FAIRNESS GUARANTEES UNDER DEMOGRAPHIC SHIFT

Anonymous authors
Paper under double-blind review

ABSTRACT

Recent studies found that using machine learning for social applications can lead to injustice in the form of racist, sexist, and otherwise unfair and discriminatory outcomes. To address this challenge, recent machine learning algorithms have been designed to limit the likelihood such unfair behavior occurs. However, these approaches typically assume the data used for training is representative of what will be encountered in deployment, which is often untrue. In particular, if certain subgroups of the population become more or less probable in deployment (a phenomenon we call *demographic shift*), prior work's fairness assurances are often invalid. In this paper, we consider the impact of demographic shift and present a class of algorithms, called `Shifty` algorithms, that provide high-confidence behavioral guarantees that hold under demographic shift. `Shifty`, the first technique of its kind, demonstrates an effective strategy for designing algorithms to overcome demographic shift's challenges. We evaluate `Shifty` using a real-world dataset of university entrance exams and subsequent student success. We show that the learned models avoid bias under demographic shift, unlike existing methods. Our experiments demonstrate that our algorithm's high-confidence fairness guarantees are valid in practice and that our algorithm is an effective tool for training models that are fair when demographic shift occurs.

1 INTRODUCTION

As machine learning (ML) algorithms continue to be used to aid decisions in socially impactful applications (Angwin et al., 2016; Goodall, 2016; Olson, 2011), it is becoming increasingly important to ensure that trained models are able to avoid bias and discrimination. Observations that the use of ML algorithms might have unexpected social implications, such as bias with respect to sex or race, have led to the creation of algorithms that provide high-confidence fairness guarantees (Thomas et al., 2019; Agarwal et al., 2018). These guarantees rely on the assumption that the data the model is trained with and the data encountered after deployment follow the same distribution. However, this assumption is false for many real-world problems (Zhuang et al., 2021), and, as we demonstrate, models often violate these guarantees and exhibit unfair bias when evaluated on data from a different distribution.

As an example, consider a model that uses university entrance exam scores to predict subsequent success. Because student demographics, such as race or sex, can shift over time, the distribution of applicants may change between model training and deployment. As a result, even if the model is trained to protect a disadvantaged group, if the learning algorithm assumes the training and deployment distributions are the same, many fairness guarantees will not hold in practice (Section 4 empirically verifies this claim). We refer to such distribution change as *demographic shift*. Specifically, demographic shift occurs when the difference between the training and deployment distributions can be explained by a shift in the marginal distribution of a single random variable, such as race or sex.

We present `Shifty`, the first strategy for designing ML algorithms that provide high-confidence guarantees that one or more user-specified fairness constraints will hold despite a demographic shift between training and deployment. We design `Shifty` algorithms to work in two scenarios based on what the user knows at training regarding the demographic shift: **(1)** the old and new demographic proportions are known, or **(2)** the demographic proportions are bounded in known intervals.

We evaluate `Shifty` on a real-world dataset of students' college entrance exam scores and their subsequent grade point average (GPA) in the Brazilian university system (da Silva, 2019). We compare `Shifty` to three families of existing algorithms: Fairness Constraints (Zafar et al., 2017), which aim

to enforce a specific definition of fairness without guarantees, and Seldonian algorithms (Thomas et al., 2019) and Fairlearn (Agarwal et al., 2018), both of which provide high-confidence fairness guarantees.

`Shifty` allows the user to provide (one or multiple simultaneous) fairness definitions from a large class appropriate to the application domain, such as demographic parity, disparate impact, equalized odds, predictive equality, and individual fairness. Unlike all prior algorithms, `Shifty` algorithms provide high-confidence guarantees that the learned model satisfies these fairness constraints even when deployed on a distribution different from the training one. We demonstrate empirically that models learned by previous algorithms do, at times, violate the desired properties under demographic shift, while `Shifty`'s models do not. If there is insufficient training data, or if the specified fairness properties are simultaneously unsatisfiable, `Shifty` algorithms are designed to return no solution, but we show that this rarely happens in practice. Finally and crucially, we demonstrate empirically that, in our evaluated domain, `Shifty` is able to learn models that exhibit no loss of accuracy compared to the models that do not guarantee fairness, as long as sufficient training data exists.

Our main contributions are: **(1)** the first classification algorithms that provide high-confidence fairness guarantees under demographic shift, **(2)** a constructive proof of the guarantees, and a method for creating such algorithms, **(3)** an evaluation on real-world data, and **(4)** an open-source `Shifty` implementation and a release of all our data.

2 Background and Related Work

We illustrate our approach on fair classification, although the methods we propose are easily extended to other problems, such as regression (Thomas et al., 2019) and contextual bandits (Metevier et al., 2019). In this setting, a data *instance* consists of a set of *features* and an associated *label*. When considering the fairness of a classifier, each instance can be augmented with a *fairness attribute*. This information is often not used for prediction (e.g., some laws prohibit the use of race or sex in hiring decisions), but is assumed available for determining if the classifier exhibits bias.

We denote features by $X \in \mathcal{X}$, labels by $Y \in \mathcal{Y}$, and the fairness attribute by $S \in \mathcal{S}$. We assume that (X, Y, S) is sampled from some joint probability distribution defined over $\mathcal{X} \times \mathcal{Y} \times \mathcal{S}$. The naïve classification setting ignores the fairness attribute, and the goal is to accurately predict the label associated with X when its true label is unknown. These predictions are generated using a *model*, $\theta : \mathcal{X} \to \mathcal{Y}$. A loss function, such as expected classification error, measures the quality of θ. To obtain an accurate classifier, one typically selects a training algorithm, a, designed to minimize the chosen loss, and supplies it with a dataset consisting of n observations sampled independently from the joint distribution—that is, $D = \{(X_i, Y_i, S_i)\}_{i=1}^n$, where $\Pr(X_i, Y_i, S_i) := \Pr(X, Y, S)$ for all $i \in \{1, ..., n\}$.

To assess the fairness of an algorithm, the user provides a function, g, which accepts a model and is calibrated so $g(\theta) > 0$ if and only if θ behaves unfairly. Typically, g depends on the fairness attribute, S. For example, to assess if a classifier is biased based on race, g might measure the difference in the classifier's accuracy for individuals of one race compared to another.

Importantly, g can be defined based on the particular fairness requirements of a given application. Here, we consider the illustrative case where $g(\theta)$ is based on conditional expected value. (Appendix D extends our approach to handle more general definitions of g.) Concretely, let $H := h(X, Y, S, \theta)$ define some choice of real-valued observable, let $\xi := c(X, Y, S, \theta)$ be some event, and let τ represent a real-valued tolerance. We then assume that g is defined by $g(\theta) := \mathbf{E}\big[\, H \,\big|\, \xi \,\big] - \tau$. For example, if for a binary classification problem, the fairness attribute is sex, and the user wants to ensure that the false-positive rate of the model is below 20% for females, one might set $g(\theta) = \mathbf{E}\big[\, \theta(X) \,\big|Y{=}0, S{=}\texttt{female}\big] - 0.2$. While this form of g is not flexible enough to represent many widely used notions of fairness, it serves to illustrate our strategies for accounting for demographic shift, which are straightforward to apply to more sophisticated definitions, as Appendix D shows.

Given a definition for g, we say that a training algorithm, a, is fair with high confidence if

$$\Pr\big(\, g(a(D)) \leq 0 \,\big) \geq 1{-}\delta, \tag{1}$$

for some confidence threshold, $\delta \in [0, 1]$. Allowing the user to set δ during training overcomes the problem that guaranteeing fairness with absolute certainty is often impossible (Thomas et al., 2019).

It is possible that for some g provided by the user, there is no model θ that satisfies $g(\theta) \leq 0$ and, consequently, no algorithm a that satisfies (1). To address this, we adopt the convention described by Thomas et al. (2019) and permit algorithms to return NO_SOLUTION_FOUND (NSF), which is assumed to be fair by definition—that is, we assume that $g(\texttt{NSF}) = 0$. Intuitively, if a fair predictive model cannot be found, such algorithms are permitted to abstain from outputting an unfair predictive model, instead alerting the user that the fairness constraints could not be enforced with the required probability using the data provided. Since the trivial algorithm, $a(D) = \texttt{NSF}$ satisfies (1) by definition, we seek algorithms that satisfy (1) but return useful predictive models as frequently as possible, and we explicitly evaluate this consideration in our experimental designs.

Fair Classification under Demographic Shift: To reason about differences between the training and deployment data distributions, we augment each data instance with a random variable representing a *demographic attribute*, denoted by $T \in \mathcal{T}$. The demographic attribute is often distinct from the other variables defining each observation, but does not need to be. For example, a user might seek a model that avoids unfair bias against individuals of a particular sex, even if the distribution of the individuals' race differs from the training distribution. In this case, S would represent the sex of an individual, while T would represent their race.

