INSPIRE: A FRAMEWORK FOR INTEGRATING INDIVIDUAL USER PREFERENCES IN RECOURSE

Anonymous authors
Paper under double-blind review

ABSTRACT

Most recourse generation approaches optimize for *indirect* distance-based metrics like diversity, proximity, and sparsity, or a shared cost function across all users to generate recourse. The latter is an unrealistic assumption because users can have diverse feature preferences which they might be willing to act upon and any changes to any undesirable feature might lead to an impractical recourse. In this work, we propose a novel framework to incorporate the *individuality* of users in both recourse generation and evaluation procedure by focusing on the cost incurred by a user when opting for a recourse. To achieve this, we first propose an objective function, *Expected Minimum Cost* (EMC) that is based on two key ideas: (1) the user should be comfortable adopting at least one solution when presented with multiple options, and (2) we can approximately optimize for users' satisfaction even when their true cost functions (i.e., costs associated with feature changes) are unknown. EMC samples multiple plausible cost functions based on diverse feature preferences in the population and then finds a recourse set with one good solution for each category of user preferences. We optimize EMC with a novel discrete optimization algorithm, *Cost-Optimized Local Search* (COLS), that is guaranteed to improve the quality of the recourse set over iterations. Our evaluation framework computes the fraction of satisfied users by simulating each user's cost function and then computing the incurred cost for the provided recourse set. Experimental evaluation on popular real-world datasets demonstrates that our method satisfies up to 25.9% more users compared to strong baselines. Moreover, human evaluation shows that our recourses are preferred more than twice as often as the strongest baseline.[1]

1 INTRODUCTION

Over the past few years, ML models have been increasingly deployed to make critical decisions related to loan approval (Siddiqi, 2012), allocation of public resources (Chouldechova et al., 2018), and hiring decisions (Ajunwa et al., 2016). These decisions have real-life consequences for the involved users. As a result, there is a growing emphasis on explaining these models' decisions (Poulin et al., 2006; Ribeiro et al., 2018) and providing *recourse* for unfavorable decisions (Voigt & dem Bussche, 2018). A *recourse* is an actionable plan that allows a user to change the decision of a deployed model to a desired alternative (Wachter et al., 2017). Recourses are often presented to users as a set of *counterfactuals* (cfs), where each cf details the changes to the *user's state vector* (i.e., their feature vector). Recourses are desired to be *actionable*, and *feasible*. *Actionable* means that only features which can be changed by the user are requested to be changed. A recourse is *feasible* if it is easy for the user to adopt, in other words, it is actionable and has a low cost for the user.

To achieve these objectives, prior work used feature distance-based objectives like *proximity*, *sparsity*, and *feature diversity*. For instance, Mothilal et al. (2020) and Wachter et al. (2017) encourage *proximity* by minimizing the distance between the user's state vector and the counterfactuals (cfs) with the assumption that proximal cfs are easier to adopt. Whereas, *sparsity* quantifies the number of features that require modification to implement a recourse (Mothilal et al., 2020). In contrast to these, *feature diversity* (Mothilal et al., 2020; Cheng et al., 2021) provides a user with multiple cfs that change diverse subsets of features assuming that users are more likely to find at least one feasible

[1] Our code is uploaded as supplementary material.

solution. These objectives capture the desired properties of recourses but do not account for individual user preferences that should be the primary objective. For instance, if a user prefers to change features f_1 and f_2, then providing them with recourses that change undesirable features make them infeasible even if they are proximal, sparse, and diverse. To address this, some recourse methods define a single cost function that is shared by all the users. A cost function $\mathcal{C}(f, i, j)$ denotes the cost of changing a feature f from value i to j. They optimize and evaluate for low-cost solutions under this function (Ustun et al., 2019; Rawal & Lakkaraju, 2020; Karimi et al., 2020c;d; Cui et al., 2015). We question this assumption and argue for the importance of user-specific cost functions as a shared cost function is likely to poorly represent different users in a diverse population. Hence, these indirect objectives and global cost functions might be necessary but are not sufficient for a feasible recourse.

In this work, we propose a novel framework, INSPIRE (INdividual uSer Preferences In RecoursE), that incorporates individual user preference via user-specific cost function to generating algorithmic recourse. INSPIRE provides each user with a recourse set that contains multiple cf options such that there is at least one feasible solution adhering to the user's personal feature preference (if possible). As noted by Rawal & Lakkaraju (2020), in most cases it is difficult for users to specify their exact feature preferences or cost functions. INSPIRE solves this issue by focusing and improving upon four major components – (1) the procedure to formalize and define individual user preferences via user-specific cost functions, (2) the recourse objective function, (3) the optimization algorithm, and (4) the evaluation procedure.

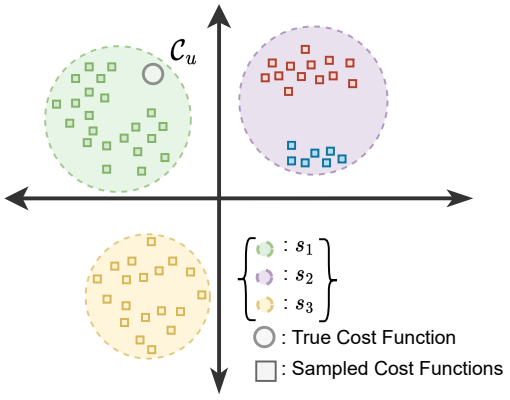


Figure 1: Diagram showing the intuition behind the Expected Minimum Cost Objective. This figure represents an abstract cost function space where squares denote cost function samples that are the same color if they are similar and form a cluster. We aim to find a recourse set where each cf (here, $\{s_1, s_2, s_3\}$) does well under a particular cluster of cost functions. The shaded big circles each represent a single cf s_i that caters to the enclosed cost functions. Here the user's hidden ground-truth cost function (grey circle) is served well by s_1.

Next, we propose a novel objective function, *Expected Minimum Cost* (EMC) that approximately optimizes for the cost incurred by the user under their cost function (which is unknown). To do this, (1) we build on Ustun et al. (2019) to propose three distributions over cost functions, $\mathcal{D}_{lin}$, $\mathcal{D}_{perc}$, and $\mathcal{D}_{mix}$, that represent diverse user preferences in a population. These distributions are based on linear and percentile changes in the feature values (§3.1). (2) Next, we compute the expected minimum cost of the generated cfs with respect to multiple sampled cost functions from one of the proposed distributions (§3.2). In order to efficiently optimize for EMC, we propose a discrete optimization method, *Cost-Optimized Local Search* (COLS) (§3.3). COLS guarantees a monotonic reduction in EMC of the recourse set, leading to large empirical reductions in the user-incurred cost. Note that, the EMC objective encourages diversity in the solution set with respect to the diverse feature preferences a user might possess by ensuring that each cf is a good cf under some particular cluster of cost functions from the sampling distribution. Hence, if the user's ground-truth cost function is well represented by any of the clusters, then we will have *some* counterfactual that is feasible (actionable and low-cost) under their cost function (shown in Figure 1).

To evaluate the effectiveness of EMC and COLS, we run experiments on two popular real-world datasets: Adult-Income (Dua & Graff, 2017) and COMPAS (Larson et al., 2016). We compare our method with multiple strong baselines methods like DICE (Mothilal et al., 2020), FACE (Poyiadzi et al., 2020), and Actionable Recourse (AR) (Ustun et al., 2019). We evaluate these methods on existing metrics from the literature like diversity, proximity, sparsity, and validity (§4.1) along with our novel cost-based evaluation framework (§3.4) and a human evaluation. In particular, we define the *fraction of satisfied users* based on whether their cost of recourse is below a certain satisfiability threshold k. We also report *coverage*, which is the fraction of users with at least one actionable recourse (Rawal & Lakkaraju, 2020). Using simulated user cost functions, we show that our method satisfies up to 25.89% more users than strong baseline methods while covering up to 22.35% more users across datasets. Furthermore, our human evaluation shows that the recourses generated by our

method are preferred by humans 57% of the time as compared to 25% for our strongest baseline (Actionable Recourse), a difference of 32%. We also perform important ablations to show what fraction of the performance can be attributed to COLS optimization method or the EMC objective. Finally, we perform a fairness analysis of all the methods across demographic subgroups to show that our method is more fair than baseline methods. Our primary contributions in this paper are listed below.

1. We conceptualize a novel framework, INSPIRE that accounts for the individuality of users while generating and evaluating recourse options. INSPIRE provides the flexibility for future researchers to further innovate on its four components.
2. We propose a new objective function, Expected Minimum Cost that approximately optimizes for a user's true cost function by using diverse plausible cost functions from a distribution.
3. We propose a discrete optimization method, *Cost-Optimized Local Search* which generates recourses that lead to higher user satisfaction. In human evaluation, we find that our recourses are preferred more than twice as often as the strongest baseline recourses while being fairer.
4. We propose a novel evaluation procedure to simulate users' hidden cost functions to asses individual user satisfaction by using our proposed metric FS@k.

2 Problem Statement

Features Types. We assume a dataset with features $\mathcal{F} = \{f_1, f_2, ...f_h\}$. Features can be mutable, conditionally mutable, or immutable, according to the causal processes that generates it. For example, *Race* is an immutable feature (Mothilal et al., 2020), *Age* and *Education* are conditionally mutable (cannot be decreased), and *number of work hours* is mutable (can both increase and decrease). Following Ustun et al. (2019), continuous features are always discretized into appropriate sized bins.

Cost Function. In this work, we assume that each user has an inherent *feature-preference* (FP) that captures the ease of changing a particular features, and different users can have different FPs. We express such differential FPs via user-specific cost-function. A cost function $\mathcal{C}(f, i, j)$ denotes the cost of changing a feature f from i to j and lies in $[0, 1] \cup \{\infty\}$. Here, 0 means that the transition has no associated cost, whereas 1 means it is maximally difficult, and ∞ means that it is infeasible.

Transition Costs. Given a cost function $\mathcal{C}$ and two feature vectors $\boldsymbol{s}_i$, $\boldsymbol{s}_j$, the cost of transition from $\boldsymbol{s}_i \rightarrow \boldsymbol{s}_j$ is the summation of the cost of changing individual features. Hence, $\text{Cost}(\boldsymbol{s}_i, \boldsymbol{s}_j; \mathcal{C}) = \sum_{f \in \mathcal{F}} \mathcal{C}(f, s_i^f, s_j^f)$, where $\boldsymbol{s}^f$ is the value of feature f in the state vector.

User Definition. A user is defined as a tuple $\boldsymbol{u} = (\boldsymbol{s}_u, \mathcal{C}_u^*)$, where $\boldsymbol{s}_u$ is the user's current state vector of length $|\mathcal{F}|$ containing their feature values and $\mathcal{C}_u^*$ is their ground-truth cost function. See Appendix Table 7 for examples of $\boldsymbol{s}_u$ and feature preferences. Next, we define the cost incurred by a user when acting on a recourse set $\mathcal{S}$ consisting of cfs $\mathcal{S}$. As a rational user will select the least costly option, the cost they will incur is the *minimum* transition cost across all cf in the recourse set, defined as,

$$\text{MinCost}(\boldsymbol{s}_u, \mathcal{S}; \mathcal{C}_u^*) = \min_{\boldsymbol{s}_j \in \mathcal{S}} \text{Cost}(\boldsymbol{s}_u, \boldsymbol{s}_j; \mathcal{C}_u^*), \text{ where } \text{Cost}(\boldsymbol{s}_u, \boldsymbol{s}_j; \mathcal{C}_u^*) = \sum_{f \in \mathcal{F}} \mathcal{C}_u^*(f, s_u^f, s_j^f). \quad (1)$$

Problem Definition. For a user $\boldsymbol{u}$, our goal is to find a recourse set $\mathcal{S}_u$ such that there exists at least one low-cost cf with the desired outcome. Hence, if the user's ground-truth cost function $\mathcal{C}_u^*$ is provided then we can provide them with a good recourse by directly optimizing for,

$$\mathcal{S}_u = \arg\min_{\mathcal{S}} \text{MinCost}(\boldsymbol{s}_u, \mathcal{S}; \mathcal{C}_u^*) \text{ s.t. } \exists\, \boldsymbol{s}_i \in \mathcal{S} \text{ s.t. } F(\boldsymbol{s}_i) = 1, \quad (2)$$

where F is the black-box ML model and 1 is the desired class. Similar to Rawal & Lakkaraju (2020), we note that in practice it is difficult for a user to precisely quantify their FP and cost function. Hence, in most practical scenarios $\mathcal{C}_u^*$ is not provided and we propose the EMC objective to approximately optimize for the user incurred cost.

3 INSPIRE: Incorporating Individuality of Users in Recourse

Given a user, we first sample some plausible cost functions from our proposed distribution that capture diverse user feature preferences (FP) in a population (§3.1). Then our COLS optimization method to generate a candidate recourse set (§3.3). Next, we compute our EMC objective for this candidate cfs with respect to multiple samples of cost functions (§3.2). We perform this iteratively to generate a final candidate recourse set which is evaluated using our evaluate procedure (§3.4).

3.1 CHARACTERISING TRANSITION COST AND COST FUNCTION DISTRIBUTIONS

Our first goal is to carefully design distributions to sample cost functions that – (1) adhere to the fundamental concepts of how the population quantifies the transition costs between two feature values, and (2) can quantify and integrate different user FPs to represent a diverse population.

Recent works like Ustun et al. (2019) argue that users in a population fundamentally quantify the transition cost of changing a feature value from x to y as being proportional to – (1) the difference in percentile of x and y; (2) the number of major steps involved to transition from x to y. For instance, when changing the education feature from *Bachelors* to *Ph.D.*, the percentile might be appropriate as very few people have Ph.D. degrees as compared to Bachelors, leading to a higher cost. While in contrast, when changing the number of working hours from 30 to 35, users might associate a fixed cost for every additional hour as opposed to percentile differences. We recognize these different underlying phenomena for quantifying transition costs and call them *percentile cost* and *linear cost*.

