Learning Fair Policies for Multi-stage Selection Problems from Observational Data

Zhuangzhuang Jia[1], **Grani A. Hanasusanto**[1], **Phebe Vayanos**[2], **Weijun Xie**[3]
[1]Department of Industrial and Enterprise Systems Engineering, University of Illinois Urbana-Champaign
[2]Center for Artificial Intelligence in Society, University of Southern California
[3]H. Milton Stewart School of Industrial and Systems Engineering, Georgia Institute of Technology
zj12@illinois.edu, gah@illinois.edu, phebe.vayanos@usc.edu, wxie@gatech.edu

Abstract

We consider the problem of learning fair policies for multi-stage selection problems from observational data. This problem arises in several high-stakes domains such as company hiring, loan approval, or bail decisions where outcomes (e.g., career success, loan repayment, recidivism) are only observed for those selected. We propose a multi-stage framework that can be augmented with various fairness constraints, such as demographic parity or equal opportunity. This problem is a highly intractable infinite chance-constrained program involving the unknown joint distribution of covariates and outcomes. Motivated by the potential impact of selection decisions on people's lives and livelihoods, we propose to focus on interpretable linear selection rules. Leveraging tools from causal inference and sample average approximation, we obtain an asymptotically consistent solution to this selection problem by solving a mixed binary conic optimization problem, which can be solved using standard off-the-shelf solvers. We conduct extensive computational experiments on a variety of datasets adapted from the UCI repository on which we show that our proposed approaches can achieve an 11.6% improvement in precision and a 38% reduction in the measure of unfairness compared to the existing selection policy.

1 Introduction

Selection problems are very common decision-making problems in many high-stakes domains, such as company hiring, college admission, and loan audit. Given a set of candidates, a decision-maker aims to select a fixed fraction of them with objectives such as hiring the most talented candidates, admitting the most qualified students, or selecting the applicants who are most likely to repay the loan. Often, selection problems are under a multi-stage setup. In hiring, for example, candidates are initially chosen for interviews based on their résumés and the final selection is subsequently made from those who have been interviewed.

Substantial evidence points to the existence of discrimination in many selection problems that involve prejudiced outcomes for individuals or groups based on *sensitive* attributes like gender, race, ethnicity, nationality, disability status, or religion. For example, job applicants with African-American names are found to receive far fewer callbacks for each résumé they send out (Bertrand and Mullainathan 2004). Also, in the Canadian labor market, there exists notable discrimination towards applicants with foreign experience or those with Indian, Pakistani, Chinese, and Greek names compared with English names (Oreopoulos 2011). In college admission, a recent study reveals that typical Asian American applicants would see their average admit rate rise by 19% if treated as white applicants (Arcidiacono, Kinsler, and Ransom 2022). Additionally, recent research finds that Black-owned businesses received loans that were approximately 50% lower than observationally similar White-owned businesses through the Paycheck Protection Program during Covid-19 (Atkins, Cook, and Seamans 2022).

With the growing availability of data and the empirical success of machine learning, data-driven decision-making is increasingly being used in many selection problems (Li et al. 2021; Ahmad et al. 2022; Marques-Silva and Ignatiev 2022). As a result, much recent research focuses on promoting fairness and mitigating discrimination toward candidates from certain groups (Barocas, Hardt, and Narayanan 2017; Green and Chen 2019; Aghaei, Azizi, and Vayanos 2019). One prevalent approach to addressing fairness concerns during the training process is either by integrating fairness constraints into the training process (Zafar et al. 2017, 2019; Wang, Nguyen, and Hanasusanto 2021; Jo et al. 2022), or penalizing discrimination using the fairness-driven regularization terms (Kamishima et al. 2012; Berk et al. 2017; Ye and Xie 2020).

Often, the models are trained and evaluated on historical datasets containing fully observed outcomes and covariates. However, this raises questions about the possible harm of the deployed models as the training data may reflect implicit biases by humans who may unconsciously be in favor of certain groups of people (Greenwald and Banaji 1995; Barocas and Selbst 2016). For example, in the hiring setup, the available dataset only contains full covariates of people who got hired, and the outcome measure is whether they are "qualified" or "not qualified"; on the other hand, we do not have access to the outcomes of candidates who were not hired. One may use a trained model based on the candidates with full covariates and outcomes to select candidates. However, when deployed to assess the qualification of future candidates, the model may exhibit significant unfairness, even if

it satisfies the fairness constraints during the training process (Kallus and Zhou 2018). This is because, in the real world, the candidates have diverse profiles, unlike the training dataset that only contains profiles of hired candidates.

This paper considers the problem of learning fair policies for multi-stage selection problems in socially sensitive domains (e.g., employment, education, finance) given a labeled observational dataset containing one (or more) protected attribute(s). The main desiderata for such a framework are: *(1)* Maximizing precision: the decision-maker wants to maximize his/her utility by hiring/admitting/approving as many "qualified" candidates among those selected as possible; *(2)* Possible to augment with arbitrary fairness notions: in different socially sensitive settings, the decision-maker needs to consider the proper legal, ethical, and social standards in choosing the appropriate fairness measure; *(3)* Applicable to the (potentially) biased real-world data: in the presence of a selection bias in the observational data, our model must be able to learn the fair policy for future candidates instead of those "recorded" candidates in the observational data. Next, we summarize the state-of-the-art in related work and highlight the need for a unifying framework that addresses these desiderata.

1.1 Related Work

Selection Problems. Kleinberg and Raghavan (2018) study selection problem with implicit bias and analyze the Rooney Rule in the selection process. They show that this rule can not only improve the representation of the disadvantaged group but also lead to higher payoffs for the decision-maker. Celis, Mehrotra, and Vishnoi (2020) investigate the ranking problem (where the selection problem can be seen as a special case) under implicit bias and obtain similar results. Khalili et al. (2021) study the possibility of using the exponential mechanism to address both privacy and unfairness issues. They show that this mechanism can be used as a post-processing step to improve the fairness and privacy of the pre-trained model. All these works focus on one-shot decision processes, whereas our proposed framework is applicable to multi-stage selection processes.

Emelianov et al. (2019) study an optimal multi-stage selection problem and propose a simple model based on a probabilistic formulation. Their model, however, assumes perfect statistical knowledge of the joint distribution of covariates and outcome labels without bias. Moreover, their policies are not consistent, as they ignore the issue that the selection probability at a stage depends on which candidates were selected in the previous stages. Khalili, Zhang, and Abroshan (2021) consider a selection problem where sequentially arriving applicants apply for a limited number of positions/jobs, and the decision-maker accepts or rejects the given applicant using a pre-trained supervised learning model at each time step. Unlike their model, we consider the setting where additional covariates can only be revealed at later stages for the subset of selected individuals, whereas they assume all covariates are observable for each applicant.

Mixed Integer Programming (MIP). There is a growing interest in using MIP to address machine learning tasks (Bertsimas and Dunn 2017; Taskesen et al. 2020; Maragno et al. 2021; Aghaei, Gómez, and Vayanos 2021; Jo et al. 2022). Aghaei, Azizi, and Vayanos (2019) introduce a versatile MIP framework for learning optimal and fair decision trees. They show that their proposed framework yields non-discriminative decisions at a lower price to overall accuracy. Ye and Xie (2020) study fair classification problems and propose a framework that can be recast as mixed-integer convex programs. Wang, Nguyen, and Hanasusanto (2021) propose a distributionally robust classification model with a fairness constraint that encourages the classifier to be fair in view of the equality of opportunity criterion. They reformulate the model as a mixed binary conic optimization problem that can be solved using off-the-shelf solvers. Note that all of the above works focus on classification problems under a one-stage setup, whereas our work considers the general multi-stage selection problems.

Inverse Probability Weighting (IPW). IPW is a common method to reduce selection bias and has been used in several fairness-related works. Kallus and Zhou (2018) study a similar setting of the censored dataset and characterize the problem of residual unfairness. They show how to use IPW to estimate and adjust fairness metrics. However, they only focus on the one-stage static classification setup. Nabi and Shpitser (2018) consider the problem of fair statistical inference involving outcome variables and use the IPW method to estimate the natural direct effect. Khademi et al. (2019) study the problem of detecting group unfairness. They introduce fair on average causal effect – a definition of group fairness grounded in causality and show how to use IPW to estimate fair on average causal effect and use the resulting estimates to detect and quantify discrimination based on specific attributes. Kilbertus et al. (2020) analyze consequential decision-making using imperfect predictive models. They use IPW to compute the expected overall profit of a given policy. To the best of our knowledge, IPW has not been used to deal with our multi-stage selection problem.

Biased Data. There are many works on the interplay between biased data and fairness in classification (Blum and Stangl 2019; Kilbertus et al. 2020; Rezaei et al. 2021; Jo et al. 2021; Liao and Naghizadeh 2023). Lakkaraju et al. (2017) study the "selective labels" problem. They develop an approach that harnesses the heterogeneity of human decision-makers. Specifically, the paper assumes that the decision-makers differ in the thresholds they use for their yes-no decisions, but the paper does not consider the fairness of the learned policy. Goel et al. (2021) established a causal framework to analyze the effect of missing data on the fairness of downstream tasks. The authors consider a multi-stage decision-making process and propose a decentralized approach, which is different from ours. Also, we do not assume the availability of the outcome label selected candidate at every stage.

1.2 Proposed Approach and Contributions

Our main contributions are summarized as follows:

1. We propose a framework for learning fair policies in multi-stage selection problems from observational data.

Our framework can be augmented with various fairness constraints, such as demographic parity or equal opportunity.