Given the demographic attribute, we let (X, Y, S, T) represent an instance observed during training and let (X', Y', S', T') represent an instance encountered once the model is deployed. To formalize the effect of demographic shift, we assume that the demographic attribute's marginal distribution may change between training and deployment, but that the pre- and post-shift joint distributions over the instances are otherwise identical. This can be summarized by the following two conditions, which we refer to as the *demographic shift assumptions*:

$$\exists\, t \in \mathcal{T} \quad \text{s.t.} \quad \Pr(T = t) \neq \Pr(T' = t), \quad \text{and} \tag{2}$$

$$\forall\, (x, y, s, t), \quad \Pr(X{=}x, Y{=}y, S{=}s \,|\, T{=}t) = \Pr(X'{=}x, Y'{=}y, S'{=}s \,|\, T'{=}t). \tag{3}$$

Because the fairness of a model—that is, the value of $g(\theta)$—implicitly depends on X, Y, and S, it follows that guarantees of fairness based on g may fail to hold after the model is deployed, which corresponds to replacing these random variables with X', Y' and S'. Formally, if $H' = h(X', Y', S', \theta)$ and $\xi' = c(X', Y', S', \theta)$, so that $g'(\theta) = \mathbf{E}\big[H' \big| \xi'\big] - \tau$ measures the prevalence of unfair behavior after θ is deployed, then the challenge presented by demographic shift is summarized by the observation that, for all training algorithms a,

$$\underbrace{\Pr\big(\, g(a(D)) \leq 0\big) \geq 1{-}\delta}_{\text{Property A}} \quad \not\Rightarrow \quad \underbrace{\Pr\big(\, g'(a(D)) \leq 0\big) \geq 1{-}\delta}_{\text{Property B}}.$$

Therefore, we address the following problem:

Problem Statement: Given a user's description of possible demographic shift that might be present between the training and deployment environments as well as one or more definitions of fairness, design an algorithm, a, that provides high-confidence fairness guarantees that a returned model will behave fairly once the model is deployed—that is, an algorithm that satisfies Property B.

In this paper, we assume the user's description of the demographic shift is defined by a set of upper and lower bounds on the marginal probability of each value of the demographic attribute after deployment. Specifically, the user provides an input, $\mathcal{Q} := \{(a_t, b_t)\}_{t \in \mathcal{T}}$, encoding the assumption that $\Pr(T'{=}t) \in [a_t, b_t]$ for all $t \in \mathcal{T}$. However, the approach used by Shifty is general and can be applied for other descriptions of demographic shift as well. Given $\mathcal{Q}$ of this form, we identify two settings in which Shifty can be applied. In the case of *known demographic shift*, which occurs when the user knows the exact post-deployment distribution over the demographic attribute, the user sets $a_t = b_t = \Pr(T'{=}t)$ for each $t \in \mathcal{T}$, and Shifty will be guaranteed not to output unfair models under the specified deployment distribution. In the second setting, the user does not know the exact demographic shift that will occur and instead specifies non-singleton intervals for each $\Pr(T'{=}t)$. We refer to this as the *unknown demographic shift* setting.

Related Work: While many strategies promote fair outcomes in ML models, most existing approaches do not offer fairness guarantees, and none provide guarantees under demographic shift. Appendix A discusses these approaches in detail. At the high level, they include methods based on enforcing hard constraints (Irani, 2015), soft constraints (Zafar et al., 2017; Smits and Kotanchek, 2005), chance-constrained programming techniques (Charnes and Cooper, 1959; Miller and Wagner,

1965; Prékopa, 1970), or by pre- or post-processing training dataset and model predictions (Verma et al., 2021; Salimi et al., 2019; Awasthi et al., 2020). Meanwhile approaches that have considered fairness under distribution shift do not provide guarantees (Lipton et al., 2018).

The closest approach to ours is Seldonian algorithms (Thomas et al., 2019), which allow the user to specify the fairness definition and provide high-probability fairness guarantees—that is, they satisfy Property A. Seldonian algorithms perform well in real-world applications given sufficient data (Thomas et al., 2019; Metevier et al., 2019), but their guarantees, unlike `Shifty`'s, become invalid under demographic shift, as our experiments will show.

3 METHODOLOGY

An overview of `Shifty` is shown in Figure 1. Motivated by the design principles of previous Seldonian algorithms (Thomas et al., 2019), which are effective at designing algorithms that satisfy Property A, `Shifty` algorithms consist of three core components: *data partitioning*, *candidate selection*, and a *fairness test*. First, the data partitioning step splits the input dataset into two parts, which are used to perform the candidate selection and fairness test steps, respectively (see Appendix B). Next, a candidate model is trained. In practice, this can be performed using any existing classification algorithm, as the candidate selection step is not responsible for establishing the fairness guarantees `Shifty` provides. Instead, once a candidate model is found, `Shifty` performs a fairness test by computing a high-confidence upper bound on the prevalence of unfair behavior when the candidate model is deployed in an environment affected by the demographic shift described by $\mathcal{Q}$. If this high-confidence upper bound is below zero, the candidate model is likely to behave fairly once deployed, and the candidate is returned. However, if the value of the high-confidence upper bound is greater than zero, then `Shifty` cannot show that the candidate model will behave fairly with the required confidence, and returns `NSF` instead. Because `Shifty` only returns candidate models if they can be shown to be fair with high confidence on the demographic-shifted deployment distribution, it is guaranteed to satisfy Property B.

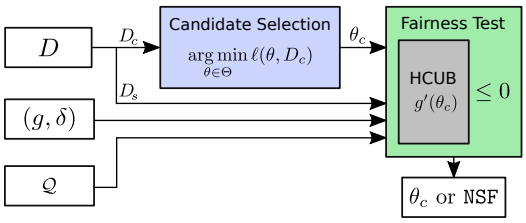

Figure 1: `Shifty` accepts a training dataset, D, one or more fairness specifications (each consisting of a definition of unfair behavior, g, and a tolerance, $\delta \in [0, 1]$), and a description of the possible demographic shifts, $\mathcal{Q}$, that might occur after deployment. It first partitions the input data into D_c and D_f, and uses D_c to select a candidate model, θ_c. Then, it uses $\mathcal{Q}$ and D_f to compute a $(1-\delta)$-probability high-confidence upper bound (HCUB) on the value of $g(\theta_c)$ after deployment, for each fairness definition. If these upper bounds are below zero, `Shifty` returns θ_c, and otherwise returns `NO_SOLUTION_FOUND` (`NSF`). Consequently, `Shifty` returns models that are unfair after deployment with probability at most δ.

Algorithm 1 `Shifty` $(D, g, \delta, \mathcal{Q})$

1: $D_c, D_f \leftarrow$ `Partition`(D)
2: $\theta_c \leftarrow$ `TrainCandidate`$(D_c, g, \delta, \mathcal{Q})$
3: $u \leftarrow$ `HighConfUpperBound`$(g, D_f, \delta, \mathcal{Q})$
4: **return** θ_c **if** $u \leq 0$ **else** `NSF`

Algorithm 1 presents high-level pseudocode for classification algorithms that provide high-confidence fairness guarantees under demographic shift. `TrainCandidate` implements the candidate selection step, and `HighConfUpperBound` implements the high-confidence upper bound on $g'(\theta_c)$ that is used to determine the output of the algorithm and to establish its theoretical guarantees. In the following sections, we describe the candidate selection and fairness test steps in detail. Because it is the fairness test that causes `Shifty` to satisfy Property B, we outline this component first.

3.1 THE SHIFTY FAIRNESS TEST

Given a candidate model, θ_c, `Shifty`'s fairness test consists of computing a $(1-\delta)$-confidence upper bound on $g'(\theta_c)$, which measures the prevalence of unfair behavior once θ_c is deployed. Therefore, the primary challenge associated with designing an algorithm that satisfies Property B is computing a

valid high-confidence upper bound on $g'(\theta_c)$ given $\mathcal{Q}$, the description of the possible demographic distributions that might be encountered upon deployment.

Here, we propose a strategy for computing high-confidence upper bounds on $g'(\theta_c)$, starting with the simpler case in which the exact demographic shift is known and then extending this approach to the case in which the shift is unknown. In both settings, we show how to compute these bounds using appropriate confidence intervals. Our strategy is general and can be applied starting with many different confidence intervals. For illustrative purposes, we derive bounds based on inversion of the Student's t-test (Student, 1908), leading to an implementation of `Shifty` which we call `Shifty-ttest`. Importantly, our use of the Student's t-test implies that the resulting high-confidence upper bounds only hold exactly if $\Pr(H'|\xi)$ is a normal distribution.

3.1.1 Known Demographic Shift

To begin, we consider the task of assessing unfair behavior when there is no demographic shift. Given n i.i.d. samples $\{(H_i, \xi_i)\}_{i=1}^n$, where $H_i := h(X_i, Y_i, S_i, \theta)$ and $\xi_i := c(X_i, Y_i, S_i, \theta)$, one can derive $U_{\texttt{ttest}}$, a function that computes a high-confidence upper bound on $\mathbf{E}[H|\xi]$, by inverting the commonly-used Student's t-Test (Student, 1908):

$$\Pr\Big(\mathbf{E}[H|\xi] \leq U_{\texttt{ttest}}(g, D, \theta, \delta)\Big) \geq 1-\delta. \tag{4}$$

Specifically, if $\mathcal{I}_\xi$ are the indices of the samples for which $\xi_i = \texttt{True}$ and $N_\xi = |\mathcal{I}_\xi|$, then

$$U_{\texttt{ttest}}(g, D, \theta, \delta) := \frac{1}{N_\xi}\sum_{i\in\mathcal{I}_\xi} H_i + \frac{\sigma(g, D, \theta)}{\sqrt{N_\xi}} t_{1-\delta, N_\xi - 1},$$

where $t_{1-\delta, N_\xi-1}$ is the $1-\delta$ quantile of the Student's t distribution with $N_\xi - 1$ degrees of freedom, and σ computes the sample standard deviation using Bessel's correction,

$$\sigma(g, D, \theta) := \sqrt{\frac{1}{N_\xi - 1}\sum_{i\in\mathcal{I}_\xi}\left(H_i - \frac{1}{N_\xi}\sum_{i\in\mathcal{I}_\xi} H_i\right)^2}.$$

As $g(\theta) = \mathbf{E}[H|\xi] - \tau$, it follows that in the absence of demographic shift, Property A is satisfied if algorithm a only returns models satisfying $U_{\texttt{ttest}}(g, D, \theta, \delta) - \tau \leq 0$, and otherwise returns `NSF`.