Even though most users quantify transition costs for features in the two ways described above, they can find it easier or harder to act upon certain features depending on their personal circumstances. We quantify these user FPs via preference-scores denoted by $\boldsymbol{p} = [p^{f_1}, \ldots, p^{f_h}]$, which sum to 1 and each $p^{f_i} \in [0, 1]$ represents the willingness of the user u to change feature f_i. We use these FPs to scale the transition costs (percentile and linear costs) of each feature f_i by $(1 - p^{f_i})$ which decreases the cost of transition for preferred features and vice-versa. This allows us to create a user-specific cost-function that accounts for their FPs. Now, given a FP scores $\boldsymbol{p}$, Algorithm 2 and 3 allow us to generate cost functions based on percentile and linear costs respectively that adhere to the FPs.

Next, we propose three distributions, $\mathcal{D}_{perc}$, $\mathcal{D}_{lin}$, $\mathcal{D}_{mix}$, that are highly flexible and can generate cost-functions that model diverse FPs to better represent a population. The distributions $\mathcal{D}_{perc}$, $\mathcal{D}_{lin}$ are based on percentile and linear transition costs. Whereas, $\mathcal{D}_{mix}$ is our most general distribution that combines both linear and percentile transition costs using a user-specific *cost-type* weight α. To generate samples from $\mathcal{D}_{mix}$, we use Algorithm 4 that first samples a FP scores $\boldsymbol{p}$ by – (1) randomly sampling a subset of preferred features for a hypothetical user that are easy to act upon, and (2) sampling their FP scores $\boldsymbol{p}$ from a Dirichlet distribution with concentration parameter with all ones for preferred features and zero otherwise. Then, we use FP score $\boldsymbol{p}$ along with a state vector $\boldsymbol{s}$ in Algorithm 2 and 3 to sample percentile and linear transition cost. We then obtain the *mixed-costs* by taking a convex combination of the percentile and linear transition costs with randomly sampled cost-type weight $\alpha \sim Unif(0, 1)$ that capture the user's fundamental way to quantify transition cost. These mixed-costs represent transition cost for a hypothetical user with FP score $\boldsymbol{p}$ and cost-type weight α. In order to capture the slight variance amongst users with similar preference, we use a Beta distribution with the mixed-costs as mean with a small noise (std = 0.01) to get the final cost function.

We highlight some desired properties of our $\mathcal{D}_{mix}$ distribution – (1) The distribution captures all possible FP score vectors $\boldsymbol{p}$ because we randomly sample the preferred features. (2) It captures a user's predilection towards linear and percentile costs by combining them with a cost-type weight. (3) The *Linear* and *Percentile* costs are monotonic, i.e., more drastic changes have higher the associated costs. (4) Following Watson et al. (2021), we provide the user with an option to specify their needs by providing their preferred features, or FP scores $\boldsymbol{p}$. These properties allow us to represent a much larger space of plausible cost functions compared to past works that assume a shared cost function with no user FPs. Hence, these distributions better represent a population.

3.2 EXPECTED MINIMUM COST (EMC) OBJECTIVE FUNCTION

As noted by Rawal & Lakkaraju (2020), in most practical scenarios the user's true cost-function $\mathcal{C}_u^*$ is hard to obtain thus we cannot exactly minimize for Equation 2. Hence, we propose the *Expected Minimum Cost* (EMC) objective function. Give a state vector $\boldsymbol{s}$, a recourse set $\mathcal{S}$, and a distribution $\mathcal{D}_{train}$ to sample cost functions, we compute EMC as follows,

$$\text{EMC}(\boldsymbol{s}, \mathcal{S}) = \mathbb{E}_{\mathcal{C}_i \sim \mathcal{D}_{train}}[\text{MinCost}(\boldsymbol{s}, \mathcal{S}; \mathcal{C}_i)] \approx \frac{1}{M} \sum_{i=1}^{M} \min_{s_j \in \mathcal{S}} \text{Cost}(\boldsymbol{s}, \boldsymbol{s}_j; \mathcal{C}_i). \tag{3}$$

We employ Monte Carlo Estimation (Robert & Casella, 2010) to approximate the expectation by sampling M cost functions $\{\mathcal{C}_i\}_{i=1}^{M}$ from $\mathcal{D}_{train}$ and then expand MinCost using Equation 1. Moreover, the distribution $\mathcal{D}_{train}$ can be any of the three distributions we proposed. Next, for the user $\boldsymbol{u}$,

we obtain the recourse set $\mathcal{S}_u$ by minimizing the EMC objective as follows,

$$\mathcal{S}_u = \arg\min_{\mathcal{S}} \text{EMC}(\boldsymbol{s}_u, \mathcal{S}). \tag{4}$$

3.3 Cost Optimized Local Search (COLS)

To generate a recourse set $\mathcal{S}_u$ for a user $\boldsymbol{u}$, we optimize for EMC as shown in Equation 4. We propose two simple, and efficient discrete search algorithms (Pirlot, 1996), namely *Cost-Optimized Local Search* (COLS) and *Parallel Cost-Optimized Local Search* (P-COLS) presented in Algorithm 1. COLS maintains a best set $\mathcal{S}^{best}$ which will be the final recourse provided to the user. At each iteration t, a candidate set $\mathcal{S}_t$ is generated by locally perturbing each element of the best set with a Hamming distance of two, i.e. making small changes to two features at once. Then it is evaluated against the EMC objective as defined in Equation 3. For efficient implementation we store the costs of all the N cfs with respect to all the M cost functions $\mathcal{C}_i$. Instead of making a direct comparison of EMC for the best-set-so-far and the candidate set, we evaluate whether any cfs from the candidate set $\mathcal{S}_t$ would improve the EMC of the best set $\mathcal{S}^{best}$ if we swapped out individual cfs. Specifically, if the benefit of replacing $s_i \in \mathcal{S}_t$ with $s_j \in \mathcal{S}^{best}$ is positive, i.e., reduces EMC of $\mathcal{S}^{best}$ then we make the replacement (see Algorithm 5). The ability to assess the benefit of each candidate cf is critical because it allows us to constantly update the best set using cfs from a candidate set instead of waiting for an entire candidate set with lower EMC. For objectives like feature diversity, evaluating the benefit of individual replacement becomes expensive (see Appendix B.1). Moreover, for COLS we can guarantee that the EMC of the best set will monotonically decrease over time, which we formally state below:

Algorithm 1 Cost-Optimized Local Search Algorithm

Input: A state vector $\boldsymbol{s}$, $\{\mathcal{C}_i\}_{i=1}^M \sim \mathcal{D}_{train}$ cost distributions
Output: $\mathcal{S}^{best}$, a set with N generated counterfactuals.
function COLS($\boldsymbol{s}$, $\{\mathcal{C}_i\}_{i=1}^M$, Budget, hamDist $= 2$)
  Initialize
    `// Perturb` $\boldsymbol{s}$, N `times.`
    $\mathcal{S}^{best} \in \mathbb{R}^{N\times d} \leftarrow$ pertubCFS($\boldsymbol{s}$, hamDist)
    `// Incurred costs for` $\mathcal{S}^{best}$.
    $\mathbf{C}^b \leftarrow$ getCostMatrix($\boldsymbol{s}, \mathcal{S}^{best}; \{\mathcal{C}_i\}_{i=1}^M$)
    $t = 0$; maxIter $=$ Budget$//N$
  while $t <$ *maxIter* **do**
    $\mathcal{S}_t \in \mathbb{R}^{N\times d} \leftarrow$ pertubCFS($\mathcal{S}^{best}$, hamDist)
    $\mathbf{C} \in \mathbb{R}^{N\times M} \leftarrow$ getCostMatrix($\boldsymbol{s}, \mathcal{S}; \{\mathcal{C}_i\}_{i=1}^M$)
    $t \mathrel{+}= 1$
    `//` $\mathbf{B}_{ij}$ `= Change in objective when` $\mathcal{S}^{best}[i] \leftarrow \mathcal{S}[j]$. `See Algorithm 5.`
    $\mathbf{B} \in \mathbb{R}^{N\times N} \leftarrow$ computeBenefits($\mathbf{C}^b, \mathbf{C}$)
    `// Greedily select which pairs to swap given` $\mathbf{B}$
    replaceIndices $\leftarrow$ getReplaceIdx($\mathbf{B}$)
    `// Swap these pairs and update` $\mathbf{C}^b$.
    forall *originalIdx*, *replaceIdx* $\in$ *replaceIndices* **do**
      $\mathcal{S}^{best}$[originalIdx] $= \mathcal{S}$[replaceIdx]
    end
    $\mathbf{C}^b \leftarrow$ getCostMatrix($\boldsymbol{s}, \mathcal{S}^{best}; \{\mathcal{C}_i\}_{i=1}^M$)
  end
  return $\mathcal{S}^{best}, \boldsymbol{C}^b$
end

Theorem 3.1 (Monotonicity of COLS Algorithm)**.** *Given the best set, $\mathcal{S}_{t-1}^{best} \in \mathbb{R}^{N\times d}$, the candidate set at iteration t, $\mathcal{S}_t \in \mathbb{R}^{N\times d}$, the matrix $\boldsymbol{C}^b \in \mathbb{R}^{N\times M}$ and $\boldsymbol{C} \in \mathbb{R}^{N\times M}$ containing the incurred cost of each counterfactual in $\mathcal{S}_{t-1}^{best}$ and $\mathcal{S}_t$ with respect to all the M sampled cost functions $\{\mathcal{C}_i\}_{i=1}^M$, there always exist a $\mathcal{S}_t^{best}$ constructed from $\mathcal{S}_{t-1}^{best}$ and $\mathcal{S}_t$ such that*

$$\mathit{EMC}(\boldsymbol{s}_u, \mathcal{S}_t^{best}; \{\mathcal{C}_i\}_{i=1}^M) \leq \mathit{EMC}(\boldsymbol{s}_u, \mathcal{S}_{t-1}^{best}; \{\mathcal{C}_i\}_{i=1}^M)$$

For the proof of the theorem, please refer to Appendix B.2.2.

P-COLS: The P-COLS method is a variant of COLS that starts multiple parallel runs of COLS with different initial sets. With a given computational budget, each run is allocated a fraction of the budget. The recourse set of the run with the least objective value is provide to the user.

3.4 Evaluation Procedure and Individual User Based Metrics

Given that users' ground-truth cost functions $\mathcal{C}_u^*$ are unknown, in order to asses user's satisfaction, we need to ask them. Generally, obtaining such feedback is challenging. Hence, for every user $\boldsymbol{u}$, we simulate a *evaluation* cost function $\mathcal{C}_u^{\#}$ from a distribution $\mathcal{D}_{test}$ and use Equation 1 to compute their

Table 1: Recourse method performance across various cost and distance metrics (Section 4.1). The numbers reported are averaged across 5 different runs. For all the metrics higher is better except for PAC.

Data	Method	Metrics						
		Cost Metrics			Distance Metrics			
		FS@1	PAC(↓)	Cov	Div	Prox	Spars	Val
Adult-Income	DICE	2.47	1.37	8.32	3.90	65.80	47.20	97.90
	Face-Eps	15.23	0.76	22.60	4.75	**92.22**	74.98	**100.0**
	Face-Knn	25.30	0.74	35.00	8.62	89.07	71.85	**100.0**
	Act. Recourse	49.93	0.55	56.85	18.38	74.68	73.57	78.67
	COLS	72.57	**0.38**	76.07	**25.77**	80.22	76.48	97.15
	P-COLS	**75.82**	0.40	**79.20**	25.57	81.67	**78.00**	94.78
COMPAS	DICE	0.40	0.54	0.40	11.30	65.00	32.00	98.90
	Face-Eps	12.20	0.29	12.20	2.50	**94.20**	60.60	**100.0**
	Face-Knn	12.20	0.29	12.20	2.60	94.10	60.60	**100.0**
	Act. Recourse	65.80	0.40	66.60	11.87	80.53	**74.07**	44.23
	COLS	82.23	**0.24**	82.23	29.32	77.82	70.05	95.48
	P-COLS	**83.73**	**0.24**	**83.73**	**29.38**	78.48	71.30	92.78

incurred cost. If the distribution $\mathcal{D}_{test}$ can model diverse users in a population then this procedure can be used to compare different recourse methods.

Proposed Metrics - FS@k and Coverage: We introduce a new cost-based metric that directly captures user satisfaction and is computed using each user's simulated cost function $\mathcal{C}_u^{\#}$. We say that a user is *satisfied* by a recourse set if the best option in that set achieves a sufficiently low cost under $\mathcal{C}_u^{\#}$. Formally, given a set of users $\mathcal{U}$ and the recourse sets $\{\mathcal{S}_u\}_{u\in\mathcal{U}}$ provided to them, we define the fraction of users satisfied at a satisfiability threshold k as:

$$FS@k(\mathcal{U}, \{\mathcal{S}_u\}_{u\in\mathcal{U}}) = \frac{1}{|\mathcal{U}|}\sum_{u\in\mathcal{U}} \mathbb{1}\{\text{MinCost}(\boldsymbol{s}_u, \mathcal{S}_u; \mathcal{C}_u^{\#}) < k\} \tag{5}$$

In reality, k can vary from user to user, we keep k fixed across users in our experiments because the goal of any method is to find low-cost recourses regardless of k. In deployment scenarios, reasonable values of k can be estimated by doing a user survey. In addition to FS@k, we also measure Population Average Cost (PAC), which is defined as PAC $= \frac{1}{|\mathcal{U}|}\sum_{u\in\mathcal{U}} \text{MinCost}(\boldsymbol{s}_u, \mathcal{S}_u; \mathcal{C}_u^{\#})$. Another important measure is the *Coverage (Cov)* which is the fraction of users to which the recourse method can provide at least one actionable recourse (Rawal & Lakkaraju, 2020), defined as $Cov(\mathcal{U}, \{\mathcal{S}_u\}_{u\in\mathcal{U}}) = FS@\infty = \frac{1}{|\mathcal{U}|}\sum_{u\in\mathcal{U}} \mathbb{1}\{\text{MinCost}(\boldsymbol{s}_u, \mathcal{S}_u; \mathcal{C}_u^{\#}) < \infty\}$.

4 Experiments

4.1 Experimental Setup

Dataset: We use the Adult-Income (Dua & Graff, 2017) and COMPAS (Larson et al., 2016) datasets which are available under Open Data Commons PDDL license. The Adult-Income dataset is based on the 1994 US Census data and contains 12 features. The model has to predict whether an individual's income is over $\$50,000$. COMPAS contains 7 features and was collected by ProPublica and contains information about the criminal history of defendants for analyzing recidivism. The model needs to decide bail based on predicting which applicants will recidivate in the next two years. These datasets are anonymized to prevent privacy. We preprocess both datasets based on a previous analysis where categorical features are binarized (Pawelczyk et al., 2021).[2] Our black-box model is an Multi-Layer Perceptron with 2-layers. Please refer to Appendix Tables 9 and 10 for experiments with logistic regression and Appendix A.1 and Table 5 for further experimental details.