2. Leveraging tools from causal inference and sample average approximation, we obtain an asymptotically consistent solution to this selection problem by solving a mixed binary conic optimization problem using standard off-the-shelf solvers.
3. We conduct experiments on synthetic and real-world datasets. The superiority of our model is observed through substantial precision improvement and unfairness reduction compared to the existing selection policy.

Notations. Vectors are printed in bold letters, while scalars are printed in regular font. For any $t \in \mathbb{N}$, we define $[t]$ as the index set $\{1, \ldots, t\}$. We denote by $\mathbf{e}$ as the vector of all ones whose dimension will be clear from the context. For any set $\mathcal{S}$, we use $|\mathcal{S}|$ to denote its cardinality. For any logical expression $\mathcal{E}$, the indicator function $\mathbb{1}(\mathcal{E})$ admits value 1 if $\mathcal{E}$ is true and 0 otherwise. We denote by $\delta_{\boldsymbol{\xi}}$ the Dirac distribution concentrating unit mass at $\boldsymbol{\xi} \in \Xi$ where Ξ is the support set of the distribution. We use $\mathbb{R}_+$ to denote the set of nonnegative real numbers and $\mathbb{R}_{++}$ to denote the set of strictly positive real numbers.

2 Multi-stage Selection Problem

We formalize the multi-stage selection problem in this part. All proofs are included in the supplemental Appendix A.

2.1 Problem Setup

We consider the multi-stage problem of learning a fair selection policy with T stages. Assume that there are n candidates from two demographic groups, distinguished based on a single sensitive attribute $A \in \mathcal{A} = \{0, 1\}$ that represents their group membership. This sensitive attribute could be information such as gender, race, or age group, which differentiates privileged and unprivileged individuals.

At stage $t \in [T-1]$, for those n_{t-1} candidates that passed stage $t-1$, the decision maker observes an extra covariate vector $\boldsymbol{X}^t \in \mathcal{X}^t \subseteq \mathbb{R}^{d_t}$ that is not available in the previous stages. In the real world, $\boldsymbol{X}^t$ can represent predictors delineating qualifications, creditworthiness, criminal history, etc. Next, based on all available features $\boldsymbol{X}^{[t]} := (\boldsymbol{X}^1, \ldots, \boldsymbol{X}^t) \in \mathbb{R}^{d_1+\cdots+d_t}$, $n_t \leq \overline{\alpha}_t n$ candidates are selected to advance to the next stage where $0 < \overline{\alpha}_t \leq 1$ denotes the predefined upper bound selection ratio by the decision-maker. This process continues until the final stage T, where all covariate vectors $\boldsymbol{X}^{[T]}$ of those who were selected at stage $T-1$ are revealed. The decision maker then selects $\underline{\alpha}_T n \leq n_T \leq \overline{\alpha}_T n$ candidates from those available at stage T, where $0 < \underline{\alpha}_T \leq \overline{\alpha}_T \leq \cdots \leq \overline{\alpha}_1 \leq 1$. Unlike the previous stages, the final stage has a lower bound selection ratio $\underline{\alpha}_T$. In the real world, $\underline{\alpha}_T$ represents the minimum hiring/admission rate by the decision-maker. For example, the university has a minimum admission rate – a level at which the school may lose money on tuition, federal aid, or not using resources like faculty and classrooms to capacity.

Additionally, we assume each candidate has a binary outcome label $Y \in \mathcal{Y} = \{0, 1\}$, which can only be observed if the candidate were selected to be hired (i.e., made it to the last stage). Without loss of generality, we use the positive response to indicate a positive (good) outcome, such as the candidate is qualified, the applicant repays the loan, or the individual does not recidivate.

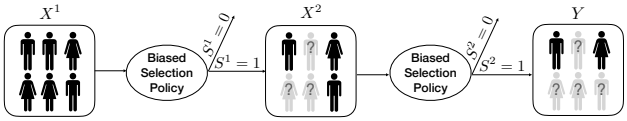

Figure 1: Two-stage selection problem: $\boldsymbol{X}^2$ is observable only for those selected in the first stage ($S^1 = 1$). Outcome label Y is only observable for those selected in both stages ($S^1 = S^2 = 1$).

Assume that we possess a training dataset containing N samples of the form $\{\hat{\boldsymbol{x}}_i^{[T]}, \hat{a}_i, \hat{y}_i\}_{i=1}^N$ that are generated from an unknown joint probability distribution $\mathbb{P}$ of the random vector $(\boldsymbol{X}^{[T]}, A, Y)$. We denote by I^t the set of selected candidates at stage $t-1$. In other words, I^t includes all candidates with covariates up to time t. Thus, we have $[N] = I^0 \supseteq I^1 \supseteq I^2 \supseteq \cdots \supseteq I^T$. The set I^T collects the indices of those candidates whose entire covariates $\boldsymbol{X}^{[T]}$ and outcome label Y are observable. Figure 1 shows the biased training dataset of a two-stage selection process.

Let $S^t \in \{0, 1\}$ be the selected outcomes where $S^t = 1$ if the candidate is selected at stage t. In this paper, we make the following assumptions.

Assumption 1 (Conditional Exchangeability)**.** *Whether a candidate is selected or not at stage $t \in [T]$ is independent of all future covariates $(\boldsymbol{X}^{t+1}, \ldots, \boldsymbol{X}^T, Y)$, and mathematically,*

$$(\boldsymbol{X}^{t+1}, \ldots, \boldsymbol{X}^T, Y) \perp\!\!\!\perp S^t \mid \boldsymbol{X}^{[t]}, S^{[t-1]} = 1 \qquad \forall \boldsymbol{X}^{[t]} \in \mathcal{X}^{[t]}.$$

Assumption 1 implies that at each stage t, whether a candidate is selected or not depends only on the available covariates $\boldsymbol{X}^{[t]}$ and that there are no unmeasured confounders that affect both $(\boldsymbol{X}^{t+1}, \ldots, \boldsymbol{X}^T, Y)$ and the selection decision S^t. In reality, the selection decision at each stage t is only based on observed covariates up to that stage. Hence, we can infer the outcome distribution for individuals who were not selected in the observational data by looking at their counterparts with the same (or similar) $\boldsymbol{X}^{[T]}$ values who got selected.

Assumption 2 (Positivity)**.** *At stage $t \in [T]$, the probability of being selected is strictly positive for any candidate's covariate values, i.e.,*

$$\mathbb{P}(S^t = 1 | \boldsymbol{X}^{[t]}) > 0 \quad \forall \boldsymbol{X}^{[t]} \in \mathcal{X}^{[t]}.$$

The positivity assumption states that any candidate should have a positive probability of being selected at any stage. Otherwise, there is no information about the distribution of

the outcomes for some covariates, and we will not be able to make inferences about it.

In the multi-stage fair selection problem, at each stage t, in view of the candidate's information $\boldsymbol{X}^{[t]}$, the decision-maker aims to find a policy $\mathcal{C}_t : \mathcal{X}^1 \times \cdots \times \mathcal{X}^t \to \mathcal{Y}$ that determines whether the candidate proceeds to the next stage or not. In the real world, those finally selected candidates can represent hired, admitted, or approved candidates. The decision-maker wants to maximize his/her utility by hiring/admitting/approving more "qualified" candidates among those selected. Hence, we use the precision $\mathbb{P}(Y = 1 \mid \mathcal{C}_T(\boldsymbol{X}^{[T]}) = 1)$ as our performance metric.

Fairness Notions In the machine learning literature, different notions of fairness can generally be classified into *individual fairness* and *group fairness*. Both perspectives have their advantages and limitations. In this paper, we concentrate on the *group fairness* due to its straightforward definition and comprehensibility for decision-makers. Furthermore, in practice, many people prioritize assessing and enforcing *group fairness* (Los Angeles Homeless Services Authority 2018).

We now briefly explain several commonly used group fairness notions based on the sensitive attribute A and introduce our unfairness measure. *Demographic Parity* requires the probability of being selected to be equal across different demographic groups (Calders, Kamiran, and Pechenizkiy 2009). *Equal Opportunity* requires the true positive rate (TPR) to be equal across different demographic groups (Hardt, Price, and Srebro 2016). *Conditional Statistical Parity* requires the probability of being selected to be equal across different demographic groups, conditional on some legitimate covariate(s) indicative of risk (Corbett-Davies et al. 2017). Additional fairness concepts include disparate impact (Feldman et al. 2015) and disparate mistreatment (Zafar et al. 2017) criteria. We refer the interested readers to Pleiss et al. (2017); Chouldechova and Roth (2020); Mehrabi et al. (2021) for extensive reviews of the literature.

In the real world, the decision-maker needs to consider the proper legal, ethical, and social context in choosing the proper fairness measure. For example, *Demographic Parity* could be chosen to address representation disparities, which can be important for promoting diversity. And the decision-maker may choose *Equal Opportunity* to ensure fairness among those "qualified" candidates. For a given fairness notion, we define the unfairness measure as follows.

Definition 2.1 (Unfairness measure)**.** *For a given policy $\mathcal{C}_T$, the unfairness measure is defined as the absolute disparity of the respective statistical metric across groups, and we denote it using $\mathbb{U}(\mathcal{C}_T(\boldsymbol{X}^{[T]}), \mathbb{P})$.*

Here, we focus on fairness in the final stage. However, it is worth noting that fairness can also be enforced at every stage by employing similar unfairness measures. The larger the value of $\mathbb{U}(\mathcal{C}_T(\boldsymbol{X}^{[T]}), \mathbb{P})$, the more unfair our selection policy is. In other words, it measures how biased the selection policy is across the privileged and unprivileged groups. Due to the page limit, we will concentrate on unfairness measures using the *Equal Opportunity* notion, which is defined as follows:

$$\mathbb{U}(\mathcal{C}_T(\boldsymbol{X}^{[T]}), \mathbb{P}) = |\mathbb{P}(\mathcal{C}_T(\boldsymbol{X}^{[T]}) = 1 \mid A = 1, Y = 1) - \mathbb{P}(\mathcal{C}_T(\boldsymbol{X}^{[T]}) = 1 \mid A = 0, Y = 1)|.$$

Nonetheless, our approach is flexible enough to accommodate other notions of fairness.

