Crab: Learning Certifiably Fair Predictive Models in the Presence of Selection Bias

ABSTRACT

A recent explosion of research focuses on developing methods and tools for building fair predictive models. However, most of this work relies on the assumption that the training and testing data are representative of the target population on which the model will be deployed. However, real-world training data often suffer from selection bias and are not representative of the target population for many reasons, including the cost and feasibility of collecting and labeling data, historical discrimination, and individual biases.

In this paper, we introduce a new framework for certifying and ensuring fairness of predictive models trained on biased data. We take inspiration from query answering over incomplete and inconsistent databases to present and formalize the problem of consistent range approximation (CRA) of answers to queries about aggregate information for the target population. We aim to leverage background knowledge about the data collection process, biased data, and limited or no auxiliary data sources to compute a range of answers for aggregate queries over the target population that are consistent with available information. We then develop methods that use CRA of such aggregate queries to build predictive models that are certifiably fair on the target population even when no external information about that population is available during training. We evaluate our methods on real data and demonstrate improvements over the state of the art. Significantly, we show that enforcing fairness using our methods can lead to predictive models that are not only fair, but more accurate on the target population.

1 INTRODUCTION

Data-driven algorithmic decision making is increasingly used to make consequential decisions about hiring employees, assigning loans and credit scores, authorizing pretrial release, and diagnosing medical conditions. Although these algorithmic systems may show an "objective" veneer, there is a growing realization that algorithms can be biased, and this bias can lead to erroneous decisions and decisions that inequitably affect certain groups, perpetuating systemic discrimination. In this regard, a growing literature in machine learning (ML), data management, and other fields investigates fairness in the context of learning predictive models from data (see [6, 36, 47], for recent surveys).

Current methods for fair predictive modeling ensure fairness either by intervening on the learning algorithm (in-processing) (see, e.g., [15, 39, 41, 50, 59, 69, 73]) or by intervening on training and testing data (pre-, post-processing) (see, e.g., [16, 28, 33, 61, 67]). However, most of these methods and tools explicitly or implicitly rely on the assumption that the training data is representative of the target population in which the ML model will be deployed [36]. However, in a real-world setting, unfairness of predictive models is often due to *data biases and quality issues* introduced during data collection and preparation that distort the underlying data distribution so that it no longer represents the target population. It has been shown and argued that in the presence of such data quality issues, current approaches may lead to predictive models that are fair and accurate during training and testing but unfair and inaccurate during deployment in the target population [31, 36].

A major data quality issue that hinder efforts to develop fair predictive models is *selection bias*, which occurs when the selection of data points in training data depends on their attributes so that training data no longer reflects the target population. In practice, available training data in sensitive domains – such as predictive policing, healthcare, finance, etc. – often suffers from selection bias [12, 23, 32, 40, 46, 49, 58] due to a variety of reasons, such as the cost and feasibility of collecting and labeling data, historical discrimination, and individual biases. Selection bias may lead to spurious conclusions about the existence of a valid statistical dependence between two variables, e.g., protected attributes such as race and gender and the training label (e.g., drug use or the likelihood of recidivism) when one does not exist, or, alternatively, about the absence of a statistical dependence between them when one is truly present in the target population. This, in turn, may lead to unfair and inaccurate predictive models.

Much work in ML address the problem of learning ML models in the presence of selection bias (see e.g., [7, 20, 20, 35, 45, 45]). However, these works rely on the existence of unbiased samples from the target population, which is often unavailable in practice, do not address unfairness of predictive models, can exhibit poor performance in practice [20, 45].

We illustrate these points with two examples.

Example 1.1 (Predictive Policing). Predictive policing is a law enforcement practice that uses ML models to predict crime. In this context, we have access only to data gathered by the police department, widely believed to harbor selection bias [14, 46, 49, 58]: police datasets include data points related only to individuals with whom the police interact, and studies show that officers are influenced by sociocultural traits in their interactions [29, 43]. Additionally, if a neighborhood is more regularly patrolled, police are more likely to detain those suspected of committing crimes [46]. Hence, police data may exhibit spurious correlations between protected attributes such as race and crime that exist in training data, but not on the underlying population, which can lead to unfair predictive models. Furthermore, using such an ML model may lead to results that are fair during training and testing, but unfair with respect to the target population.

Example 1.2 (COVID-19 Mortality Risk Prediction). Over the past two years, many publications have examined the use of predictive modeling to detect and combat COVID-19 [4, 5]. COVID studies are often conducted on data collected from individuals that have a positive PRC test or who have already been admitted to hospitals [63, 75]. These datasets are thus contaminated by selection bias since racial and ethnic minorities [3], as well as gender minorities [17], have been shown to have less access to healthcare (and thus might be less likely to go to a hospital or perform self-tests). In addition, in terms of testing for COVID-19, (1) the availability of testing sites and (2) the desire of individuals to perform the test could also influence dataset samples [37]. As [37] discusses, the

lower availability of test centers in neighborhoods with larger minority populations, coupled with the history of unfair treatment of minorities, could cause such groups to be unable or decide not to perform these tests, thereby excluding them from the datasets. As a result, the under-representation of racial and gender minorities in these datasets could prevent the models from learning their patterns, which could result in higher inaccuracies for these groups and further propagate inequalities in health care [53].

This paper introduces a new framework for **certifying and ensuring fairness of predictive models** trained in the presence of selection bias. The framework relies only on background knowledge about the underlying data collection process, which is often available in practice, and is designed to work with **any level of external knowledge about the target population**.

Our system, **C**onsistent **R**ange **A**pproximation from **B**iased Data, Crab, employs a principled **causality-based approach** to encode assumptions about the data collection process using causal diagrams. Then, it establishes conditions on the data collection process under which it is possible to train a predictive model that is **certifiably fair on a target population**. Unlike previous methods for predictive modeling in the presence of selection bias, which rely on access to unbiased samples from the target populations and do not explicitly account for unfairness, our framework can be used to train predictive models that are **guaranteed to be fair even when no information about the target population is available** during model training and testing. We show that in the presence of limited external information about the target population, Crab can lead to predictive models that are **not only fair but more accurate with respect to the target population**.

Our core technical contribution consists of an in-depth **theoretical analysis of the impact of selection bias on the fairness of predictive models.** We first establish necessary and sufficient conditions on the data collection process under which selection bias leads to unfair predictive models. We next show that in the presence of selection bias in training data, **the existing tools for fair predictive modeling may lead to ML models that are fair during testing and training, but unfair during the deployment** in the target population **(Section 3)**.

One key issue hinders the training of fair predictive models in the presence of selection bias is that using the biased data to answer queries about aggregate information on the target population may lead to biased query answers differ substantially from the true query. To address this, we introduce and formalize the problem of **Consistent Range Approximation (CRA) of aggregate queries from biased data**; given a target population and aggregate querys about the population, *CRA* aims to find approximate answers to the query when we have access only to a biased sample from the population; background knowledge about the data collection process; and no or limited auxiliary external data sources, e.g., census data, results of previous studies, open knowledge graphs, and assumptions on population parameters that could potentially reveal information about the target population.

In general, query answering from data that suffers from selection bias is akin to query answering from inconsistent and incomplete data, which has been extensively studied in databases, although not for biased data and population-level data errors [11, 27, 52]. Inspired by Consistent Query Answering in databases [11, 24], which aims to answer queries from inconsistent data, *CRA* considers the space of **all possible populations that are consistent** with the available information, i.e., all populations from which the sample could be drawn, and uses it to compute a **range for the query answers** such that the true query answers are guaranteed to lie within the range. We call this the *consistent range* **(Section 4)**.

Then, we study the problem of *CRA* for a class of aggregate queries that are sufficiently expressive to capture many existing notions of algorithmic fairness that are widely used in practice under varying levels of access to external knowledge about the target population. In this context, *CRA* enables us to **compute a consistent range for fairness of a predictive model trained on biased data** on a target population. This, in turn, lets us not only certify whether a predictive model is approximately fair on the target population, but also train models that will be certifiably fair on all possible populations consistent with the available information, and thereby on the target population **(Section 5)**.

We evaluate Crab on synthetic and real data. We find that in the presence of selection bias, the following occur. (1) The existing methods for training predictive modeling lead to unfair models. (2) In contrast, Crab develops predictive models that are guaranteed to be fair on the target population. (3) In the presence of limited auxiliary information about the target population, Crab develops predictive models that are fair and highly accurate. (4) Enforcing fairness can improve the performance of predictive models in some situations. (5) **Surprisingly, the predictive models developed by Crab in the presence of limited external data outperform those trained using current methods for learning from selection bias that have access to complete information about the target population.**

This paper makes the following contributions. We develop a new system, Crab for certifying and ensuring fairness in learning predictive models in the presence of selection bias. We presents background on fairness, causality, and selection bias (Section 2.2). We establish sufficient and necessary conditions under which fairness leads to unfair predictive models (Section 3). We introduce and study the CRA problem of aggregate queries, which aim to approximate answers from biased data, and we establish conditions under which it is possible to train certifiably fair ML models with varying access to external information about the target population (Section 4). Finally, we provide experimental evidence that Crab outperforms the state-of-the-art on methods for training fair ML models and learning from biased data (Section 5). An extended version of this paper including the missing proofs and additional experiments can be found at [2].

2 PRELIMINARIES AND BACKGROUND

2.1 Background on ML and Fairness

We now review the background on ML, causality, and selection bias. Table 1 shows the notation we use.

In this work, we focus on the problem of *binary classification*, which has been the primary focus of the literature on algorithmic fairness. Consider a population or data distribution Ω with support $\mathbf{X} \times Y$, where $\mathbf{X}$ denotes a set of discrete and continuous features and $\textsc{Dom}(Y) = \{0, 1\}$ represents some binary outcome of interest (aka the *target attribute*). A classifier $h : \textsc{Dom}(\mathbf{X}) \rightarrow \{0, 1\}$ is a

Symbol	Meaning
X, Y, Z	Attributes (features, variables)
$\mathbf{X}, \mathbf{Y}, \mathbf{Z}$	Sets of attributes
x, y, z	Attribute values
$\text{DOM}(X)$	Domain of an attribute
$x \in \text{DOM}(X)$	An attribute's value
G	A causal diagram
D_{ts}, D_{tr}	Test/train datasets
$MB(Y)$	Markov Boundary of Y
$h(\mathbf{x})$	A classifier

Table 1: Notation used in this paper.

function that predicts the unknown label y as a function of observable features $\mathbf{x}$. The quality of a classifier h can be measured using the *expected risk*, i.e., $Risk(h) = \mathbb{E}_{\Omega}[L(h(\mathbf{x}), y)]$, where $L(h(\mathbf{x}), y)$ is a *loss function* that measures the cost of predicting $h(\mathbf{x})$ when the true value is y. In this paper, we focus on the *zero-one loss*, i.e., $L(h(\mathbf{x}), y) = \mathbb{1}(h(\mathbf{x}) \neq y)$. A learning algorithm aims to find a classifier $h^* \in \mathcal{H}$ that has a minimum risk, i.e., for all classifiers $h \in \mathcal{H}$, it holds that $Risk(h^*) \leq Risk(h)$, where $\mathcal{H}$ denotes the hypothesis space. In the case of the zero-one loss, the optimal classifier h^* is called the *Bayes optimal classifier* and is given by

$$h^*(\mathbf{x}) = \begin{Bmatrix} 1, & \text{if } P(y = 1 \mid \mathbf{x}) \geq \frac{1}{2} \\ 0, & \text{otherwise} \end{Bmatrix}.$$

Since the population distribution Ω is unknown, we cannot calculate $Risk(.)$ directly. Thus, given an i.i.d. training sample $D_{tr} = \{(\mathbf{x}_i, y_i)\}_{i=1}^n$ from Ω, the *empirical risk* $\frac{1}{n}\sum_{i=1}^n L(h(\mathbf{x}), y)$ is typically used to estimate the expected risk. However, minimizing the empirical risk for the zero-one loss function is NP-hard due to its non-convexity. As a result, a convex surrogate loss function is used by learning algorithms. For instance, SVM [21] uses hinge loss and logistic regression; [22] uses logistic loss function as surrogate. A surrogate loss function is said to be *Bayes-risk consistent* if its corresponding empirical minimizer converges to the Bayes optimal classifier when the size of the training data n is sufficiently large [8]. It has been shown that the commonly used surrogate loss functions, such as logistic loss, hinge loss, and exponential loss functions, are Bayes consistent [8]. Throughout this paper, unless otherwise stated, we assume use of learning algorithms that employ a Bayes-consistent loss function.

Relevant features and Markov Boundary. Following [42], we say a feature $X \in \mathbf{X}$ is *strongly relevant* to a classification task if removal of X from $\mathbf{X}$ leads to performance deterioration of the optimal Bayes classifier on $\mathbf{X} \in \mathbf{X}$. A feature X is *weakly relevant* if there exists a subset of features $\mathbf{Z} \subseteq \mathbf{X}$ such that the performance of the Bayes classifier on $\mathbf{Z}$ is worse than the performance on $\mathbf{Z} \cup \{X\}$. A feature is *irrelevant* if it is neither strongly nor weakly relevant. It can be shown that a variable X is relevant for a classification task iff X is in the *Markov Boundary* of Y, denoted $MB(Y)$, which consists of a set of variables $MB(Y) \subseteq \boldsymbol{X}$ for which it holds that $(Y \perp\!\!\!\perp \boldsymbol{X} \backslash MB(Y) \mid MB(Y))$ [51].