To provide fairness guarantees that hold under demographic shift, we require $g'(\theta) \leq 0$ with high probability, where $g'(\theta) = \mathbf{E}[H'|\xi'] - \tau$. However, H' and ξ' are defined with respect to the demographic-shifted distribution, for which no samples are available. Thus, we seek a new random variable, $\hat{H}$, that can be computed from X, Y, and S, but which satisfies $\mathbf{E}[\hat{H}|\xi] = \mathbf{E}[H'|\xi']$. Under the demographic shift assumptions, the reweighted variable defined by $\hat{H} := \phi(T)H$ satisfies these requirements, where ϕ is an *importance weight* derived in Appendix C:

$$\phi(t) := \frac{\Pr(T'{=}t|\xi')}{\Pr(T{=}t|\xi)} = \frac{\Pr(\xi|T{=}t)\Pr(T'{=}t)}{\Pr(T{=}t|\xi)\sum\limits_{t'\in\mathcal{T}}\Pr(\xi|T{=}t')\Pr(T'{=}t')}, \tag{5}$$

for all $t \in \mathcal{T}$. Specifically, $\phi(T)$ acts as a scaling factor that reweights samples obtained during training so that their sample mean is an unbiased estimator of $\mathbf{E}[H'|\xi']$, as shown by Theorem 1.

Theorem 1. *Assume that* $\Pr(T{=}t) \geq 0$ *for all* $t \in \mathcal{T}$. *If the demographic shift properties hold, then the random variable* $\hat{H} := \phi(T)H$ *satisfies* $\mathbf{E}[\hat{H}|\xi] = \mathbf{E}[H'|\xi']$, *where* ϕ *is defined by (5).* ***Proof.*** *See Appendix C.*

Because $\hat{H}$ is defined with respect to pre-shift random variables, it is possible to generate i.i.d. samples of $\hat{H}$ during training, even when no samples from the deployment distribution are available. In particular, a set of i.i.d. observations of $\hat{H}$ is obtained by computing $\{\hat{H}_i\}_{i\in\mathcal{I}_\xi}$, where each $\hat{H}_i = \phi(T_i)h(X_i, Y_i, S_i, \theta)$. Using $\{\hat{H}_i\}_{i\in\mathcal{I}_\xi}$, we apply the inversion of the Student's t-test to derive $\hat{U}_{\texttt{ttest}}(g, D, \theta, \delta)$, which satisfies $\Pr(\mathbf{E}[\hat{H}|\xi] \leq \hat{U}_{\texttt{ttest}}(g, D, \theta, \delta)) \geq 1-\delta$. Specifically, if $\hat{\sigma}$ denotes the sample standard deviation of the reweighted observations, $\{\hat{H}_i\}_{i\in\mathcal{I}_\xi}$, then

$$\hat{U}_{\texttt{ttest}}(g, D, \theta, \delta) := \frac{1}{N_\xi}\sum_{i\in\mathcal{I}_\xi}\hat{H}_i + \frac{\hat{\sigma}(g, D, \theta)}{\sqrt{N_\xi}} t_{1-\delta, N_\xi-1}. \tag{6}$$

Since $\mathbf{E}[\hat{H}|\xi] = \mathbf{E}[H'|\xi']$ by Theorem 1, it follows that $\hat{U}_{\texttt{ttest}}$ is also a high-confidence upper bound suitable for assessing fairness after demographic shift:

$$\Pr\left(\mathbf{E}[\hat{H}|\xi] \leq \hat{U}_{\texttt{ttest}}(g, D, \theta, \delta)\right) = \Pr\left(\mathbf{E}[H'|\xi'] \leq \hat{U}_{\texttt{ttest}}(g, D, \theta, \delta)\right) \geq 1-\delta. \tag{7}$$

Recalling that $g'(\theta) := \mathbf{E}[H'|\xi'] - \tau$, it follows that $g'(\theta) \leq 0$ with high probability if $\hat{U}_{\texttt{ttest}}(g, D, \theta, \delta) - \tau \leq 0$, where each $\hat{H}_i$ implicitly depends on θ by the definition, $\hat{H}_i := \phi(T_i)h(X_i, Y_i, S_i, \theta)$. From (7), it is clear that if the pre-shift conditionals, $\Pr(\xi|T{=}t)$ and $\Pr(T{=}t|\xi)$, can be computed from the training data for all $t \in \mathcal{T}$, and if the post-shift demographic marginals, $\Pr(T'{=}t)$, are provided by the user during training, then a $(1-\delta)$-confidence upper bound $g'(\theta)$ can be computed even when data from the post-shift distribution is unavailable.

3.1.2 Unknown Demographic Shift

It is often unrealistic to assume that the post-shift marginal distribution is known exactly during training. To address this, we consider the setting in which the user provides a set, $\mathcal{Q} := \{(a_t, b_t)\}_{t\in\mathcal{T}}$, that contains non-empty intervals describing marginal distributions over T' that might be encountered after deployment. Given $\mathcal{Q}$, we compute high-confidence upper bounds on $g'(\theta)$ by determining the largest value of the high-confidence upper bound attained for a $q \in \mathcal{Q}$.

First, we parameterize the high-confidence upper bound in (7) to explicitly depend on a particular choice of post-shift demographic distribution, q. Formally, we define $\hat{U}_{\texttt{ttest}}(g, D, \theta, \delta; q)$ by replacing all occurrences of $\phi(T_i)$ in (6) with $\phi(T_i; q)$, given by

$$\phi(t; q) := \frac{\Pr(\xi|T{=}t)q_t}{\Pr(T{=}t|\xi)\sum_{t'\in\mathcal{T}}\Pr(\xi|T{=}t')q_{t'}}, \tag{8}$$

where $q_t = \Pr(T'{=}t)$ for one possible demographic-shifted demographic distribution. While the true post-shift marginal distribution, q^*, is assumed to be unknown, it is clear that if $q^* \in \mathcal{Q}$, then $\hat{U}_{\texttt{ttest}}(g, D, \theta, \delta; q^*) \leq \sup_{q\in\mathcal{Q}} \hat{U}_{\texttt{ttest}}(g, D, \theta, \delta; q)$. It follows that,

$$\Pr\left(\mathbf{E}[H'|\xi'] \leq \sup_{q\in\mathcal{Q}} \hat{U}_{\texttt{ttest}}(g, D, \theta, \delta; q)\right) \geq \Pr\left(\mathbf{E}[H'|\xi'] \leq \hat{U}_{\texttt{ttest}}(g, D, \theta, \delta; q^*)\right) \geq 1-\delta.$$

Consequently, if $g'(\theta) := \mathbf{E}[H'|\xi'] - \tau$, then an algorithm, a, can be designed to satisfy Property B by following Algorithm 1 and defining the fairness test to only return models when $\sup_{q\in\mathcal{Q}} \hat{U}_{\texttt{ttest}}(g, D, \theta, \delta; q) - \tau \leq 0$. We propose to use a numerical optimizer to approximate the supremum of $\hat{U}_{\texttt{ttest}}$ over $q \in \mathcal{Q}$. In our implementations, we use *simplicial homology optimization* (Endres et al., 2018), which converges to the global optima of non-smooth functions subject to equality and inequality constraints such as those defined by the condition $q \in \mathcal{Q}$.

3.2 Candidate Selection

The candidate selection step searches over a set of candidate models to find a model, θ_c, that achieves a small classification error. Because it is ultimately the fairness test that causes `Shifty` to satisfy Property B, this step can be implemented using any procedure for training a classifier without impacting the theoretical guarantees the algorithm provides. In practice, however, it is advantageous to select a candidate model that, in addition to minimizing error, also appears to behave fairly. In particular, if the model's accuracy is correlated with the definition of unfair behavior, a candidate selection procedure that solely optimizes accuracy will tend to select models that will fail the fairness test, causing the overall algorithm to frequently return `NSF`.

To mitigate this problem, we perform candidate selection by minimizing a loss consisting of two terms: one that measures classification error, and another that penalizes models that are predicted to fail the fairness test. Specifically, if $\mathbb{I}[\cdot]$ denotes the indicator function that returns 1 if its argument is `true` and 0 otherwise, then the candidate model is found by minimizing,

$$\ell_{\texttt{Shifty-ttest}}(g, D_c, \theta, \delta; \mathcal{Q}) := \sum_{(x,y)\in D_c} \frac{\mathbb{I}[\theta(x)\neq y]}{|D_c|} + \max\left(0, \sup_{q\in\mathcal{Q}} \hat{U}_{\texttt{ttest}}(g, D_c, \theta, \delta; q)\right). \tag{9}$$

3.3 IMPLEMENTATION DETAILS FOR SHIFTY-TTEST

Having described strategies for computing high-confidence upper bounds on the prevalence of unfair behavior under demographic shift, we now present the details for `Shifty-ttest`, an implementation `Shifty` based on the Student's t-Test. Specifically, `Shifty-ttest` is designed according to Algorithm 1, where the subroutine `TrainCandidate` minimizes (9), and `HighConfUpperBound` computes $\sup_{q \in \mathcal{Q}} \hat{U}_{\texttt{ttest}}(g, D, \theta, \delta; q) - \tau$. Consequently, `Shifty-ttest` produces binary classifiers that are fair under demographic shift by leveraging user-provided bounds on the marginal distribution of the demographic attribute after deployment.