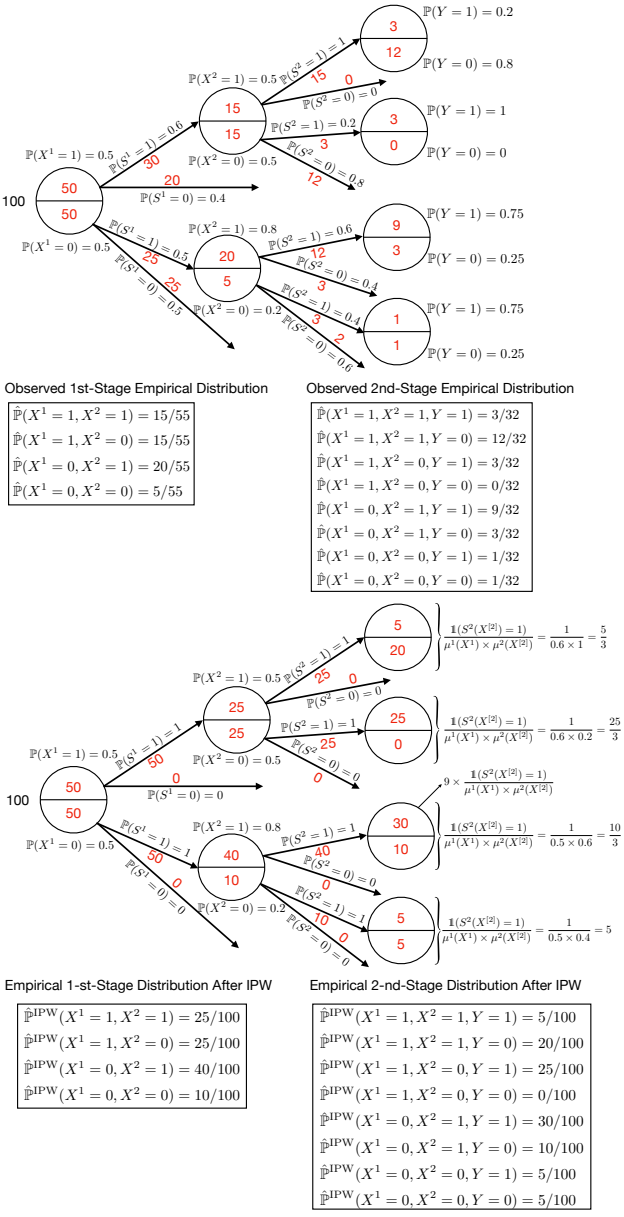

Figure 2: Evaluation of the performance of a counterfactual two-stage selection process (shown in the bottom tree) using data collected by an observed selection process (shown in the top tree) using the IPW estimator. The dataset consists of 100 candidates with binary features in each stage.

2.2 Mathematical Formulation

Based on the previous problem description, we now present our proposed infinite chance-constrained program for learn-

ing optimal multi-stage fair selection policy, as follows:

$$\begin{aligned}\max_{\{\mathcal{C}_t(\cdot)\}_{t=1}^T} \quad & \mathbb{P}(Y=1 \mid \mathcal{C}_T(\boldsymbol{X}^{[T]})=1) \\ \text{s.t.} \quad & \mathbb{P}(\mathcal{C}_t(\boldsymbol{X}^{[t]})=1) \le \overline{\alpha}_t \qquad \forall t \in [T] \\ & \mathbb{P}(\mathcal{C}_T(\boldsymbol{X}^{[T]})=1) \ge \underline{\alpha}_T \\ & \mathcal{C}_{t+1}(\boldsymbol{X}^{[t+1]}) \le \mathcal{C}_t(\boldsymbol{X}^{[t]}) \\ & \qquad \forall (\boldsymbol{X}^{[t]}, \boldsymbol{X}^{[t+1]}) \ \ \forall t \in [T-1] \\ & \mathbb{U}(\mathcal{C}_T(\boldsymbol{X}^{[T]}), \mathbb{P}) \le \eta. \end{aligned} \tag{1}$$

The objective function aims to maximize the ratio of "qualified" candidates to those who advance to the final stage, e.g., hired candidates. The first two constraints represent our selection ratio requirements. The penultimate constraint ensures the decisions are consistent – that is, candidates that were dropped at stage t can no longer be selected at the next stage $t+1$. The last constraint corresponds to the fairness constraint, where $\eta \in [0,1]$ represents the unfairness tolerance of the decision-maker.

Problem (1) is challenging because *(i)* it optimizes over functions; *(ii)* due to selection bias, we cannot observe outcome labels Y for those who did not get selected in the training data; *(iii)* we do not know the true distribution $\mathbb{P}$ of $(\boldsymbol{X}^{[T]}, A, Y)$, and even if $\mathbb{P}$ were known, the probabilistic program (1) is computationally difficult since the problem of computing the probability of an event involving multiple random variables belongs to the complexity class #P-hard (Dyer and Frieze 1988).

To address the first challenge and to also make the decisions transparent and hence accountable, we focus on the interpretable linear selection policies $\mathcal{C}_t(\boldsymbol{X}^{[t]})$ parameterized by slope parameters $\boldsymbol{W}^{[t]} = (\boldsymbol{w}_t^1, \ldots, \boldsymbol{w}_t^t) \in \mathbb{R}^{d_1+\cdots+d_t}$ and an offset $b_t \in \mathbb{R}$. The selection decision is then determined through an indicator function of the form $\mathcal{C}_t(\boldsymbol{X}^{[t]}) = \mathbb{1}(\boldsymbol{W}^{[t]} \cdot \boldsymbol{X}^{[t]} + b_t > 0)$, where $\boldsymbol{W}^{[t]} \cdot \boldsymbol{X}^{[t]} = \sum_{j=1}^t \boldsymbol{w}_t^{j\intercal} \boldsymbol{X}^j$.

For the second challenge, one may propose to include only the selected candidates in the training dataset without any modification; however, the resulting dataset is not i.i.d. due to selection bias. To tackle this challenge, we employ the IPW scheme to evaluate the performance of a *counterfactual* selection policy. Specifically, we assume that the historical selection in the data follows a *logging policy* $\{\boldsymbol{\mu}^t\}_{t=1}^T$. For a candidate with covariates $\boldsymbol{x}^{[t]} = (\boldsymbol{x}^1, \ldots, \boldsymbol{x}^t)$, we have $\boldsymbol{\mu}^t(\boldsymbol{x}^{[t]}) := \mathbb{P}(S^t = 1 | \boldsymbol{X}^{[t]} = \boldsymbol{x}^{[t]}, S^{[t-1]} = 1)$, i.e., it represents the selection probability of a candidate with covariate $\boldsymbol{x}^{[t]}$ at stage t. Horvitz and Thompson (1952) originally proposed IPW as a method to estimate causal quantities such as expected values of counterfactual outcomes, average treatment effects, and risk ratios. IPW involves reweighting the outcome of each selected candidate $i \in I^T$ by the inverse of their *propensity score*, denoted as $\boldsymbol{\mu}^t(\boldsymbol{x}_i^{[t]})$. This reweighting creates a *pseudo-population* where all candidates in the data are hypothetically selected. This allows for the estimation of the distribution of unobserved counterfactual selected outcomes for all candidates. We then estimate a counterfactual selection policy by reweighting each selected individual $i \in I^T$ at stage t by $\beta_i^t = \mathbb{1}(S_i^t = 1) / \prod_{j=1}^t \hat{\boldsymbol{\mu}}^j(\boldsymbol{x}_i^{[j]})$ as illustrated in Figure 2; see also (Bottou et al. 2013). Here, $\hat{\boldsymbol{\mu}}^t$ is an estimator of $\boldsymbol{\mu}^t$ that can be obtained using machine learning techniques like logistic regression by fitting a model to $\{\hat{\boldsymbol{x}}_i^{[t]}, S_i^t\}_{i \in I^t}$.