Algorithmic fairness. Consider a binary classifier h with a *protected* (sensitive) attribute $S \in \mathbf{X}$, such as gender or race. Without loss of generality, we interpret $h(.) = 1$ as a *favorable* (positive) prediction and $h(.) = 0$ as an *unfavorable* (negative) prediction. To simplify the exposition, we assume $DOM(s) = \{s_0, s_1\}$, where s_1 indicates a *privileged* and s_0 indicates a *protected* group (e.g., males and non-males, respectively). Algorithmic fairness aims to ensure that the classifier h makes fair predictions devoid of discrimination w.r.t. the protected attribute(s). In this paper, we focus on three of the most widely used associational notions of fairness, as follows.
Statistical parity [26] requires an algorithm to classify both the protected and privileged groups with the same probability:

$$\Pr_{\Omega}(h(\mathbf{x}) = 1 \mid s_0) = \Pr_{\Omega}(h(\mathbf{x}) = 1 \mid s_1).$$

Equal opportunity [33] requires both the protected and privileged groups to have the same true positive rate:

$$\Pr_{\Omega}(h(\mathbf{x}) = 1 \mid s_0, Y = 1) = \Pr_{\Omega}(h(\mathbf{x}) = 1 \mid s_1, Y = 1).$$

Conditional statistical parity [19] Given a set of admissible attributes $\mathbf{A}$, the effect of S on h through $\mathbf{A}$ is deemed to be fair if the model classifies both the protected and privileged groups with the same probability condition on every $\mathbf{a} \in \text{DOM}(\mathbf{A})$:

$$\Pr_{\Omega}(h(\mathbf{x}) = 1 \mid s_0, \mathbf{A} = \mathbf{a}) = \Pr_{\Omega}(h(\mathbf{x}) = 1 \mid s_1, \mathbf{A} = \mathbf{a}).$$

Here, we focus on conditional statistical parity since it generalizes both statistical parity ($\mathbf{A} = \emptyset$ implies statistical parity) and equality of opportunity ($\mathbf{A} = Y$ subsumes equality of opportunity). Note that the methods developed in this paper can be extended and adapted to other notions of fairness, e.g., causal notions [60].

2.2 Background on Causality

Causal diagrams. A *causal diagram* is a directed graph that represents the causal relationships between a collection of *observed* or *unobserved* (latent) variables $\boldsymbol{X}$ and models the underlying process that generated the observed data. Each node in a causal diagram corresponds to a variable $X \in \boldsymbol{X}$, and an edge between two nodes indicates a potential causal relationship between the two variables. A variable that has no parent is *exogenous* external and is determined only by forces outside the model; otherwise, it is *endogenous* or internal. The set of all parents of a variable X is denoted by $\mathbf{Pa}(X)$. We use bidirectional edges to represent the assumption that two variables share ancestors that are not shown in the graph; $A \leftrightarrow B$ then means that there is an unobserved variable U with directed paths to both A and B (e.g., $A \leftarrow U \rightarrow B$).
***d*-separation and collider bias.** *Causal diagrams encode a set of conditional independences* that can be read off the graph using *d-separation* [57]. A *path* is a sequence of adjacent arcs, e.g., $(R \rightarrow N \leftarrow \mathbf{X} \rightarrow Y)$ in G. Two nodes are *d-separated* by a set of variables $\mathbf{Z}$, denoted $(V_l \perp\!\!\!\perp V_r \mid_d \mathbf{Z})$ if for every path between them, *one* of the following conditions holds: (1) the path contains a *chain* $(V_l \rightarrow V_m \rightarrow V_r)$ or a *fork* $(V_l \leftarrow V_m \rightarrow V_r)$ such that $V_m \in \mathbf{Z}$, and (2) the path contains a *collider* $(V_l \rightarrow V_m \leftarrow V_r)$ such that $V_m \notin \mathbf{Z}$, and no descendants of V_m are on $\mathbf{Z}$. For example, Y and R are *d-separated* by N and $\mathbf{X}$ in G. Key to d-separation is that conditioning on a collider (common effect) can induce a spurious correlation between its parents (causes), a phenomenon known as *collider bias* [57].

Conditional Independence. A population distribution Ω is said to be *Markov compatible*, or simply compatible, with a causal diagram G if d-separation over the G implies conditional independence with respect to Ω. More formally, $(\mathbf{X} \perp\!\!\!\perp_d \mathbf{Y} \mid \mathbf{Z}) \implies (\mathbf{X} \perp\!\!\!\perp_\Omega \mathbf{Y} \mid \mathbf{Z})$, where $(\mathbf{X} \perp\!\!\!\perp_\Omega \mathbf{Y} \mid \mathbf{Z})$ means $\mathbf{X}$ is independent of $\mathbf{Y}$ conditioned on $\mathbf{Z}$ in Ω. If the converse also holds, i.e., $(\mathbf{X} \perp\!\!\!\perp_\Omega \mathbf{Y} \mid \mathbf{Z}) \implies (\mathbf{X} \perp\!\!\!\perp \mathbf{Y} \mid_d \mathbf{Z})$, Ω is considered *faithful* to G.

Markov Boundary and causal diagrams. Given a population distribution Ω compatible with a causal diagram G, it can be shown that the Markov boundary of a variable $MB(Y)$ consists of (1) Y's parents, (2) Y's children, and (3) parents of Y's children in G [57].

Example 2.1. Figure 1a shows a simplified causal model for predictive policing (Example 1.1) with variables Y: drug use; $\mathbf{W}$: variables such as income, education, job that are deemed to causally affect drug use; $\mathbf{Z}$: sociocultural traits, zip code, and race. Suppose Ω is a population distribution compatible with this causal diagram. Then, it holds that $(Race \perp\!\!\!\perp_\Omega Y)$ because *Race* and Y are d-separated by an empty set since the path $(Race \rightarrow Z \leftarrow ZIPCode \rightarrow X \rightarrow Y)$ is closed at a collider node Z.

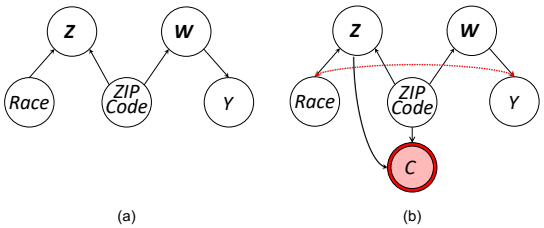

Figure 1: (a) A causal diagram for predictive policing in Examples 1.1 and 2.1. (b) Selection bias in police data in Example 2.2.

2.3 Selection Bias and Causal Diagrams

Selection bias occurs when the selection of a data point in a sample D from a population distribution Ω depends upon the attributes of the data point. In other words, the sample D is collected in a *biased process* in such a way that proper randomization is not achieved, and thereby it is not an i.i.d. sample of Ω. A principled way to model and analyze selection bias is to use causal diagrams [34, 56], where we augment the underlying causal diagram G corresponding to the population distribution Ω with a binary *selection variable* C such that $C = 1$ if a data point is selected into D, and $C = 0$ otherwise. Based on this, we use data collection diagrams to reason about selection bias, which we define next.

Data collection diagrams. Given a population distribution Ω compatible with causal diagram G, a *biased data collection process* (in which the selection of data points into a sample D from Ω depends on a set of variables $\mathbf{Z} \subseteq \boldsymbol{X}$) can be modeled using a *data collection diagram* G^c obtained by augmenting G with a selection node C in which $\mathbf{Z}$ constitutes the parents of C. In this case, D can be seen as an i.i.d sample from a *biased population distribution* Δ compatible with G^c such that $\Pr_\Delta(\mathbf{x}, y) = \Pr_\Omega(\mathbf{x}, y \mid C = 1)$. Furthermore, the biased population d distribution Δ obtained by conditioning on C may exhibit a spurious correlation between variables for which C is a collider in a path between them. We illustrate with an example.

Example 2.2. (Example 2.2 continued) Figure 1b shows a data collection diagram for the police data in Examples 2.2 and 1.1 that captures a bias in which the selection of data points into a sample depends on the neighborhood that is more regularly patrolled and individuals' sociocultural traits, indicated by arrows from the $\mathbf{Z}$ and *ZIPCode* to the selection variables C. A sample of collected D according to this biased data collection process can be seen as a sample from a biased population distribution $\Pr_\Delta(\boldsymbol{X}) = \Pr_\Omega(\boldsymbol{X} \mid C = 1)$, where $\boldsymbol{X} = \{X, Z, Y, Race, ZIPCode\}$. Furthermore, since C is a collider in the path $(Race \rightarrow Z \rightarrow C \leftarrow ZIPCode \rightarrow X \rightarrow Y)$ between *Race* and Y, conditioning on the selection variable C induces a spurious correlation between *Race* and Y that does not exist in the target population distribution Ω but exists in the biased data distribution Δ, due to the collider bias; indicated by the red bidirectional arrow between them *Race* and Y in the diagram. Specifically, while in the target population distribution Ω it holds that $(Race \perp\!\!\!\perp_\Omega Y)$, in the biased population distribution Δ $(Race \not\perp\!\!\!\perp_\Delta Y)$. Hence, a sample D from Δ may exhibit spurious correlations between *Race* and Y that may lead to spurious conclusions or unfair predictive models trained on D.

Remark. Selection bias in training data may or may not lead to an unfair model (see Section 3). Throughout this paper, we use the term *bias* to refer to selection bias in data and the term *unfairness* to refer to an unfair predictive model.

3 FAIRNESS AND SELECTION BIAS

We now analyze the effects of selection bias and identify scenarios where it can lead to unfairness of the trained classifier. Further, we show that any pre-processing bias mitigation strategy cannot ensure fairness if no external information about selection bias is available. To decouple the discrimination introduced by selection bias from other sources of bias – such as a biased data generative process, bias due to finite data, or bias due to the ML model itself – we first define a fair data generative process that does not suffer from any biases. In this formalization, we introduce selection bias to isolate its impact on the fairness of the trained classifier.

Notation. We denote a training dataset D^c_{tr} collected under selection bias $\Pr_\Omega(\mathbf{x}, y \mid C = 1) = \Pr_\Delta(\mathbf{x}, y)$ with an associated data collection diagram G^c (cf. Section 2.3). Our goal is to identify conditions under which a classifier h trained on D^c_{tr} to predict the class label Y violates conditional statistical parity with respect to a set of admissible variables $\mathbf{A}$ on an unbiased test data D_{ts}, which is a representative sample of the target population distribution Ω. We define the following notation to decouple discrimination due to selection bias from any other sources of bias.

Definition 3.1 (Fair classifier). *A classifier h trained on a training data D^c_{tr} sampled under selection bias from a biased population distribution Δ to predict the class label Y is fair if it satisfies conditional statistical parity on an (unseen) test data D_{ts} that is a representative sample of the target population distribution Ω.*

Using this notion, we define a *fair data distribution* as follows.

Definition 3.2 (Fair data distribution). *A data distribution compatible with a causal diagram G is* fair *for learning a classifier to predict the class label Y if any Bayes-consistent ML model trained on a sufficiently large sample of the target population distribution Ω learns a fair classifier h. Otherwise, we say that the data distribution is* unfair *for learning Y.*

Our goal is to identify graphical conditions on the data collection diagram G_c under which the trained classifier leads to unfair classification for the data generative process. Before we introduce our graphical condition, we prove the following proposition that shows necessary and sufficient conditions for fairness.

Proposition 3.1. *A data distribution faithful with a causal diagram G is unfair for learning a classifier that predicts class labels Y iff there exists an attribute $X \in \mathbf{X} \setminus \mathbf{A}$ such that the following conditions C1–C2 hold.*

- *C1 There is an open path from S to X in G after conditioning on the admissible attributes $\mathbf{A}$, i.e., $X \not\perp\!\!\!\perp_G S \mid \mathbf{A}$.*
- *C2 The attribute X is strongly relevant to Y.*

Proof. First we show that if $h(\mathbf{x})$ violates conditional statistical parity then $\exists X \in \mathbf{X}$ such that conditions C1–C2 hold. We consider the contrapositive: $\forall X \in \mathbf{X}$, either condition C1 or C2 do not hold, then we show that the classifier h is fair with respect to conditional statistical parity. More formally, if conditions $\neg C_1 \vee \neg C2$ hold, then $\Pr_\Omega[h(\mathbf{x}) = 1 \mid s_1, \mathbf{a}] = \Pr_\Omega[h(\mathbf{x}) = 1 \mid s_0, \mathbf{a}], \forall \mathbf{a} \in \textsc{Dom}(\mathbf{A})$.

Let $\mathbf{X_1}$ denote the set of attributes that are independent of S given the admissible variables in the underlying distribution, i.e., $X_i \perp\!\!\!\perp_G S \mid \mathbf{A}$ for all $X_i \in \mathbf{X_1}$ (which is equivalent to the set of attributes satisfying $\neg$ C1) and $\mathbf{X_2}$ correspond to the attributes that are dependent on S conditional on the admissible variables but the attribute is not relevant to the classifier, i.e., $(C1 \wedge \neg$ C2).

Since, $\neg(C1 \wedge C2) \equiv (\neg C1) \vee (C1 \wedge \neg C2)$, we know that all attributes belong $X_i \in \mathbf{X_1} \cup \mathbf{X_2}$. The attributes in $\mathbf{X_2}$ are irrelevant to the classifier h trained over the biased dataset. Therefore, the set of relevant attributes is a subset of $\mathbf{X_1}$, implying all relevant attributes (say $\mathbf{X_R} \subseteq \mathbf{X_1}$) are conditionally independent of S in the causal graph G. Since Ω is Markov compatible with G, $\mathbf{X_1} \perp\!\!\!\perp_\Omega S \mid \mathbf{A}$. Therefore, $\Pr_\Omega[h(\mathbf{x}) = 1 \mid s_1, \mathbf{a}] = \Pr_\Omega[h(\mathbf{x}) = 1 \mid s_0, \mathbf{a}] = \Pr_\Omega[h(\mathbf{x}) = 1 \mid \mathbf{a}]$.

To show the converse, i.e, proving that if conditions C1 and C2 hold, $h(\mathbf{x})$ will violate conditional statistical parity. Let $\mathbf{X}_1$ denote the features satisfying C1 and C2, then $\mathbf{X}_1 \subseteq \mathbf{X}_R$, the set of relevant attributes. Therefore, $\mathbf{X}_R \not\perp\!\!\!\perp_G S \mid \mathbf{A}$ implying that the classifier $h(\mathbf{x})$ does not satisfy conditional statistical parity (using faithfulness). □

Intuitively, the proposition states that if the protected attribute S is not in the Markov Boundary of Y and is not correlated with any attributes X in the Markov Boundary of Y after conditioning on $\mathbf{A}$, then collecting sufficiently large amounts of training data guarantees any Bayes-consistent ML model learns a fair classifier h (see Example 3.3 below). We now establish a graphical criterion on the data collection diagram G^c such that training a classifier on biased data leads to an unfair classifier.