4 EVALUATION

Our evaluation answers three research questions (RQ) regarding our algorithms' behavior for overcoming demographic shift compared to prior approaches that do not account for demographic shift.

[RQ1] Validity of fairness guarantees. In practice, do the models trained using `Shifty-ttest` or prior approaches adhere to high-probability fairness guarantees under demographic shift? **[RQ2] Model accuracy**. Is `Shifty-ttest` able to train models whose accuracy is comparable to those produced by prior approaches that do not account for demographic shift? **[RQ3] Data efficiency**. Does `Shifty-ttest` avoid returning `NSF` when reasonably sized training datasets are available?

Answering these research questions requires *multiple* pairs of datasets sampled from the same underlying distribution and exhibiting a consistent demographic shift. Therefore, we generate multiple training and deployment datasets by resampling from a fixed population using known distributions. This design ensures that failures—that is, instances in which an algorithm returns unfair models with a larger frequency than guaranteed—can be properly attributed to a failure of the algorithm instead of a violation of the user's assumptions on the demographic shift. In addition, oracle knowledge of the training and deployment distributions can be used to compute exact values for accuracy, $g(\theta)$, and $g'(\theta)$ during evaluation. We uniformly sample training datasets from the population, train models using each algorithm, and evaluate the fairness of each model after deployment by selecting a new distribution over the population that satisfies the user's assumptions about the demographic shift.

Dataset: We sample training and deployment data from a dataset of $43{,}303$ students from a university in Brazil (da Silva, 2019). The dataset consists of tuples (X, Y, S, T), where X is a vector of entrance exam scores, Y is a binary label representing if the student's GPA, between 0 and 4.0, was above 3.0, the demographic attribute, T, is the student's race, and S is the student's sex. This data allows us to train classifiers to predict academic success based on entrance exam scores, and assess whether fairness guarantees that protect against discrimination based on sex would continue to hold if demographic shift caused the marginal distribution over race to change during deployment.

4.1 EXPERIMENTS: KNOWN DEMOGRAPHIC SHIFT

First, we evaluate `Shifty-ttest`, prior Seldonian and Quasi-Seldonian algorithms (Thomas et al., 2019), Fairlearn (Agarwal et al., 2018), and Fairness Constraints (Zafar et al., 2017). We conducted experiments using these algorithms to enforce two fairness definitions—disparate impact and demographic parity. Figure 5 formalizes these definitions, which were selected as representative of real-world strategies for quantifying unfair behavior of classifiers (Verma and Rubin, 2018). To assess the impact of demographic shift, we simulated one possible choice of demographic shift: the left column of Figure 2 shows the uniformly sampled distribution used for training, and the middle column shows the demographic-shifted deployment distribution used in our experiments.

For each experiment, we conducted 15 trials while varying the amount of training data in order to identify dependencies on the training dataset's size. For each trial, we uniformly sampled a training dataset and trained models using our algorithm, standard Seldonian algorithms, Fairlearn, and Fairness Constraints. Importantly, while Fairlearn was not designed to avoid disparate impact, and Fairness Constraints was not designed to enforce either disparate impact or demographic parity constraints, we include them to test if they might empirically be fair under demographic shift, despite not being designed for this setting. After training, we recorded whether each algorithm produced a model or `NSF`, as well as the average accuracy and prevalence of unfair behavior for each trained model on both the training and demographic-shifted deployment distribution.

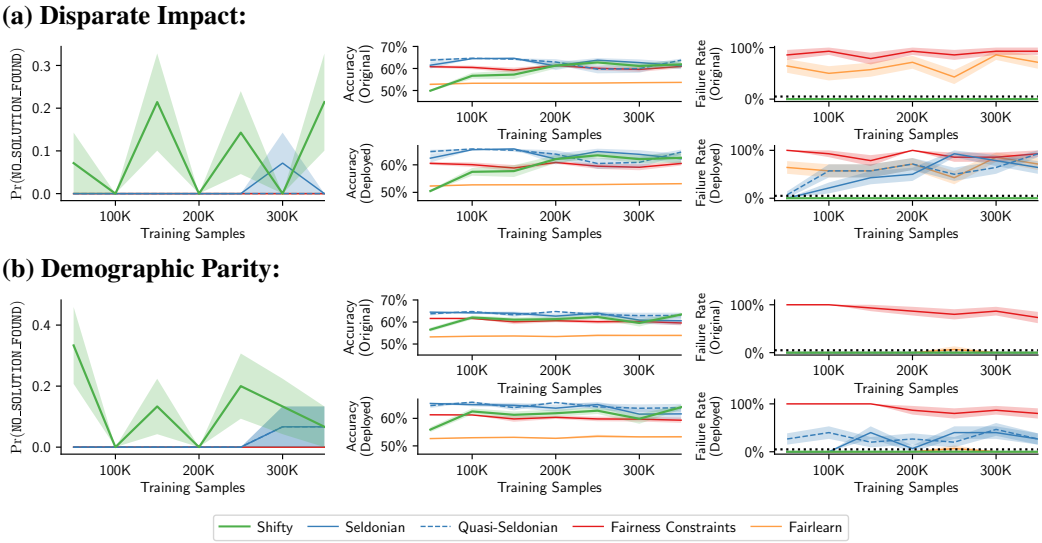

Figure 3: Results when enforcing fairness constraints under known demographic shift.

Figure 3 shows the results. For each definition of fairness, the leftmost plot shows the probability with which each algorithm returns `NSF`. To the right are two rows of plots showing evaluations of each algorithm on the training and deployment distributions, respectively. Within each row, the left plot displays the average accuracy of models returned by each algorithm when trained using various amounts of data. The right plot in each row displays the frequency with which each algorithm returns a model that is unfair, which we call the failure rate. The horizontal dashed line shows the maximum tolerance for unfair outcomes, δ=0.05, set when training `Shifty-ttest` and the Seldonian algorithms. Finally, standard error bars are shown using shaded regions.

Race	$\Pr(T{=}t)$	$\Pr(T'{=}t)$	$\mathcal{Q}$
$race_1$	0.006	0.300	(0.004, 0.304)
$race_2$	0.871	0.600	(0.610, 0.909)
$race_3$	0.054	0.050	(0.038, 0.338)
$race_4$	0.067	0.048	(0.047, 0.347)
$race_5$	0.002	0.002	(0.002, 0.301)

Figure 2: Marginal distributions over anonymized student race. The left column shows the distribution during training. The middle column shows the deployed marginal distribution used in our experiments for known demographic shift (Section 4.1). The right column shows the bounds on the marginal distribution used in our experiments for unknown demographic shift (Section 4.2).

Our experiments confirm that `Shifty` algrithms effectively avoid unfair behavior after demographic shift while prior algorithms do not. The failure rates of `Shifty-ttest` after demographic shift occurs (bottom right plots in Figure 3) are **always** below the tolerance set during training. However, while standard Seldonian algorithms were fair during training, they, along with Fairlearn and Fairness Constraints, frequently violate that fairness constraint after deployment. Interestingly, Fairlearn models were also fair when enforcing demographic parity constraints, though this was not true for disparate impact. However, Fairlearn does not provide fairness guarantees, unlike `Shifty-ttest`.

Next, when provided with a sufficient amount of training observations, `Shifty-ttest` provided guarantees of fairness without a noticeable loss in accuracy compared to the other baselines. The plots in the middle column of Figure 3 show that `Shifty-ttest` produced models that achieved the same accuracy as those trained using standard Seldonian algorithms, and outperformed both Fairlearn and Fairness Constraints for sufficiently large training datasets. This provides evidence that `Shifty-ttest` can be used to train fair models without incurring a significant loss in accuracy compared to algorithms that do not provide such guarantees.

Finally, we found that `Shifty-ttest`'s data efficiency is lower than that of alternative algorithms that ignore the impact of demographic shift. `Shifty-ttest` required slightly more training data than standard Seldonian algorithms to consistently avoid returning `NSF` (left plots in Figure 3) and to achieve comparable accuracy to those methods.

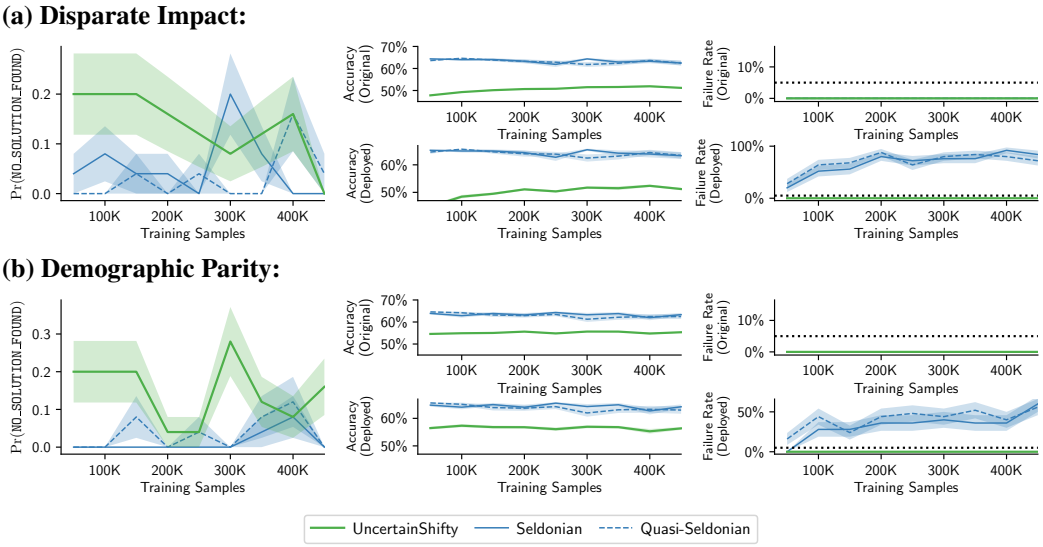

Figure 4: Results when enforcing fairness constraints under unknown demographic shift.