Finally, to tackle the last challenge, we use sample average approximation to approximate the true distribution empirically. In a data-driven setting, at stage t, we only have access to $|I^t|$ training samples generated from $\mathbb{P}$, and we define $\hat{\mathbb{P}}^{\text{IPW}}$ to be the empirical distribution supported on $\{\hat{\boldsymbol{x}}_i^{[t]}, \hat{a}_i, \hat{y}_i\}_{i \in I^T}$ after applying IPW:

$$\hat{\mathbb{P}}^{\text{IPW}} = \sum_{i=1}^{|I^t|} \frac{\beta_i^t}{\mathbf{e}^\intercal \boldsymbol{\beta}^t} \delta_{(\hat{\boldsymbol{x}}_i^{[t]}, \hat{a}_i, \hat{y}_i)}.$$

2.3 MIP Reformulation

Using the aforementioned methods, we obtain the following finite-dimensional chance-constrained program:

$$\begin{aligned}\max \quad & \hat{\mathbb{P}}^{\text{IPW}}(Y=1 \mid \mathcal{C}_T(\boldsymbol{X}^{[T]})=1) \\ \text{s.t.} \quad & \boldsymbol{W}^{[t]} \in \mathbb{R}^{d_1+\cdots+d_t},\ b_t \in \mathbb{R} \qquad \forall t \in [T] \\ & \hat{\mathbb{P}}^{\text{IPW}}(\boldsymbol{W}^{[t]} \cdot \boldsymbol{X}^{[t]} + b_t > 0) \le \overline{\alpha}_t \ \ \forall t \in [T] \\ & \hat{\mathbb{P}}^{\text{IPW}}(\boldsymbol{W}^{[T]} \cdot \boldsymbol{X}^{[T]} + b_T > 0) \ge \underline{\alpha}_T \\ & \mathcal{C}_{t+1}(\boldsymbol{X}^{[t+1]}) \le \mathcal{C}_t(\boldsymbol{X}^{[t]}) \\ & \qquad \forall (\boldsymbol{X}^{[t]}, \boldsymbol{X}^{[t+1]}) \ \ \forall t \in [T-1] \\ & \mathbb{U}(\mathcal{C}_T(\boldsymbol{X}^{[T]}), \hat{\mathbb{P}}^{\text{IPW}}) \le \eta. \end{aligned} \tag{2}$$

Program (2) allows decision-makers to explicitly bound the unfairness measure in the training set using η. Unfortunately, it remains challenging to transform (2) into an exact mixed-integer conic representable formulation that can be solved using standard MIP solvers. To see this, consider the lower bound selection ratio constraint at the final stage, which can be further represented as

$$\sum_{i=1}^{|I^T|} \beta_i^T \mathbb{1}(\boldsymbol{W}^{[T]} \cdot \boldsymbol{X}^{[T]} + b_T \le 0) \le (1 - \underline{\alpha}_T) \mathbf{e}^\intercal \boldsymbol{\beta}^T.$$

It can be verified that for $\underline{\alpha}_T \in (0,1)$, the feasible region of $(\boldsymbol{W}^{[T]}, b_T)$ with such constraint is an open set that cannot be exactly reformulated as a bounded MIP problem (Jeroslow 1987). If $1 - \underline{\alpha}_T < 1/(\mathbf{e}^\intercal \boldsymbol{\beta}^T)$, then $(\boldsymbol{W}^{[T]}, b_T)$ must satisfy $\boldsymbol{W}^{[T]} \cdot \boldsymbol{X}^{[T]} + b_T > 0 \ \forall i \in I^T$.

To address this issue, we propose a conservative approximation to (2). Firstly, for the selection ratio constraint, we change the inequality sign of the function $\mathbb{1}(\boldsymbol{W}^{[T]} \cdot \boldsymbol{X}^{[T]} + b_T \le 0)$ to a *strict* inequality. Furthermore, to ensure robustness, we modify the right-hand side to a positive quantity ϵ and use an inner approximation. Next, for the fairness constraint, leveraging the finite cardinality of $\mathcal{A}$ and $\mathcal{Y}$, we can decompose $\hat{\mathbb{P}}^{\text{IPW}}$ using its conditional measures $\hat{\mathbb{P}}_{ay}^{\text{IPW}}(\cdot) = \hat{\mathbb{P}}^{\text{IPW}}(\cdot \mid A = a, Y = y)$. We now define the ϵ-unfairness measure $\mathbb{U}_\epsilon$ as

$$\max \left\{ \begin{array}{l} \hat{\mathbb{P}}_{01}^{\text{IPW}}(\boldsymbol{W}^{[T]} \cdot \boldsymbol{X}^{[T]} + b_T > 0) \\ \quad -\hat{\mathbb{P}}_{11}^{\text{IPW}}(\boldsymbol{W}^{[T]} \cdot \boldsymbol{X}^{[T]} + b_T \ge \epsilon), \\ \hat{\mathbb{P}}_{11}^{\text{IPW}}(\boldsymbol{W}^{[T]} \cdot \boldsymbol{X}^{[T]} + b_T > 0) \\ \quad -\hat{\mathbb{P}}_{01}^{\text{IPW}}(\boldsymbol{W}^{[T]} \cdot \boldsymbol{X}^{[T]} + b_T \ge \epsilon) \end{array} \right\},$$

which is parameterized by a strictly positive value $\epsilon \in \mathbb{R}_{++}$. Lastly, for the penultimate constraint in (2), we introduce

$$C_t^\epsilon(\boldsymbol{X}^{[t]}) = \mathbb{1}(\boldsymbol{W}^{[t]} \cdot \boldsymbol{X}^{[t]} + b_t \geq \epsilon),$$

where $\epsilon \in \mathbb{R}_{++}$. This approximation yields our proposed ϵ-IPW multi-stage fair selection (ϵ-IPWMFS) model:

$$\begin{array}{ll} \max & \hat{\mathbb{P}}^{\text{IPW}}(Y = 1 \mid \mathcal{C}_T(\boldsymbol{X}^{[T]}) = 1) \\ \text{s.t.} & \boldsymbol{W}^{[t]} \in \mathbb{R}^{d_1+\cdots+d_t},\ b_t \in \mathbb{R} \qquad \forall t \in [T] \\ & \hat{\mathbb{P}}^{\text{IPW}}(\boldsymbol{W}^{[t]} \cdot \boldsymbol{X}^{[t]} + b_t > 0) \leq \overline{\alpha}_t \ \ \forall t \in [T] \\ & \hat{\mathbb{P}}^{\text{IPW}}(\boldsymbol{W}^{[T]} \cdot \boldsymbol{X}^{[T]} + b_T < \epsilon) \leq 1 - \underline{\alpha}_T \\ & \mathcal{C}_{t+1}(\boldsymbol{X}^{[t+1]}) \leq \mathcal{C}_t^\epsilon(\boldsymbol{X}^{[t]}) \\ & \qquad\qquad \forall (\boldsymbol{X}^{[t]}, \boldsymbol{X}^{[t+1]}) \ \ \forall t \in [T-1] \\ & \mathbb{U}_\epsilon(\mathcal{C}_T(\boldsymbol{X}^{[T]}), \hat{\mathbb{P}}^{\text{IPW}}) \leq \eta. \end{array} \tag{3}$$

It is worth noting that when defining the ϵ-IPWFS model (3), we have the option to use three distinct values of ϵ: one for the admission requirement constraint, one for the penultimate constraint, and another for $\mathbb{U}_\epsilon$. However, to simplify the notation and avoid the need for excessive parameter tuning, we use a single parameter ϵ.

Proposition 2.1 (Conservative approximation)**.** *Let $\{\boldsymbol{W}^{[t]^\star}, b_t^\star\}_{t=1}^T$ be an optimal solution to problem* (3)*. Then $\{\boldsymbol{W}^{[t]^\star}, b_t^\star\}_{t=1}^T$ is feasible in problem* (2)*. Moreover, let $f^\star$ and $f^\star_{opt}$ be the optimal values of problems* (3) *and* (2)*, respectively. Then, $f^\star_{opt} \geq f^\star$.*

According to Proposition 2.1, the optimal value of problem (3) provides a lower bound on the precision. We now present the main result of this section, which asserts that the problem (3) can be reformulated as a mixed binary conic optimization problem.

Theorem 2.1 (ϵ-IPWMFS reformulation)**.** *The ϵ-IPWMFS model* (3) *is equivalent to the mixed binary conic optimization problem*

$$\begin{array}{ll} \min & f \\ \text{s.t.} & \boldsymbol{W}^{[t]} \in \mathbb{R}^{d_1+\cdots+d_t},\ b_t \in \mathbb{R} \qquad \forall t \in [T] \\ & f \in \mathbb{R},\ \boldsymbol{g}_t \in \{0,1\}^{|I^T|},\ \boldsymbol{p}_t \in \{0,1\}^{|I^T|} \ \ \forall t \in [T] \\ & 1 \leq f \\ & \displaystyle\sum_{i=1}^{|I^T|} \beta_i^T (g_{Ti})^2 \leq f \sum_{i \in \mathcal{I}_1} \beta_i^T g_{Ti} \\ & \boldsymbol{g}_t^\intercal \boldsymbol{\beta}^t \leq \overline{\alpha}_t (\mathbf{e}^\intercal \boldsymbol{\beta}^t) \qquad \forall t \in [T] \\ & \boldsymbol{p}_T^\intercal \boldsymbol{\beta}^T \leq (1 - \underline{\alpha}_T)(\mathbf{e}^\intercal \boldsymbol{\beta}^T) \\ & \boldsymbol{g}_{t+1} + \boldsymbol{p}_t \leq \mathbf{e} \qquad \forall t \in [T-1] \\ & \dfrac{\sum_{i \in \mathcal{I}_{a1}} g_{Ti} \beta_i^T}{\sum_{i \in \mathcal{I}_{a1}} \beta_i^T} + \dfrac{\sum_{i \in \mathcal{I}_{a'1}} p_{Ti} \beta_i^T}{\sum_{i \in \mathcal{I}_{a'1}} \beta_i^T} - 1 \leq \eta \\ & \qquad\qquad \forall (a, a') \in \{(0,1),(1,0)\} \\ & \left.\begin{array}{l} -M(1 - g_{ti}) \leq \boldsymbol{W}^{[t]} \cdot \hat{\boldsymbol{x}}_i^{[t]} + b_t \leq M g_{ti} \\ \epsilon - \boldsymbol{W}^{[t]} \cdot \hat{\boldsymbol{x}}_i^{[t]} - b_t \leq M p_{ti} \end{array}\right\} \\ & \qquad\qquad \forall t \in [T],\ \forall i \in I^T, \end{array} \tag{4}$$

where M is the big-M parameter, $\mathcal{I}_1 = \{i \in I^T : \hat{y}_i = 1\}$, and $\mathcal{I}_{a1} = \{i \in I^T : \hat{a}_i = a, \hat{y}_i = 1\}$.

