2024 IEEE International Symposium on Information Theory (ISIT) | 979-8-3503-8284-6/24/$31.00 ©2024 IEEE | DOI: 10.1109/ISIT57864.2024.10619400

Group Fairness with Uncertain Sensitive Attributes

Abhin Shah*, Maohao Shen*, Jongha Jon Ryu*, Subhro Das†, Prasanna Sattigeri†,
Yuheng Bu§, and Gregory W. Wornell*
*Massachusetts Institute of Technology {abhin,maohao,jongha,gww}@mit.edu
†MIT-IBM Watson AI Lab, IBM Research {subhro.das@ibm.com, psattig@us.ibm.com}
†University of Florida {buyuheng@ufl.edu}

***Abstract*—Learning a fair predictive model is crucial to mitigate biased decisions against minority groups in high-stakes applications. A common approach to learn such a model involves solving an optimization problem that maximizes the predictive power of the model under an appropriate group fairness constraint. However, in practice, sensitive attributes are often missing or noisy resulting in uncertainty, and solely enforcing fairness constraints on uncertain sensitive attributes can fall significantly short of achieving the level of fairness without uncertainty. To understand this phenomenon, we consider the problem of fair learning for Gaussian data and reduce it to a quadratically constrained quadratic problem (QCQP). To ensure a strict fairness guarantee given uncertain sensitive attributes, we propose a robust QCQP, and characterize its solution with an intuitive geometric understanding. When uncertainty arises due to limited labeled sensitive attributes, our analysis identifies non-trivial regimes where uncertainty incurs no performance loss while continuing to guarantee strict fairness. As an illustrative example of our analysis, we propose a bootstrap-based algorithm that applies beyond the Gaussian case. We demonstrate the value of our analysis and algorithm on synthetic as well as real-world data.**

I. Introduction

Achieving fairness in predictive modeling, whether in classification or regression, is crucial to avoid discriminatory decisions against marginalized groups. Although various problem formulations exist for ensuring fairness in model training, a widely adopted approach is to formulate an optimization problem that maximizes the model's predictive power while satisfying a group fairness constraint. The notion of group fairness [1] stipulates a certain (conditional) independence requirement involving the model prediction and the sensitive attribute. Then, the goal is to minimize the prediction loss while ensuring that the fairness loss, which measures the degree of violation of the independence requirement, is less than a pre-defined tolerance level ϵ, i.e.,

$$\min \text{Prediction Loss} \quad \text{s.t.} \quad \text{Fairness Loss} \leq \epsilon. \tag{1}$$

Typically, it is assumed that the learner has access to true sensitive attributes for every sample in training, but in reality, labeled sensitive attributes are often missing or noisy. For instance, labeling sensitive attributes may require additional annotation of existing datasets for which such labels were not originally collected. Even if available, the sensitive attribute information can be uncertain due to various reasons, such as noisy or unreliable responses from survey participants due to fear of disclosure or discrimination [2]. Moreover, privacy/legal regulations limit the use of labeled sensitive attributes, such as race or gender, which are protected by laws, e.g., EU's General Data Protection Regulation or California's Consumer Privacy Act. In such cases, privatized sensitive attributes, which are obtained by adding noise, may be the only available option.

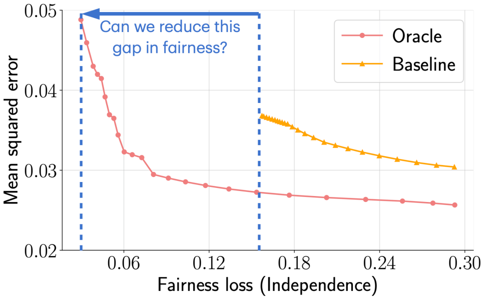


Fig. 1: Error vs. fairness on Crime data for oracle and baseline methods that enforce fairness using true and uncertain sensitive attributes, respectively. The baseline falls short of achieving the same range of fairness as the oracle.

In such scenarios, estimating the fairness loss in (1) using uncertain sensitive attributes, as if correct, can lead to a model that does not accurately capture target fairness. Figure 1 shows the trade-off between prediction (measured by mean squared error) and fairness (measured as violation of independence between predictions and sensitive attributes) obtained by varying ϵ in (1) for Crime data [3]. The oracle (in red) with access to true sensitive attributes during training, denoted by $\mathcal{D}^{\text{oracle}}$, enforces the constraint: Fairness Loss$(\mathcal{D}^{\text{oracle}}) \leq \epsilon$, and covers a wide range of fairness levels. By contrast, the baseline (in orange) with access to sensitive attributes, corrupted with Gaussian noise, during training, denoted by $\mathcal{D}^{\text{uncertain}}$, enforces the constraint: Fairness Loss$(\mathcal{D}^{\text{uncertain}}) \leq \epsilon$, but cannot achieve fairness below a threshold, i.e., it provides less control over attainable fairness compared to the oracle.

Contributions. We propose a method to learn predictive models with uncertain sensitive attributes, targeting applications where violating a fairness threshold incurs a significant cost.

1. We formulate the problem of fair learning for Gaussian data, with a focus on the independence notion of fairness. Using the principle of information bottleneck, we reduce a specific instance of this problem to a quadratically constrained quadratic problem (QCQP) when the true sensitive attributes are available. Given the uncertainty in sensitive attributes, we robustify the QCQP to provide a strict fairness guarantee and fully characterize the solution of the robust QCQP. Notably,

in certain cases of randomly missing sensitive attributes, our robust QCQP can achieve strict fairness without any performance loss, which we refer to as *free fairness*.

