Group Fairness with Uncertainty in Sensitive Attributes

Anonymous Authors[1]

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. We demonstrate that solely enforcing fairness constraints on uncertain sensitive attributes can fall significantly short in achieving the level of fairness of models trained without uncertainty. To overcome this limitation, we propose a bootstrap-based algorithm that achieves the target level of fairness despite the uncertainty in sensitive attributes. The algorithm is guided by a Gaussian analysis for the independence notion of fairness where we propose a robust quadratically constrained quadratic problem to ensure a strict fairness guarantee with uncertain sensitive attributes. Our algorithm is applicable to both discrete and continuous sensitive attributes and is effective in real-world classification and regression tasks for various group fairness notions, e.g., independence and separation.

1. Introduction

Achieving fairness in predictive modeling, whether in classification or regression tasks, 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 (Kamishima et al., 2011; Zafar et al., 2017; Agarwal et al., 2018; Verma & Rubin, 2018; Gölz et al., 2019; Mehrabi et al., 2021; Castelnovo et al., 2022). The notion of group fairness (Barocas et al., 2019) 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 group unfairness, i.e., the degree of violation of the (conditional) 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 (Krumpal, 2013). Moreover, privacy and legal regulations often limit the use of labeled sensitive attributes, such as race or gender, which are protected by laws such as the 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. In such scenarios, estimating the fairness loss in (1) using uncertain sensitive attributes, as if correct, can lead to a model that does

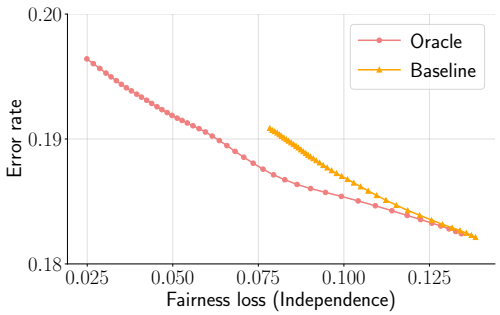


Figure 1. Error vs fairness on Adult data for oracle and baseline models that enforce fairness using true and uncertain sensitive attributes, respectively. The baseline falls short of achieving the same range of fairness as the oracle.

[1]Anonymous Institution, Anonymous City, Anonymous Region, Anonymous Country. Correspondence to: Anonymous Author <anon.email@domain.com>.

Preliminary work. Under review by the SCIS workshop.

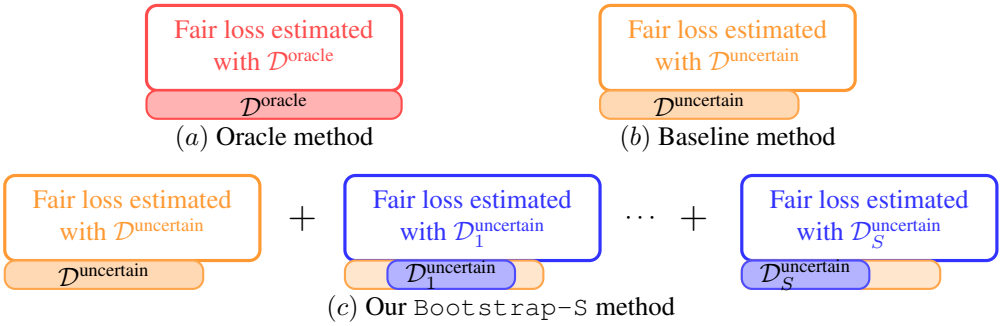


Figure 2. Comparing fairness loss estimation methods: (a) Oracle uses true sensitive attributes $\mathcal{D}^{\text{oracle}}$; (b) Baseline uses available uncertain sensitive attributes $\mathcal{D}^{\text{uncertain}}$ as if correct; (c) Bootstrap-S constrains the optimization with additional fairness losses estimated using subsamples $\mathcal{D}_i^{\text{uncertain}}, \forall i \in [\mathtt{S}]$. The horizontal bars illustrate these methods when uncertainty is due to missing sensitive attributes.

not accurately capture target fairness. Figure 1 shows trade-off between prediction (measured by error rate) and fairness (measured as violation of independence between predictions and sensitive attributes) obtained by varying ϵ in (1) for Adult data (Lantz, 2019). The oracle (in red) has access to true sensitive attributes, denoted by $\mathcal{D}^{\text{oracle}}$, enforces the fairness constraint: Fairness Loss$(\mathcal{D}^{\text{oracle}}) \leq \epsilon$, and covers a wide range of fairness levels. In contrast, the baseline (in orange) has access to a random $<1\%$ of the sensitive attributes, denoted by $\mathcal{D}^{\text{uncertain}}$, enforces the fairness constraint: Fairness Loss$(\mathcal{D}^{\text{uncertain}}) \leq \epsilon$, but is unable to achieve fairness below a threshold, i.e., the baseline provides less control over attainable fairness compared to the oracle.