4.2 Experiments: Unknown Demographic Shift

Next, we repeat our experiments from Section 4.1, but assume that the user has provided upper and lower bounds on the marginal probability of encountering individuals of each race after deployment instead of specifying the shift exactly. We generated these bounds, shown in the rightmost column of Figure 2, by interpolating between the interval $(0, 1)$ and the singleton interval containing the marginal distribution over race under the training distribution, using an interpolation factor of 0.3. Given this specification, we applied each algorithm to train models using the same procedure as described in Section 4.1. To evaluate each model's performance after demographic shift, we performed a worst-case analysis and selected the deployment distribution over the population to satisfy the user's assumptions, but otherwise *maximize* the prevalence of unfair behavior when using that model to generate predictions (See Appendix E). Figure 4 in Appendix 6 shows the results for disparate impact and demographic parity, and Appendix F provides results based on three other definitions of unfair behavior. As in Section 4.1, we include five plots that show the probability of each algorithm returning `NSF`, average accuracy, and the failure rate of each algorithm before and after deployment.

These experiments confirm that `Shifty-ttest` can provide guarantees of fair behavior after deployment, even when the post-shift marginal distribution over race is not exactly known. While standard Seldonian algorithms are fair during training (Figure 4, top right), they are consistently unfair after demographic shift occurs. However, `Shifty-ttest` was unable to produce models that achieved the same accuracy as standard Seldonian algorithms. Some loss in accuracy is expected when enforcing constraints that hold under demographic shift, which may be more limiting than standard fairness constraints. Therefore, `Shifty-ttest` is effective when fairness guarantees are critical, but these guarantees may cause decreased classification accuracy in some settings.

5 Conclusion

In this paper, we proposed `Shifty`, a strategy for designing classification algorithms that provide high-confidence fairness guarantees that hold when the distribution of demographics changes between training and deployment. This setting poses significant challenges for existing fair ML algorithms, as the fairness guarantees they provide generally assume a constant data distribution. In contrast, `Shifty` algorithms allow the proportions of demographics to change after training, provided the user has some information describing this change. `Shifty` algorithms can be used when the new demographic proportions are known, or when these proportions are bounded in known intervals. Finally, we evaluated `Shifty-ttest`, an implementation of `Shifty` based on the Student's t-test, and found that the fairness guarantees it provides are empirically valid under demographic shift, wherein models trained using existing fair algorithms consistently produced unfair outcomes.

ETHICS STATEMENT

The primary goal of this research is to identify and overcome practical challenges, namely demographic shift, that might cause current algorithms to produce unfair outcomes. Our contributions provide tools needed for data scientists and ML practitioners to use ML in conscientious, ethical ways. However, we note that, when applying algorithms such as `Shifty`, it is important to carefully select the definition of unfair behavior to be appropriate for the problem at hand. While we evaluate `Shifty` using five standard definitions of unfair behavior for illustration, many definitions have been proposed and studied (Verma and Rubin, 2018), some of which cannot be simultaneously enforced (Chouldechova, 2017; Corbett-Davies et al., 2017; Kleinberg et al., 2017). Consequently, while `Shifty` offers a valuable tool for enforcing fairness constraints, ML practitioners should carefully study their target application, ideally working with domain experts and stakeholders, to ensure that the definitions they select meaningfully capture the unfair behaviors they wish to avoid.

Next, we note that the fairness guarantees provided by `Shifty` may fail to hold if one or more assumptions made by the algorithm do not hold. Most notably, if the demographic shift assumptions, (2) and (3), do not hold for the user's choice of demographic attribute T, then the guarantees provided by `Shifty` may not hold in practice. Similarly, if the user's specification of possible demographic shift, defined by the input $\mathcal{Q}$, does not accurately represent the demographic shift that occurs, then the guarantees provided by our algorithms may be invalidated. Thus, it is important that ML practitioners study the application domain in order to ensure that the inputs they supply to `Shifty` are appropriate.

Finally, `Shifty-ttest` is based on inversion of the Student's t-test, which only holds exactly if the observations, H_i are not normal. While this approximation error becomes smaller as more samples are used for training, the high-confidence guarantees provided by `Shifty` may not hold with $1-\delta$ probability when trained on very small datasets. Therefore, in applications for which very few training observations are available, ML practitioners should employ `Shifty` algorithms based on other, non-approximate confidence intervals.

REPRODUCIBILITY STATEMENT

To support efforts to reproduce our results, all code and data used in this paper will be made publicly available upon publication. Proofs of our theoretical results can be found in Appendix C and implementation details for our algorithms can be found in Section 3 and Appendix D. In addition, experimental details can be found in Section 4, full descriptions of the fairness definitions we tested are shown in Figure 5 of Appendix F, and additional results and experimental details are included in Appendix E.

REFERENCES

Alekh Agarwal, Alina Beygelzimer, Miroslav Dudík, John Langford, and Hanna Wallach. 2018. A reductions approach to fair classification. In *International Conference on Machine Learning (ICML)*, Vol. PMLR 80. Stockholm, Sweden, 60–69.

Julia Angwin, Jeff Larson, Surya Mattu, and Lauren Kirchner. 2016. Machine Bias. *ProPublica* (23 May 2016). `https://www.propublica.org/article/machine-bias-risk-assessments-in-criminal-sentencing`

Pranjal Awasthi, Matthäus Kleindessner, and Jamie Morgenstern. 2020. Equalized odds postprocessing under imperfect group information. In *International Conference on Artificial Intelligence and Statistics (AISTATS)*. Palermo, Italy.

Lorenzo Bruzzone and Mattia Marconcini. 2009. Domain adaptation problems: A DASVM classification technique and a circular validation strategy. *IEEE Transactions on Pattern Analysis and Machine Intelligence* 32, 5 (2009), 770–787.

Toon Calders and Sicco Verwer. 2010. Three naive Bayes approaches for discrimination-free classification. *Data Mining and Knowledge Discovery* 21, 2 (2010), 277–292.

A. Charnes and W. W. Cooper. 1959. Chance-constrained programming. *Management Science* 6, 1 (1959), 73–79.

Alexandra Chouldechova. 2017. Fair prediction with disparate impact: A study of bias in recidivism prediction instruments. *Big Data* 5, 2 (2017), 153–163.

Sam Corbett-Davies, Emma Pierson, Avi Feller, Sharad Goel, and Aziz Huq. 2017. Algorithmic decision making and the cost of fairness. In *ACM Conference on Knowledge Discovery and Data Mining (KDD)*. 797–806.

Bruno Castro da Silva. 2019. UFRGS Entrance Exam and GPA Data. `https://doi.org/10.7910/DVN/O35FW8`

Miroslav Dudík, Steven J Phillips, and Robert E Schapire. 2006. Correcting sample selection bias in maximum entropy density estimation. In *Advances in Neural Information Processing Systems*. 323–330.

Cynthia Dwork, Moritz Hardt, Toniann Pitassi, Omer Reingold, and Richard Zemel. 2012. Fairness through awareness. In *Innovations in Theoretical Computer Science Conference*. ACM, 214–226.

Stefan C Endres, Carl Sandrock, and Walter W Focke. 2018. A simplicial homology algorithm for Lipschitz optimisation. *Journal of Global Optimization* 72, 2 (2018), 181–217.

Basura Fernando, Amaury Habrard, Marc Sebban, and Tinne Tuytelaars. 2014. Subspace alignment for domain adaptation. *arXiv preprint arXiv:1409.5241* (2014).

Boqing Gong, Yuan Shi, Fei Sha, and Kristen Grauman. 2012. Geodesic flow kernel for unsupervised domain adaptation. In *2012 IEEE Conference on Computer Vision and Pattern Recognition*. IEEE, 2066–2073.

Noah J. Goodall. 2016. Can you program ethics into a self-driving car? *IEEE Spectrum* 53, 6 (June 2016), 28–58. `https://doi.org/10.1109/MSPEC.2016.7473149`

Ian Goodfellow, Yoshua Bengio, and Aaron Courville. 2016. *Deep Learning*. MIT Press.

Raghuraman Gopalan, Ruonan Li, and Rama Chellappa. 2011. Domain adaptation for object recognition: An unsupervised approach. In *International Conference on Computer Vision*. IEEE, 999–1006.

Griggs v. Duke Power Co. 1971. 401 U.S. 424. `https://supreme.justia.com/cases/federal/us/401/424/`.

M. Hardt, E. Price, and N. Srebro. 2016. Equality of Opportunity in Supervised Learning. In *Conference on Neural Information Processing Systems (NIPS)*. Barcelona, Spain, 3323–3331.

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), 601–608.

Arja John Irani. 2015. *Utilizing Negative Policy Information to Accelerate Reinforcement Learning*. Ph.D. Dissertation. Georgia Institute of Technology.