2. We illustrate the usefulness of our analysis by proposing `Bootstrap-S`, an algorithm that uses a bootstrap approach to impose `S` additional constraints to the optimization in (1), for some parameter `S`. For $i \in [\mathtt{S}]$, constraint i requires Fairness Loss($\mathcal{D}_i^{\text{uncertain}}$) $\leq \epsilon$, where $\mathcal{D}_i^{\text{uncertain}}$ is a collection of a fixed number of random subsamples of the uncertain sensitive attributes $\mathcal{D}^{\text{uncertain}}$, i.e., `Bootstrap-S` aims to
$$\begin{aligned}&\min \text{ Prediction Loss} \quad \text{s.t.}\\ &\qquad \text{Fairness Loss}(\mathcal{D}^{\text{uncertain}}) \leq \epsilon \text{ and,}\\ &\qquad \text{Fairness Loss}(\mathcal{D}_i^{\text{uncertain}}) \leq \epsilon \text{ for all } i \in [\mathtt{S}].\end{aligned}$$
3. `Bootstrap-S` is applicable to various settings, e.g., classification and regression tasks, discrete and continuous sensitive attributes, and group fairness notions of independence and separation. We empirically show that `Bootstrap-S` achieves fairness comparable to the oracle without much sacrificing the predictive power of the baseline across various synthetic and real data.

Here, we focus on regression tasks, continuous sensitive attributes, independence notion of fairness, and uncertainty due to noise as well as missingness. In the longer version of this work [4], we also consider classification tasks, discrete sensitive attributes, and the separation notion of fairness. We defer all the proofs to this longer version of our work.

Related work. Several approaches have been proposed to handle noisy sensitive attributes [5, 6, 7, 8, 9, 10]. However, they focus solely on classification with discrete sensitive attributes and assume a particular noise model, e.g., the flipping noise model where the true sensitive attribute is flipped with some fixed probability. By contrast, our approach applies to regression and continuous sensitive attributes. Moreover, we do not consider a specific noise model.

II. Problem Formulation

Suppose $\mathbf{x}$ represents d-dimensional input features defined on the alphabet $\mathcal{X}$, while y and e denote 1-dimensional target and sensitive attribute defined on the alphabets $\mathcal{Y}$ and $\mathcal{E}$, respectively. Fair supervised learning seeks to find a predictor $f : \mathcal{X} \to \mathcal{Y}$ that (a) accurately estimates the target variable for new input features and (b) avoids discrimination based on the sensitive attribute. To achieve this, we are given (a) a loss function $\ell : \mathcal{Y} \times \mathcal{Y} \to \mathbb{R}_+$, where $\ell(y, f(\mathbf{x}))$ measures the disagreement between the target variable and its prediction, and (b) a fairness measure $\Phi : \mathcal{Y} \times \mathcal{Y} \times \mathcal{E} \to \mathbb{R}_+$, where $\Phi(y, f(\mathbf{x}), e)$ measures the level of discrimination of f. Given a fairness target $\epsilon \geq 0$ and a class of predictors $\mathcal{F}$, the goal is to find an $f \in \mathcal{F}$ that minimizes the expected loss ℓ, subject to the fairness measure Φ being small:

$$f^* \in \arg\min_{f \in \mathcal{F}} \mathbb{E}\big[\ell(y, f(\mathbf{x}))\big] \text{ s.t. } \Phi(y, f(\mathbf{x}), e) \leq \epsilon. \tag{2}$$

For ease of notation, hereon, we define $f \triangleq f(\mathbf{x})$. The choice of ℓ depends on the specific alphabet $\mathcal{Y}$. In this work, we focus on regression tasks, where $\mathcal{Y}$ is $\mathbb{R}$, and we use the mean squared error (MSE) loss, defined as $\ell(y, f) = (y - f)^2$.

Choice of Φ. To design Φ, it is important to establish what is meant by a perfectly fair predictor, i.e., $\epsilon = 0$ in (2). Typically, perfect fairness is described in terms of statistical independence. The independence fairness criterion, also called *demographic parity*, demands that $f \perp\!\!\!\perp e$, meaning that predictions should not reveal any information about sensitive attributes.

Achieving perfect fairness is not feasible when learning a predictor from finite training samples [11]. In practice, one often works with measures of approximate fairness by choosing $\epsilon > 0$ in (2), and varying ϵ to find a balance between fairness and accuracy. As perfect fairness measures assert that certain random variables should be independent, a natural way to measure approximate fairness is to use divergence that measures the degree of independence between these variables. Recently, χ^2-divergence has emerged as an effective measure of approximate fairness [12]. Following this, we adopt χ^2-divergence as our measure of the degree of independence, i.e., $\Phi(y, f, e) = \chi^2\left(p_{e,f} \,\|p_e p_f\right)$ where $p_{e,f}, p_e$, and p_f are marginal distributions of (e, f), e, and f, respectively. However, when the data is Gaussian, we use a related but different analytically convenient divergence (introduced later).