We remark that problem (4) can be solved using any popular off-the-shelf solver, such as Gurobi, Mosek, or CPLEX.

3 Numerical Experiments

In this section, we present the numerical experiments using both synthetic and real-world datasets. We consider the two-stage selection process to simplify the exposition. All optimization problems were implemented using Python 3.10 and solved by Gurobi 10.0.1. The experiments were run on an M1 Ultra CPU laptop with 64GB RAM.

Synthetic Data Experiments We first use a synthetic dataset to illustrate the importance of reweighting and the effectiveness of our fair optimization model. We simulate two-stage selection data with two subgroups, one being the minority (i.e., $A = 0$). Both groups have the same Gaussian distribution of true qualification: $X \sim \mathcal{N}(0, 2)$. To create a selection bias, we set $X^1 = X - 0.5B + noise_1$, where $noise_1 \sim \mathcal{N}(0, 0.5), B \sim Bernoulli(.2)$ if $A = 0$; and $B \sim Bernoulli(.1)$ if $A = 1$. Then, the candidates are selected for the next stage with probability $1/(1 + e^{-X^1})$. This selection criterion ensures that every candidate has a non-zero probability of being selected, and a higher value of X^1 corresponds to a higher probability of selection.

Next, for those who enter the second stage, we set a more biased $X^2 = X - 0.5 \times \mathbb{1}(A = 0) + 0.5 \times \mathbb{1}(A = 1) + noise_2$, where $noise_2 \sim \mathcal{N}(0, 0.25)$. Specifically, we tend to underestimate the qualifications of candidates from the minority group while overestimating those from the majority group. Then, based on both X^1 and X^2, the candidates are selected with probability $1/(1 + e^{-(0.7X^2 + 0.3X^1)})$.

Lastly, we generate the outcome label for each selected candidate as $Y = \mathbb{1}(X \geq 1)$. In other words, the label is only related to the true qualification and is not influenced by any sensitive attribute.

Using the aforementioned procedure, we conducted simulations of over 200,000 candidates. The results indicate that such an existing selection policy has an overall precision of approximately 68.57% and an overall unfairness score of 0.050. We will use these results as a benchmark to compare against our proposed methods. To implement

Training Size		Testing 1^{st}-Stage Violation Ration	Testing 2^{nd}-Stage Violation Ration	Testing Precision (mean ± standard deviation)	Testing Unfairness	Training Time
100	IPW	0%	80%	82.92% ± 12.63%	0.0447	0.0768s
	No-IPW	20%	80%	69.79% ± 29.39%	0.1224	0.3579s
200	IPW	0%	80%	90.22% ± 10.51%	0.0931	0.2004s
	No-IPW	0%	100%	77.57% ± 16.76%	0.0135	13.9079s
400	IPW	20%	40%	93.53% ± 5.93%	0.1772	26.4929s
	No-IPW	20%	80%	66.83% ± 32.31%	0.0674	30.8252s
800	IPW	0%	40%	95.00% ± 4.56%	0.1742	55.0997s
	No-IPW	40%	60%	51.28% ± 37.70%	0.0322	100.4673s
2000	IPW	0%	20%	94.72% ± 3.83%	0.1390	197.3959s
	No-IPW	60%	40%	16.05% ± 20.71%	0.0251	198.8476s
3000	IPW	0%	0%	95.94% ± 2.13%	0.1840	240.3892s
	No-IPW	20%	80%	41.91% ± 35.11%	0.0136	219.9468s

Table 1: Out-of-sample testing results. Here we set the fairness parameter $\eta = 1$.

the IPW scheme, we use logistic regression to estimate the propensity score $\boldsymbol{\beta}^t$. For the No-IPW scheme, we assign an equal weight to each selected candidate data point by setting $\boldsymbol{\beta}^t = \mathbf{e}$. We conduct out-of-sample experiments with training dataset sizes $N = 100, 200, 400, 800, 2000, 3000$. The selection ratios are set to $\overline{\alpha}_1 = 0.7$, $\overline{\alpha}_2 = 0.35$ and

$\underline{\alpha}_2 = 0.2$. The results of all experiments are averaged over five random trials. We set a time limit of 4 minutes for each trial (the final solution obtained with $N = 2000$ and 3000 samples may not be optimal). For each trial, we evaluate the performance using 10,000 independent testing samples. To ensure a fair comparison during the out-of-sample test, we randomly select candidates from previous stages if we select fewer than $\underline{\alpha}_2$ proportion candidates to meet the lower bound selection requirement. Similarly, if we select more than $\overline{\alpha}_1$ or $\overline{\alpha}_2$ proportion candidates, we randomly eliminate some candidates from those selected to meet the upper bound selection requirement.

The out-of-sample statistics in Table 1 showcase the superior performance of ϵ-IPWMFS, as evidenced by its decreasing constraint violation ratio as the training size increases. In contrast, the No-IPW scheme consistently yields a 100% violation ratio, rendering the generated policies unsuitable for implementation. Furthermore, IPW achieves higher precision as the training size increases, whereas the No-IPW scheme has low precision and a lower fairness score. This is due to the high violation ratio, so some candidates need to be randomly selected or eliminated to satisfy the constraints.

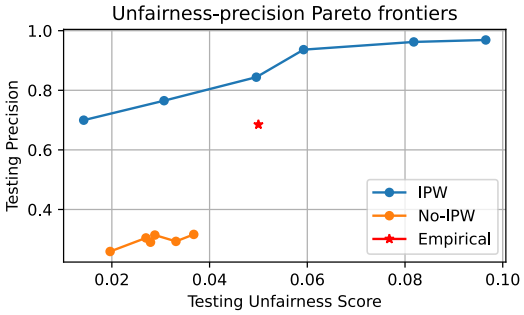

Figure 3: Pareto frontiers. The red star represents the unfairness and precision of the existing selection policy.

We plot the Pareto frontiers of the two schemes with training size $N = 800$ in Figure 3. We examine models with different values of the unfairness controlling parameter η on [0.01, 0.06] with six equidistant points, and the results were obtained from 10 independent trials. Compared to the No-IPW scheme, the IPW scheme provides higher precision for the same unfairness score. Additionally, the existing selection policy provides an unfairness score of 0.050 and a precision of 68.57% (red star in Figure 3). Using our proposed ϵ-IPWMFS method, the learned policy can achieve a much higher precision of 76.51% and a lower unfairness score of 0.031. Without IPW, the resulting learned policy is not useful as the observational data cannot represent the real-world "test" population due to selection bias.

Experiments with Real-world Data In this part, we present several semi-synthetic experiments using the Adult (Kohavi et al. 1996), COMPAS (Angwin et al. 2022), and German (Dua and Graff 2019) datasets to illustrate the effectiveness of our framework. Such semi-synthetic experiments are necessary because of the unavailability of datasets with a multi-stage selection setup. Details about the datasets are provided in the supplemental Appendix B.

For each dataset, we randomly split the covariates into two sets, one for the first-stage $\boldsymbol{X}^1$ and one for the second-stage $\boldsymbol{X}^2$. We simulate a synthetic selection process based on the following. In the first stage, we use a trained logistic regression model $f_1(\boldsymbol{X}^1) = P(Y = 1|\boldsymbol{X}^1)$ to learn the true outcome label Y, and a trained logistic regression model $g(\boldsymbol{X}^1) = P(A = 1|\boldsymbol{X}^1)$ to learn the sensitive attribute A. We create a selection bias by assigning a score to each candidate as follows: $Score_1 = 10 f_1(\boldsymbol{X}^1) - 2B + noise_1$, where $noise_1 \sim \mathcal{N}(0, 1), B \sim Bernoulli(.2)$ if $g(\boldsymbol{X}^1) = 0$; and $B \sim Bernoulli(.1)$ if $g(\boldsymbol{X}^1) = 1$. Then the candidates are selected for the next stage with probability $1/(1+e^{-Score_1})$. This selection criterion ensures that every candidate has a non-zero probability of being selected, and a higher value of $\boldsymbol{X}^1$ corresponds to a higher probability of selection.

Next, for those who get into the second stage, after we see additional covariates $\boldsymbol{X}^2$ and sensitive attribute A, we use a trained logistic regression model $f_2(\boldsymbol{X}^{[2]}) = P(Y = 1|\boldsymbol{X}^{[2]})$ to learn the true label Y. Then, we assign a score to each candidate as follows: $Score_2 = 10 f_2(\boldsymbol{X}^{[2]}) - 1.5 \times \mathbb{1}(A = 0) + 1.5 \times \mathbb{1}(A = 1) + noise_2$, where $noise_2 \sim \mathcal{N}(0, 1)$. The candidates are finally selected with probability $1/(1+e^{-(0.8 Score_2 + 0.2 Score_1)})$. We generate the selection process three times following the aforementioned procedure for each dataset. We compare the performance of two different models: the No-Fair model, where the unfairness control parameter is set to $\eta = 1$ in (4), and the Fair model, where the unfairness control parameter η is set to the respective empirical unfairness score.