Jon M. Kleinberg, Sendhil Mullainathan, and Manish Raghavan. 2017. Inherent Trade-Offs in the Fair Determination of Risk Scores. In *Innovations in Theoretical Computer Science Conference (ITCS)*, Vol. 67. Berkeley, CA, USA, 43:1–43:23. `https://doi.org/10.4230/LIPIcs.ITCS.2017.43`

Zachary C Lipton, Yu-Xiang Wang, and Alex Smola. 2018. Detecting and correcting for label shift with black box predictors. *arXiv preprint arXiv:1802.03916* (2018).

Blossom Metevier, Stephen Giguere, Sarah Brockman, Ari Kobren, Yuriy Brun, Emma Brunskill, and Philip S Thomas. 2019. Offline Contextual Bandits with High Probability Fairness Guarantees. In *Advances in Neural Information Processing Systems*. 14922–14933.

L. B. Miller and H. Wagner. 1965. Chance-constrained programming with joint constraints. *Operation Research* 13 (1965), 930–945.

Parmy Olson. 2011. The Algorithm That Beats Your Bank Manager. *CNN Money* March 15 (2011).

A. Prékopa. 1970. On probabilistic constrained programming. In *Princeton Symposium on Mathematical Programming*. 113–138.

Babak Salimi, Luke Rodriguez, Bill Howe, and Dan Suciu. 2019. Interventional Fairness: Causal Database Repair for Algorithmic Fairness. In *International Conference on Management of Data (SIGMOD)*. Association for Computing Machinery, Amsterdam, Netherlands, 793–810. https://doi.org/10.1145/3299869.3319901

Guido F Smits and Mark Kotanchek. 2005. Pareto-front exploitation in symbolic regression. In *Genetic Programming Theory and Practice II*. Springer, 283–299.

Student. 1908. The probable error of a mean. *Biometrika* (1908), 1–25.

Philip S. Thomas, Bruno Castro da Silva, Andrew G. Barto, Stephen Giguere, Yuriy Brun, and Emma Brunskill. 2019. Preventing undesirable behavior of intelligent machines. *Science* 366, 6468 (2019), 999–1004.

Sahil Verma, Michael Ernst, and René Just. 2021. Removing biased data to improve fairness and accuracy. *CoRR* abs/2102.03054 (2021). https://arxiv.org/abs/2102.03054

Sahil Verma and Julia Rubin. 2018. Fairness definitions explained. In *2018 IEEE/ACM International Workshop on Software Fairness (FairWare)*. IEEE, 1–7.

Bianca Zadrozny. 2004. Learning and evaluating classifiers under sample selection bias. In *Proceedings of the Twenty-First International Conference on Machine learning*. 114.

Muhammad Bilal Zafar, Isabel Valera, Manuel Gomez Rodriguez, and Krishna P. Gummadi. 2017. Fairness constraints: Mechanisms for fair classification. In *International Conference on Artificial Intelligence and Statistics (AISTATS)*. 797–806.

Xiang Zhang and Yann LeCun. 2017. Universum prescription: Regularization using unlabeled data. In *Thirty-First AAAI Conference on Artificial Intelligence*.

Fuzhen Zhuang, Zhiyuan Qi, Keyu Duan, Dongbo Xi, Yongchun Zhu, Hengshu Zhu, Hui Xiong, and Qing He. 2021. A Comprehensive Survey on Transfer Learning. *Proc. IEEE* 109, 1 (2021), 43–76. https://doi.org/10.1109/JPROC.2020.3004555

Daniel Zwillinger and Stephen Kokoska. 1999. *CRC standard probability and statistics tables and formulae*. CRC Press. 31 pages.

A Alternative Approaches for Achieving Fair Outcomes

Designing algorithms that meet fairness requirements with high-confidence can be seen as a type of constrained optimization problem. Algorithms that enforce *hard constraints* search for a predictive model within a feasible set of models satisfying the user's fairness constraints (Irani, 2015). Soft constraints on the objective function used to guide the search for a model can achieve fairness empirically (Zafar et al., 2017), and multi-objective methods can be used to satisfy multiple, potentially conflicting fairness objectives (Smits and Kotanchek, 2005). Unfortunately, algorithms based on hard or soft constraints alone do not provide high-confidence fairness guarantees.

Fair algorithms can be designed based on *chance-constrained programming* (Charnes and Cooper, 1959; Miller and Wagner, 1965; Prékopa, 1970), in which an objective function is optimized subject to a set of constraints. This formulation can provide high-confidence fairness guarantees but typically requires knowledge of the distribution of each variable used to quantify fairness, which is often impractical. In contrast, our methods provide assurances of fair behavior without these limiting assumptions.

We propose algorithms that produce fair models following the design principles outlined by Seldonian ML (Thomas et al., 2019) while accounting for demographic shift. Other ways to account for demographic shift include augmenting training data with synthetically generated variations or antagonistic examples, or explicitly regularizing the training objective using unlabeled data from the deployment distribution, which improves the models' generalization (Goodfellow et al., 2016; Zhang

and LeCun, 2017). However, these approaches do not tackle fairness and cannot be directly applied to provide fairness guarantees under demographic shift. If the change in distribution can be addressed by a transformation of either the features or response variables, methods can learn this transformation using access to data from the deployment distribution combined with assumptions, such as that the transformation is linear (Fernando et al., 2014; Gong et al., 2012; Gopalan et al., 2011), or by iteratively assigning predicted labels to unlabeled data from the deployment environment (Bruzzone and Marconcini, 2009). However, such approaches are ill-suited for establishing high-confidence fairness guarantees under demographic shift.

Approaches can account for the differences between the training and deployment environments by reweighting the contribution of each training observation according to the relative probability of encountering that observation. Such methods are effective at improving performance in classification (Zadrozny, 2004; Huang et al., 2006), density estimation (Dudík et al., 2006), and regression (Huang et al., 2006). While most of these approaches focus on accuracy, some have focused on safety constraints, such as fairness. Lipton et al. (2018) proposed correcting for covariate shift by using a reweighting scheme to compute intervals on a classifier's confusion rates and applying a correction step based on these intervals. This approach produces empirically fair models but does not provide high-confidence guarantees.

An algorithm's fairness can be improved by manipulating the underlying data, e.g., by removing data that violates fairness properties (Verma et al., 2021) or inserting data inferred using fairness properties (Salimi et al., 2019). These methods, however, do not provide guarantees. Fairness can also be enforced in post-processing (Awasthi et al., 2020) but, again, without guarantees.

B Data Partitioning

The data partitioning step ensures that the outcome of the candidate selection step is independent of the outcome of the fairness test, which is necessary to guarantee that the overall algorithm satisfies (1) under demographic shift.

To illustrate the requirement for this step, consider an algorithm that accepts a set of training observations, D, uses D to select a candidate model, θ_c, and finally uses D again to evaluate the high-confidence upper bound on $g'(\theta_c)$ to perform a fairness test. While this algorithm might appear to be fair, it is not guaranteed to satisfy Property B because the output of candidate selection, θ_c, is correlated with the outcome of the fairness test because the same input data, D, is used to perform both steps. Consequently, it is possible that candidate selection can consistently select models that cause the high-confidence upper bound used by the fairness test to under-estimate the value of $g'(\theta_c)$. By ensuring that candidate selection and the fairness test use independent sets of observations, this correlation is eliminated, causing the probability of the high-confidence upper bound failing to be no more than δ as required by Property B.

Finally, we note that one area of future research might consider the optimal way to split an input dataset into parts used for candidate selection, D_c, and for evaluating the fairness test, D_f. Specifically, increasing the size of D_c improves the ability of the candidate selection step to identify models that are accurate and generalize well, but reduces the size of D_f and makes the fairness test more difficult to pass. In our experiments, we split the input data evenly between D_c and D_f, but we hypothesize that there may be more effective techniques for determining the optimal splitting proportion.

C Proof of Theorem 1

Theorem 1. *Assume that* $\Pr(T{=}t) \geq 0$ *for all* $t \in \mathcal{T}$. *If the demographic shift properties hold, then the random variable* $\hat{H} := \phi(T)H$ *satisfies* $\mathbf{E}[H'|\xi'] = \mathbf{E}[\hat{H}|\xi]$, *where* ϕ *is defined by (5).*

Proof. First, we write $\mathbf{E}[H'|\xi']$ as a sum over expected values conditioned on the value of the demographic attribute by applying the law of total probability (Zwillinger and Kokoska, 1999):

$$\begin{aligned}\mathbf{E}[H'|\xi'] &= \sum_{t\in\mathcal{T}} \mathbf{E}[H'|\xi', T'{=}t]\,\Pr(T'{=}t|\xi') \\ &= \sum_{t\in\mathcal{T}} \mathbf{E}[H|\xi, T{=}t]\,\Pr(T'{=}t|\xi'). && \textit{(Using (3))}\end{aligned}$$