A. Uncertain sensitive attributes

Typically, N independent and identically distributed (i.i.d.) samples of $(\mathbf{x}, y, e)$ are assumed to be available, denoted by $\mathcal{D}^{(o)} \triangleq \{\mathbf{x}^{(i)}, y^{(i)}, e^{(i)}\}_{i \in [N]}$. Then, the objective in (2) is estimated using the subset $\mathcal{D}^{(p)} \triangleq \{\mathbf{x}^{(i)}, y^{(i)}\}_{i \in [N]}$ while the constraint is estimated using an appropriate subset of $\mathcal{D}^{(o)}$ depending on the functional form Φ. We denote these estimates by $\mathbb{E}_{\mathcal{D}^{(p)}}\big[\ell(y, f)\big]$ and $\Phi_{\mathcal{D}^{(o)}}(y, f, e)$. We assume that N is sufficiently large and ignore any errors in these estimates to focus on errors due to uncertainty in sensitive attributes.

When dealing with uncertain sensitive attributes, access to $\mathcal{D}^{(o)}$ may not be possible. To account for such uncertainty, we assume access to $\mathcal{D}^{(p)}$ as well as $n \leq N$ (potentially noisy) labeled sensitive attributes $\mathcal{D}^{(u)} \triangleq \{\mathbf{x}^{(i)}, y^{(i)}, \hat{e}^{(i)}\}_{i \in [n]}$. For $i \in [N]$, if $\hat{e}^{(i)} \neq e^{(i)}$, then sensitive attribute $\hat{e}^{(i)}$ is noisy. Further, if $n < N$, then sensitive attributes $\{e^{(i)}\}_{i=n+1}^{N}$ are missing. Then, the goal of fair learning with uncertain sensitive attributes is to solve the optimization in (2) with access to $\mathcal{D}^{(p)}$ and $\mathcal{D}^{(u)}$. While this is an intuitively appealing goal, simply computing the constraint in (2) with $\mathcal{D}^{(u)}$ may be sub-optimal as discussed in Section I. That is, a predictor f satisfying $\Phi_{\mathcal{D}^{(u)}}(y, f, e) \leq \epsilon$ may not necessarily satisfy $\Phi_{\mathcal{D}^{(o)}}(y, f, e) \leq \epsilon$. To address this issue and gain some insight, we first consider the case where $(\mathbf{x}, y, e, f)$ is jointly Gaussian.

B. Gaussian setting

For ease of exposition, consider zero-mean Gaussian variables and assume that $p_{\mathbf{x},y}$ is known or can be learned from $\mathcal{D}^{(p)}$. Let the predictor f be such that $f = \mathbb{E}[y|u]$ where $u \triangleq u(\mathbf{x})$ is a Gaussian representation of the features chosen such that $u \perp\!\!\!\perp e$. Then, $f \perp\!\!\!\perp e$ follows from the data processing inequality. We measure the degree of independence between u and e using $\overline{D}$-divergence, a second-order approximation of Kullback–Leibler divergence, introduced by [13].

Definition 1. *The $\overline{D}$-divergence between zero-mean Gaussian random vectors $\mathsf{v} \sim p_{\mathsf{v}} = \mathcal{N}(\boldsymbol{0}, \boldsymbol{\Sigma}_v)$ and $\mathsf{w} \sim p_{\mathsf{w}} = \mathcal{N}(\boldsymbol{0}, \boldsymbol{\Sigma}_w)$, with $\|\cdot\|_{\mathrm{F}}$ denoting the Frobenius norm, is given by*

$$\overline{D}(p_{\mathsf{v}}\|p_{\mathsf{w}}) \triangleq \frac{1}{2}\|\boldsymbol{\Sigma}_w^{-1/2}(\boldsymbol{\Sigma}_v - \boldsymbol{\Sigma}_w)\boldsymbol{\Sigma}_w^{-1/2}\|_{\mathrm{F}}^2.$$

For these choices, the optimization in (2) reduces to learning a Gaussian variable u such that

$$u^* \in \operatorname*{arg\,min}_{u:\ \overline{D}(p_{e,u}\|p_e p_u)\le\epsilon} \mathbb{E}\big[(y - \mathbb{E}[y|u])^2\big]. \tag{3}$$

Next, to reformulate (3) into a quadratically constrained quadratic program (QCQP), we utilize the notion of canonical correlation matrices (CCMs) defined by [13].

Definition 2. *The canonical correlation matrix (CCM) between jointly Gaussian random vectors $\mathsf{v} \sim \mathcal{N}(\boldsymbol{0}, \boldsymbol{\Sigma}_v)$ and $\mathsf{w} \sim \mathcal{N}(\boldsymbol{0}, \boldsymbol{\Sigma}_w)$ is given by $\boldsymbol{b}_{vw} \triangleq \boldsymbol{\Sigma}_{vv}^{-1/2}\boldsymbol{\Sigma}_{vw}\boldsymbol{\Sigma}_{ww}^{-1/2}$, where $\boldsymbol{\Sigma}_{vw}$ is the cross-covariance matrix between v and w.*

The $\overline{D}$-divergence is conveniently represented by CCMs. By drawing connections to information bottleneck and semi-definite programming [14], we show an equivalence between (3) and a QCQP that uses CCMs.

Theorem 1 (Gaussian Fair Learning $\iff$ QCQP)**.** *The optimization problem in (3) is equivalent to*

$$\max_{\mathbf{a}\in\mathcal{B}_d(0,1)} \big\langle \mathbf{a}, \mathbf{b}_{yx}\big\rangle^2 \quad \textit{s.t.} \quad \big\langle \mathbf{a}, \mathbf{b}_{ex}\big\rangle^2 \le \varepsilon, \tag{4}$$

where $\mathcal{B}_d(0,1)$ denotes the d-dimensional ℓ_2 ball centered at 0 with radius 1, $\langle\cdot,\cdot\rangle$ denotes the inner product, and $\mathbf{a}$ plays the role of $\mathbf{b}_{ux}$.