As a result, for high-stakes applications where violating a fairness threshold incurs a significant cost, it is essential to develop a method that can learn fair models despite uncertainty in sensitive attributes.

Contributions. In this work, we propose a solution to fair learning with uncertain sensitive attributes.

1. We propose 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

$$\begin{aligned} &\min \text{Prediction Loss} \\ \text{s.t.} \quad &\text{Fairness Loss}(\mathcal{D}^{\text{uncertain}}) \leq \epsilon \text{ and,} \\ &\text{Fairness Loss}(\mathcal{D}_i^{\text{uncertain}}) \leq \epsilon \text{ for all } i \in [\mathtt{S}]. \end{aligned} \tag{2}$$

   We illustrate Bootstrap-S in Figure 2 where it is contrasted with the oracle model that constrains the fairness loss estimated using true sensitive attributes $\mathcal{D}^{\text{oracle}}$ and the baseline model that constrains the fairness loss estimated using available uncertain sensitive attributes $\mathcal{D}^{\text{uncertain}}$ as if they are correct.
2. We analyze fair learning for Gaussian data with independence notion of fairness which serves as a motivation for Bootstrap-S. We begin by reducing a specific instance of this problem to a quadratically constrained quadratic problem (QCQP) when true sensitive attributes are available. Then, given uncertain sensitive attributes, we robustify the QCQP to provide a strict fairness guarantee. Notably, when uncertainty arises due to randomly missing sensitive attributes, the robust QCQP can achieve strict fairness without any performance loss, which we refer to as *free fairness*.
3. We showcase the practical effectiveness of Bootstrap-S in achieving fairness levels comparable to the oracle while maintaining high prediction performance (unlike the baseline) on a variety of real data, including classification and regression tasks, discrete and continuous sensitive attributes, and independence and separation notions of group fairness.

2. Problem Formulation

Suppose $\mathbf{x}$ represents d-dimensional input features defined on alphabet $\mathcal{X}$, while y and e denote 1-dimensional target and sensitive attribute defined on 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 target variable for new input features and (b) avoids discrimination based on 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 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 of fair learning is to find an $f \in \mathcal{F}$ that minimizes expected loss ℓ, subject to fairness measure Φ being small:

$$f^* \in \arg\min_{f \in \mathcal{F}} \mathbb{E}\big[\ell(y, f(\mathbf{x}))\big] \quad \text{s.t.} \quad \Phi(y, f(\mathbf{x}), e) \leq \epsilon. \tag{3}$$

For ease of notation, hereon, we define $u \triangleq f(\mathbf{x})$.

Choice of fairness measure. To design a fairness measure Φ, it is important to establish what is meant by a perfectly fair predictor, i.e., $\epsilon = 0$ in (3). Typically, perfect fairness is described in terms of statistical independence. We focus on two commonly used fairness criteria: *independence* and *separation*. The independence criterion, also called demographic parity, demands $u \perp\!\!\!\perp e$, meaning predictions should not reveal any information about sensitive attributes. The separation criterion, also known as equalized odds, requires $u \perp\!\!\!\perp e | y$, indicating predictions should not disclose any information about sensitive attributes given the knowledge of true target variables.

Achieving perfect fairness is not feasible when learning a predictor from finite training samples (Agarwal et al., 2018). Instead, in practice, one often works with measures of approximate fairness by choosing $\epsilon > 0$ in (3), and then 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 (Mary et al., 2019) and we adopt this as our measure of the degree of independence, except in cases where data is Gaussian, where we use a different analytically convenient divergence. For independence, the measure is given by $\Phi(y, u, e) = \chi^2\left(p_{e,u} \| p_e p_e\right)$ where $p_{e,u}, p_e$, and p_u are marginals of (e, u), e, and u, respectively. Likewise, for separation $\Phi(y, u, e) = \mathbb{E}_{p_y}[\chi^2\left(p_{e,u|y} \| p_{e|y} p_{e|y}\right)]$ where $p_{e,u|y}, p_{e|y}$, and $p_{u|y}$ are conditionals of (e, u), e, and u given y, respectively.