Proposition 3.2. *A data collection process faithful with a data collection diagram G^c is unfair for learning a classifier that predicts class labels Y iff either (1) the original data generative process compatible with G^c is unfair for learning a classifier, or (2) the following conditions C1 to C2 hold.*

- *C1 The selection variable $Y \in \mathbf{Pa}(C)$.*
- *C2 The selection variable C is either a child of the protected attribute S* or *there exists a variable $X \in \mathbf{X} \setminus \mathbf{A}$ such that the selection variable C is a child of X and there is an open path between X and S that is not closed after conditioning on $\mathbf{A}$.*

Proof. We begin by proving that if the case (a) or (b) hold, the classifier h would violate conditional statistical parity. If case (a) holds, the data generation process compatible with G is originally unfair. Therefore,we will have: $\exists X_i \in \boldsymbol{X} \setminus \mathbf{A}$ such that $X_i \not\perp\!\!\!\perp_G S \mid \mathbf{A}$ and $X_i \in MB(Y)$ in G (Proposition 3.1). As introducing a collider C will not remove any variable out of $MB(Y)$, we have: $X_i \in MB(Y)$ in G^c, which means X_i will be used by h for prediction. Therefore, $h(\mathbf{x}) \not\perp\!\!\!\perp_G S \mid \mathbf{A}$, the classifier h is unfair on the test data collected based the process compatible with G. If case (b) holds, based on the definition of Markov boundary, $\exists X_i \in MB(Y)$ in G^c, and for this X_i we have $X_i \not\perp\!\!\!\perp S \mid \mathbf{A}$, then h will be unfair according to conditional statistical parity.

We now prove that if h violates conditional statistical parity, then either the original data collection process, which is compatible with G, is unfair (case (a)), or the conditions C1–C2 must hold (case (b)). Since the classifier h is Bayes-risk consistent, it is guaranteed to converge to a Bayes optimal classifier in the presence of sufficient amounts of data. All attributes that are used by h, are in $MB(Y)$ in the selection graph G^c and all other attributes are not used by the learned classifier h.

Since h trained on D_{tr} (compatible with G^c) violates conditional statistical parity on D_{ts} (compatible with G), according to Proposition 3.1, there must exist an attribute $X_i \in \boldsymbol{X} \setminus \mathbf{A}$ (can be S), that is used by h and is dependent on S after conditioning on $\mathbf{A}$ in G. More formally, $\exists X_i \in MB(Y)$ in the graph G, such that $X_i \not\perp\!\!\!\perp_G S \mid \mathbf{A}$.

As introducing a collider C can expand $MB(Y)$ while having Y as a parent, there are two possible cases exist in the original data generation diagram G: $X_i \in MB(Y)$ or $X_i \notin MB(Y)$. If $X_i \in MB(Y)$ in G, h will violate conditional statistical parity before the introducing C, and this corresponds to the satisfaction of case (a) according to Proposition 3.1. And if $X_i \notin MB(Y)$, i.e. X_i is added to $MB(Y)$ after introducing C in G^c, X_i cannot be a direct child or parent of Y because C is a collider. That means X_i can only be the parent of some child of Y in G^c (condition C1). And since $X_i \notin MB(Y)$ in G, X_i will be added to $MB(Y)$ only if it is a parent of new child of Y, which is C. Since X_i has the property: $X_i \not\perp\!\!\!\perp_G S \mid \mathbf{A}$, condition C2 is also satisfied. Therefore we proved the classifier's violation of conditional statistical parity will correspond to either case (a) or (b). □

Example 3.3. We illustrate Proposition 3.2 with the data collection diagrams in Figure 2. The underlying causal diagram G that generates the test dataset D_{ts} is the same for all four sub-figures. The selection node C represents the selection procedure and is decided based on its parents. Each figure represents a different selection mechanism. We can see that since $MB_G(Y) = \{X_3, X_4\}$, and since a Bayes optimal classifier h make predictions based only on $MB(Y)$, and there is no open path from $MB(Y)$ to S, G is a fair process. The training dataset will be a sample of the generative process that is affected by selection bias.

When the set of admissible variables is empty ($\mathbf{A} = \emptyset$), Figure 2c is the only case that would introduce unfairness; this corresponds to the condition explained in Proposition 3.2, which we proved would introduce unfairness to an otherwise fair model. Specifically,

as G is a fair process, case (1) in Proposition 3.2 is not satisfied for all graphs. As for conditions in case (2), only Figure 2c and Figure 2d satisfy C1. Between these two structures, in Figure 2d, C has only 1 parent; thus, it is impossible to satisfy C2, which requires another parent $X \in \boldsymbol{X} \setminus \mathbf{A}$. In Figure 2c, we can let $X = X_2$, which becomes the child of Y after selection (C1) and is dependent on S in the original data generation process conditioning on $\mathbf{A} = \emptyset$ (C2). In the experiments, we will empirically show that Figure 2c is the only selection procedure that would introduce unfairness to this data generative process.

Remark. Proposition 3.2 requires that the data collection process is faithful with the data collection diagram G^c for proving that the conditions (1) or (2) imply unfairness of the classifier. If faithfulness is violated, unfairness of the classifier would still imply that either (1) or (2) holds. If the original data distribution is unfair, then selection bias may or may not exacerbate unfairness.

Thus far, we have discussed how selection bias can introduce unfairness. We now show that bias mitigation algorithms would be unable to perform their role if no external knowledge about the existence and effects of selection bias were available. This is because the perfect bias mitigation algorithm is the one algorithm among all algorithms that perfectly satisfies the fairness constraint/metric on the training data, which minimizes the loss (has the closest loss to the Bayes optimal classifier). Thus, if the training dataset already satisfies the fairness constraint, the perfect bias mitigation algorithm would make no changes and would simply be equal to the Bayes optimal classifier. This approach to fairness would not work if the training set became fair because of selection bias but was not so in the overall population. Our theorem follows.

Theorem 3.4. *Without external information, no fairness-aware ML algorithm trained on a biased training dataset can guarantee fairness over the unbiased test dataset.*

Proof. As proof, we construct a counterexample where no fair ML algorithm would ensure fairness. Consider the dataset shown in Figure 3, where the selection criterion chooses green records. In this sample, the protected attribute S is independent of the target variable Y. Therefore, the bias mitigation algorithm would not perform any pre-processing and simply train the best classifier. In this case, the classifier would be $h(\mathbf{x}) = X_1$. However, the dataset has a property that the data points that are not selected ($C = 0$) have a strong correlation between S and Y. Therefore, the trained classifier is highly unfair on the entire dataset. A method that has no access to external information is not aware of this property of the unselected samples, which causes unfairness of the trained classifier. In general, the unselected dataset can largely differ from the selected portion, which defeats the purpose of training a fair classifier on a biased dataset without any external information. □

In this result, we showed how selection bias can ensure that any fair ML technique cannot ensure fairness without external information. Intuitively, conditional statistical parity is ensured only when $h(\mathbf{x})$ is independent of S, conditioned on $\mathbf{A}$. Conditional independence is not closed under subset partitioning, i.e., conditional independence over a set does not guarantee conditional independence over its subsets. Since data collected under selection bias can be seen as a subset of an unbiased sample of the target population, conditional independence over the biased training set does not translate to conditional independence over the unbiased set.

4 FAIR CLASSIFICATION UNDER SELECTION BIAS USING CRAB

In the previous section, we showed that selection bias in training data can lead to a classifier that is fair during training but unfair in the target population. We now study whether it is possible to train a classifier from biased training data that is certifiably fair on the target population. To do so, we introduce and formalize the problem of consistent range approximation (CRA), which aims to approximate and bound the answers to queries about aggregate information for a target population from biased data (Section 4). Then, we study the problem of CRA of the fairness of a classifier on a target population in settings for which limited or no external information about the target population is available during classifier testing and training (Section 4.2-4.4). Finally, we show that CRA enables training of a classifier from biased data that is certifiably fair on the target population (Section 4.5).

4.1 Consistent Range Approximation for Query Answering from Biased Data

Given a target population Ω, let $\boldsymbol{Q}$ be a query about the population aggregate information. To compute the query answers $\boldsymbol{Q}(\Omega)$, suppose we have access only to a biased sample D^c that is collected from Ω under selection bias with an data collection diagram G^c. Hence, D^c can be seen as a sample from a *biased population* Δ. In general, $\boldsymbol{Q}(\Omega)$ and $\boldsymbol{Q}(\Delta)$ could be arbitrarily different; hence, using D^c to answer $\boldsymbol{Q}$ may lead to incorrect and biased insights. Our goal is to approximate $\boldsymbol{Q}(\Omega)$ using samples from Δ, information encoded in the data collection diagram G^c, and *auxiliary external data sources* $\mathcal{I}$, e.g., census data, results of previous studies, open knowledge graphs, and assumptions on population parameters that could potentially reveal information about the target population Ω. Let us denote with $\text{Poss}(\Delta, G^c, \mathcal{I})$ the set of all *possible populations* that are *consistent* with the biased data distribution Δ, the data collection diagram G^c, and the auxiliary data sources $\mathcal{I}$. Intuitively, *Poss*$(\Delta, G^c, \mathcal{I})$ *represents all possible populations from which the biased dataset* D^c *could have been drawn*; hence, the target distribution Ω is included in $\text{Poss}(\Delta, G^c, \mathcal{I})$.

Definition 4.1. *(Consistent Range Approximation (CRA).) Given a sample from a biased data distribution* Δ*, the data collection diagram* G^c*, and auxiliary data sources* $\mathcal{I}$*, we define* consistent range answers *of an aggregate* $\boldsymbol{Q}$ *as the interval* $[glb(\boldsymbol{Q}), lub(\boldsymbol{Q})]$*, where the endpoints are, respectively, the greatest lower bound (glb) and the least upper bound (lub) of the query answer* $\boldsymbol{Q}$ *on the set of consistent populations Poss*$(\Delta, G^c, \mathcal{I})$*. The problem of* consistent range approximation *of* $\boldsymbol{Q}(\Omega)$ *is the problem of computing the consistent range* $[glb(\boldsymbol{Q}), lub(\boldsymbol{Q})]$.

CRA aims find a consistent range for a query answer, such that the true answers are guaranteed to be within the range in situations where true query answers cannot be computed due to incomplete information about the target population. While CRA can be studied for any arbitrary aggregate query from biased data, in this work we are interested in CRA of the following aggregate query,

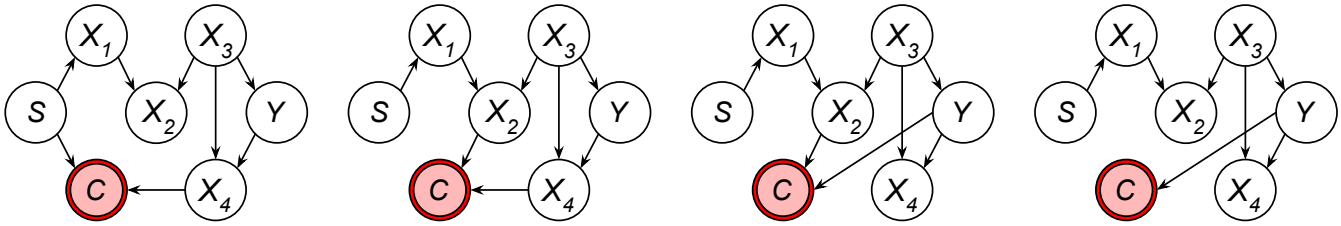

Figure 2: Examples that demonstrate our structural condition.

id	S	X_1	X_2	Y
1	0	0	0	0
2	0	1	0	1
3	1	0	1	0
4	1	1	1	1
5	1	1	0	1
6	1	0	0	1
7	1	1	0	1

$S \perp\!\!\!\perp Y \mid C=1$

Selection criterion: $(S=1 \wedge X_2=1) \vee (X_2=0)$

Figure 3: Example scenario where selection bias breaks the correlation between S and Y. Green rows have $C = 1$.

which quantifies the fairness of a classifier h wrt. conditional statistical parity (cf. Section 1):

$$\mathfrak{f}_{h,\mathbf{A}}(\Omega) = \mathbb{E}_\Omega[h(\mathbf{x}) \mid s_1, \mathbf{a}] - \mathbb{E}_\Omega[h(\mathbf{x}) \mid s_0, \mathbf{a}]. \tag{1}$$

We say that a classifier h is ϵ*−fair* on the target population Ω if $|\mathfrak{f}_{h,\mathbf{A}}(\Omega)| \leq \epsilon$. It is easy to see that if a classifier satisfies conditional statistical parity, then $\mathfrak{f}_{h,\mathbf{A}}(\Omega) = 0$.

The motivation for studying CRA for the query $\mathfrak{f}_{h,\mathbf{A}}(\Omega)$ in Eq. (1) is to enable auditing fairness of a classifier when (1) testing and training data suffer from selection bias, and (2) no or limited information about the target distribution is available during the training. Using the consistent range $[glb(Q_h), lub(Q_h)]$, one can certify that a classifier is ϵ−fair if $|lub(Q_h)| \leq \epsilon$. Furthermore, as we show in Section 4.5, the upper bound $lub(Q_h)$ can be used either as a regularizer or as a constraint to train a classifier that is certifiably ϵ−fairness on the target population.

Remark. While our focus is on *CRA* for the aggregate query $\mathfrak{f}_{h,\mathbf{A}}(\Omega)$ in Eq. 1, we note that for discrete data, $\mathfrak{f}_{h,\mathbf{A}}(\Omega)$ can be estimated via two AVG-GROUP-BY queries. The results established in the subsequent sections can be extended and adapted to Aggregate-GROUP-BY queries in general, which we defer to future work.

4.2 CRA (Presence of Suffient Auxiliary Data)

We now establish the sufficient conditions under which $\mathfrak{f}_{h,\mathbf{A}}(\Omega)$ can be estimated from biased data in settings where we have access to external data sources that reveal *sufficient information* (to be defined later) about the target population.