We note that $\mathbf{a}$ in (4) has the same dimension as $\mathbf{x}$, i.e., d. Next, we show that any d-dimensional QCQP in (4) can be mapped to a 2-dimensional QCQP. In particular, we show that the projection of any $\mathbf{a} \in \mathcal{B}_d(0,1)$, satisfying the constraint in (4), onto the subspace spanned by $\mathbf{b}_{yx}$ and $\mathbf{b}_{ex}$ preserves the value of the objective and continues to satisfy the constraint.

Proposition 1 ($d = 2$ suffices for QCQP)**.** *An optimal solution $\mathbf{a}^\star$ of the QCQP in (4) lies in the subspace spanned by the vectors $\mathbf{b}_{yx}$ and $\mathbf{b}_{ex}$.*

In the longer version [4], we also characterize an optimal $\mathbf{a}^\star$ in (4) as a function of $\mathbf{b}_{yx}$, $\mathbf{b}_{ex}$, and ε using the resulting geometry for $d = 2$. Theorem 1 shows that the uncertainty in the canonical correlation matrix b_{ex} sufficiently captures the uncertainty in sensitive attributes, which we explore next.

III. Analysis

In this section, we provide a characterization of fair learning for Gaussian data, given some uncertainty in sensitive attributes. Specifically, we study how to *robustify* the QCQP in (4) to ensure a strict fairness guarantee with high probability. Then, to understand how this robustification affects the optimal objective and to reach a computationally inexpensive robust QCQP, we perform a series of constraint relaxations.

A. Robust QCQP

Let $\widehat{\mathbf{b}}_{ex}$ be an estimate of $\mathbf{b}_{ex}$, say obtained from $\mathcal{D}^{(u)}$, such that $\|\mathbf{b}_{ex} - \widehat{\mathbf{b}}_{ex}\|_2 \le \tau$ (with probability $1-\delta$), for some $\tau \ge 0$,[1]. We denote this by $\mathbf{b}_{ex} \in \mathcal{B}_d(\widehat{\mathbf{b}}_{ex}, \tau)$. To achieve fairness as in (4) with probability $1-\delta$, in the worst case, $\langle \mathbf{a}, \mathbf{b}\rangle^2 \le \varepsilon$ should hold for all $\mathbf{b} \in \mathcal{B}_d(\widehat{\mathbf{b}}_{ex}, \tau)$. Then, the following robust optimization maximizes the desired objective while achieving fairness as in (4) (with probability $1-\delta$) without the precise knowledge of $\mathbf{b}_{ex}$:

$$\max_{\mathbf{a}\in\mathcal{B}_d(0,1)} \langle \mathbf{a}, \mathbf{b}_{yx}\rangle^2 \text{ s.t. } \langle \mathbf{a}, \mathbf{b}\rangle^2 \le \varepsilon,\ \forall \mathbf{b} \in \mathcal{B}_d(\widehat{\mathbf{b}}_{ex}, \tau). \tag{5}$$

Next, similar to Proposition 1, we show that any d-dimensional QCQP in (5) can be mapped to a 2-dimensional QCQP.

Proposition 2 ($d = 2$ suffices for robust QCQP)**.** *An optimal solution $\mathbf{a}^\star$ of the robust QCQP in (5) lies in the subspace spanned by the vectors $\mathbf{b}_{yx}$ and $\widehat{\mathbf{b}}_{ex}$.*

B. Relaxed Robust QCQP

Now, to characterize the solution of the robust QCQP in (5), we focus on $d = 2$ and use polar coordinates. To analyze the corresponding feasible space, we relax the uncertainty space $\mathcal{B}_2(\widehat{\mathbf{b}}_{ex}, \tau)$ from a ball to an annular sector. Formally, let $\widehat{\mathbf{b}}_{ex} \triangleq \widehat{r}_e(\cos\widehat{\theta}_e, \sin\widehat{\theta}_e)$ be the estimate of $\mathbf{b}_{ex} \triangleq r_e(\cos\theta_e, \sin\theta_e)$ such that $|r_e - \widehat{r}_e| \le \Delta$ and $|\theta_e - \widehat{\theta}_e| \le \phi$ with probability $1-\delta$ where $\Delta \triangleq \tau \ge 0$ and $\phi \triangleq \sin^{-1}(\tau/\|\widehat{\mathbf{b}}_{ex}\|_2) \in [0, \pi/2]$. In other words, given $\widehat{r}_e, \widehat{\theta}_e, \Delta$, and ϕ, with probability $1-\delta$,

$$\mathbf{b}_{ex} \in \mathcal{A}(\Delta,\phi) \triangleq \{\mathbf{b} = r(\cos\theta, \sin\theta) : \\ |r - \widehat{r}_e| \le \Delta \text{ and } |\theta - \widehat{\theta}_e| \le \phi\},$$