Baselines: We compare our methods COLS and P-COLS with DICE (Mothilal et al., 2020), FACE-Knn and FACE-Epsilon (Poyiadzi et al., 2020), and Actionable Recourse (Ustun et al., 2019). Importantly, **we control for compute across methods** by restricting the number of forward passes to the black-box model, which are needed to decide if a counterfactual produces the desired class. For most big models, this is the rate-limiting step for each method. We ran our experiments on a local

[2] The code for the Actionable Recourse method (Ustun et al., 2019) requires binary categorical variables.

Table 2: Ablation results with Search algorithms trained on different objectives.

Search Alg.	Objective	Cost Metrics			Distance Metrics		
		FS@1	PAC(↓)	Cov	Div	Prox	Spars
LS	Sparsity	10.1	1.304	29.0	42.7	66.2	55.8
LS	Proximity	9.7	1.275	27.0	42.1	67.5	55.0
LS	Diversity	0.0	2.393	7.6	**53.3**	50.8	35.6
LS	EMC	49.8	0.597	59.1	37.8	73.3	67.5
COLS	EMC	**68.8**	**0.391**	**72.6**	27.1	**77.5**	**73.5**

Table 3: Percentage of times each method was preferred by human annotators (Fleiss kappa=0.74 and p=1e-4).

No Preference	Actionable Recourse	Ours
18%	25%	**57%**

Table 4: Fairness analysis of recourse methods for Gender based subgroups. **DIR**: Disparate Impact Ratio; **M**: Male, **F**: Female.

Method	Gender	FS@1	Cov	DIR-FS	DIR-Cov
DICE	F	0.0	0.0	-	-
	M	4.7	15.6		
Face-Eps	F	12.5	22.1	1.504	1.118
	M	18.8	24.7		
Face-Knn	F	29.9	36.3	0.719	0.89
	M	21.5	32.3		
Act. Recourse	F	53.8	58.7	0.881	0.959
	M	47.4	56.3		
Random	F	7.8	34.6	0.859	0.792
	M	6.7	27.4		
COLS	F	72.7	76.2	0.994	0.992
	M	72.3	75.6		
P-COLS	F	76.5	80.2	**1.004**	**1.0**
	M	76.8	80.2		

server using a single Nvidia 1080 Ti GPU. We set a fixed budget of 5000 model queries, a set size $|\mathcal{S}| = 10$, and number of cost function sample $M = 1000$ for all methods. For a description of the objective function and other details of these baselines refer to Appendix B.1.1, B.2.3.

Evaluation Details: For our main experiments, we set $\mathcal{D}_{train}$ and $\mathcal{D}_{test}$ to our most general distribution $\mathcal{D}_{mix}$ as it is able to generate samples that are diverse and captures a wide variety of user FPs. This is analogous to supervised learning where train and test data comes from the same distribution. Note that, in contrast to past works (see §5), the cost samples $\{\mathcal{C}_i\}_{i=1}^{M}$ used during training to evaluate EMC are different from $\mathcal{C}_u^{\#}$ which is used for evaluation. For completeness, in Section 4.2 Q5, and Appendix A.2 Q5, Q6, we provide additional results in cases where D_{train} and D_{test} are different to show that our method is robust to choices of these distributions.

Distance Based Recourse Metrics: To compare with past work, we evaluate methods on distance-based metrics like feature diversity, proximity, sparsity, and validity that lie between $[0, 1]$ with higher values being better. We report the average of these metrics, in percentage across all users. For a single user, Proximity is defined as $prox(\boldsymbol{x}, \mathcal{S}) = 1 - \frac{1}{|\mathcal{S}|}\sum_{i=1}^{|\mathcal{S}|} dist(\boldsymbol{x}, \mathcal{S}_i)$, where $\mathcal{S}_i$ is a counterfactual. Sparsity (Mothilal et al., 2020) is defined as $spar(\boldsymbol{x}, \mathcal{S}) = 1 - \frac{1}{|\mathcal{S}|*d}\sum_{i=1}^{|\mathcal{S}|}\sum_{j=1}^{|\boldsymbol{x}|} \mathbb{1}_{\{x_j \neq \mathcal{S}_{ij}\}}$. Feature diversity (Mothilal et al., 2020) is defined as $div(\mathcal{S}) = \frac{1}{Z}\sum_{i=1}^{|\mathcal{S}|-1}\sum_{j=i+1}^{|\mathcal{S}|} dist(\mathcal{S}_i, \mathcal{S}_j)$, where Z is the number of terms in the double summation. Validity is defined as $val(Y) = \frac{|\{\text{unique } s_i \in \mathcal{S} : f(s_i) = +1\}|}{|\mathcal{S}|}$.

4.2 Research Questions

Q1. Which Recourse Method Satisfies the Most Users?

In this experiment, we compare different recourse methods on our cost-based evaluation framework and distance-based metrics. We report the average performance over five random seeds in Table 1. We observe that COLS and P-COLS, that optimize for EMC, achieve 22.64% and 25.89% higher user satisfaction while covering 19.28% and 22.42% more users compared to the strongest baseline on Adult-Income and COMPAS, respectively. Meanwhile, other methods that optimize for a combination of distance-based metrics, perform worse on user cost-based metrics that directly model user satisfaction. Interestingly, we find that COLS and P-COLS solutions exhibit very high feature diversity, proximity, and sparsity. This implies that – (1) the $\mathcal{D}_{mix}$ distribution is generating cost functions that model diverse FPs and COLS along with EMC allows us to obtain the highest diversity even compared to other methods that directly optimize for it, and (2) proximity, sparsity, and diversity emerge as necessary metrics even under our cost-based evaluation framework but they are not sufficient to satisfy users with preferences as shown by other methods performance on cost-metrics.

Q2. Is the Performance Improved by the COLS Optimization Method or the EMC Objective?

We perform an ablation study to understand the impact of COLS optimization method and the EMC objective. To do so, we run a basic local search (LS) to optimize objectives used by other methods like feature diversity, proximity, and sparsity along with validity. We use a basic local search because

there are no simple and efficient way to guarantee reductions in the diversity objective by swapping out single elements from the solution set that is required for using COLS (see Appendix B.1). To quantify the usefulness of COLS, we also optimize EMC using a basic local search.

The results in Table 2 suggest that optimizing for distance-metrics is sub-optimal. For proximity, sparsity, and feature diversity objectives, the FS score and coverage are very low, while they perform well on their respective metrics. The low FS score for distance metrics is expected as they ignore user preferences and hence can edit features that are not preferred making the generated recourses infeasible under the user's cost function. We find that EMC with LS outperforms all distance objectives not only on FS, PAC but also proximity, and sparsity suggesting that the EMC is a better objective. Meanwhile, the 19% difference in the performance of EMC with LS and COLS can be attributed to our cost optimization (§3.3) that allows COLS to efficiently search the solution space.

Q3. Do Recourse Methods Provide Fair Solutions Across Subgroups?

Next, We assess if the recourse methods provide equitable solutions across subgroups based on demographic features like *Gender* and *Race*. This is important because we want to ensure that recourse methods are not further inducing bias towards any particular group because it directly affects the life of users. We adapt existing fairness metrics for disparate impact across population subgroups (Feldman et al., 2015) for the recourse outcomes we study, which we denote by the Disparate Impact Ratio (DIR). Given a metric $\mathcal{M}$, DIR is a ratio between metric scores across two subgroups. DIR-$\mathcal{M}$ $= \mathcal{M}$(S=1)/$\mathcal{M}$(S=0). We use either Cov or FS@1 as $\mathcal{M}$. Under the DIR metric, the maximum fairness score that can be achieved is 1, though this might not be achievable depending on the black-box model. We run experiments on the Adult-Income dataset, with a budget of 5000 model queries and $|\mathcal{S}| = 10$.

We present the *gender* and *race* based subgroup results in Table 4 and Appendix Table 6 respectively. We observe that **our methods are typically more fair than baselines on both Gender and Race-based subgroups while providing recourse to a larger fraction of people in both subgroups**. In particular, we find that our method achieves a score very close to 1 on DIR-FS and DIR-Cov implying a very high degree of fairness. We attribute the fairness of our method to (1) the fact that our method does not depend on the data distribution, and (2) the use of diverse cost functions to generate recourse. Condition (2) is important since there are other individualized methods that do not rely on the data distribution, such as Face, which can generate less fair solutions than COLS.

Q4. Which Recourse Method Do Humans Prefer?

Next, we are interested in whether humans would consider recourses to be reasonable for our synthetic users. We designed a small study where we provided human annotators with state vectors $\boldsymbol{s}_u$ and the sampled FP scores $\boldsymbol{p}_u$ for 100 users from the Adult-Income dataset (see Appendix A.1.4 for details). We presented the recourse generated by COLS and Actionable Recourse (strongest baseline) to the annotators while anonymizing each method's name and asked them two questions: (1) Acting as if they were the user with the provided preferences and state vector, which recourse would they prefer to adopt? (2) Does the recourse generated by each method seem reasonable to them? We collect three annotations for each sample and take a majority vote for each response; we allow for users to indicate "no preference" between the two proposed recourses, and if there was a tie in annotation we record the majority vote as "no preference." We found that **our method was preferred 57% of the time, while AR was preferred only 25% of the time**, a difference of 32% (+/- 16 points variance, Fleiss' kappa=0.74, and p=1e-4). Furthermore, human annotators found 60% of the recourses generated by COLS to be reasonable as compared to 33% for AR, a 27% difference with p<1e-4. This study shows that our method is preferred by humans over the baseline.

Q5. Robustness to Distribution Shifts And Other Research Questions

In this experiment, we test our evaluation framework's robustness in cases where the train and test distributions, $\mathcal{D}_{train}$ and $\mathcal{D}_{test}$ are different. In Appendix Figure 2, the top-left and bottom-right corners shows cases where $\mathcal{D}_{train} = \mathcal{D}_{lin}$ while the users' evaluation cost functions $\mathcal{C}_u^{\#}$ are drawn from $\mathcal{D}_{test} = \mathcal{D}_{per}$ (and vice versa). This is a complete distribution shift and our method performs equally well for these cases demonstrating that our method is robust to distribution shift during test time. For full experimental design and conclusions please refer to Appendix Section A.2 and Figures 2 and 3. **We also provide several additional research questions in the Appendix A.2**, which we summarize here: (1) We can make use of a larger compute budget to scale up the performance (Fig-

ure 4); (2) The recourse sets provide *high quality solutions to users using as few as 3 counterfactuals* (Figure 5); (3) We can *achieve high user satisfaction with as few as 20 Monte Carlo samples*, rather than 1000 (Figure 6); (4) Our method works for other classification models as well (Table 10); and (5) We present the computational complexity and runtimes in Appendix A.1.2. We also show some qualitative examples of recourses provided by our method in Table 7.

5 Related Work

Here, we distinguish our approach based on our recourse objectives, optimizer, and evaluation. We point readers to Venkatasubramanian & Alfano (2020) for a philosophical basis for algorithmic recourse and to Karimi et al. (2020b) for a comprehensive survey of the existing recourse methods.

Objectives: The most prominent family of objectives for recourse includes distance-based objectives (Wachter et al., 2017; Karimi et al., 2020a; Dhurandhar et al., 2018; Mothilal et al., 2020; Rasouli & Yu, 2021). These methods primarily seek recourses that are close to the original data point. In DICE, Mothilal et al. (2020) provide users with a set of counterfactuals while trading off between proximity and feature diversity. A second category of methods uses other heuristics based on the data distribution (Aguilar-Palacios et al., 2020; Gomez et al., 2020) to come up with counterfactuals. FACE constructs a graph from the given data and then tries to find a high-density path between points in order to generate counterfactuals (Poyiadzi et al., 2020). Lastly, the works closest to ours are the cost-based objectives, which capture feasibility in terms of the cost of recourse: (1) Cui et al. (2015) define a cost function specifically for tree-based classifiers, which compares the different paths that two data points follow in a tree to obtain a classifier-dependent measure of cost. (2) Karimi et al. (2020c;d) take a causal intervention perspective on the task and define cost in terms of the normalized distance between the user state and the counterfactual. (3) Ustun et al. (2019) define cost in terms of the number of changed features and frame recourse generation as an Integer Linear Program. (4) Rawal & Lakkaraju (2020) infer global cost function from pairwise comparisons of features that are drawn from simulated users. However, they take a different approach to the recourse generation problem, which is to find a list of rules that can apply to any user to obtain a recourse, rather than specially generating recourses for each user as in this work. Importantly, all of these works assume there is a known and single cost function that is shared by all users.

Optimization: Several recourse methods uses gradient-based optimization to generate counterfactuals close to a user's data point (Wachter et al., 2017; Mothilal et al., 2020). Some recent approaches use tree-based techniques (Rawal & Lakkaraju, 2020; von Kügelgen et al., 2020; Kanamori et al., 2020) or kernel-based methods (Dandl et al., 2020; Gomez et al., 2020; Ramon et al., 2020), while others employ some heuristic (Poyiadzi et al., 2020; Aguilar-Palacios et al., 2020) to generate counterfactuals. A few works use autoencoders to generate recourses (Pawelczyk et al., 2020; Joshi et al., 2019), while Karimi et al. (2020a) and Ustun et al. (2019) utilize SAT and ILP solvers, respectively.

Evaluation: Besides ensuring that recourses are classified as the desired outcome by a model (validity), the most prominent approaches to evaluate recourses rely on distance-based metrics. In DICE, Mothilal et al. (2020) evaluate recourses according to their proximity, sparsity, and feature diversity. Meanwhile, several works directly consider the cost of the recourses, using a single known cost function as a metric, meaning that all users share the same cost function. In contrast, Rawal & Lakkaraju (2020) estimate a cost function from simulated pairwise feature comparisons. For all these method, a single cost function is used for both recourse generation and evaluation, i.e. the solutions are optimized and tested on the same cost function (Cui et al., 2015; Karimi et al., 2020c;d; Rawal & Lakkaraju, 2020). In contrast, we evaluate recourse methods by simulating user-specific cost functions that can vary greatly across users to capture their preference.

