficulty of taking a set of actions $\mathbf{a}_0$ by an individual represented by $\mathbf{x}_0$ and let $\mathcal{A}(\mathbf{x}_0)$ be the set of feasible actions.

Measuring Plausibility

Recourse actions are traditionally specified by the cost of changes and actionability constraints (see e.g., *feasibility sets* of Karimi, Schölkopf, and Valera 2021). In this study, we intend to maximize the overall feasibility in terms of a *proximity score* $\text{prox}(\mathbf{x}_0 + \mathbf{a}_0, \mathcal{S}_0)$ for an individual $\mathbf{x}_0$ with respect to a user-specified *Exemplar Set* $\mathcal{S}_0$.

An *exemplar set* contains all clusters of predefined individuals with certain robust properties, including prevalence and model agnostic adversarial robustness. Given a classifier f, a set of feasibility constraints $\mathcal{A}(\mathbf{x}_0)$, we recover an action by solving the optimization problem:

$$\begin{aligned} \min_{\mathbf{a}_0} \quad & \text{cost}(\mathbf{x}_0, \mathbf{a}_0) \\ \text{s.t.} \quad & \text{prox}(\mathbf{x}_0 + \mathbf{a}_0, \mathcal{S}_0) \geq \rho, \\ & f(\mathbf{x}_0 + \mathbf{a}_0) = +1, \\ & \mathbf{a}_0 \in \mathcal{A}(\mathbf{x}_0). \end{aligned} \tag{2}$$

Here:

- $\text{prox}(\mathbf{x}_0 + \mathbf{a}_0, \mathcal{S}_0) : \mathcal{X} \to \mathbb{R}^+$ is a *proximity score* for the post-action features $\mathbf{x}_0 + \mathbf{a}_0$ to an *Exemplar Set* $\mathcal{S}_0$.
- ρ measures the minimum required proximity for $\mathbf{x}_0 + \mathbf{a}_0$ to be feasible and can vary for each group.

ρ can be specifically configured for every group based on the variance within $\mathcal{S}_0$. This ensures that $\mathbf{a}_0$ ensures underlying group characteristics. ρ ensures that $\mathbf{x}_0 + \mathbf{a}_0$ gets closer to $\mathcal{S}_0$. Configuring $\rho = 0$ returns a traditional low-cost action and $\rho > 0$ leads $\mathbf{x}_0 + \mathbf{a}_0$ to be within a specified width of $\mathcal{S}_0$, for example, an ε-ball around $\mathcal{S}_0$.

Let $\hat{\mathbf{x}}_0 = \mathbf{x}_0 + \mathbf{a}_0$ be the post action feature profile of $\mathbf{x}_0$. $\text{prox}(\hat{\mathbf{x}}_0, \mathcal{S}_0)$ estimates a plausibility score by capturing the proximity of $\hat{\mathbf{x}}_0$ to the closest exemplar set $\mathcal{S}_0$. Our choice is motivated by Karimi et al. (2020a)'s definition of: (i) domain-consistency; (ii) density-consistency; and (iii) prototypical-consistency.

Group Plausibility. For a the post-action feature profile $\hat{\mathbf{x}}_0$ for an individual $\mathbf{x}_0$ from a group g, we characterize *plausibility score* using the proximity nature of $\text{prox}(\hat{\mathbf{x}}_0, \mathcal{S}_0)$ as $\text{plaus}(\hat{\mathbf{x}}_0, \mathcal{S}_g)$ of a post-action feature profile $\hat{\mathbf{x}}_0$ with respect to any corresponding (approved) exemplar set $\mathcal{S}_g \in \mathcal{H}_g^+$, using:

$$\begin{aligned} &\text{plaus}(\hat{\mathbf{x}}_0, \mathcal{S}_g) \propto \text{density of } \mathcal{S}_g, \text{ and} \\ &\text{plaus}(\hat{\mathbf{x}}_0, \mathcal{S}_g) \propto \text{similarity with } \mathcal{S}_g \end{aligned} \tag{3}$$

We now define group plausibility using the patch proximity index (Gustafson and Parker 1994) used to quantify the spatial context of a patch in relation to its neighbors. In our context, we define the proximity of $\hat{\mathbf{x}}_0$ with respect to any resulting neighbors set $\mathcal{S}_g^{(i)} \in \mathcal{S}_g$.