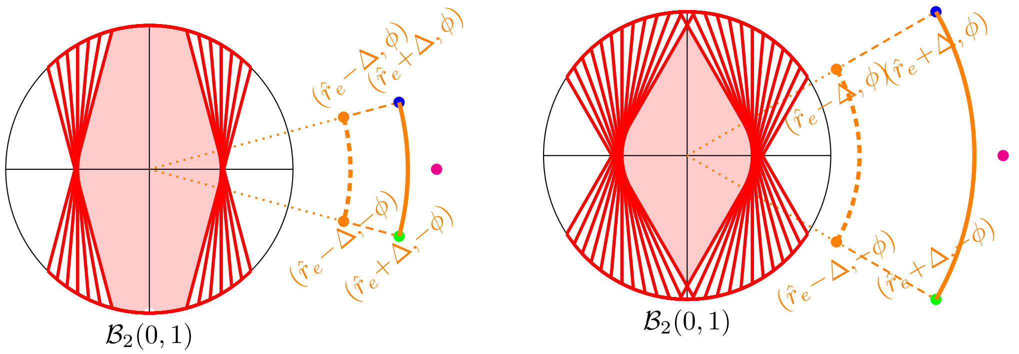
i.e., $\mathcal{A}(\Delta,\phi) (\supset \mathcal{B}_2(\widehat{\mathbf{b}}_{ex}, \tau))$ denotes the smallest annular sector around $\widehat{\mathbf{b}}_{ex}$ capturing our uncertainty in knowing $\mathbf{b}_{ex}$ (see Figure 2 where $\mathcal{A}(\Delta,\phi)$ is the shown in orange). Now, to achieve fairness as in (4) (with probability $1-\delta$), we constrain the robust QCQP in (5) as follows:

$$\max_{\mathbf{a}\in\mathcal{B}_2(0,1)} \langle \mathbf{a}, \mathbf{b}_{yx}\rangle^2 \text{ s.t. } \big\langle \mathbf{a}, \mathbf{b}\big\rangle^2 \le \varepsilon,\ \forall \mathbf{b} \in \mathcal{A}(\Delta,\phi). \tag{6}$$

In the longer version [4], we characterize the optimal $\mathbf{a}$ in (6) as a function of $\widehat{\mathbf{b}}_{ex}$, $\mathbf{b}_{yx}$, Δ, ϕ, and ε using the geometry for $d = 2$. As we show below, the constraint in (6) is equivalent to ensuring $\langle \mathbf{a}, \mathbf{b}\rangle^2 \le \varepsilon$ for all $\mathbf{b} \in \overline{\mathcal{A}}(\Delta,\phi)$ where $\overline{\mathcal{A}}(\Delta,\phi)$ is the arc on the boundary of the angular sector $\mathcal{A}(\Delta,\phi)$ with maximum radius (shown in solid orange in Figure 2). Intuitively, the feasible space corresponding to $(\widehat{r}_e + \Delta)(\cos\theta, \sin\theta)$ on the arc is equivalent to the intersection of the feasible spaces corresponding to $\{r(\cos\theta, \sin\theta)\}_{r\in[\widehat{r}_e-\Delta, \widehat{r}_e+\Delta]}$ in the sector.

Theorem 2 (Relaxed robust QCQP with infinite constraints)**.** *Let $\overline{\mathcal{A}}(\Delta,\phi) \triangleq \{\mathbf{b} : \mathbf{b} = (\widehat{r}_e + \Delta)(\cos\theta, \sin\theta)$ and $|\theta - \widehat{\theta}_e| \le$*

[1] For ease of the exposition, we assume $\tau \le \|\mathbf{b}_{ex}\|_2$.


$(\widehat{r}_e-\Delta,\phi)$
$(\widehat{r}_e+\Delta,\phi)$
$(\widehat{r}_e-\Delta,-\phi)$
$(\widehat{r}_e+\Delta,-\phi)$
$\mathcal{B}_2(0,1)$
$(\widehat{r}_e-\Delta,\phi)$
$(\widehat{r}_e+\Delta,\phi)$
$(\widehat{r}_e-\Delta,-\phi)$
$(\widehat{r}_e+\Delta,-\phi)$
$\mathcal{B}_2(0,1)$

(a) Feasible space when $\sqrt{\varepsilon} > (\widehat{r}_e + \Delta)\sin\phi$. (b) Feasible space when $\sqrt{\varepsilon} < (\widehat{r}_e + \Delta)\sin\phi$.

Fig. 2: Visualizing feasible space of the relaxed robust QCQP in Theorem 2 for $\varepsilon = 0.9$, $\widehat{r}_e = 1.6$, and $\widehat{\theta}_e = 0$. We set $\Delta = 0.2$ and $\phi = \pi/12$ for panel (a), and $\Delta = 0.4$ and $\phi = \pi/6$ for panel (b). Each point is shown in polar coordinates, i.e., a point (r, θ) denotes $(r\cos\theta, r\sin\theta)$. The annular sector $\mathcal{A}(\Delta, \phi)$ is the region enclosed by dashed lines, dashed arc, and solid arc in orange. The arc $\overline{\mathcal{A}}(\Delta, \phi)$ is the solid arc in orange. The shaded region is the feasible space. The points $\mathbf{b}_{ex}^{(1)}$, $\mathbf{b}_{ex}^{(2)}$, and $\mathbf{b}_{ex}^{(3)}$ from Theorem 3 are in magenta, blue and green, respectively.