Here, the second line follows from the second demographic shift assumption, which states that for all $t \in \mathcal{T}$ and $x, y, s \in \mathcal{X} \times \mathcal{Y} \times \mathcal{S}$, $\Pr(X'{=}x, Y'{=}y, S'{=}s|T'{=}t) = \Pr(X{=}x, Y{=}y, S{=}s|T{=}t)$. Next, we multiply each term by $\Pr(T{=}t|\xi)/\Pr(T{=}t|\xi) = 1$, reorganize terms, and write the sum over $t \in \mathcal{T}$ as a single expected value:

$$\begin{aligned}\mathbf{E}[H'|\xi] &= \sum_{t\in\mathcal{T}} \mathbf{E}[H|\xi, T{=}t]\left(\frac{\Pr(T{=}t|\xi)}{\Pr(T{=}t|\xi)}\right)\Pr(T'{=}t|\xi') \\ &= \sum_{t\in\mathcal{T}} \mathbf{E}[H|\xi, T{=}t]\left(\frac{\Pr(T'{=}t|\xi')}{\Pr(T{=}t|\xi)}\right)\Pr(T{=}t|\xi) \\ &= \mathbf{E}\left[\phi(T)H|\xi\right],\end{aligned}$$

where

$$\phi(t) = \frac{\Pr(T'{=}t|\xi')}{\Pr(T{=}t|\xi)}.$$

Finally, we rewrite $\phi(t)$ to depend on the post-shift marginal distribution, $\Pr(T'{=}t)$, and the pre-shift conditional distributions, $\Pr(T{=}t|\xi)$ and $\Pr(C|T{=}t)$, for each $t \in \mathcal{T}$:

$$\begin{aligned}\phi(t) &= \frac{\Pr(T'{=}t|\xi')}{\Pr(T{=}t|\xi)} \\ &= \frac{\Pr(\xi'|T'{=}t)\Pr(T'{=}t)}{\Pr(T{=}t|\xi)\Pr(\xi')} && \textit{(Using Bayes' Theorem)} \\ &= \frac{\Pr(\xi|T{=}t)\Pr(T'{=}t)}{\Pr(T{=}t|\xi)\Pr(\xi')} && \textit{(Using (3))} \\ &= \frac{\Pr(\xi|T{=}t)\Pr(T'{=}t)}{\Pr(T{=}t|\xi)\sum_{t'\in\mathcal{T}}\Pr(\xi'|T'{=}t')\Pr(T'{=}t')} \\ &= \frac{\Pr(\xi|T{=}t)\Pr(T'{=}t)}{\Pr(T{=}t|\xi)\sum_{t'\in\mathcal{T}}\Pr(\xi|T{=}t')\Pr(T'{=}t').} && \textit{(Using (3))}\end{aligned}$$

□

D Enforcing Complex Fairness Definitions

In Section 3, we assume that the user-specified definition of unfair behavior, g, is defined by,

$$g(\theta) \coloneqq \mathbf{E}\big[\, H \,\big|\, \xi \,\big] - \tau. \tag{10}$$

However common fairness definitions such as the 80% rule (Griggs v. Duke Power Co., 1971) and equalized odds (Hardt et al., 2016) do not have the same form as (10). Nonetheless, these definitions can often be described by known *expressions* of multiple terms, where each term is a conditional or marginal expected value. Consider, for example, the definition of bias codified by *disparate impact* (Griggs v. Duke Power Co., 1971; Chouldechova, 2017; Zafar et al., 2017):

$$g_{\text{DI}}(\theta) \coloneqq 0.8 - \min\left\{\frac{\mathbf{E}[\theta(X)|S{=}\texttt{female}]}{\mathbf{E}[\theta(X)|S{=}\texttt{male}]}, \frac{\mathbf{E}[\theta(X)|S{=}\texttt{male}]}{\mathbf{E}[\theta(X)|S{=}\texttt{female}]}\right\}.$$

This definition is clearly not of the form described by (10); furthermore, it is challenging to estimate because i.i.d. samples of $\mathbf{E}[\theta(X)|S{=}\texttt{male}]$ and $\mathbf{E}[\theta(X)|S{=}\texttt{female}]$ cannot be simultaneously

computed given a single observation. Regardless, it is possible to compute valid high-confidence upper bounds on $g_{\text{DI}}(\theta)$ by leveraging the recursive bound-propagation described by Metevier et al. (2019). Importantly, we evaluate our proposed algorithms using complex but realistic fairness definitions by leveraging this strategy, and as a consequence, our results are influenced by this decision.

While we refer the reader to Metevier et al. (2019) for a more complete discussion of this strategy, we present a brief intuition here. First, we assume that the user's definition of g can be written as,

$$g(\theta) = f\big(\phi_1(\theta), ..., \phi_k(\theta)\big),$$

where each ϕ_i for $i \in \{1, ..., k\}$ denotes a *parameter* of the joint distribution of (X, Y, S), and f is some function of k arguments specified by the user. In our work, we assume that these parameters are each expressed as a conditional expected value analogous to (10) and that the expression defining f is provided by the user as text.

Following (Metevier et al., 2019), the expression for f is parsed into a tree structure representing the recursive application of various predefined operations. To construct a high-confidence upper bound on $g(\theta)$—or in this work, $g'(\theta)$—we first construct a set of confidence intervals on each parameter using the methods described in Section 3. Importantly, we apply the union bound to ensure that the set of confidence intervals on the parameters hold jointly with probability $1-\delta$. Next, these intervals are recursively propagated through the expression for f, where at each node of the computation tree, the interval for that node is computed by applying *interval arithmetic*, which describes the image of certain mathematical operations given intervals as their arguments. Since the root of the computation tree denotes the quantity $g'(\theta)$, the result of this recursive system is an interval containing the true value of $g'(\theta)$ with at least probability $1-\delta$.

A drawback of this approach is that it assumes that the intervals describing each parameter are independent, which is often false when considering demographic shift. This can lead to inflation of the resulting confidence interval on $g'(\theta)$. For example, suppose g is defined to measure bias based on sex, and consider two races. If, for individuals of one race, a certain sex is encountered much more often than other sexes, while for individuals of the second race, all sexes are encountered equally often, then a demographic shift that makes the first race more likely may cause certain parameters defining fairness to be highly correlated with others. By ignoring these dependencies, the approach presented in (Metevier et al., 2019) may produce significantly larger confidence intervals for $g'(\theta)$ compared to alternative approaches that leverage this dependency. For this reason, we consider this problem to be a strong candidate for future work, as it has the potential to improve the data efficiency of our methods as well as those proposed in existing work (Metevier et al., 2019; Thomas et al., 2019).

E Simulating and Evaluating Bounded Demographic Shift

Here, we describe our procedure for simulating the impact of demographic shift given a fixed population dataset when the exact deployment distribution is unknown. Intuitively, after generating a training dataset, we antagonistically select a new, non-uniform distribution over the population that satisfies the user's demographic shift assumptions—that is, that the marginal distribution over demographics is contained in $\mathcal{Q}$—but otherwise maximizes the prevalence of unfair behavior. Since the population and sampling distributions are known during evaluation, this oracle knowledge can be used to compute exact values for various statistics, such as expected classification accuracy and the value of $g'(\theta)$ for all models θ.

To make this procedure formal, let the population dataset be denoted by $\mathcal{D}_{pop}$:

$$\mathcal{D}_{pop} \coloneqq \{(x_i, y_i, s_i, t_i)\}_{i=1}^{n}.$$

Note that we do not refer to this set using the standard notation for random variables because in our experimental context the population is treated as a fixed, non-random population. To generate a random training dataset, D, we sample observations uniformly from $\mathcal{D}_{pop}$ with replacement. Specifically, if P denotes the uniform distribution over the observations in $\mathcal{D}_{pop}$, then training datasets are defined by $D \coloneqq \{(X_j, Y_j, S_j, T_j)\}_{j=1}^{n_0}$, where each $(X_j, Y_j, S_j, T_j) \sim P$.

Next, we generate a new distribution over the population that satisfies the user's assumptions but otherwise maximizes the prevalence of unfair behavior for a given model, which we denote by Q.

However, to comply with the user's assumptions about the demographic shift, Q must be selected carefully. The following theorem provides the conditions that Q must satisfy to achieve this.

Theorem 2. *Let P denote a uniform distribution over $\mathcal{D}_{pop} \coloneqq \{(x_i, y_i, s_i, t_i)\}_{i=1}^{n}$. Assume that the demographic attribute takes values in some set $\mathcal{T}$ and that each demographic $t \in \mathcal{T}$ occurs at least once in the population. Next, let each $q \in \mathcal{Q}$ denote a marginal distribution over $\mathcal{T}$, where q_t denotes the probability of encountering demographic t. Finally, let $\mathbb{N}_{\mathcal{D}_{pop}}[x, y, s, t]$ denote the number of observations in $\mathcal{D}_{pop}$ that are equal to (x, y, s, t) and let $\mathbb{N}_{\mathcal{D}_{pop}}[t]$ denote the number of observations that have demographic attribute equal to t. It follows that Q, defined below, is a distribution over $\mathcal{D}_{pop}$ that satisfies both of the demographic shift assumptions, and has a marginal distribution over demographics that is contained in $\mathcal{Q}$:*

$$Q(X{=}x, Y{=}y, S{=}s, T{=}t) = \frac{\mathbb{N}_{\mathcal{D}_{pop}}[x, y, s, t]}{\mathbb{N}_{\mathcal{D}_{pop}}[t]} q_t.$$

Proof. To show this result, we derive an expression for Q that has these properties by construction.