6 Discussion and Conclusion

Our novel framework INSPIRE provides a way to incorporate the individuality of the user in the recourse generation and evaluation process. INSPIRE lays a foundation for the future works to build more complex distributions to better represent the population by designing non-linear transition costs or modifying the COLS procedure to account for the causal relationships between features while accounting for individual user preferences. We show that our method achieves much higher rates of user satisfaction than comparable baselines and observe that diversity, proximity, and sparsity emerge as important metrics even in our framework but are not sufficient for user satisfaction.

ETHICS STATEMENT

We hope that our recourse method is adopted by institutions seeking to provide reasonable paths to users for achieving more favorable outcomes under the decisions of black-box machine learning models or other inscrutable models. We see this as a "robust good," similar to past commentators Venkatasubramanian & Alfano (2020). Below, we comment on a few other ethical aspects of the algorithmic recourse problem.

First, we suggest that fairness is an important value that recourse methods should always be evaluated along, but we note that evaluations will depend heavily on the model, training algorithm, and training data. For instance, a sufficiently biased model might not even allow for suitable recourses for certain subgroups. As a result, any recourse method will fail to identify an equitable set of solutions for the population. That said, recourse methods can still be designed to be more or less fair. This much is evident from our varying results on fairness metrics using a number of recourse methods. What will be valuable in future work is to design experiments that separate the effects on the fairness of the model, training algorithm, training data, and recourse algorithm. Until then, we risk blaming the recourse algorithm for the bias of a model, or vice versa.

Additionally, there are possible dual-use risks from developing stronger recourse methods. For instance, malicious actors may use recourse methods when developing models in order to *exclude* certain groups from having available recourse, which is essentially a reversal of the objective of training models for which recourse is guaranteed (Ross et al., 2021). We view this use case as generally unlikely, but pernicious outcomes are possible. We also note that these kinds of outcomes may be difficult to detect, and actors may make bad-faith arguments about the fairness of their deployed models based on other notions of fairness (like whether or not a model has access to protected demographic features) that distract from an underlying problem in the fairness of recourses.

REPRODUCIBILITY STATEMENT

To encourage reproducibility, we provide our source code, including all the data pre-processing, model training, recourse generation, and evaluation metric scripts as supplementary material. The details about the datasets and the pre-processing are provided in Appendix A.1.1. We also provide clear and concise Algorithms 4, 2, 3 for our cost sampling procedures and our optimization method COLS in Algorithm 1. Additionally, we also provide formal proof of the Theorem 3.1 stated in the main paper in Appendix B.2.2 along with the constructive procedure for the proof provided in Algorithm 1.

REFERENCES

Carlos Aguilar-Palacios, Sergio Muñoz-Romero, and José Luis Rojo-Álvarez. Cold-start promotional sales forecasting through gradient boosted-based contrastive explanations. *IEEE Access*, 2020.

Ifeoma Ajunwa, Sorelle Friedler, Carlos E Scheidegger, and Suresh Venkatasubramanian. Hiring by algorithm: Predicting and preventing disparate impact. *Available at SSRN*, 2016.

Furui Cheng, Yao Ming, and Huamin Qu. Dece: Decision explorer with counterfactual explanations for machine learning models. *IEEE Transactions on Visualization and Computer Graphics*, 27: 1438–1447, 2021.

Alexandra Chouldechova, Diana Benavides-Prado, Oleksandr Fialko, and Rhema Vaithianathan. A case study of algorithm-assisted decision making in child maltreatment hotline screening decisions. In *Conference on Fairness, Accountability and Transparency*, pp. 134–148, 2018.

Zhicheng Cui, Wenlin Chen, Yujie He, and Yixin Chen. Optimal action extraction for random forests and boosted trees. In *Proceedings of the 21th ACM SIGKDD international conference on knowledge discovery and data mining*, pp. 179–188, 2015.

Susanne Dandl, Christoph Molnar, Martin Binder, and Bernd Bischl. Multi-objective counterfactual explanations. In Thomas Bäck, Mike Preuss, André Deutz, Hao Wang, Carola Doerr, Michael

Emmerich, and Heike Trautmann (eds.), *Parallel Problem Solving from Nature – PPSN XVI*, pp. 448–469, Cham, 2020. Springer International Publishing. ISBN 978-3-030-58112-1.

Amit Dhurandhar, Pin-Yu Chen, Ronny Luss, Chun-Chen Tu, Pai-Shun Ting, Karthikeyan Shanmugam, and Payel Das. Explanations based on the missing: Towards contrastive explanations with pertinent negatives. In *Advances in Neural Information Processing Systems*, 2018.

Dheeru Dua and Casey Graff. UCI machine learning repository, 2017. URL https://archive.ics.uci.edu/ml/datasets/adult.

Michael Feldman, Sorelle A. Friedler, John Moeller, C. Scheidegger, and S. Venkatasubramanian. Certifying and removing disparate impact. *Proceedings of the 21th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, 2015.

Oscar Gomez, Steffen Holter, Jun Yuan, and Enrico Bertini. Vice: visual counterfactual explanations for machine learning models. In *Proceedings of the 25th International Conference on Intelligent User Interfaces*, pp. 531–535, 2020.

Shalmali Joshi, Oluwasanmi Koyejo, Warut D. Vijitbenjaronk, Been Kim, and Joydeep Ghosh. Towards realistic individual recourse and actionable explanations in black-box decision making systems. *International Conference on Learning Representations; Workshop on Safe ML*, abs/1907.09615, 2019.

Kentaro Kanamori, Takuya Takagi, Ken Kobayashi, and Hiroki Arimura. Dace: Distribution-aware counterfactual explanation by mixed-integer linear optimization. In Christian Bessiere (ed.), *Proceedings of the Twenty-Ninth International Joint Conference on Artificial Intelligence, IJCAI-20*, pp. 2855–2862. International Joint Conferences on Artificial Intelligence Organization, 7 2020. doi: 10.24963/ijcai.2020/395. URL https://doi.org/10.24963/ijcai.2020/395. Main track.

Amir-Hossein Karimi, G. Barthe, B. Balle, and Isabel Valera. Model-agnostic counterfactual explanations for consequential decisions. *International Conference on Artificial Intelligence and Statistics*, abs/1905.11190, 2020a.

Amir-Hossein Karimi, Gilles Barthe, Bernhard Schölkopf, and Isabel Valera. A survey of algorithmic recourse: definitions, formulations, solutions, and prospects. *CoRR*, abs/2010.04050, 2020b. URL https://arxiv.org/abs/2010.04050.

Amir-Hossein Karimi, Bernhard Schölkopf, and Isabel Valera. Algorithmic recourse: from counterfactual explanations to interventions. *FAccT*, 2020c.

Amir-Hossein Karimi, Julius von Kügelgen, Bernhard Schölkopf, and Isabel Valera. Algorithmic recourse under imperfect causal knowledge: a probabilistic approach. *Advances in Neural Information Processing Systems*, 2020d.

A. Kulesza and B. Taskar. Determinantal point processes for machine learning. *Found. Trends Mach. Learn.*, 5:123–286, 2012.

Jeff Larson, Surya Mattu, Lauren Kirchner, and Julia Angwin. https://github.com/propublica/compas-analysis, 2016.

Hans Mittleman. Mixed integer linear programming benchmarks (miplib 2010). http://plato.asu.edu/ftp/milpc.html, 2018.

Ramaravind K Mothilal, Amit Sharma, and Chenhao Tan. Explaining machine learning classifiers through diverse counterfactual explanations. In *Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency*, pp. 607–617, 2020.

Martin Pawelczyk, Klaus Broelemann, and Gjergji Kasneci. Learning model-agnostic counterfactual explanations for tabular data. In *WWW*, pp. 3126–3132, New York, NY, USA, 2020. Association for Computing Machinery. ISBN 9781450370233. doi: 10.1145/3366423.3380087. URL https://doi.org/10.1145/3366423.3380087.

Martin Pawelczyk, Sascha Bielawski, Johan Van den Heuvel, Tobias Richter, and Gjergji. Kasneci. CARLA: A python library to benchmark algorithmic recourse and counterfactual explanation algorithms. In *Thirty-fifth Conference on Neural Information Processing Systems Datasets and Benchmarks Track (Round 1)*, 2021. URL https://openreview.net/forum?id=vDilkBNNbx6.

Marc Pirlot. General local search methods. *European Journal of Operational Research*, 92(3):493–511, 1996. ISSN 0377-2217. doi: https://doi.org/10.1016/0377-2217(96)00007-0. URL https://www.sciencedirect.com/science/article/pii/0377221796000070.

Brett Poulin, Roman Eisner, Duane Szafron, Paul Lu, Russell Greiner, David S Wishart, Alona Fyshe, Brandon Pearcy, Cam MacDonell, and John Anvik. Visual explanation of evidence with additive classifiers. In *Proceedings Of The National Conference On Artificial Intelligence*, volume 21, pp. 1822. Menlo Park, CA; Cambridge, MA; London; AAAI Press; MIT Press; 1999, 2006.

Rafael Poyiadzi, Kacper Sokol, Raul Santos-Rodriguez, Tijl De Bie, and Peter Flach. Face: Feasible and actionable counterfactual explanations. In *Proceedings of the AAAI/ACM Conference on AI, Ethics, and Society*, AIES '20, pp. 344–350, New York, NY, USA, 2020. Association for Computing Machinery. ISBN 9781450371100. doi: 10.1145/3375627.3375850. URL https://doi.org/10.1145/3375627.3375850.

Yanou Ramon, David Martens, Foster Provost, and Theodoros Evgeniou. Counterfactual explanation algorithms for behavioral and textual data. *Advances in Data Analysis and Classification*, 2020.

Peyman Rasouli and Ingrid Chieh Yu. Care: Coherent actionable recourse based on sound counterfactual explanations, 2021.

Kaivalya Rawal and Himabindu Lakkaraju. Beyond individualized recourse: Interpretable and interactive summaries of actionable recourses. In H. Larochelle, M. Ranzato, R. Hadsell, M. F. Balcan, and H. Lin (eds.), *Advances in Neural Information Processing Systems*, volume 33, pp. 12187–12198. Curran Associates, Inc., 2020. URL https://proceedings.neurips.cc/paper/2020/file/8ee7730e97c67473a424ccfeff49ab20-Paper.pdf.

Marco Tulio Ribeiro, Sameer Singh, and Carlos Guestrin. Anchors: High-precision model-agnostic explanations. In *AAAI Conference on Artificial Intelligence*, 2018.

Christian P. Robert and George Casella. *Monte Carlo Statistical Methods*. Springer Publishing Company, Incorporated, 2010. ISBN 1441919392.

Alexis Ross, Himabindu Lakkaraju, and Osbert Bastani. Learning models for algorithmic recourse. In *Advances in Neural Information Processing Systems*, 2021.

Naeem Siddiqi. *Credit Risk Scorecards: Developing and Implementing Intelligent Credit Scoring*, volume 3. John Wiley & Sons, 2012.

Berk Ustun, Alexander Spangher, and Yang Liu. Actionable recourse in linear classification. In *Proceedings of the Conference on Fairness, Accountability, and Transparency*, FAT* '19, pp. 10–19, New York, NY, USA, 2019. Association for Computing Machinery. ISBN 9781450361255. doi: 10.1145/3287560.3287566. URL https://doi.org/10.1145/3287560.3287566.

Suresh Venkatasubramanian and Mark Alfano. The philosophical basis of algorithmic recourse. In *Proceedings of the 2020 conference on fairness, accountability, and transparency*, pp. 284–293, 2020.

Paul Voigt and Axel Von dem Bussche. 2018 reform of EU data protection rules. 2018. URL https://ec.europa.eu/commission/sites/beta-political/files/data-protection-factsheet-changes_en.pdf.

Julius von Kügelgen, Amir-Hossein Karimi, Umang Bhatt, Isabel Valera, Adrian Weller, and Bernhard Schölkopf. On the fairness of causal algorithmic recourse. *NeurIPS AFCI Workshop*, 2020.

Table 5: Table containing data statistics and black-box model details. The binary version of the datasets are take from (Pawelczyk et al., 2021) whereas the non-binary version are taken from (Mothilal et al., 2020).

	Adult-Income Binary	COMPAS Binary	Adult-Income	COMPAS
# Continuous features	3	4	2	3
# Categorical features	9	3	10	12
Undesired class	$\leq$ 50k	Will Recidivate	$\leq$ 50k	Will Recidivate
Desired class	$>$ 50k	Won't Recidivate	$>$ 50k	Won't Recidivate
Train/val/test	20088/2338/749	1415/229/491	13172/1569/748	5491/705/444
Model Type	ANN(2, 20)	ANN(2, 20)	ANN(2, 20)	ANN(2, 20)
Val Accuracy	82%	69%	81%	61%

Sandra Wachter, B. Mittelstadt, and Chris Russell. Counterfactual explanations without opening the black box: Automated decisions and the gdpr. *Cybersecurity*, 2017.

David S. Watson, Limor Gultchin, Ankur Taly, and Luciano Floridi. Local explanations via necessity and sufficiency: unifying theory and practice. In Cassio de Campos and Marloes H. Maathuis (eds.), *Proceedings of the Thirty-Seventh Conference on Uncertainty in Artificial Intelligence*, volume 161 of *Proceedings of Machine Learning Research*, pp. 1382–1392. PMLR, 27–30 Jul 2021. URL https://proceedings.mlr.press/v161/watson21a.html.

A Appendix for INSPIRE: A Framework for Integrating Individual User Preferences in Recourse

A.1 Experimental Setup

A.1.1 Datasets and Black-Box Model

In our experiments, we have two versions of the dataset, one with binary categorical features, whereas the other with non-binary categorical features. In the main paper, we show results on the binarized version (Table 1) as an important baseline, Actionable Recourse (Ustun et al., 2019), operates with binary categorical features.[3] The data statistics for all the datasets can be found in Table 5. In our experiments, for all the datasets, the features gender and race are considered to be immutable (Mothilal et al., 2020), since we perform subgroup analysis with these variables that would be rendered meaningless if users could switch subgroups. Other features can either be mutable or conditionally mutable depending on semantics. These constraints can be incorporated into the methods by providing a schema of feature mutability criterion. Our black-box model is a multi-layer perceptron model with 2 hidden layers trained on the trained set and validated on the dev set. The accuracy numbers are shown in Table 5. The test set which is used in the counterfactual generation experiments only contains users which are classified to the undesired class by the trained black-box model. Note that our frameworks can operate with any type of model, the only requirement is the ability to query the model for outcome given a user's state vector.