Definition 1 (Group Plausibility)**.** *For any individual* $\mathbf{x}_0$ *in group* $g \in \mathcal{G}$*, we measure the* group-level recourse plausibility $\text{plaus}(\hat{\mathbf{x}}_0, \mathcal{S}_g)$ *of post-action features* $\hat{\mathbf{x}}_0$ *using:*

$$\begin{aligned} \text{plaus}(\hat{\mathbf{x}}_0, \mathcal{S}_g) := \max\Big\{ &\text{coverage}\left(\mathcal{S}_g^{(i)}\right) \\ &\times \text{similarity}\left(\hat{\mathbf{x}}_0, \mathcal{S}_g^{(i)}\right) : \mathcal{S}_g^{(i)} \in \mathcal{S}_g \Big\} \end{aligned} \tag{4}$$

where $\text{coverage}\left(\mathcal{S}_g^{(i)}\right)$ *measures the fraction of data points covered by* $\mathcal{S}_g^{(i)}$ *and* $\text{similarity}\left(\hat{\mathbf{x}}_0, \mathcal{S}_g^{(i)}\right)$ *provides a score of how similar* $\hat{\mathbf{x}}_0$ *is with respect to* $\mathcal{S}_g^{(i)}$*, respectively.*

We maximize the proximity score of the resulting post-action features with respect to any $\mathcal{S}_g^{(i)} \in \mathcal{S}_g$. The resulting $\hat{\mathbf{x}}_0$ must be closer to any of the exemplar profile clusters irrespective of the proximity score with other clusters.

Alternatively, $mean$ based proximity score $\sum_{S_g^{(i)} \in S_g} \text{coverage}\left(S_g^{(i)}\right) \times \text{similarity}\left(\hat{\mathbf{x}}_0, S_g^{(i)}\right)$ fails in the following scenario in our formulation.

Let $\text{plaus}(\hat{\mathbf{x}}_0, \mathcal{S}_0) = 2.0$ with two clusters having $\text{coverage}\left(\mathcal{S}_g^{(1)}\right) \times \text{similarity}\left(\hat{\mathbf{x}}_0, \mathcal{S}_g^{(1)}\right) = 2.0$ and $\text{coverage}\left(\mathcal{S}_g^{(2)}\right) \times \text{similarity}\left(\hat{\mathbf{x}}_0, \mathcal{S}_g^{(2)}\right) = 2.0$. Here, the resulting profile is not specifically closer to any of the exemplar sets.

Equalizing Recourse across Plausible Groups

In this section, we introduce *exemplar* set, our proposed metric to measure the plausibility of a post-action feature profile, and introduce a notion of plausibility bias. Then, we propose an optimization based model training technique to alleviate such bias caused at the group level.

Specifying an Exemplar Set

Action plausibility does not rely on the traditional cost of actions due to its prototypical nature (Karimi et al. 2020a). This is unlike the traditional model decision boundary based low-cost actions. This provides degrees of freedom to capture individual action costs. For example, a low-density cluster signals profiles that are more likely to be outliers, which are possible to attain but *peculiar or atypical* for most individuals from that group.

The proposed plausibility metric captures the individual's group-level desirability of the actions. Identification of $\mathcal{G}$ should be done with care to ensure that it will not lead to inadvertent discrimination across protected groups.

Our study is motivated by the fact that an individual is more likely to enact actions that have led to approval for individuals in their exemplar group.

We define groups based on the prevalence of feature values.

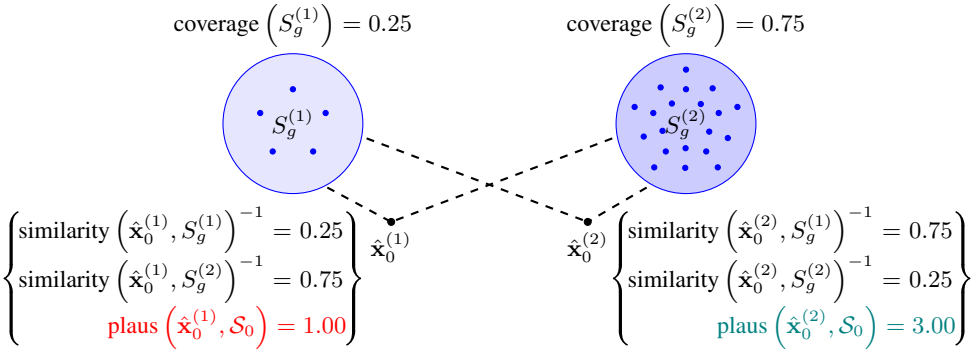


Figure 2: Demonstration of the effectiveness of plaus $(\hat{\mathbf{x}}_0, \mathcal{S}_0)$. $\hat{\mathbf{x}}_0^{(2)}$ has a high plaus $\left(\hat{\mathbf{x}}_0^{(2)}, \mathcal{S}_0\right)$ due to its high coverage $\left(S_g^{(2)}\right)$ and similarity $\left(\hat{\mathbf{x}}_0^{(2)}, S_g^{(1)}\right)^{-1}$, unlike $\hat{\mathbf{x}}_0^{(1)}$ which has a low plaus $\left(\hat{\mathbf{x}}_0^{(1)}, \mathcal{S}_0\right)$.