Dataset	Metric	No-Fair	Fair	Empirical
Adult	Precision	**0.52 ± 0.05**	0.51 ± 0.04 (1.92%)	0.28
	Unfairness (U)	0.26 ± 0.19	**0.17 ± 0.13 (34.61%)**	0.22
	Precision	**0.45 ± 0.05**	0.44 ± 0.05 (2.22%)	0.27
	Unfairness (U)	0.21 ± 0.15	**0.16 ± 0.12 (23.81%)**	0.23
	Precision	**0.39 ± 0.05**	0.37 ± 0.06 (5.13%)	0.27
	Unfairness (U)	0.17 ± 0.11	**0.14 ± 0.12 (17.65%)**	0.24
German	Precision	**0.78 ± 0.12**	0.78 ± 0.13 (0%)	0.76
	Unfairness (U)	0.08 ± 0.07	**0.05 ± 0.05 (37.5%)**	0.16
	Precision	**0.82 ± 0.11**	0.76 ± 0.13 (7.32%)	0.76
	Unfairness (U)	0.11 ± 0.09	**0.05 ± 0.03 (54.55%)**	0.17
	Precision	**0.73 ± 0.09**	0.72 ± 0.08 (1.37%)	0.72
	Unfairness (U)	0.12 ± 0.06	**0.09 ± 0.07 (25%)**	0.13
COMPAS	Precision	**0.67 ± 0.04**	0.60 ± 0.09 (10.45%)	0.52
	Unfairness (U)	0.14 ± 0.04	**0.09 ± 0.06 (35.71%)**	0.13
	Precision	**0.57 ± 0.07**	0.53 ± 0.04 (7.02%)	0.51
	Unfairness (U)	0.07 ± 0.07	**0.05 ± 0.05 (28.57%)**	0.12
	Precision	**0.59 ± 0.05**	0.58 ± 0.06 (1.69%)	0.53
	Unfairness (U)	0.09 ± 0.05	**0.06 ± 0.03 (33.33%)**	0.16

Table 2: Out-of-sample testing results (mean ± standard deviation). The numbers inside the parentheses represent the average reduction compared to the No-Fair model.

We conduct out-of-sample experiments with a training dataset of size $N = 200$. The selection ratios are set to $\overline{\alpha}_1 = 0.7, \overline{\alpha}_2 = 0.35$ and $\underline{\alpha}_2 = 0.2$. The results of all experiments are averaged over 20 random trials. Table 2 demonstrates the superior performance of ϵ-IPWMFS. As we can see, both the No-Fair and Fair models achieve high precision compared with the existing selection policy. Besides, the Fair model can achieve much lower unfairness scores with a negligible decrease in precision compared with the No-Fair model.

Acknowledgments

Z. Jia and G. Hanasusanto are funded in part by the National Science Foundation under grants 2342505 and 2343869. P. Vayanos is funded in part by the National Science Foundation under grant 2046230. W. Xie is funded in part by the National Science Foundation under grants 2246414 and 2246417. They thank the four anonymous referees whose reviews helped substantially improve the quality of the paper.

References

Aghaei, S.; Azizi, M. J.; and Vayanos, P. 2019. Learning optimal and fair decision trees for non-discriminative decision-making. In *Proceedings of the AAAI conference on artificial intelligence*, volume 33(01), 1418–1426.

Aghaei, S.; Gómez, A.; and Vayanos, P. 2021. Strong optimal classification trees. *arXiv preprint arXiv:2103.15965*.

Ahmad, S. F.; Alam, M. M.; Rahmat, M. K.; Mubarik, M. S.; and Hyder, S. I. 2022. Academic and administrative role of artificial intelligence in education. *Sustainability*, 14(3): 1101.

Angwin, J.; Larson, J.; Mattu, S.; and Kirchner, L. 2022. Machine bias. In *Ethics of data and analytics*, 254–264. Auerbach Publications.

Arcidiacono, P.; Kinsler, J.; and Ransom, T. 2022. Asian American discrimination in Harvard admissions. *European Economic Review*, 144: 104079.

Atkins, R.; Cook, L.; and Seamans, R. 2022. Discrimination in lending? Evidence from the paycheck protection program. *Small Business Economics*, 1–23.

Barocas, S.; Hardt, M.; and Narayanan, A. 2017. Fairness in machine learning. *Nips tutorial*, 1: 2017.

Barocas, S.; and Selbst, A. D. 2016. Big data's disparate impact. *California law review*, 671–732.

Berk, R.; Heidari, H.; Jabbari, S.; Joseph, M.; Kearns, M.; Morgenstern, J.; Neel, S.; and Roth, A. 2017. A convex framework for fair regression. *arXiv preprint arXiv:1706.02409*.

Bertrand, M.; and Mullainathan, S. 2004. Are Emily and Greg more employable than Lakisha and Jamal? A field experiment on labor market discrimination. *American economic review*, 94(4): 991–1013.

Bertsimas, D.; and Dunn, J. 2017. Optimal classification trees. *Machine Learning*, 106: 1039–1082.

Blum, A.; and Stangl, K. 2019. Recovering from biased data: Can fairness constraints improve accuracy? *arXiv preprint arXiv:1912.01094*.

Bottou, L.; Peters, J.; Quiñonero-Candela, J.; Charles, D. X.; Chickering, D. M.; Portugaly, E.; Ray, D.; Simard, P.; and Snelson, E. 2013. Counterfactual Reasoning and Learning Systems: The Example of Computational Advertising. *Journal of Machine Learning Research*, 14(11).

Calders, T.; Kamiran, F.; and Pechenizkiy, M. 2009. Building classifiers with independency constraints. In *2009 IEEE international conference on data mining workshops*, 13–18. IEEE.

Celis, L. E.; Mehrotra, A.; and Vishnoi, N. K. 2020. Interventions for ranking in the presence of implicit bias. In *Proceedings of the 2020 conference on fairness, accountability, and transparency*, 369–380.

Chouldechova, A.; and Roth, A. 2020. A snapshot of the frontiers of fairness in machine learning. *Communications of the ACM*, 63(5): 82–89.

Corbett-Davies, S.; Pierson, E.; Feller, A.; Goel, S.; and Huq, A. 2017. Algorithmic decision making and the cost of fairness. In *Proceedings of the 23rd acm sigkdd international conference on knowledge discovery and data mining*, 797–806.

Dua, D.; and Graff, C. 2019. UCI Machine Learning Repository. University of California, School of Information and Computer Science, Irvine, CA (2019).

Dyer, M. E.; and Frieze, A. M. 1988. On the complexity of computing the volume of a polyhedron. *SIAM Journal on Computing*, 17(5): 967–974.

Emelianov, V.; Arvanitakis, G.; Gast, N.; Gummadi, K.; and Loiseau, P. 2019. The price of local fairness in multistage selection. *arXiv preprint arXiv:1906.06613*.

Feldman, M.; Friedler, S. A.; Moeller, J.; Scheidegger, C.; and Venkatasubramanian, S. 2015. Certifying and removing disparate impact. In *proceedings of the 21th ACM SIGKDD international conference on knowledge discovery and data mining*, 259–268.

Goel, N.; Amayuelas, A.; Deshpande, A.; and Sharma, A. 2021. The importance of modeling data missingness in algorithmic fairness: A causal perspective. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 35(9), 7564–7573.

Green, B.; and Chen, Y. 2019. The principles and limits of algorithm-in-the-loop decision making. *Proceedings of the ACM on Human-Computer Interaction*, 3(CSCW): 1–24.

Greenwald, A. G.; and Banaji, M. R. 1995. Implicit social cognition: attitudes, self-esteem, and stereotypes. *Psychological review*, 102(1): 4.

Hardt, M.; Price, E.; and Srebro, N. 2016. Equality of opportunity in supervised learning. *Advances in neural information processing systems*, 29.

Horvitz, D. G.; and Thompson, D. J. 1952. A generalization of sampling without replacement from a finite universe. *Journal of the American statistical Association*, 47(260): 663–685.

Jeroslow, R. G. 1987. Representability in mixed integer programmiing, I: Characterization results. *Discrete Applied Mathematics*, 17(3): 223–243.

Jo, N.; Aghaei, S.; Gómez, A.; and Vayanos, P. 2021. Learning optimal prescriptive trees from observational data. *arXiv preprint arXiv:2108.13628*.

Jo, N.; Aghaei, S.; Gómez, A.; and Vayanos, P. 2022. Learning Optimal Fair Classification Trees: Trade-offs Between Interpretability, Fairness, and Accuracy. *arXiv preprint arXiv:2201.09932*.

Kallus, N.; and Zhou, A. 2018. Residual unfairness in fair machine learning from prejudiced data. In *International Conference on Machine Learning*, 2439–2448. PMLR.

Kamishima, T.; Akaho, S.; Asoh, H.; and Sakuma, J. 2012. Fairness-aware classifier with prejudice remover regularizer. In *Machine Learning and Knowledge Discovery in Databases: European Conference, ECML PKDD 2012, Bristol, UK, September 24-28, 2012. Proceedings, Part II 23*, 35–50. Springer.

Khademi, A.; Lee, S.; Foley, D.; and Honavar, V. 2019. Fairness in algorithmic decision making: An excursion through the lens of causality. In *The WWW Conference*, 2907–2914.

Khalili, M. M.; Zhang, X.; and Abroshan, M. 2021. Fair sequential selection using supervised learning models. *Advances in Neural Information Processing Systems*, 34: 28144–28155.

Khalili, M. M.; Zhang, X.; Abroshan, M.; and Sojoudi, S. 2021. Improving fairness and privacy in selection problems. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 35(9), 8092–8100.

Kilbertus, N.; Rodriguez, M. G.; Schölkopf, B.; Muandet, K.; and Valera, I. 2020. Fair decisions despite imperfect predictions. In *International Conference on Artificial Intelligence and Statistics*, 277–287. PMLR.

Kleinberg, J.; and Raghavan, M. 2018. Selection problems in the presence of implicit bias. *arXiv preprint arXiv:1801.03533*.