$\phi\}$ be the arc on the boundary of $\mathcal{A}(\Delta, \phi)$ with maximum radius. Then, (6) *is equivalent to*

$$\max_{\mathbf{a}\in\mathcal{B}_2(0,1)} \langle \mathbf{a}, \mathbf{b}_{yx}\rangle^2 \ \textit{s.t.}\ \langle \mathbf{a}, \mathbf{b}\rangle^2 \le \varepsilon,\ \forall \mathbf{b} \in \overline{\mathcal{A}}(\Delta, \phi). \tag{7}$$

There is a phase transition in the nature of the feasible space of the QCQP in (7). Figure 2(a) and (b) illustrate the space for $\sqrt{\varepsilon} \ge (\widehat{r}_e + \Delta)\sin\phi$ and $\sqrt{\varepsilon} \le (\widehat{r}_e + \Delta)\sin\phi$, respectively.

C. Computationally feasible robust QCQP

While Theorem 2 simplifies the optimization in (6), the resulting optimization still has infinite constraints. In high dimensions, the projection onto the feasible space becomes costly. Below, we provide an approximation to the feasible space in (7) such that it has finitely many constraints. We note that alternative approximations are possible.

Theorem 3 (Computationally feasible robust QCQP with 3 constraints)**.** *Let $\mathbf{b}_{ex}^{(1)} = \frac{(\widehat{r}_e+\Delta)}{\cos\phi}(\cos\widehat{\theta}_e, \sin\widehat{\theta}_e)$, $\mathbf{b}_{ex}^{(2)} = (\widehat{r}_e + \Delta)(\cos(\widehat{\theta}_e + \phi), \sin(\widehat{\theta}_e + \phi))$, and $\mathbf{b}_{ex}^{(3)} = (\widehat{r}_e + \Delta)(\cos(\widehat{\theta}_e - \phi), \sin(\widehat{\theta}_e - \phi))$. Then, the feasible space of the optimization below is a subset of the feasible space of the optimization in* (7)*:*

$$\max_{\mathbf{a}\in\mathcal{B}_2(0,1)} \langle \mathbf{a}, \mathbf{b}_{yx}\rangle^2 \ \textit{s.t}\ \langle \mathbf{a}, \mathbf{b}_{ex}^{(i)}\rangle^2 \le \varepsilon \ \textit{ for all } i \in [3]. \tag{8}$$

We visualize $\mathbf{b}_{ex}^{(1)}$, $\mathbf{b}_{ex}^{(2)}$, and $\mathbf{b}_{ex}^{(3)}$ in Figure 2 (magenta, blue, and green, respectively), and note that $\mathbf{b}_{ex}^{(2)}$ and $\mathbf{b}_{ex}^{(3)}$ are the extreme points of the arc $\overline{\mathcal{A}}(\Delta, \phi)$. Then, intuitively, Theorem 3 uses $\mathbf{b}_{ex}^{(1)}$ to approximate the effect of points in-between $\mathbf{b}_{ex}^{(2)}$ and $\mathbf{b}_{ex}^{(3)}$ on $\mathcal{A}(\Delta, \phi)$. In the longer version [4], we provide a characterization of the optimal $\mathbf{a}$ in (8) as a function of $\widehat{\mathbf{b}}_{ex}$, $\mathbf{b}_{yx}$, Δ, ϕ, and ε as well as compare it with the optimal $\mathbf{a}$ in (7) using the geometry for $d = 2$. We note that solving the QCQP in (8) for $d = 2$ is straightforward using a standard convex optimization solver, e.g., the CVXPY library [15].

D. Sensitive attributes missing completely at random.

In the longer version [4], we express the uncertainty parameters Δ and ϕ as a function of n, and consider understanding how the optimal objective in (8) changes when the uncertainty set $\mathcal{A}(\Delta, \phi)$ changes. For concreteness, we consider uncertainty due to sensitive attributes missing completely at random. We analyze the power of each new labeled sensitive attribute, and characterize scenarios where the optimal performance of the robust QCQP (8) matches the optimal performance in (4) without having to collect any extra labeled sensitive attributes.

Corollary 1 (Free fairness)**.** *There exist problem instances of the robust QCQP in* (8) *where the uncertainty incurs no performance loss while achieving a strict fairness guarantee without requiring additional labeled sensitive attributes.*

IV. Bootstrap-S: An Illustrative Application

Next, we leverage our analysis to propose a generic algorithm that handles high-dimensional features and non-Gaussian data while accounting for uncertainty. At its core, the robust QCQP (Theorem 3) constructs an uncertainty set around the estimated canonical correlation matrix $\widehat{\mathbf{b}}_{ex} \triangleq \mathbf{b}_{ex}^{(0)}$ by imposing additional constraints to effectively addresses the unknown true $\mathbf{b}_{ex}$. An alternative perspective is to view the robust QCQP in (8) as

$$\max_{\mathbf{a}\in\mathcal{B}_2(0,1)} \langle \mathbf{a}, \mathbf{b}_{yx}\rangle^2 \ \text{s.t}\ \langle \mathbf{a}, \mathbf{b}_{ex}^{(i)}\rangle^2 \le \varepsilon \ \text{ for all } i \in \{0\} \cup [3].$$

Here, the constraint with $\mathbf{b}_{ex}^{(0)}$ is redundant due to the constraint with $\mathbf{b}_{ex}^{(1)}$, and $\{\mathbf{b}_{ex}^{(i)}\}_{i\in[3]}$ can be viewed as multiple estimates of $\widehat{\mathbf{b}}_{ex}$. For non-Gaussian data, we use a similar idea non-parametrically following the *bootstrap* procedure [16]. We refer to this method as `Bootstrap-S`.