PROPOSITION 4.1. *Consider a data collection process compatible with a data collection diagram G^c and a classifier h. If there exists a set of variables $\mathbf{U} \subseteq \mathbf{X}$ such that $(\mathbf{X} \perp\!\!\!\perp C \mid S, \mathbf{U}, \mathbf{A})$, where for all $u \in Dom(\mathbf{U})$, $\mathbf{a} \in Dom(\mathbf{A})$, and $s \in \{s_0, s_1\}$, $\Pr_\Omega(\mathbf{u} \mid s, \mathbf{a})$ can be computed using auxiliary data sources $\mathcal{I}$, then $\mathfrak{f}_{h,\mathbf{A}}(\Omega)$ can be computed using the following equation:*

$$\mathfrak{f}_{h,\mathbf{A}}(\Omega) = \sum_{u} \mathbb{E}_\Delta[h(\mathbf{x}) \mid s_1, \mathbf{u}, \mathbf{a}] \Pr_\Omega(\mathbf{u} \mid s_1, \mathbf{a}) - \mathbb{E}_\Delta[h(\mathbf{x}) \mid s_0, \mathbf{u}, \mathbf{a}] \Pr_\Omega(\mathbf{u} \mid s_0, \mathbf{a}) \tag{2}$$

PROOF. The following equations are obtained from the law of total expectation, the independence $(\mathbf{X} \perp\!\!\!\perp C \mid S, \mathbf{U}, \mathbf{A})$. For each $s \in s_0, s_1$, it holds that:

$$\mathbb{E}_\Omega[h(\mathbf{x}) \mid s, \mathbf{a}] \tag{3}$$

$$= \sum_{\mathbf{u} \in Dom(\mathbf{U})} \mathbb{E}_\Omega[h(\mathbf{x}) \mid s, \mathbf{u}, \mathbf{a}] \Pr_\Omega(\mathbf{u} \mid s, \mathbf{a}) \tag{4}$$

$$= \sum_{\mathbf{u} \in Dom(\mathbf{U})} \mathbb{E}_\Omega[h(\mathbf{x}) \mid s, \mathbf{u}, \mathbf{a}, C] \Pr_\Omega(\mathbf{u} \mid s, \mathbf{a}) \tag{5}$$

$$= \sum_{\mathbf{u} \in Dom(\mathbf{U})} \mathbb{E}_\Delta[h(\mathbf{x}) \mid s, \mathbf{u}, \mathbf{a}] \Pr_\Omega(\mathbf{u} \mid s, \mathbf{a}) \tag{6}$$

Eq (2) immediately follows from the preceding equations. □

Note that in Eq (2), the expression $\mathbb{E}_\Delta[h(\mathbf{x}) \mid s, \mathbf{u}, \mathbf{a}]$ can be estimated on the biased data, but the auxiliary data source $\mathcal{I}$ should contain sufficient information for computing $\Pr_\Omega(\mathbf{u} \mid s, \mathbf{a})$.

Remark. The independence condition $(\mathbf{X} \perp\!\!\!\perp C \mid S, \mathbf{U}, \mathbf{A})$ is central for applying Proposition 4.1. It is easy to see that to satisfy this condition, it is sufficient to set $\mathbf{U} = \mathbf{Pa}(C)$ since $\mathbf{Pa}(C)$ forms the Markov Boundary for C. Therefore, Proposition 4.1 is applicable in situations where the underlying causal model is unknown and one has access only to information about $\mathbf{pa}(C)$. However, in the presence of the causal diagram, one can select a minimal subset of $\mathbf{U}$ that satisfies the independence condition and for which external information in $\mathcal{I}$ is available. Furthermore, as we show next, even if $\mathbf{Pa}C$ includes the training label Y, it is still possible to find a set of variable U that does not contain Y and satisfies the condition $(\mathbf{X} \perp\!\!\!\perp C \mid S, \mathbf{U}, \mathbf{A})$; thereby, it is still possible to apply Proposition 4.1 even if no information about Y is available in $\mathcal{I}$, which is often the case in practice. In both real-world scenarios in Example 1.1 and 1.2 these assumptions are feasible and can ensured without access to the underlying causal diagram.

PROPOSITION 4.2. *Given a data collection diagram G^c, there always exists a set of variables $\mathbf{U} \subseteq \mathbf{Pa}(C) \cup MB(Y) \setminus \{C, Y\}$ for which the independence $(\mathbf{X} \perp\!\!\!\perp C \mid \mathbf{U}, S, \mathbf{A})$ holds. (Recall from Section 1 that $\mathbf{Pa}(C)$ and $MB(Y)$ denote the parents of C and the Markov Boundary of Y in G^c, respectively.)*

Example 4.3. Consider the data collection diagram in Figure 2b. The independence $(\boldsymbol{X} \perp\!\!\!\perp C \mid S, \boldsymbol{U}, \mathbf{A})$ holds for $\boldsymbol{U} = \mathbf{pa}(C) = \{X_2, X_4\}$. Hence, $\mathfrak{f}_{h,\mathbf{A}}(\Omega)$ can be estimated from a biased sample collected according to the diagram in Figure 2b and from access to auxiliary information to estimate probabilities of the form $\Pr_\Omega(\mathbf{u} \mid \mathbf{s}, \mathbf{a})$. Similarly, in Figure 2b, the independence condition holds for $\boldsymbol{U} = \mathbf{pa}(C) = \{X_2, Y\}$; however, if no external information about the training labels Y is available, one can choose $\mathbf{U} = \{X_2, X_3\}$, for which the independence condition also holds. Hence, $\mathfrak{f}_{h,\mathbf{A}}(\Omega)$ can be computed via Proposition 4.1.

4.3 CRA (Absence of Auxiliary Data Sources)

We next consider the case where no auxiliary external data sources are available to compute $\mathfrak{f}_{h,\mathbf{A}}(\Omega)$. In such cases, we prove a sufficient condition under which one can estimate a non-trivial consistent range $\mathfrak{f}_{h,\mathbf{A}}(\Omega)$ that could be estimated merely using the biased data. Since in the context of algorithmic fairness we are interested only in worst case analysis, from now on, we focus on establishing an upper bound on $\mathfrak{f}_{h,\mathbf{A}}(\Omega)$. Before we proceed, we show a simple sufficient condition under which $\mathfrak{f}_{h,\mathbf{A}}(\Omega) = \mathfrak{f}_{h,\mathbf{A}}(\Delta)$, i.e., the unfairness of a classifier on the target population can be accurately quantified from biased data.

Proposition 4.4. *Given a data collection diagram G^c and a classifier h, if $(\boldsymbol{X} \perp\!\!\!\perp C \mid \mathbf{A}, S)$ holds in G^c, then $\mathfrak{f}_{h,\mathbf{A}}(\Omega) = \mathfrak{f}_{h,\mathbf{A}}(\Delta)$.*

Proof. It is immediately implied by the independence $(\boldsymbol{X} \perp\!\!\!\perp C \mid \mathbf{A}, S)$ in the following steps.

$$\begin{aligned}
\mathfrak{f}_{h,\mathbf{A}}(\Omega) &= \mathbb{E}_\Omega[h(\mathbf{x}) \mid s_1, \mathbf{a}] - \mathbb{E}_\Omega[h(\mathbf{x}) \mid s_0, \mathbf{a}] \\
&= \mathbb{E}_\Omega[h(\mathbf{x}) \mid s_1, \mathbf{a}, C = 1] - \mathbb{E}_\Omega[h(\mathbf{x}) \mid s_0, \mathbf{a}, C = 1] \\
&= \mathbb{E}_\Delta[h(\mathbf{x}) \mid s_1, \mathbf{a}] - \mathbb{E}_\Delta[h(\mathbf{x}) \mid s_0, \mathbf{a}] \\
&= \mathfrak{f}_{h,\mathbf{A}}(\Delta)
\end{aligned}$$

□

The independence condition in Proposition 4.4 does not hold in general, and $\mathfrak{f}_{h,\mathbf{A}}(\Omega) \neq \mathfrak{f}_{h,\mathbf{A}}(\Delta)$. However, we now show that $\mathfrak{f}_{h,\mathbf{A}}(\Omega)$ can be upper bounded by the biased training data.

Proposition 4.5. *Given a data collection diagram G^c and a classifier h, the following holds for any set of variables $\boldsymbol{U} \subseteq \boldsymbol{X}$ such that $(\boldsymbol{X} \perp\!\!\!\perp C \mid \boldsymbol{U}, S, \mathbf{A})$:*

$$\begin{aligned}
\mathfrak{f}_{h,\mathbf{A}}(\Omega) \leq & \max_{\mathbf{u} \in Dom(\mathbf{U})} \mathbb{E}_\Delta[h(\mathbf{x}) \mid s_1, \mathbf{a}, \mathbf{u}] \\
& - \min_{\mathbf{u} \in Dom(\mathbf{U})} \mathbb{E}_\Delta[h(\mathbf{x}) \mid s_0, \mathbf{a}, \mathbf{u}]. \qquad (7)
\end{aligned}$$

Proof. The proposition can be proved by upper bounding the expression $\mathbb{E}_\Omega[h(\mathbf{x}) \mid s_1, \mathbf{a}]$ and lower bounding $\mathbb{E}_\Omega[h(\mathbf{x}) \mid s_0, \mathbf{a}]$ in the definition of $\mathfrak{f}_{h,\mathbf{A}}(\Omega)$ in Eq. 1 using the independence assumption $(\boldsymbol{X} \perp\!\!\!\perp C \mid \boldsymbol{U}, S, \mathbf{A})$ as follows:

$$\begin{aligned}
&\mathbb{E}_\Omega[h(\mathbf{x}) \mid s_1, \mathbf{a}] \\
&= \sum_{\mathbf{u} \in Dom(\mathbf{U})} \mathbb{E}_\Omega[h(\mathbf{x}) \mid s_1, \mathbf{u}, \mathbf{a}] \Pr_\Omega(\mathbf{u} \mid s_1, \mathbf{a}) \\
&= \sum_{\mathbf{u} \in Dom(\mathbf{U})} \mathbb{E}_\Omega[h(\mathbf{x}) \mid s_1, \mathbf{u}, \mathbf{a}, C = 1] \Pr_\Omega(\mathbf{u} \mid s_1, \mathbf{a}) \\
&\leq \sum_{\mathbf{u} \in Dom(\mathbf{U})} \Big(\max_{\mathbf{u}^* \in Dom(\mathbf{U})} \mathbb{E}_\Omega[h(\mathbf{x}) \mid s_1, \mathbf{u}^*, \mathbf{a}, C = 1] \Big) \Pr_\Omega(\mathbf{u} \mid s_1, \mathbf{a}) \\
&= \max_{\mathbf{u}^* \in Dom(\mathbf{U})} \mathbb{E}_\Omega[h(\mathbf{x}) \mid s_1, \mathbf{u}^*, \mathbf{a}, C = 1] \sum_{\mathbf{u} \in Dom(\mathbf{U})} \Pr_\Omega(\mathbf{u} \mid s_1, \mathbf{a}) \\
&= \max_{\mathbf{u}^* \in Dom(\mathbf{U})} \mathbb{E}_\Omega[h(\mathbf{x}) \mid s_1, \mathbf{u}^*, \mathbf{a}, C = 1] \\
&= \max_{\mathbf{u}^* \in Dom(\mathbf{U})} \mathbb{E}_\Delta[h(\mathbf{x}) \mid s_1, \mathbf{u}^*, \mathbf{a}] \qquad (8)
\end{aligned}$$

$$\begin{aligned}
&\mathbb{E}_\Omega[h(\mathbf{x}) \mid s_0, \mathbf{a}] \\
&= \sum_{\mathbf{u} \in Dom(\mathbf{U})} \mathbb{E}_\Omega[h(\mathbf{x}) \mid s_0, \mathbf{u}, \mathbf{a}] \Pr_\Omega(\mathbf{u} \mid s_0, \mathbf{a}) \\
&= \sum_{\mathbf{u} \in Dom(\mathbf{U})} \mathbb{E}_\Omega[h(\mathbf{x}) \mid s_0, \mathbf{u}, \mathbf{a}, C = 1] \Pr_\Omega(\mathbf{u} \mid s_0, \mathbf{a}) \\
&\geq \sum_{\mathbf{u} \in Dom(\mathbf{U})} \Big(\min_{\mathbf{u}^* \in Dom(\mathbf{U})} \mathbb{E}_\Omega[h(\mathbf{x}) \mid s_0, \mathbf{u}^*, \mathbf{a}, C = 1] \Big) \Pr_\Omega(\mathbf{u} \mid s_0, \mathbf{a}) \\
&= \min_{\mathbf{u}^* \in Dom(\mathbf{U})} \mathbb{E}_\Omega[h(\mathbf{x}) \mid s_0, \mathbf{u}^*, \mathbf{a}, C = 1] \sum_{\mathbf{u}} \Pr_\Omega(\mathbf{u} \mid s_0, \mathbf{a}) \\
&= \min_{\mathbf{u}^* \in Dom(\mathbf{U})} \mathbb{E}_\Omega[h(\mathbf{x}) \mid s_0, \mathbf{u}^*, \mathbf{a}, C = 1] \\
&= \min_{\mathbf{u}^* \in Dom(\mathbf{U})} \mathbb{E}_\Delta[h(\mathbf{x}) \mid s_0, \mathbf{u}^*, \mathbf{a}] \qquad (9)
\end{aligned}$$

Proposition (4.5) immediately follows from Eq. (1), (8) and (9). □

Notice that the bound in Proposition 4.5 can be computed *merely* using the biased training data.

Remark. The independence assumption $(\boldsymbol{X} \perp\!\!\!\perp C \mid \boldsymbol{U}, S, \mathbf{A})$ in Proposition 4.5 in similar to that in Proposition 4.1. As discussed in Section 4.2, it is always possible to find variables $\mathbf{U}$ that satisfy this assumption, and thereby it is always possible to compute an upper bound on $\mathfrak{f}_{h,\mathbf{A}}(\Omega)$, even when sufficient information for computing the exact value of $\mathfrak{f}_{h,\mathbf{A}}(\Omega)$ is not available.

4.4 CRA (Presence of Limited Auxiliary Data)

Sections 4.2 and Section 4.3 examined different ends of the spectrum; i.e., having access to sufficient auxiliary information and to no external information for CRA of $\mathfrak{f}_{h,\mathbf{A}}(\Omega)$. Here, we investigate the middle of the spectrum, where auxiliary information about the target population is available but it is insufficient for exact computation of $\mathfrak{f}_{h,\mathbf{A}}(\Omega)$ as in Section 4.2. We show that any level of external information can be used to compute a tighter upper bound for $\mathfrak{f}_{h,\mathbf{A}}(\Omega)$ than that established in Proposition 4.5. Specifically, we consider similar assumptions to Proposition 4.1, but situations in which auxiliary data sources have only partial information about the population statistics $\Pr_\Omega(\mathbf{u} \mid s_0, \mathbf{a})$ required to compute $\mathfrak{f}_{h,(}(\Omega)\mathbf{A})$ via Proposition 4.1.