Kohavi, R.; et al. 1996. Scaling up the accuracy of naive-bayes classifiers: A decision-tree hybrid. In *Kdd*, volume 96, 202–207.

Lakkaraju, H.; Kleinberg, J.; Leskovec, J.; Ludwig, J.; and Mullainathan, S. 2017. The selective labels problem: Evaluating algorithmic predictions in the presence of unobservables. In *Proceedings of the 23rd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, 275–284.

Li, L.; Lassiter, T.; Oh, J.; and Lee, M. K. 2021. Algorithmic hiring in practice: Recruiter and HR Professional's perspectives on AI use in hiring. In *Proceedings of the 2021 AAAI/ACM Conference on AI, Ethics, and Society*, 166–176.

Liao, Y.; and Naghizadeh, P. 2023. Social bias meets data bias: The impacts of labeling and measurement errors on fairness criteria. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 37(7), 8764–8772.

Los Angeles Homeless Services Authority. 2018. Report and Recommendations of the Ad Hoc Committee on Black People Experiencing Homelessness.

Maragno, D.; Wiberg, H.; Bertsimas, D.; Birbil, S. I.; Hertog, D. d.; and Fajemisin, A. 2021. Mixed-integer optimization with constraint learning. *arXiv preprint arXiv:2111.04469*.

Marques-Silva, J.; and Ignatiev, A. 2022. Delivering Trustworthy AI through formal XAI. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 36(11), 12342–12350.

Mehrabi, N.; Morstatter, F.; Saxena, N.; Lerman, K.; and Galstyan, A. 2021. A survey on bias and fairness in machine learning. *ACM computing surveys (CSUR)*, 54(6): 1–35.

Nabi, R.; and Shpitser, I. 2018. Fair inference on outcomes. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 32(1).

Oreopoulos, P. 2011. Why do skilled immigrants struggle in the labor market? A field experiment with thirteen thousand resumes. *American Economic Journal: Economic Policy*, 3(4): 148–171.

Pleiss, G.; Raghavan, M.; Wu, F.; Kleinberg, J.; and Weinberger, K. Q. 2017. On fairness and calibration. *Advances in neural information processing systems*, 30.

Rezaei, A.; Liu, A.; Memarrast, O.; and Ziebart, B. D. 2021. Robust fairness under covariate shift. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 35(11), 9419–9427.

Taskesen, B.; Nguyen, V. A.; Kuhn, D.; and Blanchet, J. 2020. A distributionally robust approach to fair classification. *arXiv preprint arXiv:2007.09530*.

Wang, Y.; Nguyen, V. A.; and Hanasusanto, G. A. 2021. Wasserstein robust classification with fairness constraints. *arXiv preprint arXiv:2103.06828*.

Ye, Q.; and Xie, W. 2020. Unbiased subdata selection for fair classification: A unified framework and scalable algorithms. *arXiv preprint arXiv:2012.12356*.

Zafar, M. B.; Valera, I.; Gomez Rodriguez, M.; and Gummadi, K. P. 2017. Fairness beyond disparate treatment & disparate impact: Learning classification without disparate mistreatment. In *Proceedings of the 26th international conference on world wide web*, 1171–1180.

Zafar, M. B.; Valera, I.; Gomez-Rodriguez, M.; and Gummadi, K. P. 2019. Fairness constraints: A flexible approach for fair classification. *The Journal of Machine Learning Research*, 20(1): 2737–2778.

Appendix

A Proofs in the main text

In this section, we provide the proofs for Proposition 2.1 and Theorem 2.1 in Section 2.3.

A.1 Proof of Proposition 2.1

Proof. Consider any feasible solution $(\boldsymbol{W}^{[t]}, b_t)_{t=1}^T$ to (3) and any distribution $\hat{\mathbb{P}}^{\text{IPW}}$. We first prove the feasibility of the fairness constraint in (2). For any $\epsilon \in \mathbb{R}_{++}$, we have

$$\begin{aligned}
&\mathbb{U}(\mathcal{C}_T(\boldsymbol{X}^{[T]}), \hat{\mathbb{P}}^{\text{IPW}}) \\
=&\max \left\{ \begin{array}{l} \hat{\mathbb{P}}^{\text{IPW}}_{01}(\boldsymbol{W}^{[T]} \cdot \boldsymbol{X}^{[T]} + b_T > 0) \\ -\hat{\mathbb{P}}^{\text{IPW}}_{11}(\boldsymbol{W}^{[T]} \cdot \boldsymbol{X}^{[T]} + b_T > 0), \\ \hat{\mathbb{P}}^{\text{IPW}}_{11}(\boldsymbol{W}^{[T]} \cdot \boldsymbol{X}^{[T]} + b_T > 0) \\ -\hat{\mathbb{P}}^{\text{IPW}}_{01}(\boldsymbol{W}^{[T]} \cdot \boldsymbol{X}^{[T]} + b_T > 0) \end{array} \right\} \\
\leq&\max \left\{ \begin{array}{l} \hat{\mathbb{P}}^{\text{IPW}}_{01}(\boldsymbol{W}^{[T]} \cdot \boldsymbol{X}^{[T]} + b_T > 0) \\ -\hat{\mathbb{P}}^{\text{IPW}}_{11}(\boldsymbol{W}^{[T]} \cdot \boldsymbol{X}^{[T]} + b_T \geq \epsilon), \\ \hat{\mathbb{P}}^{\text{IPW}}_{11}(\boldsymbol{W}^{[T]} \cdot \boldsymbol{X}^{[T]} + b_T > 0) \\ -\hat{\mathbb{P}}^{\text{IPW}}_{01}(\boldsymbol{W}^{[T]} \cdot \boldsymbol{X}^{[T]} + b_T \geq \epsilon) \end{array} \right\} \\
=&\mathbb{U}_\epsilon(\mathcal{C}_T(\boldsymbol{X}^{[T]}), \hat{\mathbb{P}}^{\text{IPW}}) \leq \eta.
\end{aligned}$$

Therefore, $(\boldsymbol{W}^{[t]}, b_t)_{t=1}^T$ are also feasible to the fairness constraint in (2). Next, by definition, we have

$$\begin{aligned}
\mathcal{C}_{t+1}(\boldsymbol{X}^{[t+1]}) &\leq C_t^\epsilon(\boldsymbol{X}^{[t]}) \\
&= \mathbb{1}(\boldsymbol{W}^{[t]} \cdot \boldsymbol{X}^{[t]} + b_t \geq \epsilon) \\
&\leq \mathbb{1}(\boldsymbol{W}^{[t]} \cdot \boldsymbol{X}^{[t]} + b_t > 0) = C_t(\boldsymbol{X}^{[t]}).
\end{aligned}$$

This proves the feasibility of the penultimate constraint to (2). Lastly, we have

$$\begin{aligned}
&\hat{\mathbb{P}}^{\text{IPW}}(\boldsymbol{W}^{[T]} \cdot \boldsymbol{X}^{[T]} + b_T \leq 0) \\
\leq&\hat{\mathbb{P}}^{\text{IPW}}(\boldsymbol{W}^{[T]} \cdot \boldsymbol{X}^{[T]} + b_T < \epsilon) \\
\leq&1 - \underline{\alpha}_T,
\end{aligned}$$

which proves the feasibility of the selection ratio requirement to (2). Consequently, the feasible set of problem (3) is an *inner* approximation of the feasible set of problem (2) for any $\epsilon \in \mathbb{R}_{++}$. Therefore, an optimal solution $(\boldsymbol{W}^{[t]^*}, b_t^*)_{t=1}^T$ of problem (3) is also feasible for problem (2). By plugging in the optimal solution $(\boldsymbol{W}^{[t]^*}, b_t^*)_{t=1}^T$ to (2), we have $f_{opt}^\star \geq f^\star$. This completes the proof. □

A.2 Proof of Theorem 2.1

Proof. We start with the reformulation for the objective function of (3). By definition of conditional probability, we can rewrite the precision as

$$\begin{aligned}
&\hat{\mathbb{P}}^{\text{IPW}}(Y = 1 \mid \mathcal{C}_T(\boldsymbol{X}^{[T]}) = 1) \\
=& \frac{\hat{\mathbb{P}}^{\text{IPW}}(Y = 1, \mathcal{C}_T(\boldsymbol{X}^{[T]}) = 1)}{\hat{\mathbb{P}}^{\text{IPW}}(\mathcal{C}_T(\boldsymbol{X}^{[T]}) = 1)}.
\end{aligned}$$

Using the fact that the final stage selection constraint requires $0 < \underline{\alpha}_T \leq \hat{\mathbb{P}}^{\text{IPW}}(\mathcal{C}_T(\boldsymbol{X}^{[T]}) = 1)$, instead of maximizing the precision, we can minimize the reciprocal of it. By introducing an epigraphical variable f, which we minimize, and moving the reciprocal of the precision into the constraint, we have

$$\begin{aligned}
&\frac{\hat{\mathbb{P}}^{\text{IPW}}(\mathcal{C}_T(\boldsymbol{X}^{[T]}) = 1)}{\hat{\mathbb{P}}^{\text{IPW}}(Y = 1, \mathcal{C}_T(\boldsymbol{X}^{[T]}) = 1)} \\
=& \left\{ \begin{array}{ll} \min & f \\ \text{s.t.} & f \in \mathbb{R} \\ & 1 \leq f \\ & \hat{\mathbb{P}}^{\text{IPW}}(\mathcal{C}_T(\boldsymbol{X}^{[T]}) = 1) \\ & \quad \leq f \hat{\mathbb{P}}^{\text{IPW}}(Y = 1, \mathcal{C}_T(\boldsymbol{X}^{[T]}) = 1), \end{array} \right.
\end{aligned}$$

where the first constraint comes from the fact that the reciprocal of a conditional probability is at least one. For any $t \in [T]$ and $i \in I^T$, we define

$$\begin{aligned}
g_{ti} &= \mathbb{1}(C_t(\hat{\boldsymbol{x}}_i^{[t]}) = 1) = \mathbb{1}(\boldsymbol{W}^{[t]} \cdot \hat{\boldsymbol{x}}_i^{[t]} + b_t > 0) \\
p_{ti} &= \mathbb{1}(C_t^\epsilon(\hat{\boldsymbol{x}}_i^{[t]}) = 1) = \mathbb{1}(\boldsymbol{W}^{[t]} \cdot \hat{\boldsymbol{x}}_i^{[t]} + b_t < \epsilon).
\end{aligned}$$