First, we expand the post-shift joint distribution using the laws of conditional probability:

$$Q(X{=}x, Y{=}y, S{=}s, T{=}t) = Q(X{=}x, Y{=}y, S{=}s|T{=}t)Q(T{=}t).$$

Next, we apply the second demographic shift assumption:

$$Q(X{=}x, Y{=}y, S{=}s, T{=}t) = P(X{=}x, Y{=}y, S{=}s|T{=}t)Q(T{=}t).$$

Then, we represent the conditional $P(X, Y, S|T{=}t)$ as a ratio using laws of conditional probability:

$$Q(X{=}x, Y{=}y, S{=}s, T{=}t) = \frac{P(X{=}x, Y{=}y, S{=}s, T{=}t)}{P(T{=}t)} Q(T{=}t).$$

Because P is a uniform distribution over $\mathcal{D}_{pop}$, it follows that the value of $P(X{=}x, Y{=}y, S{=}s, T{=}t)$ is simply the number of occurrences of (x, y, s, t) in $\mathcal{P}$ divided by the total number of samples in the population, n. Similarly, $P(T{=}t)$ is equal to the number of observations that have demographic attribute equal to t, divided by n. Since we assume that each demographic is observed in the population, it follows that $P(T{=}t) > 0$ for all $t \in \mathcal{T}$. Let $\mathbb{N}_{\mathcal{D}_{pop}}[x, y, s, t]$ denote the number of observations in $\mathcal{D}_{pop}$ that are equal to (x, y, s, t) and let $\mathbb{N}_{\mathcal{D}_{pop}}[t]$ denote the number of observations that have demographic attribute equal to t. It follows that for all observations, $(x, y, s, t) \in \mathcal{D}_{pop}$, we have

$$Q(X{=}x, Y{=}y, S{=}s, T{=}t) = \frac{\mathbb{N}_{\mathcal{D}_{pop}}[x, y, s, t]}{\mathbb{N}_{\mathcal{D}_{pop}}[t]} Q(T{=}t).$$

Finally, we define the marginal distribution of Q over demographics to be given by q:

$$Q(X{=}x, Y{=}y, S{=}s, T{=}t) = \frac{\mathbb{N}_{\mathcal{D}_{pop}}[x, y, s, t]}{\mathbb{N}_{\mathcal{D}_{pop}}[t]} q_t.$$

Since Q has the same conditional distribution given the demographic as P by construction, it satisfies the demographic shift assumptions. Furthermore, since the marginal distribution of Q over demographics is defined to be given by q, which satisfies $q \in \mathcal{Q}$, it also satisfies the user's assumptions about the demographic shift. □

Theorem 2 shows that, given a $q \in \mathcal{Q}$, it is straightforward to construct a distribution over the population which satisfies the demographic shift assumptions. Therefore, to select a distribution that maximizes the prevalence of unfair behavior for a given model, θ, we numerically optimize $g'(\theta)$ over $q \in \mathcal{Q}$ using simplical homology optimization (Endres et al., 2018) to determine the maximizing marginal distribution, q^*, and then define the final distribution over $\mathcal{D}_{\text{pop}}$ using Theorem 2.

Theorem 2 can also be used to compute exact values for various statistics of interest during evaluation, such as expected classification accuracy or the value of $g'(\theta)$ for all models, θ. For example, consider estimating the post-deployment classification accuracy of a model, θ, given by $\mathbf{E}_Q[\mathbb{I}[\theta(X){=}Y]]$. If $\bar{\mathcal{D}}_{pop}$ denotes the set of unique observations in $\mathcal{D}_{pop}$, then we have

$$\begin{aligned}
&\mathbf{E}_Q\big[\,\mathbb{I}[\theta(X){=}Y]\,\big] \\
&\quad= \sum_{(x,y,s,t)\in\bar{\mathcal{D}}_{pop}} \mathbb{I}[\theta(x){=}y]\, Q(X{=}x, Y{=}y, S{=}s, T{=}t) \\
&\quad= \sum_{(x,y,s,t)\in\bar{\mathcal{D}}_{pop}} \mathbb{I}[\theta(x){=}y] \tfrac{\mathbb{N}_{\mathcal{D}_{pop}}[x,y,s,t]}{\mathbb{N}_{\mathcal{D}_{pop}}[t]}\, Q(T{=}t).
\end{aligned}$$

Analogous expressions can be used to find exact values for post-shift model accuracy as well as the value of $g'(\theta)$ for all models θ.

F ADDITIONAL RESULTS FOR BOUNDED DEMOGRAPHIC SHIFT

Results for our experiments on enforcing fairness under unknown demographic shift, according to the principles of equal opportunity, equalized odds, and predictive equality, are shown in Figure 6.

Demographic Parity (Dwork et al., 2012; Calders and Verwer, 2010)

$$g_{\text{DP}}(\theta) \coloneqq \Big| \, \mathbf{E}[\theta(X)|S{=}\texttt{female}] - \mathbf{E}[\theta(X)|S{=}\texttt{male}] \, \Big| - 0.1$$

Disparate Impact (Griggs v. Duke Power Co., 1971; Chouldechova, 2017; Zafar et al., 2017)

$$g_{\text{DI}}(\theta) \coloneqq 0.8 - \min\left\{ \frac{\mathbf{E}[\theta(X)|S{=}\texttt{female}]}{\mathbf{E}[\theta(X)|S{=}\texttt{male}]}, \frac{\mathbf{E}[\theta(X)|S{=}\texttt{male}]}{\mathbf{E}[\theta(X)|S{=}\texttt{female}]} \right\}$$

Equal Opportunity (Hardt et al., 2016; Chouldechova, 2017)

$$g_{\text{EOp}}(\theta) \coloneqq \Big| \, \mathbf{E}[\theta(X)|Y{=}0, S{=}\texttt{female}] - \mathbf{E}[\theta(X)|Y{=}0, S{=}\texttt{male}] \, \Big| - 0.1$$

Equalized Odds (Hardt et al., 2016)

$$g_{\text{EOd}}(\theta) \coloneqq \begin{aligned} &\Big| \, \mathbf{E}[\theta(X)|Y{=}0, S{=}\texttt{female}] - \mathbf{E}[\theta(X)|Y{=}0, S{=}\texttt{male}] \, \Big| + \\ &\Big| \, \mathbf{E}[1{-}\theta(X)|Y{=}1, S{=}\texttt{female}] - \mathbf{E}[1{-}\theta(X)|Y{=}1, S{=}\texttt{male}] \, \Big| - 0.1 \end{aligned}$$

Predictive Equality (Chouldechova, 2017; Corbett-Davies et al., 2017)

$$g_{\text{PE}}(\theta) \coloneqq \Big| \, \mathbf{E}[1{-}\theta(X)|Y{=}1, S{=}\texttt{female}] - \mathbf{E}[1{-}\theta(X)|Y{=}1, S{=}\texttt{male}] \, \Big| - 0.1$$

Figure 5: Definitions of fairness used in the additional experiments presented in this section. These definitions were specified as text input and bounded with a recursive technique used by prior Seldonian algorithms (Metevier et al., 2019).

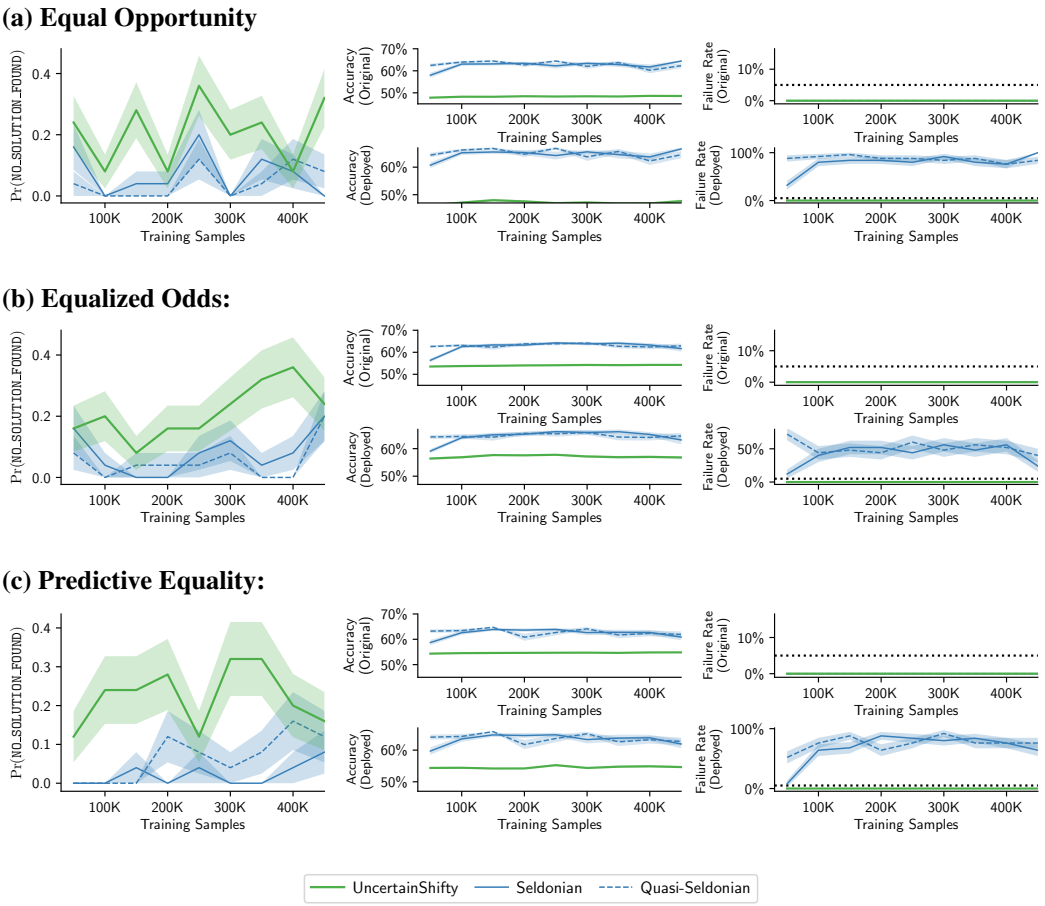

Figure 6: Results for experiments enforcing fairness constraints when the future marginal demographic distribution is unknown.