We start by clustering the approved profiles of the group g from the training dataset into c clusters $S_g = \left\{\mathcal{S}_g^{(1)}, \ldots, \mathcal{S}_g^{(c)}\right\}$, where c is a hyperparameter selected by a domain expert. The details of the main procedure are as follows:

1. We estimate the *density* of each cluster $\mathcal{S}_g^{(i)} \in \mathcal{S}_g$ using the training dataset. We cluster approved data samples from the training dataset and associate a coverage score coverage $\left(\mathcal{S}_g^{(i)}\right)$ to each cluster. The choice of clusters must satisfy:
   1) *Positive Coverage:* coverage $\left(\mathcal{S}_g^{(i)}\right) > 0 \; \forall \; \mathcal{S}_g^{(i)} \in \mathcal{S}_g$,
   2) *Total coverage:* $\sum_{\mathcal{S}_g^{(i)} \in \mathcal{S}_g}$ coverage $\left(\mathcal{S}_g^{(i)}\right) = 1$.
2. The number of clusters c is domain dependent and can influence the average plaus $(\cdot)$ score. Please note that any choice of c should be identical across all the groups for consistency of plaus $(\cdot)$.
   For both the special cases of $c = 1$, and of $c = |\mathcal{D}_g^+|$ where $|\mathcal{D}_g^+| = |\mathcal{D}_{g'}^+| : \forall g, g' \in \mathcal{G}$, we have plaus $(\cdot) \propto$ similarity $(\cdot)$. In the former scenario, we have 1 cluster per group, and in the latter scenario, we have $|\mathcal{D}_g^+|$ clusters for every $g \in \mathcal{G}$.
3. Similarity score similarity $\left(\hat{\mathbf{x}}_0, \mathcal{S}_g^{(i)}\right)$ of the post-action feature profile $\hat{\mathbf{x}}_0$ with respect $\mathcal{S}_g^{(i)}$ can be approximated using any ℓ_p norm based distance metric. We choose ℓ_2 norm-based distance metrics to estimate the similarity score for our experiments.

Measuring Plausibility Bias

Our formulation of plausibility draws on group level information, which requires a closer look at differences across groups. Existing literature focuses on equalizing recourse costs across groups (Gupta et al. 2019).

However, fairness in terms of the traditional *cost* function, which is approximated using a distance metric from the factual profile, may not capture the unfairness in plausibility. To address this blind spot, we propose to capture a straightforward notion of *group-level* unfairness in plausibility.

We start with a measure of the group-level plausibility-based unfairness measure for a classifier f.

Definition 2 (Expected plausibility)**.** *The expected plausibility of recourse for a classifier $f : \mathcal{X} \to \{\pm 1\}$ over $\mathcal{H}^-$ is:* $\overline{\text{plaus}}_{\mathcal{H}^-}(f) = \mathbb{E}_{\mathcal{H}^-, \mathcal{D}^+}[\text{plaus}(\hat{\mathbf{x}}_0, \mathcal{S}_0)]$*, where $\hat{\mathbf{x}}_0$ is the post-action feature profile resulting from solving the optimization problem in (2).*

Definition 3 (Group plausibility bias)**.** *The group-level plausibility unfairness of a classifier f for a dataset $\mathcal{D}$ is measured as:* $\Delta_{\mathcal{P}} := \max_{g,g' \in \mathcal{G}} \left| \overline{\text{plaus}}_{\mathcal{H}_g^-}(f) - \overline{\text{plaus}}_{\mathcal{H}_{g'}^-}(f) \right|$.

where $\overline{\text{plaus}}_{\mathcal{H}_g^-}(f)$ is the group average of $\overline{\text{plaus}}(\hat{\mathbf{x}}_0, f) : \forall \, \mathbf{x}_0 \in \mathcal{H}_g^-$.

Our work advocates for equalized plausibility across protected groups. We propose an optimization-based modeling procedure we call "Fair Feasible Training" (FFT) to train a model with an additional bias constraint.