Uncertain sensitive attributes. Typically, N independent and identically distributed (i.i.d.) samples of the tuple $(\mathbf{x}, y, e)$ are available, denoted by $\mathcal{D}^{(o)} \triangleq \{\mathbf{x}^{(i)}, y^{(i)}, e^{(i)}\}_{i \in [N]}$. Then, in the optimization in (3), the objective 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 $\mathbb{E}_{\mathcal{D}^{(p)}}\left[\ell(y, f(\mathbf{x}))\right]$ and $\Phi_{\mathcal{D}^{(o)}}(y, f(\mathbf{x}), e)$, respectively, for brevity. We assume 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 (3) with access to $\mathcal{D}^{(p)}$ and $\mathcal{D}^{(u)}$. While this is an intuitively appealing goal, simply computing the constraint in (3) with $\mathcal{D}^{(u)}$ may be sub-optimal as discussed in Section 1. In other words, a predictor u satisfying $\Phi_{\mathcal{D}^{(u)}}(y, u, e) \leq \epsilon$ may not necessarily satisfy $\Phi_{\mathcal{D}^{(o)}}(y, u, e) \leq \epsilon$.

In Appendix B, we consider the case where $(\mathbf{x}, y, e, u)$ is jointly Gaussian. Under this scenario, we fully characterize the optimization problem in (3) and guarantee strict fairness despite uncertainty in sensitive attributes. Then, we describe the motivation for our general-purpose algorithm by building on this analysis to develop.

3. A General-Purpose Algorithm

Now, we propose a generic algorithm, `Bootstrap-S`, to accounting for uncertainty in a non-parametric fashion following the *bootstrap* procedure (Efron, 1992). Specifically, given uncertain sensitive attribute data $\mathcal{D}^{(u)} = \{\mathbf{x}^{(i)}, y^{(i)}, \hat{e}^{(i)}\}_{i \in [n]}$, a fairness measure $\Phi(y, u, e)$, and a parameter $\mathtt{S}$: we draw uniformly $\mathtt{S}$ subsets $\mathcal{D}_1^{(u)}, \ldots, \mathcal{D}_{\mathtt{S}}^{(u)}$ of some size $k \in [n]$ from $\mathcal{D}^{(u)}$ 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 $\mathtt{S}$ constraints $\{\Phi_{\mathcal{D}_i^{(u)}}(y, u, e) \leq \epsilon\}_{i \in [\mathtt{S}]}$ together with the constraint $\Phi_{\mathcal{D}^{(u)}}(y, u, e) \leq \epsilon$. In summary, we aim to solve:

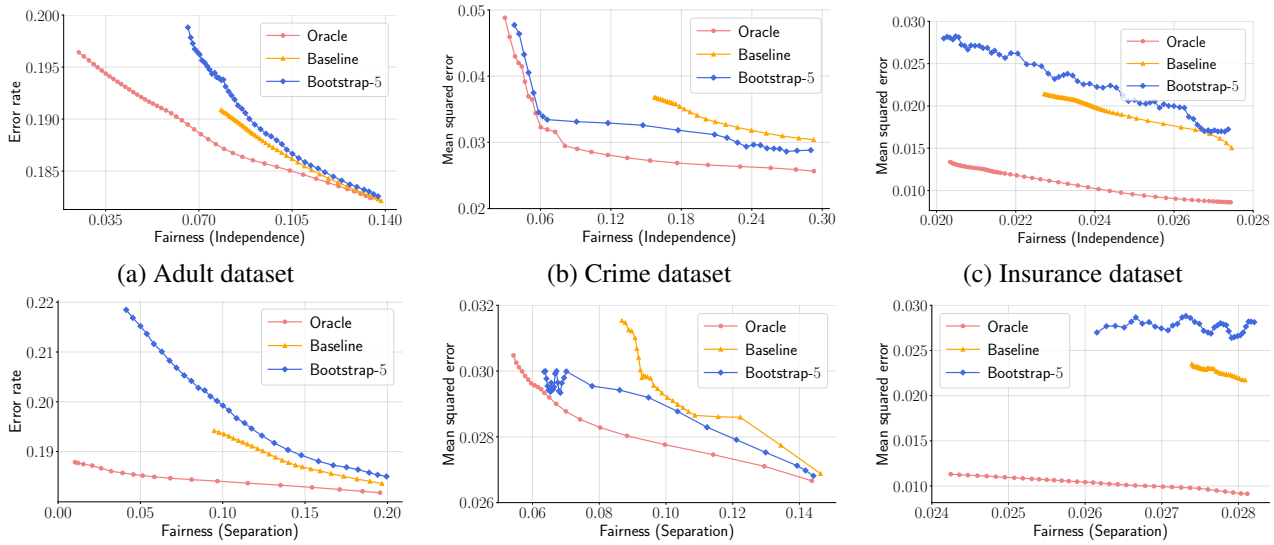
$$\min_u \mathbb{E}_{\mathcal{D}^{(p)}}\left[\ell(y, u)\right] \text{ s.t } \Phi_{\mathcal{D}^{(u)}}(y, u, e) \leq \epsilon \text{ and, } \quad \Phi_{\mathcal{D}_i^{(u)}}(y, u, e) \leq \epsilon \text{ for all } i \in [\mathtt{S}]. \tag{4}$$