A.1.2 Computational Complexity:

COLS is a local search-based method and runs for $\mathcal{O}(\frac{B}{|S|})$ iterations for each user to generate the recourse set, where B is the budget (see section 5.1 - Baselines). The complexity of the cost optimization step in COLS is $\mathcal{O}(|S|^2 * M)$ per iteration. Values of $|S|$ and M as low as 3 and 10 respectively work well in practice (see Appendix B.2 and Figure 5, 6). Finally for the current implementation the wall clock time on the adult dataset for each user with $|S|$ = 10, M = 100, B = 5000 setting is COLS = 20s, Random = 7.5s, DICE = 7.5s, AR = 11s, Face-knn = 7s, Face-Eps = 6s (can be parallelized across users). Cost function samples can be pre-computed once and saved for all experiments, this typically takes a few minutes ($<$ 5 min) across all users.

[3]The binary datasets can be downloaded from https://github.com/carla-recourse/cf-data, whereas the non-binary data can be found on https://github.com/interpretml/DiCE.

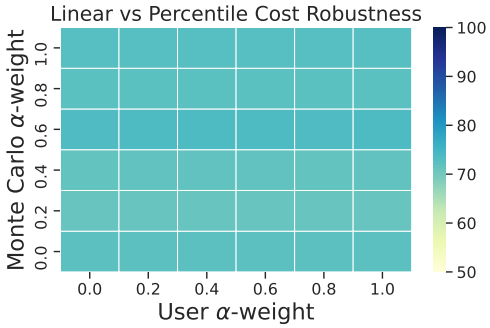


Figure 2: This figure shows the performance of the method on FS@k when recourses are generated with Monte Carlo cost samples from a distribution with α-weight varying between 0 and 1, where the user costs follow different α-weight values varying between 0, 1. Performance is robust to misspecification of α. Refer to Section A.2 for more details.

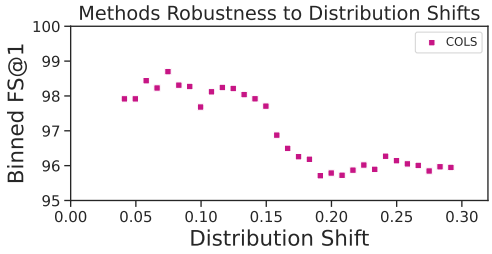


Figure 3: In this plot we show the fraction of users satisfied vs the distance between the train and test distributions. The results demonstrate that as the distance increases the performance drops a bit and then plateaus, which means that the method is robust to this kind of distribution shift. Please refer to Section A.2 for more details.

A.1.3 Recourse Generation and Evaluation Pipeline

To approximate the expectation in equation 3, our algorithm samples a set of random cost functions $\{\mathcal{C}_i\}_{i=1}^{M} \sim \mathcal{D}_{train}$, which are used at the generation time to optimize for the user's hidden cost function. In the generation phase, we use Equation 4 as our objective. Note that, this objective promotes that the generated counterfactual set contains at least one good counterfactual for each of the cost samples, hence this set satisfies a large variety of samples from $\mathcal{D}_{train}$. This is achieved via minimizing the mean of the minimum cost incurred for each of the Monte Carlo samples (Robert & Casella, 2010). Equivalently, the objective is minimized by a set of counterfactuals $\mathcal{S}$ where for each cost function there exists an element in $\mathcal{S}$ which incurs the least possible cost. In practice the size of set $\mathcal{S}$ is restricted, hence we may not achieve the absolute minimum cost but the objective tries to ensure that the counterfactuals which belong to the set have a low cost at least with respect to one Monte Carlo cost sample. The generation phase outputs a set of counterfactuals $\mathcal{S}$ which is to be provided to the users as recourse options. Given this set $\mathcal{S}_u$, in the evaluation phase, we use the users simulated cost functions which are hidden in the generation phase, to compute the cost incurred by the user $\text{MinCost}(\boldsymbol{s}_u, \mathcal{S}; \mathcal{C}_u^{\#})$ and calculate the metrics defined in the Section 4.1.

A.1.4 Details of Human Evaluation

For our human evaluation experiments, we had three undergraduate research assistants with a background in computer science. They were provided with a set of instructions on how to interpret and perform the task. Specifically, in virtual meetings, we provided them with an overview of the dataset along with the feature descriptions, a description of the task, and an overview of the recourse generation problem. Before testing, we conducted a small understanding quiz including example problems, and we corrected any misunderstandings of the study procedure. For each data point, they were asked to assume that they were a hypothetical user with the given state vector and preference scores in the sample and then were provided with the recourses generated by our method and Actionable Recourse Ustun et al. (2019) (in a blind format with randomized ordering). In total, we collected three annotations each for 100 samples from the Adult-Income dataset. We don't see any participant risks from doing the study, as the participant were asked to assume the identity of hypothetical user and asked to guess which recourse are better. The hourly rate for for the annotators was 12.5$/hours and it took around 2.5 hours for the whole study which leads to a total cost of 93.75$. The instructions provided to the participants of the human study are shown in Figure 7 and a screenshot of how the study was conducted is provided in Figure 8.

A.2 Additional Research Questions

Q5. Robustness to Misspecification in population's true and proposed cost function distribution?

Design: Our $\mathcal{D}_{mix}$ distribution samples cost by taking an α-weighted combination of linear and percentile costs. These two cost have different underlying assumptions about the how users view the cost of transition between the states. We want to test the robustness of our method in terms of misspecification in users disposition to these types of cost. We perform a robustness analysis where the users cost function has a different α mixing weight as compared to the Monte Carlo samples we use to optimize for EMC. This creates a distribution shift in the user cost function distribution ($\mathcal{D}_{test}$) and the Monte Carlo sampling distribution ($\mathcal{D}_{train}$) used in EMC. We vary the user and Monte Carlo distributions α-weights within the range of 0 to 1 in steps of 0.2. At the extremes values of $\alpha = 0, 1$, the shifts are very drastic as the underlying distribution changes completely. In the case when monte carlo α weight is 0 and user α weight is 1 then $\mathcal{D}_{train} = \mathcal{D}_{perc}$ and $\mathcal{D}_{test} = \mathcal{D}_{lin}$, simlarly for the other case we get $\mathcal{D}_{train} = \mathcal{D}_{lin}$ and $\mathcal{D}_{test} = \mathcal{D}_{perc}$. Please note that the distribution $\mathcal{D}_{lin}$ and $\mathcal{D}_{perc}$ have completely different underlying principles and are two completely different distributions. Hence, the corners of the heatmap represent drastic distribution shifts.

Results: In Figure 2, we show a heatmap plot to which demonstrates the robustness of our method. The color of the block corresponding to Monte Carlo alpha, $\alpha_{mc} = x$ and the users alpha, $\alpha_{user} = y$ represents the fraction of users that were satisfied when $\alpha_{mc} = x$ and $\alpha_{user} = y$. This means that if the user thought of costs only in terms of Linear step involved but the recourse method used samples with only percentile based cost, still the recourse set can satisfy almost the same number of users. In Figure 3, the corners correspond to these extreme cases described above, the user satisfaction for the top left corner ($\mathcal{D}_{train} = \mathcal{D}_{perc}$ and $\mathcal{D}_{test} = \mathcal{D}_{lin}$) is similar to the bottom left corner ($\mathcal{D}_{train} = \mathcal{D}_{lin}$ and $\mathcal{D}_{test} = \mathcal{D}_{lin}$). Similarly things happen for the opposite case which is denoted by the top-right ($\mathcal{D}_{train} = \mathcal{D}_{perc}$ and $\mathcal{D}_{test} = \mathcal{D}_{perc}$) and bottom-right ($\mathcal{D}_{train} = \mathcal{D}_{lin}$ and $\mathcal{D}_{test} = \mathcal{D}_{perc}$) corners. This means that even when a complete distribution shift occurs the performance user satisfaction remains similar. This can be attributed to the hierarchical step for user preference sampling in the procedure because the preferences values can be arbitrary and they scale the raw percentile and linear cost hence the distribution designed this way to model extremely diverse types of transition costs.

This means that **our methods is robust to misspecification in the train and test distributions.** The almost consistent color of the grid means that **there is very slight variation in the Fraction of Satisfied users when the model is tested on out of distribution user cost types.**

Q6. Are Solutions Robust to Misspecified Cost Distributions?

Design: In our cost sampling procedure, we make minimal assumptions about the user's feature preferences if they are not provided by the user. When finding recourses, we select a random subset of features along with their preference score for each user. However, there are situations where user preferences may be relatively homogeneous for certain features where people usually share common preferences. For example, to increase their *income*, many users might prefer to edit their *occupation type* or *education level* rather than their *work hours* or *marital status*. Given the possibility of this kind of distribution shift in feature preferences, we want to measure how robust our method is to distribution shift between our sampling distribution and the actual cost distribution followed by users.

In this experiment, we test a case of this kind of distribution shift over cost functions. For users in the Adult-Income data, we generate recourse sets using Monte Carlo samples from our standard distribution $\mathcal{D}_{mix}$ (Algorithm 4). To obtain hidden user cost functions that differ from this distribution, we first generate 500 different feature subsets indicating which features are editable, where each subset corresponds to a binary vector $concentration$ representing a user having specific preferences for some features over others (see Sec. 3.1 and Alg. 4). Since having different editable features induces a different distribution over cost functions, we obtain a measure of distribution shift for each of the 500 $concentration$ vectors by taking an l_2 distance between the vector and *its nearest neighbor in the space of* concentration *vectors used to generate the recourses*. We use the nearest neighbor because the most outlying $concentration$ vectors are least likely to be satisfied by the recourse set.

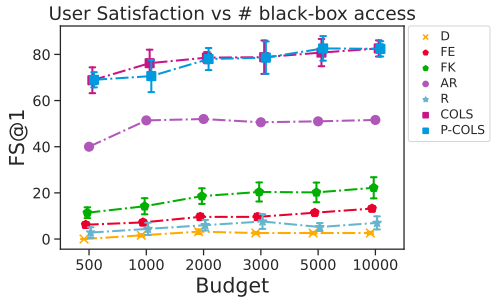


Figure 4: Figure showing the performance of different recourse methods as the Budget is increased. These are the average number across 5 different runs along with the standard deviation error bars. For some methods the standard deviation is very low hence not visible as bars in the plot. It can be seen that as the budget increases the performance of COLS and P-COLS increases. Please refer to Section A.2 for more details.

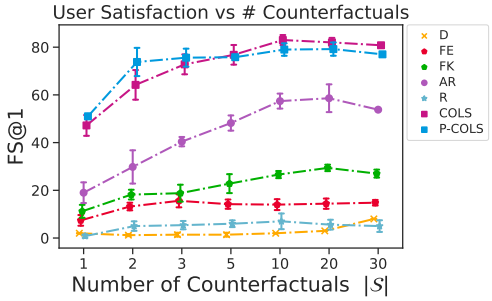


Figure 5: Figure showing the performance of different recourse methods as the the number of counterfacuals to be generated is increased. These are the average number across 5 different runs along with standard deviation error bars. We see that there is a monotonic increase in the fraction of users satisfied as the size of the set increases. We also observe that most of the performance can be obtained with a small set size. Please refer to Section A.2 for more details.

Table 6: Fairness analysis of recourse methods for subgroups with respect to Race. **DIR**: Disparate Impact Ratio; **W**: White, **NW**: Non-White (Section 4.2).

Method	Race	FS@1	Cov	DIR-FS	DIR-Cov
DICE	**NW**	0.0	0.0	-	-
	W	3.1	10.4		
Face-Eps	**NW**	7.7	12.7	2.312	2.047
	W	17.8	26.0		
Face-Knn	**NW**	12.7	25.4	2.228	1.425
	W	28.3	36.2		
Act. Recourse	**NW**	46.5	54.9	1.101	**1.056**
	W	51.2	58.0		
Random	**NW**	4.9	28.9	1.571	1.076
	W	7.7	31.1		
COLS	**NW**	67.6	71.1	1.089	1.082
	W	73.6	76.9		
P-COLS	**NW**	72.5	74.6	**1.07**	1.092
	W	77.6	81.5		

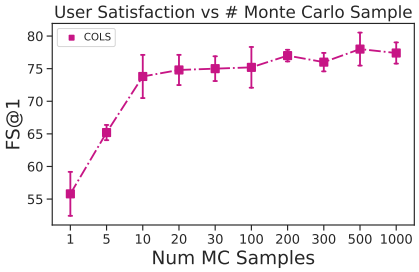
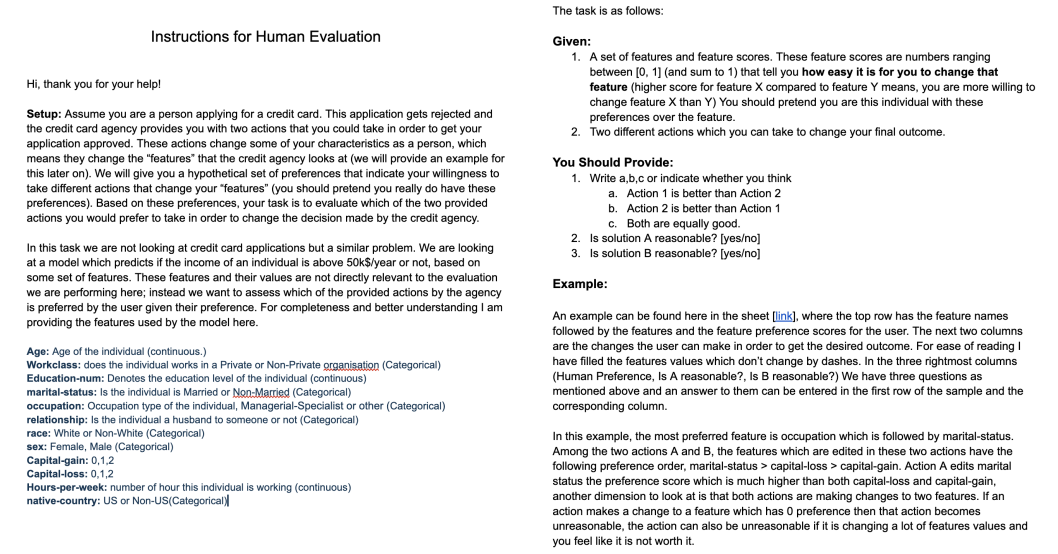


Figure 6: Figure showing the performance of the COLS method as the number of Monte Carlo samples increase. These are the average number across 5 different runs along with standard deviation error bars. There is a steep increase and then the performance saturates. This implies that in practice we do not need a large number of samples to converge to the higher user satisfaction. Refer to Section A.2 for more details.