We now alleviate the effects of plausibility bias. Gupta et al. (2019) equalizes recourse action costs across groups, while we propose to train models that equalize recourse across latent groups by including $\Delta_{\mathcal{P}}$ as part of the model training procedure.

Definition 4 (Fair Feasible Training)**.** *Given a dataset $\mathcal{D} = \left\{\left(\mathbf{x}^{(i)}, \mathbf{y}^{(i)}\right)\right\}_{i=1}^n$ and $\epsilon > 0$, we train a feasibly fair classifier f by solving the following optimization problem:*

$$\begin{aligned} \min \quad & L(x, y) \\ \textit{s.t.} \quad & \max_{g,g' \in \mathcal{G}} \left| \overline{\text{plaus}}_{\mathcal{H}_g^-}(f) - \overline{\text{plaus}}_{\mathcal{H}_{g'}^-}(f) \right| \leq \epsilon. \end{aligned} \tag{5}$$

where $L(x, y)$ is overall loss aggregated across $\mathcal{D}$ and we approximate $\overline{\text{plaus}}_{\mathcal{H}_g^-}(f)$ using $\overline{\text{plaus}}_{\mathcal{D}_g^-}(f)$ during the training process. $\overline{\text{plaus}}_{\mathcal{D}_g^-}(f)$ measures the mean distance of denied individuals of group g to their approved group counterparts, using the training dataset.

Model	Method	Standard Training						Proposed Training					
		Succ. Rate	Avg. Tim.	Con. Vio.	Red.	Pro.	Spa.	Succ. Rate	Avg. Tim.	Con. Vio.	Red.	Pro.	Spa.
N.N.	GS	1.00	0.03	0.00	2.63	1.14	4.97	1.00	0.03	0.00	2.10	1.24	5.05
	Wachter	1.00	0.05	2.00	3.20	1.25	6.94	1.00	0.05	2.00	1.77	1.42	6.95
	AR$_{\text{(-LIME)}}$	0.51	1.72	0.00	0.00	1.34	1.60	0.76	1.94	0.00	0.00	1.31	1.50
	CCHVAE	1.00	0.11	3.73	7.82	3.11	8.64	1.00	0.28	3.74	7.80	3.13	8.64
	FACE	1.00	4.37	4.81	6.63	4.35	7.84	1.00	4.46	4.66	6.39	4.34	7.79
L.R.	GS	1.00	0.02	0.00	2.30	1.50	5.32	1.00	0.02	0.00	2.19	1.74	5.59
	Wachter	1.00	0.05	2.00	2.00	1.38	6.94	1.00	0.06	2.00	1.43	1.69	6.92
	AR	0.80	1.84	0.00	0.00	1.81	1.98	0.80	2.14	0.00	0.00	1.52	1.64
	CCHVAE	1.00	0.17	3.74	8.78	3.77	9.29	1.00	0.22	3.71	3.33	3.91	9.41
	FACE	1.00	4.29	4.72	6.57	4.42	7.87	1.00	5.83	4.69	6.11	4.49	7.71

Table 1: Overview of recourse actions for models trained using baseline methods and our approach on the Adult Income dataset. Reference– Succ. Rate: Success Rate, Avg. Tim.: Average Time, Con. Vio.: Constraint Violations, Red.: Redundancy, Pro.:Proximity, Spa.: Sparsity.

The main idea for this approximation is to equalize the spread between approved and denied sub-populations across groups during model training. With the proposed optimization, any existing recourse methodologies can be used to achieve equalized group-level plausibility across groups. An alternate approach of post-training based technique carries the risk of increased recourse costs for disadvantaged groups.

Experiments

In this section, we present empirical results to show that the traditional approaches for recourse provision lead to plausibility bias and that our proposed approach (FFT) can mitigate these effects.

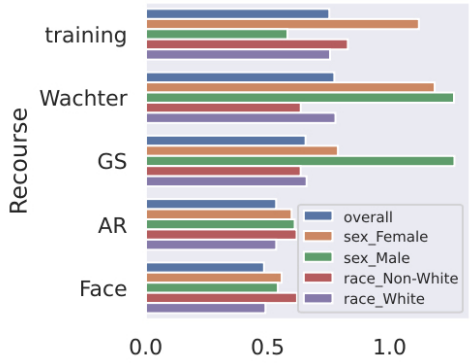
Setup

We train two kinds of classification models on the *Adult Income* dataset: Neural Networks (NN) and Logistic Regression (LR). For each model class, we fit a model using a baseline algorithm that optimizes cross-entropy loss and another using our proposed risk minimization in (4) utilizing the `Male` and `Female` sub-populations as the constraining groups.