Proposition 4.6. *Consider a data collection process compatible with a data collection diagram G^c and a classifier h. Let $\boldsymbol{U} \subseteq \boldsymbol{X}$ be a subset of variables for which $(\boldsymbol{X} \perp\!\!\!\perp C \mid S, \boldsymbol{U}, \mathbf{A})$. Furthermore, suppose auxiliary data sources $\mathcal{I}$ only admit computation of the population statistics $\Pr_\Omega(\mathbf{u}' \mid s, \mathbf{a})$ for all $u' \in Dom(\mathbf{U}')$ and $s \in \{s_0, s_1\}$ for some subset $\mathbf{U}' \subsetneq \mathbf{U}$. Then, the following bound can be computed for $\mathfrak{f}_{h,\mathbf{A}}(\Omega)$:*

$$\begin{aligned}
&\mathfrak{f}_{h,\mathbf{A}}(\Omega) \\
&\leq \sum_{\mathbf{u}' \in \mathbf{U}'} \Big(\Pr_\Omega(\mathbf{u}' \mid s_1, \mathbf{a}) \max_{\mathbf{u}^* \in Dom(\boldsymbol{U} \setminus \mathbf{U}')} (\mathbb{E}_\Delta[h(\mathbf{x}) \mid s_1, \mathbf{u}', \mathbf{u}^*, \mathbf{a}]) \\
&\quad - \Pr_\Omega(\mathbf{u}' \mid s_0, \mathbf{a}) \min_{\mathbf{u}^* \in Dom(\boldsymbol{U} \setminus \mathbf{U}')} (\mathbb{E}_\Delta[h(\mathbf{x}) \mid s_0, \mathbf{u}', \mathbf{u}^*, \mathbf{a}]) \Big) \quad (10)
\end{aligned}$$

Proof. Let's partition $\mathbf{U}$ into $\mathbf{U}'$ and $\mathbf{U} \setminus \mathbf{U}'$ in Eq. 2 the following inequality obtained from applying the Fréchet inequality to $\Pr_\Omega(\mathbf{u}', \mathbf{u}^* \mid s, \mathbf{a})$:

$$\begin{aligned}
&\mathfrak{f}_{h,\mathbf{A}}(\Omega) \\
&= \sum_{\mathbf{u}' \in \mathbf{U}', \mathbf{u}^* \in Dom(\boldsymbol{U} \setminus \mathbf{U}')} \Big(\mathbb{E}_\Delta[h(\mathbf{x}) \mid s_1, \mathbf{u}', \mathbf{u}^*, \mathbf{a}] \Pr_\Omega(\mathbf{u}', \mathbf{u}^* \mid s_1, \mathbf{a}) \\
&\quad - \mathbb{E}_\Delta[h(\mathbf{x}) \mid s_0, \mathbf{u}', \mathbf{u}^*, \mathbf{a}] \Pr_\Omega(\mathbf{u}', \mathbf{u}^* \mid s_0, \mathbf{a}) \Big) \\
&\leq \sum_{\mathbf{u}' \in \mathbf{U}', \mathbf{u}^* \in Dom(\boldsymbol{U} \setminus \mathbf{U}')} \Big(\mathbb{E}_\Delta[h(\mathbf{x}) \mid s_1, \mathbf{u}', \mathbf{u}^*, \mathbf{a}] \Pr_\Omega(\mathbf{u}' \mid s_1, \mathbf{a}) \\
&\quad - \mathbb{E}_\Delta[h(\mathbf{x}) \mid s_0, \mathbf{u}', \mathbf{u}^*, \mathbf{a}] \Pr_\Omega(\mathbf{u}' \mid s_0, \mathbf{a}) \Big) \quad (11)
\end{aligned}$$

Then we upper bound the expression $\mathbb{E}_\Delta[h(\mathbf{x}) \mid s_1, \mathbf{u}', \mathbf{u}^*, \mathbf{a}]$ and lower bound $\mathbb{E}_\Delta[h(\mathbf{x}) \mid s_0, \mathbf{u}', \mathbf{u}^*, \mathbf{a}]$ in Eq. (11) using the independence $(\boldsymbol{X} \perp\!\!\!\perp C \mid \mathbf{U}, S, \mathbf{A})$ in as similar steps as in the proof of proposition 4.5:

$$\mathbb{E}_\Delta[h(\mathbf{x}) \mid s_1, \mathbf{u}', \mathbf{u}^*, \mathbf{a}] \leq \max_{\mathbf{u}^* \in Dom(\boldsymbol{U} \setminus \mathbf{U}')} (\mathbb{E}_\Delta[h(\mathbf{x}) \mid s_1, \mathbf{u}, \mathbf{a}]) \quad (12)$$

$$\mathbb{E}_\Delta[h(\mathbf{x}) \mid s_0, \mathbf{u}', \mathbf{u}^*, \mathbf{a}] \geq \min_{\mathbf{u}^* \in Dom(\boldsymbol{U} \setminus \mathbf{U}')} (\mathbb{E}_\Delta[h(\mathbf{x}) \mid s_0, \mathbf{u}, \mathbf{a}]) \quad (13)$$

Eq. (10) immediately follows from Eq. (11), (12) and (13). □

Proposition 4.6 computes an upper bound for $\mathfrak{f}_{h,\Omega}(\Omega)$ in the presence of any level of auxiliary information about marginals of the population statistics $\Pr_\Omega(\mathbf{u} \mid s, \mathbf{a})$.

Next, we show that under some assumptions about the data collection process, it is still possible to compute $\mathfrak{f}_{h,\mathbf{A}}(\Omega)$ from biased data even if no auxiliary information about the admissible variables $\mathbf{A}$ is available. A major motivation for considering this setting is a need to deal with fairness definitions based on error rate balance, e.g., equality of odds, in which $\mathbf{A}$ includes training labels Y and for which external information is often unavailable.

Proposition 4.7. *Consider a data collection process compatible with a data collection diagram G^c and a classifier h. If $\mathbf{A} \cap \mathbf{Pa}(C) = \emptyset$ in G^c, i.e., if data selection does not directly depend on the admissible variables $\mathbf{A}$, then for a set of variables $\boldsymbol{U} \subseteq \boldsymbol{X}$ that satisfies the independence $(\mathbf{X} \perp\!\!\!\perp C \mid S, \boldsymbol{U}, \mathbf{A})$, if the population statistics $\Pr_\Omega(s, \mathbf{u})$ for all $\mathbf{u} \in Dom(\mathbf{U})$ and $s \in \{s_0, s_1\}$ can be computed from auxiliary data sources $\mathcal{I}$, then $\mathfrak{f}_{h,\mathbf{A}}(\Omega)$ can be computed as follows:*

$$\begin{aligned}
\mathfrak{f}_{h,\mathbf{A}}(\Omega) = &\sum_{\mathbf{u} \in Dom(\mathbf{U})} \mathbb{E}_\Delta[h(\mathbf{x}) \mid s_1, \mathbf{u}, \mathbf{a}] \; w(\mathbf{u}, s_1, \mathbf{a}) \\
&- \mathbb{E}_\Delta[h(\mathbf{x}) \mid s_0, \mathbf{u}, \mathbf{a}] \; w(\mathbf{u}, s_0, \mathbf{a}), \quad (14)
\end{aligned}$$

where $w(\mathbf{u}, s, \mathbf{a}) = \frac{\sum_{\mathbf{x} \in Dom(\boldsymbol{X})} \Pr_\Delta(\mathbf{a}, \mathbf{x} \mid s, \mathbf{u}) \Pr_\Omega(s, \mathbf{u})}{\sum_{\mathbf{u}^* \in Dom(\mathbf{U})} \sum_{\mathbf{x} \in Dom(\boldsymbol{X})} \Pr_\Delta(\mathbf{a}, \mathbf{x} \mid s, \mathbf{u}^*) \Pr_\Omega(s, \mathbf{u}^*)}$

Proof. Since $\mathbf{A} \cap \mathbf{Pa}(C) = \emptyset$, the independence independence $\mathbf{A}, \boldsymbol{X} \perp\!\!\!\perp C \mid S, \boldsymbol{U}$ holds for $\boldsymbol{U} = \mathbf{Pa}(C)$ such that $\mathbf{A} \cap \boldsymbol{U} = \emptyset$. The following equations are obtained from the above the independence assumption, law of total probability, the Bayes' Theorem and marginalization on $\boldsymbol{X}$. For each $s \in \{s_0, s_1\}$, we can estimate $\Pr_\Omega(\mathbf{u} \mid s, \mathbf{a})$ in Eq. 6 as follows:

$$\begin{aligned}
\Pr_\Omega(\mathbf{u} \mid s, \mathbf{a}) &= \frac{\sum_{x \in \mathrm{Dom}(\boldsymbol{X})} \Pr_\Omega(\mathbf{a}, \mathbf{x} \mid s, \mathbf{u}) \Pr_\Omega(s, \mathbf{u})}{\sum_{x \in \mathrm{Dom}(\boldsymbol{X})} \Pr_\Omega(\mathbf{a}, \mathbf{x} \mid s) \Pr_\Omega(s)} \\
&= \frac{\sum_{x \in \mathrm{Dom}(\boldsymbol{X})} \Pr_\Omega(\mathbf{a}, \mathbf{x} \mid s, \mathbf{u}) \Pr_\Omega(s, \mathbf{u})}{\sum_{u^* \in \mathrm{Dom}(\boldsymbol{U})} \sum_{x \in \mathrm{Dom}(\boldsymbol{X})} \Pr_\Omega(\mathbf{a}, \mathbf{x} \mid s, \mathbf{u}^*) \Pr_\Omega(s, \mathbf{u}^*)} \\
&= \frac{\sum_{x \in \mathrm{Dom}(\boldsymbol{X})} \Pr_\Omega(\mathbf{a}, \mathbf{x} \mid s, \mathbf{u}, C = 1) \Pr_\Omega(s, \mathbf{u})}{\sum_{u^* \in \mathrm{Dom}(\boldsymbol{U})} \sum_{x \in \mathrm{Dom}(\boldsymbol{X})} \Pr_\Omega(\mathbf{a}, \mathbf{x} \mid s, \mathbf{u}^*, C = 1) \Pr_\Omega(s, \mathbf{u}^*)} \\
&= \frac{\sum_{x \in \mathrm{Dom}(\boldsymbol{X})} \Pr_\Delta(\mathbf{a}, \mathbf{x} \mid s, \mathbf{u}) \Pr_\Omega(s, \mathbf{u})}{\sum_{u^* \in \mathrm{Dom}(\boldsymbol{U})} \sum_{x \in \mathrm{Dom}(\boldsymbol{X})} \Pr_\Delta(\mathbf{a}, \mathbf{x} \mid s, \mathbf{u}^*) \Pr_\Omega(s, \mathbf{u}^*)} \quad (15)
\end{aligned}$$

Eq. (4.7) immediately follows from Eq. (1), (3)-(6), and (15). □

4.5 Fair ML with CRA

The results established for CRA of $\mathfrak{f}_{h,\mathbf{A}}(\Omega)$ from biased data can be used to train predictive models from biased data that are certifiably fair on the target population when no or limited information about the target population is available.

We build upon the previous approaches in fair predictive modeling, e.g., [9, 39], that enforce fairness by adding a regularization term to the objective function of a learning algorithm to control for the *degree of unfairness* of the model, and solving the following regularized optimization problem:

$$\operatorname*{argmin}_{h \in \mathcal{H}} \mathbb{E}_\Delta[L(h(\mathbf{x}), y)] + \lambda \; \max\Big(lub\big(\mathfrak{f}_{h,\mathbf{A}}(\Omega)\big), \tau\Big), \quad (16)$$

where $lub(\mathfrak{f}_{h,\mathbf{A}}(\Omega))$ is the consistent upper bound of the degree of fairness of h that can be computed from biased training data via the results established in Section 4.2-4.4, λ is the regularizer coefficient, and τ is a threshold that can be jointly tuned with λ to trade fairness with accuracy. For a differentiable loss function L, a differentiable classification model h parameterized by θ, and a discrete set of admissible variables $\mathbf{A}$, the regularized optimization problem Eq.16 can be solved using standard numerical optimization techniques by *merely* using biased data. This is because the regularization term $lub(\mathfrak{f}_{h,\mathbf{A}}(\Omega), \tau)$ reduced to a linear combination of a set of differentiable equations on the parameters θ.

5 EXPERIMENTS

We now evaluate the impact of selection bias on the fairness of the trained classifier and validate the effectiveness of Crab to ensure fairness while maintaining high prediction quality. All code and input files used to generate the results are available on this at [2].

Table 2: Average runtime in seconds for CRAB and ORIG.

Dataset	Att. [#]	Rows[#]	CRAB-M∅	CRAB-MX	Orig
Adult	8	45k	2.5s	1.8s	0.8s
Law	11	18k	5.7s	3.4s	0.6s
HMDA	8	3.2m	2m30s	16.8s	6.7s
Syn	5	200k	40.0s	31.8s	1.8s

We aim to address the following questions. **Q1**: Can CRAB leverage varied amounts of external data (including the absence of external data) to guarantee fairness when the training dataset suffers from selection bias? How does CRAB compare to the state-of-the-art fair classification methods and current techniques for learning in the presences of selection bias? (Section 5.2). **Q2**: When does selection bias introduce unfairness in scenarios where the unbiased data generative process is fair? Can we improve classifier quality by enforcing fairness? (Section 5.3). **Q3**: How does CRAB adapt to other fairness metrics and classification techniques? (Section 5.4)

5.1 Setup

Datasets and metrics. We consider the following real-world and synthetic datasets in our evaluation: **Adult** [44]: Contains demographic and financial information of individuals. The goal is to predict if the individual's income exceeds $50K, and gender is considered as the sensitive attribute. **Law** [66]: Contains demographics and academic performance of law school students. The target attribute characterizes the outcome of the exam as Pass/Fail, and the protected attribute is race. **HMDA** [1]: Contains US mortgage applications, including information about an individual's financials, the house, their mortgage conditions, etc. The goal is to predict whether the loan is granted or denied, and applicant race is considered as the protected attribute. **Syn**: Contains synthetic data generated based on a fair data collection diagram compatible with the graph in Figure 2 without the selection variable C, where S denotes the protected attribute.