Specifically, given uncertain sensitive attribute data $\mathcal{D}^{(u)}$, a fairness measure Φ, and a parameter $\mathtt{S}$: we draw $\mathtt{S}$ subsets $\mathcal{D}_1^{(u)}, \ldots, \mathcal{D}_{\mathtt{S}}^{(u)}$ of some size $k \in [n]$ from $\mathcal{D}^{(u)}$ uniformly

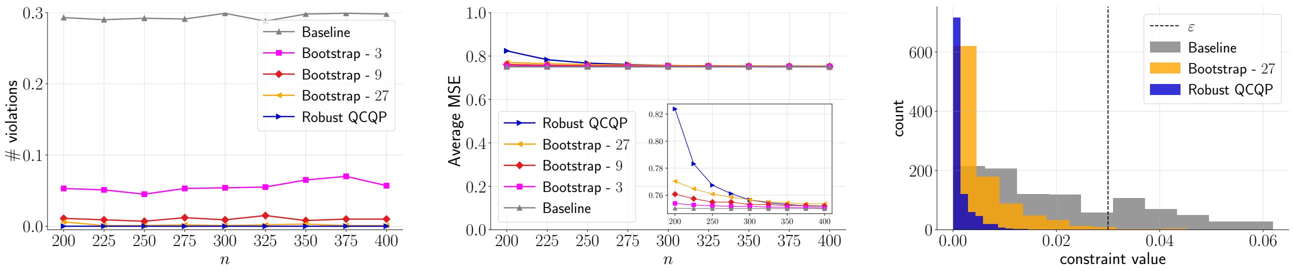


Fig. 3: The performance of robust QCQP in (8), `Bootstrap-S` with $\mathtt{S}\in\{3,9,27\}$, and `Baseline` for $d=2$ and $\varepsilon=0.025$. In the left column, we plot the fraction of violations of the true fairness constraint $\langle\mathbf{a},\mathbf{b}_{ex}\rangle^2\le\varepsilon$ vs. n; in the middle column, we plot average MSE vs. n; in the right column, we plot the histogram of the value of $\langle\mathbf{a},\mathbf{b}_{ex}\rangle^2$ over 1,000 trials for $n=250$.

at random with replacement. Then, we estimate the fairness measure using each of these subsets as well as $\mathcal{D}^{(u)}$, and impose the collection of S constraints $\{\Phi_{\mathcal{D}_i^{(u)}}(y,f,e)\le\epsilon\}_{i\in[\mathtt{S}]}$ together with $\Phi_{\mathcal{D}^{(u)}}(y,f,e)\le\epsilon$. In summary, we aim to solve:

$$\min_f \mathbb{E}_{\mathcal{D}^{(p)}}\left[\ell(y,f)\right] \quad \text{s.t} \quad \Phi_{\mathcal{D}^{(u)}}(y,f,e)\le\epsilon \tag{9}$$
$$\text{and } \Phi_{\mathcal{D}_i^{(u)}}(y,f,e)\le\epsilon \text{ for all } i\in[\mathtt{S}].$$

The high level idea is similar to bootstrap confidence intervals [17] allowing construction of better uncertainty set as number of subsamples S increase. Notice that (9) is a constrained optimization problem, which is non-trivial to solve in practice, especially for neural network training. Typically, this is addressed by adding the constraints as regularizers with hyperparameter to control the trade-off during optimization, i.e., min Prediction Loss + $\lambda\times$Fairness Loss. However, the performance can be sub-optimal as it depends on choice of λ. Instead, we consider the Lagrangian dual of (9) and optimize the resulting objective over the dual variables as in [18],

$$\min_f \max_{\lambda,\lambda_1,\cdots,\lambda_{\mathtt{S}}\ge0} \mathbb{E}_{\mathcal{D}^{(p)}}\left[\ell(y,f)\right]+\lambda\big(\Phi_{\mathcal{D}^{(u)}}(y,f,e)-\epsilon\big)$$
$$+\sum_{i\in[\mathtt{S}]}\lambda_i\big(\Phi_{\mathcal{D}_i^{(u)}}(y,f,e)-\epsilon\big). \tag{10}$$

V. Empirical Evaluation

Synthetic data. First, we show the efficacy of the robust QCQP in (8) in achieving strict fairness on synthetic Gaussian data. We generate data using a zero-mean Gaussian distribution with $d=2$ and covariance Σ_2^{fair} detailed in the longer version [4]. We present our results when uncertainty is due to sensitive attributes missing completely at random, and provide many additional experiments in the longer version [4]

We estimate $\widehat{\mathbf{b}}_{ex}$ using n samples of $(\mathbf{x},e)$ for various choices of n. Then, we compare the robust QCQP and `Bootstrap-S` applied to the QCQP in (4) (for various S) against `Baseline`, which solves the QCQP in (4) using $\widehat{\mathbf{b}}_{ex}$.

The results, averaged over 1000 random trials, are shown in Figure 3 (the error bars are too small to see). Robust QCQP always ensures no fairness violations, and its performance (in terms of average MSE) monotonically improves as n increases. Importantly, it does not incur any significant loss in the performance, demonstrating the free-fairness phenomenon in Corollary 1, say, $n\approx350$ onwards. `Bootstrap-S` well approximates the performance of robust QCQP and outperforms `Baseline` in terms of fairness violations. As alluded to earlier, `Bootstrap-S` achieves a better fairness criterion as we increase S by forming a more accurate uncertainty set, albeit with an increased computation.

Real data. Next, we test `Bootstrap-S` on the Crime data [3], a regression task to predict the number of crimes per 100K population in the U.S. The sensitive attribute is the percentage of people belonging to a particular race in the community. We detail the pre-processing steps in the longer version [4].