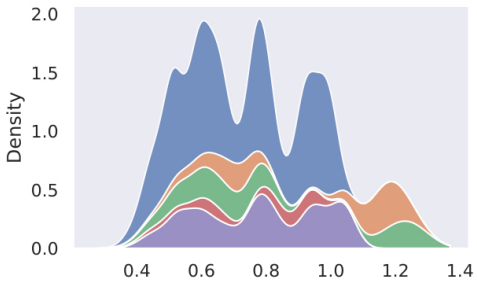
The NN models contain three layers of $[18, 9, 3]$ nodes with *ReLU* activation functions, a standard drawn from the CARLA (Pawelczyk et al. 2021) recourse package.

All models achieved comparable accuracy on the holdout set: the standard and constrained NN models denoted by θ_{nn}^{std}, θ_{nn}^{fft} saw 78.8% and 79.4% accuracy, respectively. While the standard and constrained LR models denoted by θ_{lr}^{std} and θ_{lr}^{fft} saw 79.2% and 78.6% accuracy, respectively. We chose *sex_Female* as our protected group for our experiments.

Recourse methods. Although our experiments focus on one protected group, we note that the selection of groups can be parameterized to capture all the necessary groups. For all

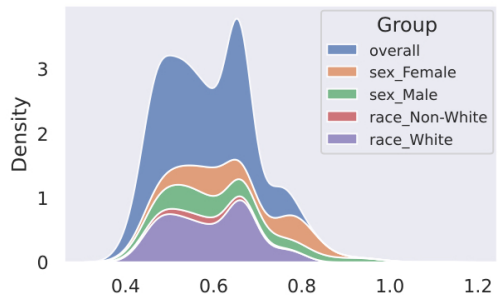


(a) Comparison chart of $\overline{\text{plaus}}_{(\cdot)}(f)$

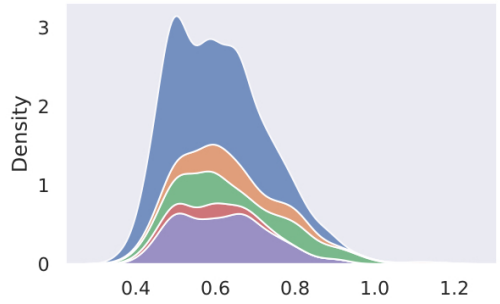


(b) Stacked distribution of $\overline{\text{plaus}}_{\mathcal{D}^+_{(\cdot)}}(f)$

Figure 3: $\overline{\text{plaus}}_{(\cdot)}(f)$ of various recourse techniques for *gender* and *race* groups. For reference, $\overline{\text{plaus}}_{\mathcal{D}^+_{(\cdot)}}(f)$ for training data is also shown in image (a). Image (b) visualizes distributional differences of $\overline{\text{plaus}}_{\mathcal{D}^+_{(\cdot)}}(f)$ across immutable groups.



(b) Standard training



(c) Fair Feasible Training

Figure 4: Stacked distribution of $\overline{\text{plaus}}_{(\cdot)}(f)$ illustrates the distribution of plausibility scores across groups.

models, we then calculated a variety of recourse options on a sample of 500 adversely impacted individuals.

Recourse Methods used for our experiments are: Wachter (Wachter, Mittelstadt, and Russell 2017), Growing Spheres (GS) (Laugel et al. 2017), Actionable Recourse (AR) (Ustun, Spangher, and Liu 2019), Feasible and Actionable Counterfactual Explanations (FACE) (Poyiadzi et al. 2020) and CCHVAE (Pawelczyk, Broelemann, and Kasneci 2020).

Results & Discussion

We provide evidence of several forms of plausibility bias. For instance, we identify that a particular *feature* distribution of a population categorized by strategically identifying protected groups shows idiosyncrasies across these groups.

On Group Level Effects Firstly, we show that feature distributions vary significantly at the immutable feature level. The distribution of `age`, `education-num` and `hours-per-week` for the *Adult Income* (Dua and Graff 2017) dataset, when stratified by group shows the distributional uniqueness of individual protected groups (corresponding figures are included in the Appendix). For example, we observe twin peaks for *single woman* in `education-num`, which suggests that any recourse that lands the individual in the low-frequency region may not be actionable. The similar small second peak for *single woman* can be observed for *hours-per-week* feature.