In other words, the likelihood that a user is satisfied depends on the minimum distance between their *concentration* vector and its nearest neighbor in the cost samples used at recourse generation time. Therefore, when the minimum distance increases, there is a greater distribution shift between the user's cost functions and those obtained from $\mathcal{D}_{mix}$. Finally, we measure how many users are satisfied for a given degree of distribution shift.

Results: In Figure 3, we show a binned plot of FS@1 against our measure of distribution shift. We observe that as the distance between the distributions increases, the fraction of users satisfied decreases slightly and then plateaus. Even at the maximum distance we obtain, performance has only dropped about 3 points. This implies that **our method is robust to distribution shift in the cost distribution in terms of which features people prefer to edit.** We attribute this result to the fact that our method (1) assumes random feature preferences which subsumes these skewed preferences and (2) provides multiple recourse options, each of which can cater to different kinds of preferences. As a result, we achieve a good covering of the cost function space (see experiments with respect to varying recourse set size and number of sampled cost functions in the Appendix A.2).

Q7. Does Method Performance Scale with Available Compute?



Instructions for Human Evaluation

Hi, thank you for your help!

Setup: Assume you are a person applying for a credit card. This application gets rejected and the credit card agency provides you with two actions that you could take in order to get your application approved. These actions change some of your characteristics as a person, which means they change the "features" that the credit agency looks at (we will provide an example for this later on). We will give you a hypothetical set of preferences that indicate your willingness to take different actions that change your "features" (you should pretend you really do have these preferences). Based on these preferences, your task is to evaluate which of the two provided actions you would prefer to take in order to change the decision made by the credit agency.

In this task we are not looking at credit card applications but a similar problem. We are looking at a model which predicts if the income of an individual is above 50k$/year or not, based on some set of features. These features and their values are not directly relevant to the evaluation we are performing here; instead we want to assess which of the provided actions by the agency is preferred by the user given their preference. For completeness and better understanding I am providing the features used by the model here.

Age: Age of the individual (continuous.)
Workclass: does the individual works in a Private or Non-Private organisation (Categorical)
Education-num: Denotes the education level of the individual (continuous)
marital-status: Is the individual is Married or Non-Married (Categorical)
occupation: Occupation type of the individual, Managerial-Specialist or other (Categorical)
relationship: Is the individual a husband to someone or not (Categorical)
race: White or Non-White (Categorical)
sex: Female, Male (Categorical)
Capital-gain: 0,1,2
Capital-loss: 0,1,2
Hours-per-week: number of hour this individual is working (continuous)
native-country: US or Non-US(Categorical)

The task is as follows:

Given:

1. A set of features and feature scores. These feature scores are numbers ranging between [0, 1] (and sum to 1) that tell you **how easy it is for you to change that feature** (higher score for feature X compared to feature Y means, you are more willing to change feature X than Y) You should pretend you are this individual with these preferences over the feature.
2. Two different actions which you can take to change your final outcome.

You Should Provide:

1. Write a,b,c or indicate whether you think
   a. Action 1 is better than Action 2
   b. Action 2 is better than Action 1
   c. Both are equally good.
2. Is solution A reasonable? [yes/no]
3. Is solution B reasonable? [yes/no]

Example:

An example can be found here in the sheet [link], where the top row has the feature names followed by the features and the feature preference scores for the user. The next two columns are the changes the user can make in order to get the desired outcome. For ease of reading I have filled the features values which don't change by dashes. In the three rightmost columns (Human Preference, Is A reasonable?, Is B reasonable?) We have three questions as mentioned above and an answer to them can be entered in the first row of the sample and the corresponding column.

In this example, the most preferred feature is occupation which is followed by marital-status. Among the two actions A and B, the features which are edited in these two actions have the following preference order, marital-status > capital-loss > capital-gain. Action A edits marital status the preference score which is much higher than both capital-loss and capital-gain, another dimension to look at is that both actions are making changes to two features. If an action makes a change to a feature which has 0 preference then that action becomes unreasonable, the action can also be unreasonable if it is changing a lot of features values and you feel like it is not worth it.

Figure 7: Instructions for Human Evaluation. Please refer to Section A.1.4 for more details.

		age	workclass	education-num	marital-status	occupation	relationship	race	sex	capital-gain	capital-loss	hours-per-week	native-country	Human_Preference	Is A reasonable?	Is B reasonabe?
29	preference_scores	**0.24**	**0**	**0**	**0.1**	**0.14**	**0.35**	**0**	**0**	**0**	**0.1**	**0.06**	**0**			
30	features	70	Non-Private	5	Non-Married	Other	Non-Husband	White	Male	0	0	45	US			
31	A	-	-	-	Married	Managerial-Specialist	-	-	-	-	1	-	-			
32	B	-	-	-	Married	-	-	-	-	-	1	-	-			
33																
34	preference_scores	**0**	**0.02**	**0**	**0.05**	**0.09**	**0.15**	**0**	**0**	**0.02**	**0.25**	**0.42**	**0**			
35	features	38	Private	10	Non-Married	Other	Non-Husband	White	Female	1	0	72	US			
36	A	-	-	-	-	-	-	-	-	-	-	82	-			
37	B	-	-	-	-	-	-	-	-	2	1	-	-			
38																
39	preference_scores	**0**	**0**	**1**	**0**	**0**	**0**	**0**	**0**	**0**	**0**	**0**	**0**			
40	features	34	Private	5	Non-Married	Managerial-Specialist	Non-Husband	White	Female	0	0	38	US			
41	A	-	-	-	Married	-	-	-	-	1	1	-	-			
42	B	84	-	13	-	-	-	-	-	-	-	-	-			
43																
44	preference_scores	**0**	**0.15**	**0.24**	**0.22**	**0**	**0.1**	**0**	**0**	**0**	**0.29**	**0**	**0**			
45	features	20	Private	10	Non-Married	Managerial-Specialist	Non-Husband	White	Female	0	0	35	US			
46	A	-	-	-	Married	-	-	-	-	-	-	-	-			
47	B	-	-	-	Married	-	-	-	-	-	-	-	-			
48																
49	preference_scores	**0.17**	**0**	**0.22**	**0**	**0**	**0.09**	**0**	**0**	**0**	**0.08**	**0.44**	**0**			
50	features	22	Private	10	Non-Married	Managerial-Specialist	Non-Husband	White	Female	0	0	35	US			
51	A	-	-	-	-	-	-	-	-	-	2	62	-			
52	B	-	-	-	Married	-	-	-	-	-	1	-	-			
53																

Figure 8: Screenshots of how the human evaluation test was conducted. Please refer to Section A.1.4 for more details.

Table 7: Table providing qualitative examples for two users from the dataset. We show each users state vector, the features that user is willing to edit, the preference scores for those editable features, the recourses provided and the cost of the generated recourses. In the first example we see that user highly prefers the feature *capital loss* and the recourse which suggests edit to that has the lowest cost for the user. Whereas, the recourse which makes changes to both *Occupation* and *Capital Loss* has the highest cost as its changing multiple features. For the second user, we see that the most preferred feature is *Education-Num* but the changes suggested in the recourse requires three steps 7-8-9-10, hence the cost for that recourse is not the lowest but still relatively low. Whereas, the recourse suggesting smaller changes to *Capital Loss* which is the second most preferred feature has the lowest cost for the user.

Feature Name	State Vector	Editable Features	Preference scores	Recourses	Cost
Age	24	No	0		
Workclass	Private	No	0	(Capital Loss: 0 → 1)	0.009
Education-Num	10	No	0		
Martial-Status	Married	No	0		
Occupation	Other	Yes	0.055		
Relationship	Husband	No	0	(Occupation: Other → Manager)	0.378
Race	White	No	0		
Gender	Male	No	0		
Capital Gain	0	No	0		
Capital Loss	0	Yes	0.944	(Occupation: Other → Manager Capital Loss: 0 → 1)	0.387
# Work Hours	40	No	0		
Country	US	No	0		
Age	45	No	0		
Workclass	Private	No	0	(Capital Loss: 0 → 1)	0.071
Education-Num	7	Yes	0.537		
Martial-Status	Married	No	0		
Occupation	Other	No	0	(Capital Gain: 0 → 1)	0.106
Relationship	Non-Husband	No	0		
Race	White	No	0		
Gender	Female	No	0	(Education-Num: 7 → 10)	0.187
Capital Gain	0	Yes	0.078		
Capital Loss	0	Yes	0.240		
# Work Hours	32	Yes	0.142	(# Work Hours: 32 → 70)	0.695
Country	US	No	0		

Design: In this experiment on the Adult-Income dataset, we measure the change in performance of all the models as the number of access to the black-box model (budget) increases. Ideally, a good recourse method should be able to exploit these extra queries and use it to satisfy more users. We vary the allocated budget in the set $\{500, 1000, 2000, 3000, 5000, 10000\}$ and report the FS@1. We run the experiment on a random subset of 100 users for 5 independent runs and then report the average performance with standard deviation-based error bars in Figure 4.

Results: In Figure 4, we can see that **as the allocated budget increases the performance of the COLS and P-COLS increases** and then saturates. This suggests that our method can exploit the additional black-box access to improve the performance. Other methods like AR and Face-Knn also show performance improvement but our method COLS and P-COLS consistently upper-bound their performance. **Our method satisfies approximately 70% of the user with a small budget of 500** and quickly starts to saturate around a budget of 1000. This suggests that **our methods are suitable even under tight budget constraints** as they can achieve good performance rapidly. For example, in a real-world scenario where the recourse method is deployed and has to cater to a large population, in such cases there might be budget constraints imposed onto the method where achieving good quality solution quickly is required. Lastly, for DICE and Random search the performance on the FS@1 increase by a very small margin and then stays constant as these methods are trying to optimize for different objectives which don't align well with user satisfaction as demonstrated in Section 4.2.

Q8. Does providing more options to users help?

Design: In this experiment, we measure the effect of having flexibility to provide the user with more options, i.e. a bigger set $\mathcal{S}$. The question here is that can the methods effectively exploit this advantage and provide lower cost solution sets to the user such that the overall user satisfaction is improved. In this experiment on the Adult-Income dataset, we take a random subset of 100 users and fix the budget to 5000, Monte Carlo cost sample is set to 1000 and then vary the size of the set $\mathcal{S}$ in the set $\{1, 2, 3, 5, 10, 20, 30\}$. We restrict the size of the set to a maximum of 30 as beyond a point it becomes hard for users to evaluate all the recourse options and decide which one to act upon. We run 5 independent runs for all the data points and plot the mean performance along with

Table 8: Table comparing different recourse methods across various cost and distance metrics on Non-Binary versions of the datasets (Section A.1.1).The numbers reported are averaged across 5 different runs. Variance values have been as 89% of them were lower than 0.05, with the maximum being 0.86. FS@1: Fraction of users satisfied at $k = 1$. PAC: Population Average Cost. Cov: Population Coverage. For all the metrics higher is better except for PAC where lower is better.

Data	Method	Metrics						
		Cost Metrics			Distance Metrics			
		FS@1	PAC(↓)	Cov	Div	Prox	Spars	Val
Adult-Income - NB	DICE	6.28	1.45	27.01	53.01	57.02	47.80	86.20
	Random	0.08	2.42	17.41	**70.35**	33.32	22.45	75.71
	COLS	72.67	**0.36**	**74.60**	29.27	**79.06**	**76.64**	**97.85**
	P-COLS	**70.03**	0.39	72.81	29.85	78.45	76.29	92.30
COMPAS - NB	DICE	14.86	1.02	25.45	27.88	82.38	69.44	**99.86**
	Random	1.31	1.87	21.76	**49.07**	54.10	42.34	67.82
	COLS	67.34	**0.31**	68.11	20.53	85.47	82.34	95.97
	P-COLS	**70.86**	0.35	**72.03**	21.03	**85.48**	**82.88**	91.93

standard deviation error bars. In Figure 5, we plot the fraction of users satisfied @1 as the size of set $\mathcal{S}$ is increased.

Result: We observed that **COLS and P-COLS monotonically increase the FS@1 metric as $|\mathcal{S}|$ increases** from 1 to 30. This is consistent with the intuition behind our methods (See Figure 1, section 3.2, A.1.3 for more details). It is a fundamental property of our objective that as $|\mathcal{S}|$ increases towards M which is 1000 in this case, then the quality of the solution set should increase and reach the best possible value that can be provided under the user's cost function. We note empirically that **smaller set size $|\mathcal{S}|$ between 3 to 10 is enough in most practical cases** to reach close to maximum performance. Additionally, even with $|\mathcal{S}| \in \{1, 2, 3\}$ our methods significantly outperform all the other methods in terms of the number of users satisfied. This property is useful in real-world scenarios where the deployed recourse method can provide as little as 3 options while still satisfying a large fraction of users. Additionally, we also see improvement in the case of AR and Face-Knn methods as $|\mathcal{S}|$ increases. Note that Randoms Search's performance doesn't change as we increase the set size because the method doesn't take local steps from the best set and samples random points from a very large space, hence it is much harder to end up with low-cost counterfactuals.

Q9. Does increase the number of Monte Carlo samples help with user satisfaction?

Design: In this experiment, we want to demonstrate the effect of increasing the number of Monte Carlo samples on the performance of our COLS method. We take a random subset of 100 users, a budget of 5000, $|\mathcal{S}| = 10$. We vary the number of Monte Carlo samples (M) in the set $\{1, 5, 10, 20, 30, 100, 200, 300, 500, 1000\}$ and compute the user satisfaction. We ran 5 different runs with different Monte Carlo samples and show the average FS@1 along with the standard deviation in the Figure 6.