We simulated six different mechanisms to introduce selection bias by adding the selection variable as a child of different sets of attributes and varying the selection probability. Selection bias was introduced in the training and validation sets of each dataset, and the test data remained untouched. All experiments were run five times, and we report the average and standard deviation for each technique to demonstrate variation. We evaluate both statistical parity and equal opportunity as the fairness measures in the experiments. A lower value of $\mathsf{f}_{h,Y}(\Omega)$ and a higher F1 Score denote ideal performance of any technique.

Classifier. We train a logistic regression (LR) Support Vector Machine (SVM) with a linear kernel and neural network (NN) to compare CRAB with various baselines. All techniques were implemented in Python with PyTorch [55]. Unless specified, we used logistic regression as the default choice of the classifier.

Auxiliary data sources. We test CRAB under four settings of access to auxiliary data to evaluate the CRA of $\mathsf{f}_{h,\mathbf{A}}(\Omega)$, presented in Section 4. (1) CRAB-M∅: sufficient external data to evaluate Proposition 4.1. (2) CRAB-MX: no external data to evaluate Proposition 4.5. (3) CRAB-MU: limited external data to evaluate Proposition 4.6. (4) CRAB-MA: no external data about admissible attributes (training label Y in our case) to evaluate Proposition 4.7.

Baselines. We compare CRAB with the following representative baselines: ZEMEL [71] and KAMIRAN [38] are pre-processing methods that modify the training data to obfuscate information about the protected attributes. ZHANG [72] and ZAFAR [68, 70] are in-processing methods that maximize model quality while minimizing fairness violation. CORTES is an inverse propensity score weighting an (IPW) based pre-processing method that utilizes external data to recover from selection bias [20]. This baseline is used only in the presence of external data to estimate selection probability for different subgroups in the dataset.

We used IBM's AI Fairness 360 [10] to run the pre-processing methods. All in-processing baseline methods were implemented from scratch in our framework to facilitate fair comparison.

5.2 Solution Quality and Fairness Comparison

We now evaluate the performance of CRAB and baselines on varying level of access to external data. In all figures, the green background denotes the region that satisfies fairness requirements, and the red region corresponds to the unfair region.

5.2.1 Absence of external data. This experiment compares CRAB-MX with the fair classification baselines. Since CRAB allows the user to specify the fairness requirement, we varied the fairness requirement τ between 0 and θ, where θ is the unfairness of the classifier trained on the biased dataset. We introduce selection node C as a child of S and another variable X_2 (shown in Figure 2a) and vary the selection criterion, $\Pr_\Omega(C = 1 \mid \mathbf{Pa}(C))$, to simulate two scenarios (S1 and S2). Specifically, the biased dataset has much lower unfairness than the test data in the first scenario (S1); therefore, ensuring fairness over the training data while ignoring the selection bias would not guarantee fairness over the test data. The biased dataset is as fair as the unbiased test in the second scenario ($\mathsf{f}_{h,Y}(\Omega) \approxeq \mathsf{f}_{h,Y}(\Delta)$); therefore, enforcing fairness on the biased training data ensures fairness on the unbiased test data in this case.

Solution quality. Figure 4 compares the fairness of the trained classifier and F1 Score. We observe that CRAB-MX learns a fair classifier (EO < 0.05 for $\tau = 0$ corresponding to the point with lowest y-coordinate) for all datasets across both scenarios. In fact, CRAB-MX achieves perfect fairness while incurring a very low loss in F1 Score for the Law and HMDA datasets. In contrast, most of the baseline methods fail to completely remove the model's discrimination in most of the cases. In certain cases, baseline techniques like ZEMEL improve fairness (HMDA-G1-S1), but the same technique returns a highly unfair classifier for HMDA-G1-S2. This is because the selection mechanism creates a false sense of fairness on training data ($\mathsf{f}_{h,Y}(\Delta) \approx 0$), while $\mathsf{f}_{h,Y}(\Omega)$ remains high. This is the reason that most of the baselines achieve worse fairness for the first scenario compared to the second one. It shows that the two scenarios behave differently even though the selection bias is a function of the same set of attributes (causal diagram does not change), with the only difference being the selection probabilities. However, CRAB-MX upper bounds $\mathsf{f}_{h,}(\Omega)$, which helps to enforce fairness across both settings for all datasets.

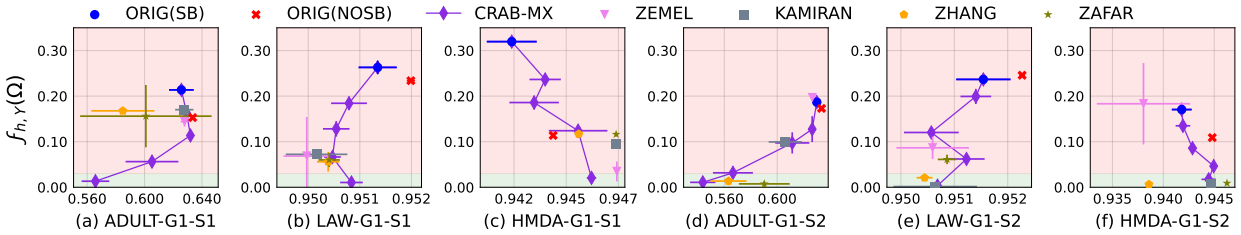

Figure 4: Equal opportunity (y-axis)-F1 Score (x-axis) comparison for Crab-MX and baseline methods in the absence of external data. (a to c) correspond to scenario 1, where $f_{h,Y}(\Delta)$ largely deviates from $f_{h,Y}(\Omega)$; (d to f) correspond to scenario 2, where $f_{h,Y}(\Delta) \approxeq f_{h,Y}(\Omega)$. Selection variable C is placed as a child of S and another attribute $X \in \mathbf{X}$ (same as Figure 2a).

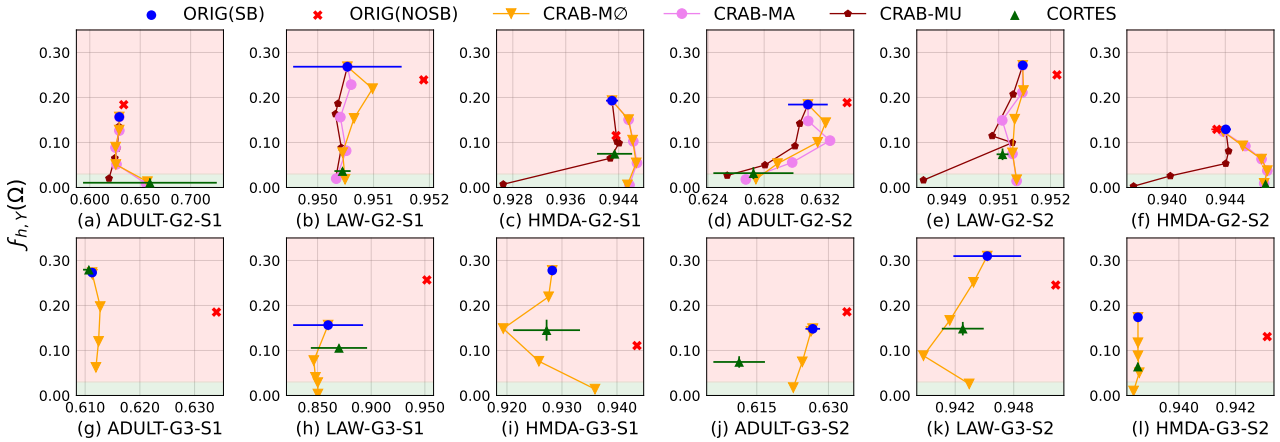

Figure 5: Equal Opportunity (y-axis)-F1 Score (x-axis) comparison for Crab-MØ, Crab-MU, Crab-MA, and Cortes given the existence of external data. (a to f) have a selection variable similar to Figure 2b, and (g to l) have a selection variable similar to Figure 2c. S1 corresponds to the scenario where $f_{h,Y}(\Delta)$ largely deviates from $f_{h,Y}(\Omega)$, and S2 denotes the scenario where $f_{h,Y}(\Delta) \approxeq f_{h,Y}(\Omega)$. Error Bars for Crab-MØ, Crab-MA and Crab-MU are omitted for clarity.

> *Key Takeaway.* Crab-MX (with fairness requirement threshold $\tau = 0$) achieves perfect fairness for all scenarios and datasets, while prior pre-processing and in-processing methods demonstrate aberrant behavior.

Comparison between Orig (NoSB) and Crab-MX. Figure 4 shows that Crab-MX achieves a fairer model compared to Orig (NoSB) whenever it achieves an F-score comparable to that of Orig (NoSB) (red point has higher equal opportunity when compared with Crab-MX point at same x-coordinate). The only case where Orig (NoSB) has marginally higher accuracy than Crab-MX is Law-G1-S2 is where the difference in accuracy is not significant ($< 1\%$). This shows that Crab-MX is capable of achieving an F-score comparable to that of Orig (NoSB) even without any external data to recover from selection bias.

Quality vs. fairness tradeoff. Unlike prior fair ML techniques, Crab allows the user to specify a fairness requirement τ, which is varied to simulate varying needs. Figure 4 shows that Crab-MX achieves the same fairness and accuracy as Orig when τ is set greater than the fairness bound in Theorem 4.5. On reducing τ, Crab-MX's fairness improves consistently with a minor or no loss in F-score until $\tau > 0.1$. Further reducing τ to 0 worsens the F-score for the Adult dataset by around 8% but less than 1% for all other datasets. We also observe that Crab-MX and baseline techniques achieve similar F-scores when Crab-MX is configured to achieve similar fairness. This shows that Crab-MX not only achieves performance comparable to the baselines but also allows perfect fairness by achieving zero equal opportunity difference.

> *Key Takeaway.* Crab-MX considerably improves fairness of the trained classifier, with only a minor loss in F1 score.

5.2.2 Availability of external data. In this experiment, we evaluate the performance of Crab in settings where varied amounts of external data are available. We consider three different cases. (1) **Sufficient information:** this setting is applicable when external information about some statistics of the unbiased distribution are available to recover from selection bias. Crab-MØ uses estimates of $\Pr_{\Omega}(\mathbf{u} \mid s, \mathbf{a})$ for model training. (2) **Missing A**: this setting is applicable when external information about **A** is not available. For example, when **A** is the prediction target $\mathbf{A} = \{Y\}$, we cannot access labels for the unbiased distribution. (3) **Missing U**: this setting considers availability of partial external information about a set of attributes **U**. For a fair comparison, we compare Crab under these settings with Cortes, a method that uses external data to estimate the propensity score to recover from selection bias. Note that Cortes requires additional information compared to Crab-MØ, which relies only on the estimates of $\Pr_{\Omega}(\mathbf{u} \mid s, \mathbf{a})$ computed from external unbiased data or any other source, e.g., census data.

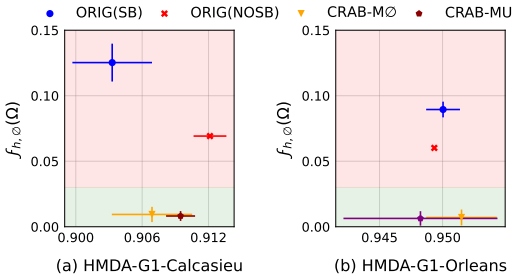

Figure 6: Statistical Parity-F1 score plots for two parishes in Louisiana where Crab-MØ and Crab-MU use real-world census to provide external information for training.

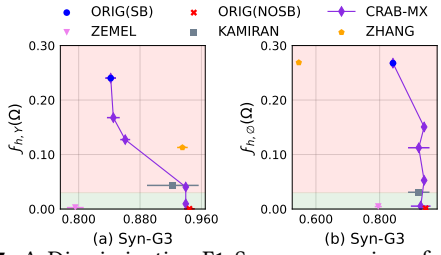

Figure 7: A Discrimination-F1 Score comparison for Crab-MX and baseline methods evaluated on Syn. Zafar was omitted due to its extremely low F1 Score in this experiment.

Real-world case study: Louisiana. First, we evaluate the impact of selection bias in HMDA dataset for two different parishes (equivalent to counties in other states) in the state of Louisiana (Calcasieu and Orleans). We leveraged statistics from public census data as external information to recover from selection bias and run Crab-MØ, Crab-MA, and Crab-MU. τ was set to 0 to achieve equal opportunity. Selection bias was introduced based on age and race of individuals, and we used $\Pr_\Omega(age \mid race)$ from the public census data of these parishes [64]. Even though the real-world ratios were not rigorously consistent with those in the unbiased dataset, Crab successfully trained a fair model (Figure 6). Further, we notice that the F1 Score of the trained model is higher than that of Orig in Calcasieu. This experiment demonstrates the potential of Crab for using publicly available census data as compensation for the lack of unbiased external data.

We now compare the performance of Crab and other baselines with varied settings of external data.

Solution quality. Figure 5 compares Crab with Cortes and Orig (NoSB) under different settings of access to auxiliary data source. We observe that all Crab methods achieve ≈ 0 equal opportunity for $\tau = 0$ without much loss in F1-Score. In fact, the quality of the most fair classifier is higher than the fairness-agnostic classifier trained on the original data (HMDA-G2-S2 and HMDA-G3-S1). We tested the fairness-accuracy tradeoff in detail in Figure 7 and show that ensuring fairness can indeed improve overall classifier performance. In contrast, the Cortes baseline requires additional external information to recover from selection bias but remains unstable with respect to fairness. Cortes helps to ensure fairness in Figure 5 (a to d), but the trained classifier is highly unfair in Figure 5 (g to l). Cortes relies on the estimation of propensity scores, which are highly dependent on the quality of classifiers learned to estimate $\Pr_\Omega(C = 1 \mid \boldsymbol{X} = x)$ and $\Pr_\Omega(C = 1 \mid Y = 1, \boldsymbol{X} = x)$. Noisy estimation of these probabilities affects Cortes performance.

Since Crab-MU is an upper bound-based method, enforcing a low upper bound of $\mathfrak{f}_{h,}(\Omega)$ might *overly* enforce fairness to an unnecessarily low level when the gap between the bound and $\mathfrak{f}_{h,}(\Omega)$ is large. In general, we cannot conclude whether the bound-based or estimation-based approach would result in a model with better performance. However, if the gap between the upper bound and $\mathfrak{f}_{h,}(\Omega)$ is large, a bound-based approach like Crab-MU would worsen classifier quality to ensure fairness ((e.g., Figure 5 (f))).