Using big-M parameter M, we can reformulate the above indicator function as

$$\left\{ \begin{array}{l} \boldsymbol{g}_t \in \{0,1\}^{|I^T|},\ \boldsymbol{p}_t \in \{0,1\}^{|I^T|} \\ -M(1 - g_{ti}) \leq \boldsymbol{W}^{[t]} \cdot \hat{\boldsymbol{x}}_i^{[t]} + b_t \leq M g_{ti} \\ \epsilon - \boldsymbol{W}^{[t]} \cdot \hat{\boldsymbol{x}}_i^{[t]} - b_t \leq M p_{ti}. \end{array} \right.$$

Recalling the definition of $\hat{\mathbb{P}}^{\text{IPW}}$,

$$\hat{\mathbb{P}}^{\text{IPW}} = \sum_{i=1}^{|I^t|} \frac{\beta_i^t}{\mathbf{e}^\intercal \boldsymbol{\beta}^t} \delta_{(\hat{\boldsymbol{x}}_i^{[t]}, \hat{a}_i, \hat{y}_i)}.$$

we can thus rewrite the above minimization problem as

$$\left\{ \begin{array}{ll} \min & f \\ \text{s.t.} & f \in \mathbb{R},\ \boldsymbol{g}_T \in \{0,1\}^{|I^T|} \\ & 1 \leq f \\ & \sum_{i=1}^{|I^T|} \beta_i^T (g_{Ti})^2 \leq f \sum_{i \in \mathcal{I}_1} \beta_i^T g_{Ti}, \end{array} \right.$$

where $\mathcal{I}_1 = \{i \in I^T : \hat{y}_i = 1\}$, M is the big-M parameter. Next, we provide the reformulation for the upper bound selection ratio requirement constraint of (3) at $t \in [T]$. By definition, we have

$$\begin{aligned}
&\hat{\mathbb{P}}^{\text{IPW}}(\boldsymbol{W}^{[t]} \cdot \boldsymbol{X}^{[t]} + b_t > 0) \leq \overline{\alpha}_t \\
\Longleftrightarrow& \frac{\sum_{i=1}^{|I^T|} \beta_i^t \mathbb{1}(\boldsymbol{W}^{[t]} \cdot \hat{\boldsymbol{x}}_i^{[t]} + b_t > 0)}{\mathbf{e}^\intercal \boldsymbol{\beta}^t} \leq \overline{\alpha}_t \\
\Longleftrightarrow& \left\{ \begin{array}{l} \boldsymbol{g}_t \in \{0,1\}^{|I^T|} \\ \boldsymbol{g}_t^\intercal \boldsymbol{\beta}^t \leq \overline{\alpha}_t (\mathbf{e}^\intercal \boldsymbol{\beta}^t). \end{array} \right.
\end{aligned}$$

For the lower bound selection ratio requirements at the terminal stage, we have

$$\begin{aligned}
&\hat{\mathbb{P}}^{\text{IPW}}(\boldsymbol{W}^{[T]} \cdot \boldsymbol{X}^{[T]} + b_T < \epsilon) \leq 1 - \underline{\alpha}_T \\
=& \left\{ \begin{array}{l} \boldsymbol{p}_T \in \{0,1\}^{|I^T|} \\ \boldsymbol{p}_T^\intercal \boldsymbol{\beta}^T \leq (1 - \underline{\alpha}_T)(\mathbf{e}^\intercal \boldsymbol{\beta}^T). \end{array} \right.
\end{aligned}$$

The admission requirement constraint of (3) can be written as

$$\begin{aligned}
&\mathcal{C}_{t+1}(\boldsymbol{X}^{[t+1]}) \leq \mathcal{C}_t^{\epsilon}(\boldsymbol{X}^{[t]}) \\
\Longleftrightarrow &\mathbb{1}(\boldsymbol{W}^{[t+1]} \cdot \hat{\boldsymbol{x}}_i^{[t+1]} + b_{t+1} > 0) \\
&\qquad \leq \mathbb{1}(\boldsymbol{W}^{[t]} \cdot \hat{\boldsymbol{x}}_i^{[t]} + b_t \geq \epsilon) \\
\Longleftrightarrow &\mathbb{1}(\boldsymbol{W}^{[t+1]} \cdot \hat{\boldsymbol{x}}_i^{[t+1]} + b_{t+1} > 0) \\
&\qquad + \mathbb{1}(\boldsymbol{W}^{[t]} \cdot \hat{\boldsymbol{x}}_i^{[t]} + b_t < \epsilon) \leq 1.
\end{aligned}$$

which can be reformulated as

$$\begin{cases}
\boldsymbol{g}_{t+1} \in \{0,1\}^{|I^T|},\ \boldsymbol{p}_t \in \{0,1\}^{|I^T|} \\
\boldsymbol{g}_{t+1} + \boldsymbol{p}_t \leq \mathbf{e}.
\end{cases}$$

Finally, we provide the reformulation for the fairness constraint of (3). Define the index sets $\mathcal{I}_{a1} = \{i \in [N] : \hat{a}_i = a, \hat{y}_i = 1\}\ \forall a \in \mathcal{A}$. For any fixed pair $(a, a') \in \{(0,1),(1,0)\}$, we have

$$\begin{aligned}
&\hat{\mathbb{P}}_{a1}^{\text{IPW}}(\boldsymbol{W}^{[T]} \cdot \boldsymbol{X}^{[T]} + b_T > 0) \\
&\quad -\hat{\mathbb{P}}_{a'1}^{\text{IPW}}(\boldsymbol{W}^{[T]} \cdot \boldsymbol{X}^{[T]} + b_T \geq \epsilon) \\
=\ &\hat{\mathbb{P}}_{a1}^{\text{IPW}}(\boldsymbol{W}^{[T]} \cdot \boldsymbol{X}^{[T]} + b_T > 0) \\
&\quad +\hat{\mathbb{P}}_{a'1}^{\text{IPW}}(\boldsymbol{W}^{[T]} \cdot \boldsymbol{X}^{[T]} + b_T < \epsilon) - 1 \\
=\ &\mathbb{E}_{\hat{\mathbb{P}}^{\text{IPW}}}\Big[\hat{p}_{a1}^{-1}\mathbb{1}(\boldsymbol{W}^{[T]} \cdot \boldsymbol{X}^{[T]} + b_T > 0)\mathbb{1}((A,Y) = (a,1)) \\
&\quad +\hat{p}_{a'1}^{-1}\mathbb{1}(\boldsymbol{W}^{[T]} \cdot \boldsymbol{X}^{[T]} + b_T < \epsilon)\mathbb{1}((A,Y) = (a',1))\Big] - 1,
\end{aligned}$$

where $\hat{p}_{a1} = \hat{\mathbb{P}}^{\text{IPW}}(A = a, Y = 1)$ and $\hat{p}_{a'1} = \hat{\mathbb{P}}^{\text{IPW}}(A = a', Y = 1)$. Using the definition of $\hat{\mathbb{P}}^{\text{IPW}}$, we can rewrite the about expression as

$$\begin{cases}
\boldsymbol{g}_T \in \{0,1\}^{|I^T|},\ \boldsymbol{p}_T \in \{0,1\}^{|I^T|} \\
\dfrac{1}{\sum_{i \in \mathcal{I}_{a1}} \beta_i^T} \displaystyle\sum_{i \in \mathcal{I}_{a1}} \beta_i^T g_{Ti} + \dfrac{1}{\sum_{i \in \mathcal{I}_{a'1}} \beta_i^T} \displaystyle\sum_{i \in \mathcal{I}_{a'1}} \beta_i^T p_{Ti} - 1.
\end{cases}$$

Combining all the above analysis, we have the result in Theorem 2.1. □

B Real Datasets

In this section, we provide the details for the datasets in Section 3.

Dataset	Covariates d	Protected Attribute	Number of Samples
Adult	12	Gender	32561, 12661
German	19	Gender	1000
COMPAS	10	Ethnicity	6172

Table 3: Summary of Datasets.

The Adult dataset, also known as "Census Income" dataset, is taken from a 1994 Census database. The goal is to determine whether a person's annual income exceeds $50,000 or not. It contains 13 features concerning demographic characteristics of 45,222 instances, and we consider gender as the sensitive attribute.

The German Credit Risks dataset consists of samples of bank account holders with good or bad credit. The data are collected from 1,000 individuals, and we take gender as the sensitive attribute.

The COMPAS dataset consists of the variables of a commercial algorithm called COMPAS (Correctional Offender Management Profiling for Alternative Sanctions), used to predict a convicted criminal's likelihood of recidivism within two years. The data contains 6,172 data points, and we use ethnicity as the sensitive attribute.