Results: We observe that **as the number of Monte Carlo samples increases, the performance of the method on the FS@1 metric monotonically increases.** This supports the intuition underlying our method (see Figure 1). That is, given a user with a cost function $\mathcal{C}_u^*$ as we get more and more samples from the cost distribution $\mathcal{D}_{train}$ the probability of having a cost sample similar to $\mathcal{C}_u^*$ increases and hence the fraction of satisfied users increase. It is important to note that **empirically the method's performance approaches maximum user satisfaction with as low as 20 Monte Carlo samples.** In real-world scenarios, where the deployed model is catering to a large population this can lead to small recourse generation time, hence making it more practical.

Q10. Qualitative examples of the recourses generated for some of the users.

In Table 7, we show a few examples of users along with their state vector, their editable features, their preference scores along with the recourses provided to them and their cost.

Q11. Comparison of methods on Non Binary Dataset?

Table 9: Table comparing different recourse methods across various cost and distance metrics for a black-box model with different seed but belonging to the same model family. The numbers reported are averaged across 5 different runs.

Data	Method	Metrics						
		Cost Metrics			Distance Metrics			
		FS@1	PAC(↓)	Cov	Div	Prox	Spars	Val
Adult-Income	DICE	2.70	1.24	7.10	3.80	66.20	47.30	97.80
	Face-Eps	13.32	0.79	19.88	5.43	91.97	74.80	**100.00**
	Face-Knn	21.78	0.83	34.13	8.67	88.68	71.43	**100.00**
	Act. Recourse	46.55	0.58	53.82	19.07	74.33	73.25	80.72
	Random	5.71	1.42	28.24	**48.93**	55.10	39.30	78.73
	COLS	75.12	**0.36**	77.40	25.43	81.00	77.70	98.28
	P-COLS	**75.76**	0.38	**79.14**	25.54	**81.84**	**78.38**	95.10
COMPAS	DICE	0.90	0.88	1.50	12.50	63.90	30.70	99.30
	Face-Eps	6.80	0.29	6.80	2.40	**95.00**	60.40	**100.00**
	Face-Knn	6.80	0.29	6.80	2.40	94.90	60.30	**100.00**
	Act. Recourse	56.24	0.45	58.48	9.72	80.12	**73.62**	39.10
	Random	27.44	0.78	35.70	**41.76**	58.14	33.06	49.34
	COLS	77.08	**0.24**	77.90	29.33	76.90	68.87	95.78
	P-COLS	**78.32**	**0.24**	**79.02**	29.02	77.88	70.08	92.10

In Table 8, we show the results on the non-binary version of the dataset. We observe similar performance on and trends in these results as well. COLS and P-COLS performs the best in terms of user satisfaction.

Q12. Robustness to black-box model architecture families and randomness?

In this experiment we demonstrate the result of our model when we train the same ANN architecture with different random seed (Table 9) and when we change the model family to a logistic regression classifier (Table 10). These obtained results have similar trends and demonstrate the effectiveness and robustness of our methods COLS and P-COLS which consistently satisfy cover and satisfy more users with low average population costs. In Table 9, we show the results when we train another black-box model with a different seed to see the effect of having a different trained model from the same model family.

Q13. Additional results for different values of k in FS@k
In Table 11, we report the fraction of satisfied user metric FS@k for four different values of $k \in \{0.5, 1, 2, 3\}$. These results are an extension of the results presented in Table 1.

B Appendix - Objective and Optimization

B.1 Proposed Method

B.1.1 Other Objectives

To obtain feasible a counterfactual set, past works have used various objective terms. We list objectives below from methods we compare with.

1. DICE (Mothilal et al., 2020) optimizes for a combination of Distance Metrics like *diversity* and *proximity*. They model diversity via Determinantal Point Processes (Kulesza & Taskar, 2012) adopted for solving subset selection problems with diversity constraints. They use determinant of the kernel matrix given by the counterfactuals as their diversity objective as defined below.

$$dpp\_diversity(\mathcal{S}) = det(\mathbf{K}), \text{ where} \mathbf{K}_{ij} = \frac{1}{1 + dist(\boldsymbol{s}_i, \boldsymbol{s}_j)}$$

Here, $dist(\boldsymbol{s}_i, \boldsymbol{s}_j)$ is the normalized distance metric as defined in Wachter et al. (2017) between two state vectors. *Proximity* is defined in terms of the distance between the original state vector and the counterfacutals, $prox(\boldsymbol{x}, \mathcal{S}) = 1 - \frac{1}{N}\sum_{i=1}^{|\mathcal{S}|} dist(\boldsymbol{x}, \mathcal{S}_i)$, where $\mathcal{S}_i$ is a counterfactual.

Table 10: Table comparing different recourse methods across various cost and distance metrics for a logistic regression black-box model. The numbers reported are averaged across 5 different runs.

Data	Method	Metrics						
		Cost Metrics			Distance Metrics			
		FS@1	PAC(↓)	Cov	Div	Prox	Spars	Val
Adult-Income	D	1.30	1.47	7.10	6.50	64.80	49.20	76.60
	FE	2.82	0.99	5.46	9.44	83.08	65.90	**100.00**
	FK	17.04	0.90	28.08	7.22	83.98	67.80	**100.00**
	AR	44.40	0.62	52.62	21.74	74.00	72.92	87.58
	R	3.60	1.59	24.10	**48.14**	54.76	38.94	79.46
	COLS	67.93	**0.39**	69.97	27.83	78.40	74.43	99.13
	P-COLS	**69.17**	0.40	**71.57**	27.20	**79.30**	**76.70**	95.67
COMPAS	D	0.00	-	0.00	11.10	63.10	29.20	100.00
	FE	6.30	**0.16**	6.30	3.60	**95.00**	60.50	**100.00**
	FK	6.30	**0.16**	6.30	3.60	**95.00**	60.50	**100.00**
	AR	74.32	0.31	74.32	15.66	80.98	74.26	53.66
	R	28.76	0.77	36.96	**43.22**	56.22	32.00	82.10
	COLS	87.88	0.18	87.88	31.92	76.93	71.63	89.33
	P-COLS	**89.25**	0.17	**89.25**	28.17	81.13	**74.73**	91.62

Table 11: Cost metrics for additional k values in FS@k for the results presented in the main Table 1.

Data	Method	Cost Metrics			
		FS@0.5	FS@1	FS@2	FS@3
Adult-Income	DICE	0.4	2.47	6.9	8.23
	Face-Eps	7.32	15.23	22.57	22.6
	Face-Knn	11.08	25.3	34.82	35
	Act. Recourse	28.03	49.93	56.78	56.85
	Random	0.82	6.27	28.5	31.68
	COLS	53.4	72.57	76.05	76.07
	P-COLS	**53.42**	**75.82**	**79.18**	**79.2**
COMPAS	DICE	0.2	0.4	0.4	0.4
	Face-Eps	10.4	12.2	12.2	12.2
	Face-Knn	10.4	12.2	12.2	12.2
	Act. Recourse	42.47	65.8	66.6	66.6
	Random	7.85	29.95	39.2	39.2
	COLS	73.07	82.23	82.23	82.23
	P-COLS	**74.33**	**83.73**	**83.73**	**83.73**

2. Actionable Recourse (Ustun et al., 2019) work under the assumption that all features have equal preference scores for all the users. They define cost function based on the log-percentile shift is given by,

$$cost(\boldsymbol{s}+a;\boldsymbol{s}) = \sum_{j \in \mathcal{J}_A} \log \frac{1-Q_j(\boldsymbol{s}_j+a_j)}{1-Q_j(\boldsymbol{s}_j)}$$

where $Q_j(.)$ is the cumulative distribution function of $\boldsymbol{s}_j$ in the target population, $\mathcal{J}_A$ is the set of actionable features and a_j is the action performed on the feature j.

B.2 Optimization Methods

Notation: We assume that we have a dataset with features $\mathcal{F} = \{f_1, f_2, ...f_k\}$. Each feature can either be continuous $\mathcal{F}^{con} \subset \mathcal{F}$ or categorical $\mathcal{F}^{cat} \subset \mathcal{F}$. Each continuous feature f_i^{con} takes values in the range $[r_i^{min}, r_i^{max}]$, which we discretize to integer values. For a continuous feature f_i, we define the range $Q^{(f_i)} = \{k \in \mathbb{Z} : k \in [r_i^{min}, r_i^{max}]\}$ and for a categorical feature f_i, we define it as $Q^{(f_i)} = \{q_1^{f_i}, q_2^{f_i}, ..., q_{d_i}^{f_i}\}$, where $q_{(.)}^{f_i}$ are the states that feature f_i can take. Features can either be mutable ($\mathcal{F}^m$), conditionally mutable ($\mathcal{F}^{cm}$), or immutable ($\mathcal{F}^{\oslash}$), according to the real-world causal

Algorithm 2 Sampling procedure for Percentile Transition Costs

Input: State vector $\boldsymbol{s}$, **Optional:** Feature Preference Scores $\boldsymbol{p}$
Output: Percentile based Transition Cost functions $\mathcal{C}$.
function PerCost($\boldsymbol{s}, \boldsymbol{p}$ = None)
  forall $f_i \in \mathcal{F}$ **do**
    // s_i value of feature f_i in s.
    if $\boldsymbol{p}^{f_i} = 0$ **then**
      $\mathcal{C}(f_i, s_i, .) = \infty$
      $\mathcal{C}(f_i, s_i, s_i) = 0$
    else
      if f_i *is ordered* **then**
        if f_i *can only increase* **then**

$$\mathcal{C}(f_i, s_i, x) = \begin{cases} |getPercentile(x) - getPercentile(s_i)| & \forall x > s_i \\ 0 & \forall x = s_i \\ \infty & \forall x < s_i \end{cases}$$

        else if f_i *can only decrease* **then**

$$\mathcal{C}(f_i, s_i, x) = \begin{cases} |getPercentile(s_i) - getPercentile(x)| & \forall x < s_i \\ 0 & \forall x = s_i \\ \infty & \forall x > s_i \end{cases}$$

        else if f_i *can both increase or decrease* **then**

$$\mathcal{C}(f_i, s_i, x) = \begin{cases} |getPercentile(x) - getPercentile(s_i)| & \forall x > s_i \\ 0 & \forall x = s_i \\ |getPercentile(s_i) - getPercentile(x)| & \forall x < s_i \end{cases}$$

      else if f_i *is unordered* **then**
        $\mathcal{C}(f_i, s_i, .) = Uniform(0, 1)$
  end
  if $\boldsymbol{p}$ *is not None* **then**
    $\mathcal{C}(f_i, s_i, .) \leftarrow \mathcal{C}(f_i, s_i, .) * (1 - p^{f_i})$
  return $\mathcal{C}$
end

processes that generate the data. Mutable features can transition from between any pair of states in $Q^{(f_i)}$; conditionally mutable features can transition between pairs of states only when permitted by certain conditions; and immutable features cannot be changed under any circumstances. For example, *Race* is an immutable feature (Mothilal et al., 2020), *Age* and *Education* are conditionally mutable (cannot be decreased under any circumstances), and *number of work hours* is mutable (can both increase and decrease). Lastly, while continuous features inherently define an ordering in its values, categorical features can either be ordered or unordered based on its semantic meaning. For instance, *Age* is an ordered feature that is conditionally mutable (can only increase).

B.2.1 Hierarchical Cost Sampling Procedure

To optimize for EMC, we need a plausible distribution which can model users' cost functions. We propose a hierarchical cost sampling distribution which provides cost samples that are a linear combination of *percentile shift cost* (Ustun et al., 2019) and *linear cost*, where the weights of this combination are user-specific. *Percentile shift cost* for ordered features is proportional to the change in a feature's percentile associated with the change from an old feature value to a new one. E.g., if a user is asked to increase the number of work hours from 40 to 70, then given the whole dataset, we can estimate the percentile of users working 40 and 70 hours a week. The cost incurred is then proportional to the difference in these percentiles. The *Linear cost* for ordered features is proportional to the number of intermediate states a user will have to go through while transitioning from their current state to the final state. E.g., if a user is asked to change their education level from *High-school* to *Masters* then there are two steps involved in the process. First, they need to get a *Bachelors* degree and then a *Masters* degree in which case, the user's cost is proportional to 2 because of the two steps involved in the process.

Algorithm 3 Sampling procedure for Linear Transition Costs

Input: State vector $\boldsymbol{s}$, **Optional:** Feature Preference Scores $\boldsymbol{p}$
Output: Linear change based Transition Cost functions $\mathcal{C}$.
function LinCost($\boldsymbol{s}, \boldsymbol{p} = $ None)
  if $\boldsymbol{p}^{f_i} = 0$ **then**
    $\mathcal{C}(f_i, s_i, .) = \infty$
    $\mathcal{C}(f_i, s_i, s_i) = 0$
  else
    if f_i *is ordered* **then**
      if f_i *can only increase* **then**

$$\mathcal{C}(f_i, s_i, x) = \begin{cases} \frac{|\{y \mid y > s_i \wedge y \leq x\}|}{|\{y \mid y > s_i\}|} & \forall x > s_i \\ 0 & \forall x = s_i \\ \infty & \forall x < s_i \end{cases}$$

      else if f_i *can only decrease* **then**

$$\mathcal{C}(f_i, s_i, x) = \begin{cases} \frac{|\{y \mid y < s_i \wedge y \geq x\}|}{|\{y \mid y < s_i\}|} & \forall x < s_i \\ 0 & \forall x = s_i \\ \infty & \forall x > s_i \end{cases}$$

      else if f_i *can both increase or decrease* **then**

$$\mathcal{C}(f_i, s_i, x) = \begin{cases} \frac{|\{y \mid y > s_i \wedge y \leq x\}|}{|\{y \mid y > s_i\}|} & \forall x > s_i \\ 0 & \forall x = s_i \\ \frac{|\{y \mid y < s_i \wedge y \geq x\}|}{|\{y \mid y < s_i\}|} & \forall x < s_i \end{cases}$$

    else if f_i *is unordered* **then**
      $\mathcal{C}(f_i, s_i, .) = Uniform(0, 1)$
  if $\boldsymbol{p}$ *is not None* **then**
    $\mathcal{C}(f_i, s_i, .) \leftarrow \mathcal{C}(f_i, s_i, .) * (1 - p^{f_i})$
  return $\mathcal{C}$
end

Algorithm 4 Sampling Cost Functions from $\mathcal{D}_{mix}$.