Recourse Performance Metrics. Our results in Table 1 show that performance is remarkably consistent for FFT. Although FFT often incurs longer recourse generation times (seeing an average 24.6% increase in run time across recourse methods), it consistently identifies recourse that shows lower redundancy (an average 31.7% reduction). This is somewhat surprising; although we hypothesize that FFT learns more separable data representations, which may impact the ultimate redundancy of generated recourse. We observe that overall proximity costs are not significantly affected by FFT. Rather, FFT constrains the ultimate recourse to be feasibly fair to protected groups. Although recourse proximity fairness is not explicitly included in the cost function, we suspect the ultimate gains in proximity fairness result from learning a max-margin classifier on underlying fair representations.

Standard Training Exhibits Plausibility Bias. Figure 3 (a) shows $\overline{\text{plaus}}_{(\cdot)}(f)$ for the recourse actions generated by Wachter, GS, AR, FACE. Figure 3 (b) further shows the distributional differences of $\overline{\text{plaus}}_{(\cdot)}(f)$ at an individual level for the raw dataset.

FFT moderates Plausibility Bias. We compare the plaus $(\cdot)$ distributional differences across individuals based on their prediction, group, true label, and model. We observe from Figure 4 that the proposed training induces a consistent uni-modal plaus $(\cdot)$ distribution across groups, while standard training results in bimodal feasibility scores where female individuals in particular, see higher feasibility costs. To assess the fairness performance of FFT, we compare:

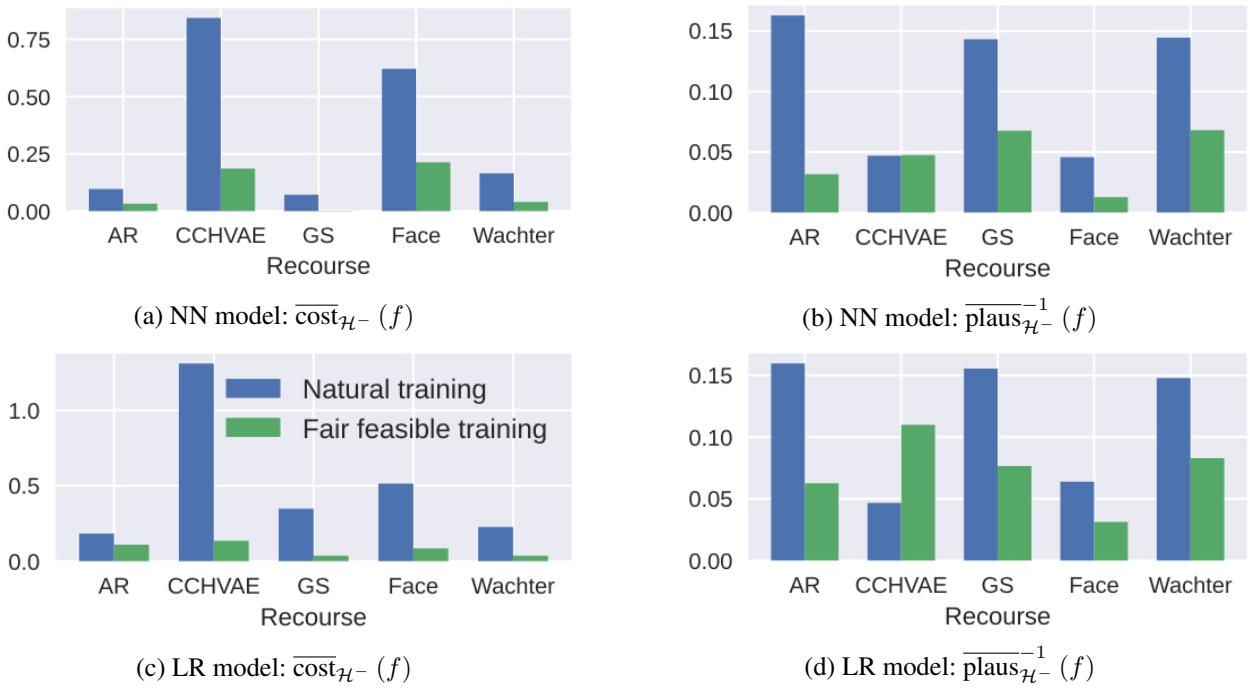
- *Expected recourse cost of a classifier (Ustun, Spangher, and Liu 2019),* $\overline{\text{cost}}_{\mathcal{H}^-}(f)$: measured as the average ℓ_2 distance $\hat{\mathbf{x}}_0$ and $\mathbf{x}_0$, of the protected groups used to constrain the training process.
- *Expected Plausibility of a classifier (Definition 2),* $\overline{\text{plaus}}_{\mathcal{H}^-}(f)$: measured as the inverse average ℓ_2 distance of $\hat{\mathbf{x}}_0$ and corresponding exemplar set $\mathcal{S}_0$ of its associated positive group.