Effect of varying external data. Among the three Crab methods that require different settings of access to auxiliary data source for classifier training, Crab-MU requires the least amount of data, followed by Crab-MA and Crab-MØ. However, Crab-MA and Crab-MU are applicable only in cases where C is not a child of Y.Therefore, we omit these two methods in Figure 5 (g–l) (graph G3), where the selection criterion is a function of the target variable. In Figure 5 (a to f), Crab-MØ and Crab-MA achieve slightly higher F1 Scores for the same fairness requirements compared to Crab-MU. Unlike Crab-MØ and Crab-MA, Crab-MU removes the discrimination by limiting the upper bound of $\mathfrak{f}_{h,}(\Omega)$ rather than its estimation; thus, the skyline of Crab-MU appears to be non-monotonic in Law-G2-S2. However, Crab ensures fairness in all scenarios across all settings of access to auxiliary data.

> *Key Takeaway.* Crab-MØ trains the most accurate model with zero equal opportunity. Even though Cortes uses more external data, it does not consistently learn a fair classifier.

Running time. Table 2 compares the running time of Crab-MX and Crab-MØ with Orig for the logistic regression classifier. We observe that Crab trains a fair classifier in less than 10 seconds for small-scale datasets like Adult and Law and in less than 3 minutes for million scale datasets like HMDA. Among baselines, Zemel was the least efficient, requiring around 60 seconds for the Adult dataset and more than one hour for the HMDA dataset. All other baselines required a similar execution time as Crab.

5.3 Effect of Selection Bias on Fairness and F1 Score

In this section, we empirically evaluate the results established in Section 3 and demonstrate conditions under which biased data collection processes may lead to unfairness. We then evaluate whether enforcing fairness can enhance model performance, as well.

Unfairness due to selection bias. In this experiment, we consider the Syn dataset and introduce selection bias in five different ways (R denoting randomly and G1 to G4 corresponding to the procedures in Figure 2(a to d)) to evaluate its impact on fairness. The test distribution of the Syn dataset satisfies $Y \perp\!\!\!\perp S$ ($\mathfrak{f}_{h,Y}(\Omega) = 0.001$), but the independence does not hold in the biased training dataset for scenarios G1 to G3, where $Y \not\perp\!\!\!\perp SC$. We trained the logistic regression classifier on the biased dataset and evaluated its accuracy and unfairness on the unbiased test set (Table 3). The only case where selection bias causes a notable increase in unfairness is when the selection node is a child of Y and another variable X such that $S \not\perp\!\!\!\perp X$. We observe similar results for SVM classifier. This empirical observation conforms with our analysis in Proposition 3.2.

Logistic Regression				
Scenario	$Y \in \mathsf{Pa}(C)$	$\mathsf{Pa}(C) \perp\!\!\!\perp S$	F1 Score	$f_{h,Y}(\Omega)$
R	No	No	0.957	0.001
G1	No	Yes	0.957	0.017
G2	No	Yes	0.957	0.015
G3	Yes	Yes	0.850	**0.235**
G4	Yes	No	0.957	0.002
SVM				
R	No	No	0.957	0.002
G1	No	Yes	0.957	0.008
G2	No	Yes	0.957	0.004
G3	Yes	Yes	0.838	**0.180**
G4	Yes	No	0.957	0.002

Table 3: Comparison of F1 Score and unfairness when varying the selection bias procedure. R denotes a random procedure, and G1 to G4 correspond to the graphs in Figure 2.

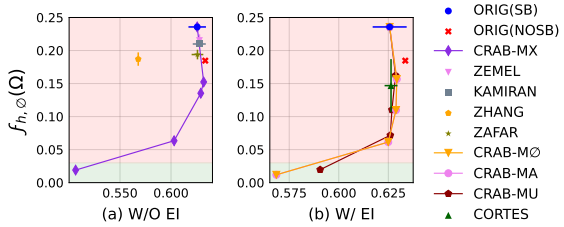

Figure 8: Statistical Parity (y-axis)-F1 Score (x-axis) comparison for CRAB and baseline methods for Adult-G1-S1.

Can fairness improve classifier quality? Figure 4 (c) and Figure 5 (f) showed that the classifier trained by CRAB can sometimes achieve higher F1 Scores than ORIG and ORIG (NOSB) while ensuring fairness ($f_{h,Y}(\Omega) = 0$). To validate this observation, we evaluated CRAB extensively on the Syn dataset where the unbiased dataset is fair ($f_{h,Y}(\Delta) = 0$). Figure 7 shows that ensuring fairness can indeed improve classifier F1 Scores by more than 10%. However, most baselines remain either unfair or achieve much lower F1 Scores. This evaluation demonstrates that ensuring fairness can on occasion improve classifier quality, and CRAB achieves the best F1 score and fairness.

Key Takeaway. Ensuring fairness can sometimes improve classifier quality.

5.4 Sensitivity to Parameters

We now evaluate the behavior of CRAB with access to different settings of auxiliary data and other baselines when varying the fairness metric, classification algorithm, and external data size.

Fairness metric: statistical parity. In this experiment, we consider the statistical parity fairness metric for evaluation. Figure 8 compares CRAB on both settings, presence and absence of external data for the Adult dataset. In figure 8(a), all baselines achieve statistical parity of 0.15 except CRAB-MX, which achieves zero statistical parity. Further, we observe that CRAB-MX achieves the maximum F1 Score of 0.63 while achieving a statistical parity of less than 0.17. All other techniques, including the classifier trained on the unbiased original dataset, achieve statistical parity that exceeds 0.17.

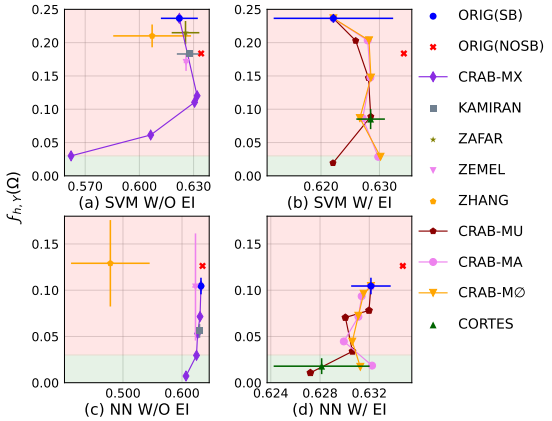

Figure 9: Equal Opportunity-F1 Score (x-axis) comparison for CRAB and baseline methods on the Adult dataset with the selection mechanism described in Figure 2b.

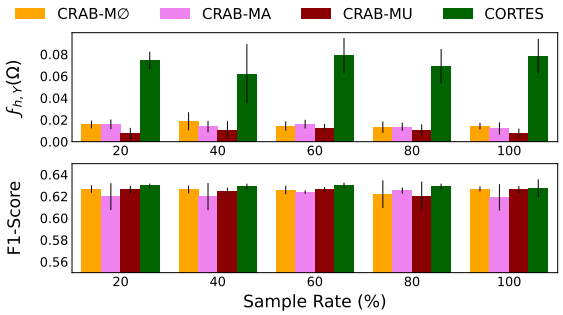

Figure 10: The upper plot shows the Equal Opportunity-Sample Rate; the lower plot shows the F1 Score-Sample Rate.

This demonstrates the effectiveness of CRAB-MX at achieving fairness even in the absence of any external information. Figure 8 compares CRAB under three different settings of access to external data (CRAB-MA, CRAB-MU, CRAB-M∅). We observe that CRAB (with τ is set to zero) achieves fairness (statistical parity ≈ 0) across all three settings of access to auxiliary data. Further, the CORTES baseline does not achieve fairness even though it has access to a maximum amount of external information. Comparing figures 8(a) and (b), we observe that external information helps CRAB-M∅ achieve a 0.59 F1 Score with zero statistical parity as opposed to 0.52 in the absence of external data. We also observe that CRAB-MA and CRAB-MU exceed a 0.57 F1 Score, showing that any amount of external data can boost the quality of a trained classifier.

Varying classification methods. CRAB can be integrated into numerous classification algorithms by modifying the loss function. Figure 9 demonstrates the generalizability of CRAB to run with SVM and NN classifiers. In all cases, all CRAB methods achieve close to zero equal opportunity, while most of baselines remain unfair (EO $>$ 0.05). Further, CRAB-MX achieves the maximum F1 Score with the maximum fairness (EQ $<$ 0.12 in Figure 9(a)), while all baselines perform worse (EO $>$ 0.17). Comparing Figure 9 (a) and (b), we observe that CRAB-M∅ achieves higher F1 Scores than CRAB-MX while maintaining EO$<$ 0.03. We observe similar trends for the neural network (Figure 9 (c), (d)).

Varying external data size. Previous experiments used the unbiased training data to provide external information (whenever needed). In practice, the amount of external information might be much smaller, which is likely to affect the quality of ratios estimated by Crab and the trained probability estimators used by Cortes. Therefore, we constructed randomly chosen samples of the unbiased training data, with sampling rate ranging from 20% to 100%, and tested the fairness and F1 Scores of Crab-MØ, Crab-MA, Crab-MU and Cortes. These techniques used the sampled data as external information to train a fair classifier. The threshold τ for Crab was set to 0.01. The results are reported in Figure 10.

In Figure 10, all techniques have comparable F1 Scores but Crab (under all settings of access to auxiliary data) has a considerably lower $f_{h,Y}(\Omega)$ than Cortes. In fact, Crab achieves the desired fairness for all sampling rates and accesses to external data. Further, Cortes has the maximum standard deviation (denoting instability of performance) compared to other methods. As the sample size decreases, the standard deviation of both equal opportunity and F1 Scores for Crab methods increases slightly, indicating reduced stability when the quality of estimated ratios degrades.

6 RELATED WORK AND CONCLUSION

Our work is related to the existing work in ML for learning in the presence of selection bias. In this context, reweighting is one of the most commonly used approaches for dealing with selection bias. Reweighting methods modify the cost of an error on each training point according to some computed weights. These weights are often computed by estimation techniques using an unbiased sample of the population [7, 20, 35, 45]. However, studies have shown that these estimations can have errors that severely influence model performance in downstream tasks [20, 45], which is confirmed by the experimental results in this paper. This approach has also been used for data debiasing in the context of query answering from biased data in [54].

Our work is also related to existing methods for training fair predictive algorithms, which can be categorized into pre-, post- and in-processing; pre-processing approaches, most relevant to our research (see [18] for recent surveys), often rely on the assumption that training and testing data are representative of that target population; hence, they cannot capture discrimination due to selection bias. We acknowledge that some post-processing approaches that are designed to modify an algorithm's decisions based on individual's sensitive attribute during model deployment are unaffected by selection bias. However, these methods not only require access to information about sensitive attributes during deployment, but they also can be applied only to individuals that form a representative sample of the population. Therefore, they are very restrictive and have limited applicability in practice.

Selection bias and its effects on fairness have received recent attention in the fairness literature [13, 25, 30, 48, 65]. The study most closely related to ours is [30], which examines the problem from a causal perspective and discusses how selection bias can nullify given fairness guarantees based on training data. That study explores different mechanisms corresponding to real-world scenarios where selection bias occurs and discusses whether the probability distributions needed for different fairness metrics would be recoverable in those scenarios. Another work [65] studies the impact of selection bias on fairness and proposes a method to address the issue based on existing methods for learning in the presence of selection bias. Unlike these works, we rigorously study the impact of selection bias on fairness and establish conditions under which one can learn fair ML models in the presence of selection bias with varying levels of external knowledge about the target population.

Our work is also related to theoretical studies that investigate the impact of sample under-representation on fairness. [13] shows under certain assumptions that enforcing fairness constraints can help improve fairness and accuracy simultaneously. The empirical results obtained in this work confirm this intuition even in the more general setting considered in [13]. [48] analyzes a slightly different problem of robustness of decision trees in the presence of various data biases, including selection bias. However, their solution relies on the availability of strong assumption detail of the impact of selection bias on data distribution and cannot go beyond decision tress. [25] combines the issues of robustness and fairness when there are issues with sample selection bias; however, they assume that fairness of a model on the biased data implies fairness on the target population. Finally, taking inspiration from the transfer learning domain and how fairness would transfer across domains, [62] and [74] establish theoretical bounds on fairness of a model on the target population. However, these bound can be calculated only under the assumption of access to samples from the target distribution and hence cannot be used to train fair predictive models.

REFERENCES

[1] 2019. HMDA Data Publication. https://ffiec.cfpb.gov/data-publication/2019.
[2] 2022. CRAB Code. https://anonymous.4open.science/r/Crab-B1A3/.
[3] ACP. [n.d.]. Racial and Ethnic Disparities in Health Care. https://www.acponline.org/acp_policy/policies/racial_ethnic_disparities_2010.pdf.
[4] Israel Edem Agbehadji, Bankole Osita Awuzie, Alfred Beati Ngowi, and Richard C Millham. 2020. Review of big data analytics, artificial intelligence and nature-inspired computing models towards accurate detection of COVID-19 pandemic cases and contact tracing. *International journal of environmental research and public health* 17, 15 (2020), 5330.
[5] Norah Alballa and Isra Al-Turaiki. 2021. Machine learning approaches in COVID-19 diagnosis, mortality, and severity risk prediction: A review. *Informatics in Medicine Unlocked* 24 (2021), 100564.
[6] Agathe Balayn, Christoph Lofi, and Geert-Jan Houben. 2021. Managing bias and unfairness in data for decision support: a survey of machine learning and data engineering approaches to identify and mitigate bias and unfairness within data management and analytics systems. *The VLDB Journal* 30, 5 (2021), 739–768.
[7] Elias Bareinboim and Judea Pearl. 2012. Controlling selection bias in causal inference. In *Artificial Intelligence and Statistics*. PMLR, 100–108.
[8] Peter L Bartlett, Michael I Jordan, and Jon D McAuliffe. 2006. Convexity, classification, and risk bounds. *J. Amer. Statist. Assoc.* 101, 473 (2006), 138–156.
[9] Yahav Bechavod and Katrina Ligett. 2017. Learning fair classifiers: A regularization-inspired approach. (2017).
[10] Rachel KE Bellamy, Kuntal Dey, Michael Hind, Samuel C Hoffman, Stephanie Houde, Kalapriya Kannan, Pranay Lohia, Jacquelyn Martino, Sameep Mehta, Aleksandra Mojsilovic, et al. 2018. AI Fairness 360: An extensible toolkit for detecting, understanding, and mitigating unwanted algorithmic bias. *arXiv preprint arXiv:1810.01943* (2018).
[11] Leopoldo Bertossi. 2006. Consistent query answering in databases. *ACM Sigmod Record* 35, 2 (2006), 68–76.
[12] Jelke Bethlehem. 2010. Selection bias in web surveys. *International statistical review* 78, 2 (2010), 161–188.
[13] Avrim Blum and Kevin Stangl. 2019. Recovering from biased data: Can fairness constraints improve accuracy? *arXiv preprint arXiv:1912.01094* (2019).
[14] Laura Bronner. 2020. *Why Statistics Don't Capture The Full Extent Of The Systemic Bias In Policing.* https://fivethirtyeight.com/features/why-statistics-dont-capture-the-full-extent-of-the-systemic-bias-in-policing/
[15] Toon Calders and Sicco Verwer. 2010. Three naive bayes approaches for discrimination-free classification. *Data mining and knowledge discovery* 21, 2 (2010), 277–292.