Input: State vector $\boldsymbol{s}$, *Optional:* Preferred features $\mathcal{F}_p$, feature preference scores $\boldsymbol{p}$, cost-type mixing weight α
Output: Preference scores $\boldsymbol{p}$ and the cost functions $\mathcal{C}$.
function sampleCost($\boldsymbol{s}, \alpha = $ None, $\mathcal{F}_p = \{\}, p = $ None)
  if $\mathcal{F}_p$ *is {}* **then**
    $\mathcal{F}_p \sim RandomSubset(\mathcal{F}_{mutable})$
  if p *is None* **then**
    $concentration = [1 \; if \; f \in \mathcal{F}_p \; else \; 0 \; for \; f \; in \; \mathcal{F}]$
    $p \sim Dirichlet(concentration)$
  if α *is None* **then**
    $\alpha \sim Uniform(0, 1)$
  $\mathcal{C} = \{\}$
  ▷ Get means and variance for costs.
  $\mathcal{C}^{(Lin)} \leftarrow LinCost(\boldsymbol{s}, p^{(f_i)}, f_i, \mathcal{F}_p)$
  $\mathcal{C}^{(Perc)} \leftarrow PerCost(\boldsymbol{s}, p^{(f_i)}, f_i, \mathcal{F}_p)$
  $\mathcal{C}^{(Mix)} \longleftarrow \alpha * \mathcal{C}^{(Lin)} + (1 - \alpha) * \mathcal{C}^{(Perc)}$
  ▷ Beta parametrized with mean and variance
  $\mathcal{C}_{\boldsymbol{p}} \longleftarrow Beta(\mathcal{C}^{(Mix)}, \sigma = 0.01)$
  return p, $\mathcal{C}_p$
end

B.2.2 Merging Counterfactual Sets

When searching for a good solution set, it would be useful to have the option of improving on the best set we have obtained so far using individual counterfactuals in the next candidate set we see, rather than waiting for a new, higher-scoring set to come along. While optimizing for objectives like diversity, which operate over all pairs of elements in the set, it is computationally complex to

Algorithm 5 Algorithm for Theorem 3.1

Input: $\mathbf{C}^b, \mathbf{C} \in \mathbb{R}^{N \times M}$ matrices containing the costs with respect to all cost samples..
Output: $\mathbf{B} \in \mathbb{R}^{N \times N}$, matrix containing the benefits of replacing pairs from $\mathcal{S}_{t-1}^{best} \times \mathcal{S}_t$

function computeBenefits($\mathbf{C}^b, \mathbf{C}$)
  Initialize
    $\mathbf{B} \in \mathbb{R}^{N \times N} \leftarrow \mathbf{0}$
  // Find the indices of the best and second best counterfactual in $\mathcal{S}^{best}$ for each of the M cost function.
  $\boldsymbol{b}^1 \in \mathbb{R}^M = \arg\max_i \mathbf{C}^b_{ij}$
  $\boldsymbol{b}^2 \in \mathbb{R}^M = \arg \text{second } \max_i \mathbf{C}^b_{ij}$
  // Iterate over all pairs of counterfactuals.
  forall $p, q \in [N] \times [N]$ **do**
    // Iterate over cost functions for which p^{th} counterfactual in $\mathcal{S}^{best}$ has the minimum cost.
    forall $r \in \{i \in [M] \mid \boldsymbol{b}^1_i = p\}$ **do**
      if $\boldsymbol{C}^b_{pr} > \boldsymbol{C}_{qr}$ **then**
        // This replacement reduces the cost of $\mathcal{S}^{best}$ by $\mathbf{C}^b_{pr} - \mathbf{C}_{qr}$.
        $\mathbf{B}_{pq} += \mathbf{C}^b_{pr} - \mathbf{C}_{qr}$
      else
        // $\mathbf{C}^b_{\boldsymbol{b}^2_r, r}$ = cost of second best counterfactual in $\mathcal{S}^{best}$ for r^{th} cost function.
        $\mathbf{B}_{pq} += \mathbf{C}^b_{pr} - min(\mathbf{C}_{qr}, \mathbf{C}^b_{\boldsymbol{b}^2_r, r})$
    end
  end
  return $\boldsymbol{B}$
end

evaluate the change in the objective function if one element of the set is replaced by a new one. To evaluate the change in objective in such cases, we need iterate over all pairs of element in the best and the candidate set and then evaluate the objective for the whole set again. The iteration over both the sets here is not the hard part but the computation that needs to be done within. For our objective, we can compute costs for individual recourses rather than sets, meaning we can do a trivial operation to compute the benefits of each pair replacement. But, if we wanted to do this with diversity then for each pair of replacement we need to compute additional $\mathcal{S}$ distances for each replacement because the distance of the new replace vector needs to be computed with respect to all the other vectors, for each iteration of the nested loop. This quickly makes it infeasible to improve the best set by replacing individual candidates with the best set elements. However, for metrics where it is easy to evaluate the effect of individual elements on the objective function, we can easily merge the best set and any other set $\mathcal{S}_t$ from time t to monotonically increase the objective function value.

In our objective function, EMC, we can compute the goodness of individual counterfactuals with respect to all the Monte Carlo samples (Robert & Casella, 2010). Given a set of counterfactuals we can obtain a matrix of incurred cost $\mathbf{C} \in \mathbb{R}^{N\times M}$, which specifies the cost of each counterfactual for each of the Monte Carlo samples. We can use this to update the best set $\mathcal{S}^{best}$ using elements from the perturbed set $\mathcal{S}_t$ at time t. This procedure is defined in algorithm 5. It iterates over all pairs of element in $\boldsymbol{s}_i \in \mathcal{S}^{best}$ and $\boldsymbol{s}_j \in \mathcal{S}_t$ and computes the change that will occur in the objective function by replacing $\boldsymbol{s}_i \rightarrow \boldsymbol{s}_j$. Note that we are not recomputing the costs here. Given $\mathcal{S}^{best}$, $\mathcal{S}_t$, $\mathbf{C}^b$ and $\mathbf{C}$, we can guarantee that we will update the best set $\mathcal{S}_{best}$ in a way to improve the mean of the minimum costs incurred for all the Monte Carlo samples. This is shown in algorithm 5 and the monotonicity of the EMC objective under this case can be formally stated as,

Theorem B.1 (Monotonicity of Cost-Optimized Local Search Algorithm)**.** *Given the best set, $\mathcal{S}_{t-1}^{best} \in \mathbb{R}^{N\times d}$, the candidate counterfactual at iteration t, $\mathcal{S}_t \in \mathbb{R}^{N\times d}$, the matrix $\boldsymbol{C}^b \in \mathbb{R}^{N\times M}$ and $\boldsymbol{C} \in \mathbb{R}^{N\times M}$ containing the incurred cost of each counterfactual in $\mathcal{S}_{t-1}^{best}$ and $\mathcal{S}_t$ with respect to all the M sampled cost functions $\{\mathcal{C}_i\}_{i=1}^M$, there always exist a $\mathcal{S}_t^{best}$ constructed from $\mathcal{S}_{t-1}^{best}$ and $\mathcal{S}_t$ such that*

$$\textit{EMC}(\boldsymbol{s}_u, \mathcal{S}_t^{best}; \{\mathcal{C}_i\}_{i=1}^M) \leq \textit{EMC}(\boldsymbol{s}_u, \mathcal{S}_{t-1}^{best}; \{\mathcal{C}_i\}_{i=1}^M)$$

Proof. To prove this theorem, we construct a procedure that ensures that the EMC is monotonic. For this procedure, we prove that the monotonicity of EMC holds. Check algorithm 5 for a constructive procedure for this proof, which is more intuitive to understand.

We start off by noting that each element of $\mathbf{C}_{ij}^b$ is the cost of the i^{th} counterfactual $\boldsymbol{s}_i^b$ in the best set $\mathcal{S}_{t-1}^{best}$ with respect to the cost function $\mathcal{C}_j$ given by $\text{Cost}(\boldsymbol{s}_u, \boldsymbol{s}_i^b; \mathcal{C}_j)$. Similarly $\mathbf{C}_{ij} = \text{Cost}(\boldsymbol{s}_u, \boldsymbol{s}_i; \mathcal{C}_j)$ where $\boldsymbol{s}_i$ is the i^{th} candidate counterfactual. Note that, the EMC is the average of the MinCost with respect to all the sampled cost function $\mathcal{C}_j$. What this means is that given a pair of counterfactual from $\mathcal{S}_{t-1}^{best} \times \mathcal{S}_t$ and for each $\mathcal{C}_j$, we can compute the change in the MinCost which we describe later. These replacements can lead to an increase in the cost with respect to certain cost function but the overall reduction depend on the aggregate change over all the cost functions. Given this, for each replacement candidate pair in $\mathcal{S}_{t-1}^{best} \times \mathcal{S}_t$, we can compute the change in EMC by summing up the changes in the MinCost across all cost functions $\mathcal{C}_j$; this is called the cost-benefit for this replacement pair. The cost benefit can be negative for certain replacements as well if the candidate counterfactual increases the cost across all the cost functions. The pairs with the highest positive cost benefits are replaced to construct the set $\mathcal{S}_t^{best}$, if no pair has a positive benefit then we keep set $\mathcal{S}_{t-1}^{best} = \mathcal{S}_t^{best}$. Hence, this procedure monotonically reduces EMC. We now specify how the change in MinCost can be computed to complete the proof.

To compute the change in MinCost for a single cost function $\mathcal{C}_i$, first we find the counterfactual in $\mathcal{S}_{t-1}^{best}$ with the lowest and second lowest cost which we denote by $\boldsymbol{s}_{l_1}^b$ and $\boldsymbol{s}_{l_2}^b$. These are the counterfactuals which can affect the MinCost with respect to a particular cost function $\mathcal{C}_i$. This is true because when we replace the counterfactual $\boldsymbol{s}_{l_1}^b$ which has the lowest cost for $\mathcal{C}_j$ with a new candidate counterfactual $\boldsymbol{s}_i$, there are two cases. Either, $\mathbf{C}_{l_1 j}^b > \mathbf{C}_{ij}$ or $\mathbf{C}_{l_1 j}^b \leq \mathbf{C}_{ij}$. In case when the candidate $\boldsymbol{s}_i$ has lower cost for $\mathcal{C}_j$ than $\mathbf{C}_{l_1 j}^b$, i.e. $\mathbf{C}_{l_1 j}^b > \mathbf{C}_{ij}$, then the replacement reduces the cost by $\mathbf{C}_{l_1 j}^b - \mathbf{C}_{ij}$. In case when the candidate cost for $\mathcal{C}_j$, $\mathbf{C}_{ij}$, is higher than the lowest cost in the best set $\mathbf{C}_{l_1 j}^b$, i.e. $\mathbf{C}_{l_1 j}^b \leq \mathbf{C}_{ij}$, it means that this replacement will increase the cost for $\mathcal{C}_i$ by

$\mathbf{C}^b_{l_1 j} - min(\mathbf{C}_{ij}, \mathbf{C}^b_{l_2 j})$. Here, $\mathbf{C}^b_{l_2 j}$ is the second lowest cost counterfactual for $\mathcal{C}_i$. Note that the change in this case will be negative and also depend on the second best counterfactual because once the $\boldsymbol{s}^b_{l_1}$ is removed from the set, the best cost for $\mathcal{C}_i$ will either be for $\boldsymbol{s}^b_{l_2}$ or $\boldsymbol{s}_i$, hence we take the minimum of those two and then take the difference as the increase in cost. Please refer to Algorithm 5 for a cognitively easier way to understand the proof. □

B.2.3 Other Methods

In this section, we describe some of the optimization methods used by relevant baselines.

1. DICE (Mothilal et al., 2020) perform gradient-based optimization in this continuous space while optimizing for objective defined in Section B.1.1. Their final objective function is defined as

$$C(\boldsymbol{x}) = \underset{\boldsymbol{c}_1,\ldots,\boldsymbol{c}_k}{\arg\min} \frac{1}{k}\sum_{i=1}^{k} \text{loss}(f(\boldsymbol{c}_i), y) + \frac{\lambda_1}{k}\sum_{i=1}^{k} \text{dist}(\boldsymbol{c}_i, \boldsymbol{x}) - \lambda_2 \text{ dpp\_diversity}(\boldsymbol{c}_1, \ldots, \boldsymbol{c}_k)$$

where $\boldsymbol{c}_i$ is a counterfactual, k is the number of counterfactuals, f(.) is the black box ML model, yloss(.) is the metric which minimizes the distance between models prediction and the desired outcome y. dpp_diversity(.) is the diversity metric as defined in Section B.1.1 and λ_1 and λ_2 are hyperparameters to balance the components in the objective. Please refer to Mothilal et al. (2020) for more details.

2. FACE (Poyiadzi et al., 2020) operates under the idea that to obtain actionable counterfactuals they need to be connected to the user state via paths that are probable under the original data distribution aka high-density path. They construct two different types of graphs based on nearest neighbors (Face-knn) and the ϵ-graph (Face-Eps). They define geodesic distance which trades-off between the path length and the density along this path. Lastly, they use the Shortest Path First Algorithm (Dijkstra's algorithm) to get the final counterfactuals. Please refer to (Poyiadzi et al., 2020) for more details.

3. Actionable Recourse (Ustun et al., 2019) tries to find an action set $\boldsymbol{a}$ for a user such that taking the action changes the black-box models decision to the desired outcome class, denoted by $+1$. They try to minimize the cost incurred by the user while restricting the set of actions within an action set $A(x)$. The set $A(x)$ imposes constraints related to feasibility and actionability with respect to features. They optimize the log-percentile shift objective (see Section B.1.1). Their final optimization equation is

$$\min cost(\boldsymbol{a}; \boldsymbol{x}) \;\; s.t. \;\; f(\boldsymbol{x} + \boldsymbol{a}) = +1, \;\; \boldsymbol{a} \in A(\boldsymbol{x})$$

which is cast as an Integer Linear Program (Mittleman, 2018) to provide users with recourses. Their publicly available implementation is limited to a binary case for categorical features,[4] hence we demonstrate results on the binarized version of the dataset.

[4]Please refer to the this example where they mention about these restricted abilities https://github.com/ustunb/actionable-recourse/blob/master/examples/ex_01_quickstart.ipynb