Our findings are shown in Figure 5. We observe that for both model families, FFT consistently provides recourse that is fairer in terms of plausibility and the overall cost.

Concluding Remarks

In this work, we outlined a new approach to account for latent groups in applications where we wish to provide recourse. In particular, we developed machinery to identify such groups from data and studied the implicit disparity in plausibility across these groups. For example, suggesting naive and arguably famous recourse action of increasing the working hours to a *single parent* is not feasible. We proposed a method to train classifiers to mitigate these effects and demonstrated their capacity in practice.

Limitations. Group-level plausibility may not ensure individual actionability (see e.g., Kothari et al. 2023). Our proposed approach may also exacerbate the cost of recourse. Our study raises the question of whether it is sufficient for a recourse to change the model's decision or whether a recourse improves the affected individual's overall group-level profile.



(a) NN model: $\overline{\text{cost}}_{\mathcal{H}^-}(f)$

(b) NN model: $\overline{\text{plaus}}^{-1}_{\mathcal{H}^-}(f)$

(c) LR model: $\overline{\text{cost}}_{\mathcal{H}^-}(f)$

(d) LR model: $\overline{\text{plaus}}^{-1}_{\mathcal{H}^-}(f)$

Figure 5: Feasibility performance metrics for NN and LR models across a variety of recourse methods.

Related Work. Our work is related to a previous study (Karimi, Schölkopf, and Valera 2021) where the authors referred to it as *believability or realizability* of recourse and refers to the likeness of the counterfactual profile resulting from the suggested set of actions. (Karimi et al. 2020a) refers plausibility as (i) domain-consistency; (ii) density-consistency; and (iii) prototypical-consistency. Providing recourse based on *manifold learning* (Pawelczyk, Broelemann, and Kasneci 2020) motivates us to utilize underlying group distributions for suggesting group-level data-dependent recourse that accounts for group-level actionability patterns (Yetukuri 2023). Manifold-based CCHVAE (Pawelczyk, Broelemann, and Kasneci 2020) generates high-density counterfactuals using a latent space model. However, there is often no guarantee that the *what-if* scenarios identified are attainable. Another line of research (Karimi et al. 2020a) leverages causal knowledge (Karimi, Schölkopf, and Valera 2021) to identify *recourse via minimal interventions*. Taking causal knowledge is beneficial for identifying a recourse; however, the true underlying structural causal model is often unavailable (Karimi et al. 2020b).

Density-based *soft constraints* are essential for capturing group-level feasibility signals. FACE (Poyiadzi et al. 2020) follows high-density paths to produce *feasible counterfactual explanations*, establishing the necessary condition of density for a feasible recourse. However, such *feasible paths* may not exist for certain groups if the approved and denied subpopulations are significantly farther apart than other groups. Other studies that learn from the dataset's underlying structure include REVISE (Joshi et al. 2019) and CRUDS (Downs et al. 2020). However, existing literature does not consider the distributional differences across groups while suggesting a recourse leading to *plausibility bias* across groups. We differ from existing literature, which prioritizes distance to the decision boundary by evaluating the actionability of recourse with respect to the distance to $\mathcal{H}^+$.

Acknowledgments

This work is partially supported by the National Science Foundation (NSF) under grants IIS-2313105, IIS 2040880, IIS-2143895, IIS-2040800, CCF-2023495, IIS-2214141, IIS-1905558, IIS-1939677 and Amazon.

References

Ajunwa, I.; Friedler, S.; Scheidegger, C. E.; and Venkatasubramanian, S. 2016. Hiring by algorithm: predicting and preventing disparate impact. *Available at SSRN*.

Chouldechova, A.; Benavides-Prado, D.; Fialko, O.; and Vaithianathan, R. 2018. A case study of algorithm-assisted decision making in child maltreatment hotline screening decisions. In *Conference on Fairness, Accountability and Transparency*, 134–148. PMLR.

DeBrusk, C. 2018. The risk of machine-learning bias (and how to prevent it). *MIT Sloan Management Review*.

Downs, M.; Chu, J. L.; Yacoby, Y.; Doshi-Velez, F.; and Pan, W. 2020. Cruds: Counterfactual recourse using disentangled subspaces. *ICML WHI*, 2020: 1–23.

Dua, D.; and Graff, C. 2017. UCI Machine Learning Repository.