[16] Flavio Calmon, Dennis Wei, Bhanukiran Vinzamuri, Karthikeyan Natesan Ramamurthy, and Kush R Varshney. 2017. Optimized Pre-Processing for Discrimination Prevention. In *Advances in Neural Information Processing Systems 30*, I. Guyon, U. V. Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, and R. Garnett (Eds.). Curran Associates, Inc., 3992–4001. http://papers.nips.cc/paper/6988-optimized-pre-processing-for-discrimination-prevention.pdf

[17] Kenzie A Cameron, Jing Song, Larry M Manheim, and Dorothy D Dunlop. 2010. Gender disparities in health and healthcare use among older adults. *Journal of women's health* 19, 9 (2010), 1643–1650.

[18] Simon Caton and C. Haas. 2020. Fairness in Machine Learning: A Survey. *ArXiv* abs/2010.04053 (2020).

[19] Sam Corbett-Davies, Emma Pierson, Avi Feller, Sharad Goel, and Aziz Huq. 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.

[20] Corinna Cortes, Mehryar Mohri, Michael Riley, and Afshin Rostamizadeh. 2008. Sample selection bias correction theory. In *International conference on algorithmic learning theory*. Springer, 38–53.

[21] Corinna Cortes and Vladimir Vapnik. 1995. Support-vector networks. *Machine learning* 20, 3 (1995), 273–297.

[22] David R Cox. 1958. The regression analysis of binary sequences. *Journal of the Royal Statistical Society: Series B (Methodological)* 20, 2 (1958), 215–232.

[23] Aron Culotta. 2014. Reducing sampling bias in social media data for county health inference. In *Joint Statistical Meetings Proceedings*. Citeseer, 1–12.

[24] Akhil A Dixit and Phokion G Kolaitis. 2021. Consistent answers of aggregation queries using SAT solvers. *arXiv preprint arXiv:2103.03314* (2021).

[25] Wei Du and Xintao Wu. 2021. Robust fairness-aware learning under sample selection bias. *arXiv preprint arXiv:2105.11570* (2021).

[26] Cynthia Dwork, Moritz Hardt, Toniann Pitassi, Omer Reingold, and Richard Zemel. 2012. Fairness through awareness. In *Proceedings of the 3rd innovations in theoretical computer science conference*. 214–226.

[27] Wenfei Fan and Floris Geerts. 2012. Foundations of data quality management. *Synthesis Lectures on Data Management* 4, 5 (2012), 1–217.

[28] Michael Feldman, Sorelle A Friedler, John Moeller, Carlos Scheidegger, and Suresh Venkatasubramanian. 2015. Certifying and removing disparate impact. In *Proceedings of the 21th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*. ACM, 259–268.

[29] Andrew Gelman, Jeffrey Fagan, and Alex Kiss. 2007. An analysis of the New York City police department's "stop-and-frisk" policy in the context of claims of racial bias. *Journal of the American statistical association* 102, 479 (2007), 813–823.

[30] Naman Goel, Alfonso Amayuelas, Amit Deshpande, and Amit Sharma. 2020. The Importance of Modeling Data Missingness in Algorithmic Fairness: A Causal Perspective. *arXiv preprint arXiv:2012.11448* (2020).

[31] Naman Goel, Alfonso Amayuelas, Amit Deshpande, and Amit Sharma. 2021. The Importance of Modeling Data Missingness in Algorithmic Fairness: A Causal Perspective. In *Proceedings of the AAAI Conference on Artificial Intelligence*, Vol. 35. 7564–7573.

[32] Zerrin Asan Greenacre et al. 2016. The importance of selection bias in internet surveys. *Open Journal of Statistics* 6, 03 (2016), 397.

[33] Moritz Hardt, Eric Price, and Nati Srebro. 2016. Equality of opportunity in supervised learning. *Advances in neural information processing systems* 29 (2016), 3315–3323.

[34] Miguel A Hernán, Sonia Hernández-Díaz, and James M Robins. 2004. A structural approach to selection bias. *Epidemiology* (2004), 615–625.

[35] Jiayuan Huang, Arthur Gretton, Karsten Borgwardt, Bernhard Schölkopf, and Alex Smola. 2006. Correcting sample selection bias by unlabeled data. *Advances in neural information processing systems* 19 (2006).

[36] Maliha Tashfia Islam, Anna Fariha, Alexandra Meliou, and Babak Salimi. 2022. Through the data management lens: Experimental analysis and evaluation of fair classification. In *Proceedings of the 2022 International Conference on Management of Data*. 232–246.

[37] HV Jagadish, Julia Stoyanovich, and Bill Howe. 2021. The many facets of data equity. In *theWorkshop Proceedings of the EDBT/ICDT 2021 Joint Conference*.

[38] Faisal Kamiran and Toon Calders. 2012. Data preprocessing techniques for classification without discrimination. *Knowledge and information systems* 33, 1 (2012), 1–33.

[39] Toshihiro Kamishima, Shotaro Akaho, Hideki Asoh, and Jun Sakuma. 2012. Fairness-aware classifier with prejudice remover regularizer. In *Joint European conference on machine learning and knowledge discovery in databases*. Springer, 35–50.

[40] Michelle E Kho, Mark Duffett, Donald J Willison, Deborah J Cook, and Melissa C Brouwers. 2009. Written informed consent and selection bias in observational studies using medical records: systematic review. *Bmj* 338 (2009).

[41] Niki Kilbertus, Mateo Rojas-Carulla, Giambattista Parascandolo, Moritz Hardt, Dominik Janzing, and Bernhard Schölkopf. 2017. Avoiding discrimination through causal reasoning. *arXiv preprint arXiv:1706.02744* (2017).

[42] Ron Kohavi and George H John. 1997. Wrappers for feature subset selection. *Artificial intelligence* 97, 1-2 (1997), 273–324.

[43] James E Lange, Mark B Johnson, and Robert B Voas. 2005. Testing the racial profiling hypothesis for seemingly disparate traffic stops on the New Jersey Turnpike. *Justice Quarterly* 22, 2 (2005), 193–223.

[44] M. Lichman. 2013. UCI machine learning repository.

[45] Anqi Liu and Brian Ziebart. 2014. Robust classification under sample selection bias. *Advances in neural information processing systems* 27 (2014).

[46] Kristian Lum and William Isaac. 2016. To predict and serve? *Significance* 13, 5 (2016), 14–19.

[47] Ninareh Mehrabi, Fred Morstatter, Nripsuta Saxena, Kristina Lerman, and Aram Galstyan. 2021. A survey on bias and fairness in machine learning. *ACM Computing Surveys (CSUR)* 54, 6 (2021), 1–35.

[48] Anna Meyer, Aws Albarghouthi, and Loris D'Antoni. 2021. Certifying Robustness to Programmable Data Bias in Decision Trees. *Advances in Neural Information Processing Systems* 34 (2021).

[49] Clayton J Mosher, Terance D Miethe, and Timothy C Hart. 2010. *The mismeasure of crime*. Sage Publications.

[50] Razieh Nabi and Ilya Shpitser. 2018. Fair inference on outcomes. In *Proceedings of the... AAAI Conference on Artificial Intelligence. AAAI Conference on Artificial Intelligence*, Vol. 2018. NIH Public Access, 1931.

[51] Richard E Neapolitan et al. 2004. *Learning bayesian networks*. Vol. 38. Pearson Prentice Hall Upper Saddle River, NJ.

[52] Felix Neutatz, Binger Chen, Ziawasch Abedjan, and Eugene Wu. 2021. From Cleaning before ML to Cleaning for ML. *IEEE Data Eng. Bull.* 44, 1 (2021), 24–41.

[53] Natalia Norori, Qiyang Hu, Florence Marcelle Aellen, Francesca Dalia Faraci, and Athina Tzovara. 2021. Addressing bias in big data and AI for health care: A call for open science. *Patterns* 2, 10 (2021), 100347.

[54] Laurel Orr, Magdalena Balazinska, and Dan Suciu. 2020. Sample debiasing in the themis open world database system. In *Proceedings of the 2020 ACM SIGMOD International Conference on Management of Data*. 257–268.

[55] Adam Paszke, Sam Gross, Francisco Massa, Adam Lerer, James Bradbury, Gregory Chanan, Trevor Killeen, Zeming Lin, Natalia Gimelshein, Luca Antiga, et al. 2019. Pytorch: An imperative style, high-performance deep learning library. *Advances in neural information processing systems* 32 (2019).

[56] Judea Pearl. 1995. Causal diagrams for empirical research. *Biometrika* 82, 4 (1995), 669–688.

[57] Judea Pearl. 2009. *Causality*. Cambridge university press.

[58] Rashida Richardson, Jason M Schultz, and Kate Crawford. 2019. Dirty data, bad predictions: How civil rights violations impact police data, predictive policing systems, and justice. *NYUL Rev. Online* 94 (2019), 15.

[59] Ricardo Salazar, Felix Neutatz, and Ziawasch Abedjan. 2021. Automated feature engineering for algorithmic fairness. *Proceedings of the VLDB Endowment* 14, 9 (2021), 1694–1702.

[60] Babak Salimi, Luke Rodriguez, Bill Howe, and Dan Suciu. 2019. Capuchin: Causal database repair for algorithmic fairness. *arXiv preprint arXiv:1902.08283* (2019).

[61] Babak Salimi, Luke Rodriguez, Bill Howe, and Dan Suciu. 2019. Interventional fairness: Causal database repair for algorithmic fairness. In *Proceedings of the 2019 International Conference on Management of Data*. 793–810.

[62] Candice Schumann, Xuezhi Wang, Alex Beutel, Jilin Chen, Hai Qian, and Ed H Chi. 2019. Transfer of machine learning fairness across domains. *arXiv preprint arXiv:1906.09688* (2019).

[63] Patrick Schwab, August DuMont Schütte, Benedikt Dietz, Stefan Bauer, et al. 2020. Clinical predictive models for COVID-19: systematic study. *Journal of medical Internet research* 22, 10 (2020), e21439.

[64] https://statisticalatlas.com/United-States/Overview. 2018. the demographic statistical atlas of the united states - statistical atlas 2018. *Statisticalatlas* (2018).

[65] Yanchen Wang and Lisa Singh. 2021. Analyzing the impact of missing values and selection bias on fairness. *International Journal of Data Science and Analytics* 12, 2 (2021), 101–119.

[66] Linda F Wightman. 1998. LSAC National Longitudinal Bar Passage Study. LSAC Research Report Series. (1998).

[67] Blake Woodworth, Suriya Gunasekar, Mesrob I. Ohannessian, and Nathan Srebro. 2017. Learning Non-Discriminatory Predictors. In *Proceedings of the 2017 Conference on Learning Theory (Proceedings of Machine Learning Research, Vol. 65)*, Satyen Kale and Ohad Shamir (Eds.). PMLR, Amsterdam, Netherlands, 1920–1953. http://proceedings.mlr.press/v65/woodworth17a.html

[68] Muhammad Bilal Zafar, Isabel Valera, Manuel Gomez Rodriguez, and Krishna P Gummadi. 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.

[69] Muhammad Bilal Zafar, Isabel Valera, Manuel Gomez Rogriguez, and Krishna P. Gummadi. 2017. Fairness Constraints: Mechanisms for Fair Classification. In *Proceedings of the 20th International Conference on Artificial Intelligence and Statistics (Proceedings of Machine Learning Research, Vol. 54)*, Aarti Singh and Jerry Zhu (Eds.). PMLR, Fort Lauderdale, FL, USA, 962–970. http://proceedings.mlr.

press/v54/zafar17a.html
[70] Muhammad Bilal Zafar, Isabel Valera, Manuel Gomez Rogriguez, and Krishna P Gummadi. 2017. Fairness constraints: Mechanisms for fair classification. In *Artificial intelligence and statistics*. PMLR, 962–970.
[71] Rich Zemel, Yu Wu, Kevin Swersky, Toni Pitassi, and Cynthia Dwork. 2013. Learning fair representations. In *International conference on machine learning*. PMLR, 325–333.
[72] Brian Hu Zhang, Blake Lemoine, and Margaret Mitchell. 2018. Mitigating unwanted biases with adversarial learning. In *Proceedings of the 2018 AAAI/ACM Conference on AI, Ethics, and Society*. 335–340.
[73] Hantian Zhang, Xu Chu, Abolfazl Asudeh, and Shamkant B Navathe. 2021. Omnifair: A declarative system for model-agnostic group fairness in machine learning. In *Proceedings of the 2021 International Conference on Management of Data*. 2076–2088.
[74] Yiliang Zhang and Qi Long. 2021. Assessing Fairness in the Presence of Missing Data. *Advances in Neural Information Processing Systems* 34 (2021).
[75] Zirun Zhao, Anne Chen, Wei Hou, James M Graham, Haifang Li, Paul S Richman, Henry C Thode, Adam J Singer, and Tim Q Duong. 2020. Prediction model and risk scores of ICU admission and mortality in COVID-19. *PloS one* 15, 7 (2020), e0236618.