We train a two-layer neural network and use χ^2-divergence to impose the independence criterion (Section II). We induce uncertainty in every sensitive attribute by adding independent $\mathcal{N}(0,0.25)$ noise. Given a fairness target ϵ, we train a model over 50 independent trials of this noise and report average values. We sweep over 500 ϵ from 0.001 to 0.5, and plot the trade-off frontier using a moving average over 5 entries. We provide more details & experiments in the long version [4].

We report predictive power via MSE (lower is better) and fairness loss via χ^2-divergence (lower the better) on a held-out test set in Figure 4. We compare with `Baseline` which solves (10) with $\lambda_1,\cdots,\lambda_{\mathtt{S}}$ fixed to 0. For reference, we also compare with `Oracle` that has access to all the true sensitive attributes. `Bootstrap-S` (with $\mathtt{S}=5$) achieves fairness levels that are comparable to `Oracle` while maintaining a relatively high level of predictive power, similar to Corollary 1.

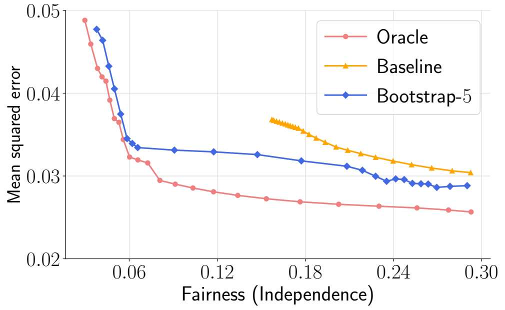


Fig. 4: Performance of `Baseline` and `Bootstrap-S` on the Crime data (the error bars are too small to see).

References

[1] S. Barocas, M. Hardt, and A. Narayanan, *Fairness and Machine Learning: Limitations and Opportunities*. fairmlbook.org, 2019, http://www.fairmlbook.org.

[2] I. Krumpal, "Determinants of social desirability bias in sensitive surveys: a literature review," *Quality & quantity*, vol. 47, no. 4, pp. 2025–2047, 2013.

[3] M. Redmond and A. Baveja, "A data-driven software tool for enabling cooperative information sharing among police departments," *European Journal of Operational Research*, vol. 141, no. 3, pp. 660–678, 2002.

[4] A. Shah, M. Shen, J. J. Ryu, S. Das, P. Sattigeri, Y. Bu, and G. W. Wornell, "Group fairness with uncertainty in sensitive attributes," *arXiv preprint arXiv:2302.08077*, 2023.

[5] A. Lamy, Z. Zhong, A. K. Menon, and N. Verma, "Noise-tolerant fair classification," *Advances in Neural Information Processing Systems*, vol. 32, 2019.

[6] P. Awasthi, M. Kleindessner, and J. Morgenstern, "Equalized odds postprocessing under imperfect group information," in *International Conference on Artificial Intelligence and Statistics*. PMLR, 2020, pp. 1770–1780.

[7] H. Mozannar, M. Ohannessian, and N. Srebro, "Fair learning with private demographic data," in *International Conference on Machine Learning*. PMLR, 2020, pp. 7066–7075.

[8] S. Wang, W. Guo, H. Narasimhan, A. Cotter, M. Gupta, and M. Jordan, "Robust optimization for fairness with noisy protected groups," *Advances in Neural Information Processing Systems*, vol. 33, pp. 5190–5203, 2020.

[9] L. E. Celis, L. Huang, V. Keswani, and N. K. Vishnoi, "Fair classification with noisy protected attributes: A framework with provable guarantees," in *International Conference on Machine Learning*. PMLR, 2021, pp. 1349–1361.

[10] L. E. Celis, A. Mehrotra, and N. Vishnoi, "Fair classification with adversarial perturbations," *Advances in Neural Information Processing Systems*, vol. 34, pp. 8158–8171, 2021.

[11] A. Agarwal, A. Beygelzimer, M. Dudík, J. Langford, and H. Wallach, "A reductions approach to fair classification," in *International Conference on Machine Learning*. PMLR, 2018, pp. 60–69.

[12] J. Mary, C. Calauzenes, and N. El Karoui, "Fairness-aware learning for continuous attributes and treatments," in *International Conference on Machine Learning*. PMLR, 2019, pp. 4382–4391.

[13] S.-L. Huang, A. Makur, G. W. Wornell, and L. Zheng, "On universal features for high-dimensional learning and inference," *arXiv preprint arXiv:1911.09105*, 2019.

[14] Y. Bu, T. Wang, and G. W. Wornell, "SDP methods for sensitivity-constrained privacy funnel and information bottleneck problems," in *2021 IEEE International Symposium on Information Theory (ISIT)*. IEEE, 2021, pp. 49–54.

[15] S. Diamond and S. Boyd, "Cvxpy: A python-embedded modeling language for convex optimization," *The Journal of Machine Learning Research*, vol. 17, no. 1, pp. 2909–2913, 2016.

[16] B. Efron, "Bootstrap methods: another look at the jackknife," in *Breakthroughs in statistics*. Springer, 1992, pp. 569–593.

[17] L. Wasserman, *All of nonparametric statistics*. Springer Science & Business Media, 2006.

[18] M. Lee, T. B. Hashimoto, and P. Liang, "Learning autocomplete systems as a communication game," *arXiv preprint arXiv:1911.06964*, 2019.