Espinosa, L. L.; Turk, J. M.; Taylor, M.; and Chessman, H. M. 2019. Race and ethnicity in higher education: A status report.

Gupta, V.; Nokhiz, P.; Roy, C. D.; and Venkatasubramanian, S. 2019. Equalizing Recourse across Groups. *ArXiv*, abs/1909.03166.

Gustafson, E. J.; and Parker, G. R. 1994. Using an index of habitat patch proximity for landscape design. *Landscape and urban planning*, 29(2-3): 117–130.

Joshi, S.; Koyejo, O.; Vijitbenjaronk, W.; Kim, B.; and Ghosh, J. 2019. Towards Realistic Individual Recourse and Actionable Explanations in Black-Box Decision Making Systems. *Safe Machine Learning workshop at ICLR*.

Karimi, A.-H.; Barthe, G.; Schölkopf, B.; and Valera, I. 2020a. A survey of algorithmic recourse: definitions, formulations, solutions, and prospects. *arXiv preprint arXiv:2010.04050*.

Karimi, A.-H.; Schölkopf, B.; and Valera, I. 2021. Algorithmic Recourse: From Counterfactual Explanations to Interventions. In *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency*, FAccT '21, 353–362. New York, NY, USA. ISBN 9781450383097.

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

Khosla, A.; Zhou, T.; Malisiewicz, T.; Efros, A. A.; and Torralba, A. 2012. Undoing the damage of dataset bias. In *European Conference on Computer Vision*, 158–171. Springer.

Kothari, A.; Kulynych, B.; Weng, T.-W.; and Ustun, B. 2023. Prediction without Preclusion: Recourse Verification with Reachable Sets. *arXiv preprint arXiv:2308.12820*.

Laugel, T.; Lesot, M.-J.; Marsala, C.; Renard, X.; and Detyniecki, M. 2017. Inverse Classification for Comparison-based Interpretability in Machine Learning. *stat*, 1050: 22.

Madras, D.; Creager, E.; Pitassi, T.; and Zemel, R. 2019. Fairness through Causal Awareness: Learning Causal Latent-Variable Models for Biased Data. In *Proceedings of the Conference on Fairness, Accountability, and Transparency*, FAT* '19, 349–358. New York, NY, USA: Association for Computing Machinery. ISBN 9781450361255.

Mehrabi, N.; Morstatter, F.; Saxena, N.; Lerman, K.; and Galstyan, A. 2021. A Survey on Bias and Fairness in Machine Learning. *ACM Comput. Surv.*, 54(6).

Pawelczyk, M.; Bielawski, S.; van den Heuvel, J.; Richter, T.; and Kasneci, G. 2021. CARLA: A Python Library to Benchmark Algorithmic Recourse and Counterfactual Explanation Algorithms. In *Proceedings of the Neural Information Processing Systems Track on Datasets and Benchmarks 1, NeurIPS Datasets and Benchmarks 2021, December 2021*.

Pawelczyk, M.; Broelemann, K.; and Kasneci, G. 2020. Learning Model-Agnostic Counterfactual Explanations for Tabular Data. In *Proceedings of The Web Conference 2020*, WWW '20, 3126–3132. New York, NY, USA: Association for Computing Machinery. ISBN 9781450370233.

Poyiadzi, R.; Sokol, K.; Santos-Rodriguez, R.; De Bie, T.; and Flach, P. 2020. FACE: feasible and actionable counterfactual explanations. In *Proceedings of the AAAI/ACM Conference on AI, Ethics, and Society*, 344–350.

Siddiqi, N. 2012. *Credit risk scorecards: developing and implementing intelligent credit scoring*, volume 3. John Wiley & Sons.

Ustun, B.; Spangher, A.; and Liu, Y. 2019. Actionable recourse in linear classification. In *Proceedings of the conference on fairness, accountability, and transparency*, 10–19.

Von Kügelgen, J.; Karimi, A.-H.; Bhatt, U.; Valera, I.; Weller, A.; and Schölkopf, B. 2022. On the fairness of causal algorithmic recourse. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 36, 9584–9594.

Wachter, S.; Mittelstadt, B.; and Russell, C. 2017. Counterfactual explanations without opening the black box: Automated decisions and the GDPR. *Harvard Journal of Law and Technology*, 31: 841.

Yetukuri, J. 2023. Individual and Group-Level Considerations of Actionable Recourse. In *Proceedings of the 2023 AAAI/ACM Conference on AI, Ethics, and Society*, AIES '23, 1008–1009. New York, NY, USA: Association for Computing Machinery. ISBN 9798400702310.