At a high level, the idea is similar to bootstrap confidence intervals (Wasserman, 2006) allowing construction of better uncertainty set with a larger number of subsamples $\mathtt{S}$.

Notice that (4) is a constrained optimization problem, which is non-trivial to solve in practice, especially for neural network training. Typically, this problem is addressed by simply adding the fairness constraints as regularizers with hyperparameters 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 the choice of λ. Instead, we follow the approach of Lee et al. (2019) by considering the Lagrangian dual of (4) and optimizing

Algorithm 1 `Bootstrap-S`

Input: $\mathcal{D}^{(p)}$, $\mathcal{D}^{(u)}$, subsample size $k \in [n]$, number of subsamples $\mathtt{S} \geq 1$.
for $i = 1$ **to** $\mathtt{S}$ **do**
  Draw a subsample $\mathcal{D}_i^{(u)}$ of size k from $\mathcal{D}^{(u)}$ at random with replacement.
end for
Then, solve the optimization problem in (5).



(a) Adult dataset (b) Crime dataset (c) Insurance dataset

(d) Adult dataset (e) Crime dataset (f) Insurance dataset

Figure 3. Performance of `Bootstrap-S` and `Baseline` for independence (top row) and separation (bottom row). `Bootstrap-S` achieves much better fairness levels compared to `Baseline` throughout.

the resulting objective over the duality variables, i.e.,

$$\min_{u} \max_{\lambda,\lambda_1,\cdots,\lambda_S} \mathbb{E}_{\mathcal{D}^{(p)}}\left[\ell(y,u)\right] + \lambda\left(\Phi_{\mathcal{D}^{(u)}}(y,u,e) - \epsilon\right) + \sum_{i\in[S]} \lambda_i\left(\Phi_{\mathcal{D}_i^{(u)}}(y,u,e) - \epsilon\right) \quad (5)$$

We summarize `Bootstrap-S` in Algorithm 1.

4. Empirical Evaluation

We test `Bootstrap-S` on real-world classification and regression tasks for group fairness notions of independence and separation using 3 datasets: Adult, Crime, & Insurance. We provide an overview of these in Table 1 with detailed descriptions and pre-processing steps in Appendix I.2.

Table 1. Overview of datasets.

Dataset	Task	Sensitive Attribute
Adult	Classification	Sex (binary)
Crime	Regression	Race (continuous)
Insurance	Regression	Sex (binary)

Training. We train a two-layered neural network and use the log loss for classification, and MSE for regression. We use the χ^2-divergence to impose the independence (conditional independence) for independence (separation). For continuous sensitive attributes, we induce uncertainty in every sensitive attribute by adding independent $\mathcal{N}(0,\sigma^2)$ noise ($\sigma = 0.5$ for Crime). For binary sensitive attributes, we induce uncertainty by keeping only n out of N sensitive attributes ($n = 100$ for Adult and $n = 10$ Insurance). For `Bootstrap-S`, we set $S = 5$. Given a fairness target ϵ, we train a model over 50 independent trials of random missingness (for Adult and Insurance) or random noise (for Crime), and report the average performance (the error bars are too small to see). We sweep over 500 different ϵ from 0.001 to 0.5, and plot the prediction-fairness trade-off frontier by using a simple moving average over 5 entries. We provide more details and experiments in Appendix I.2

Evaluation. On a held-out test set, we report predictive power using error rate (lower is better) for classification and MSE (lower is better) for regression. We evaluate (a) independence (lower is better) using demographic parity, i.e., $|P(\hat{y}|e=1) - P(\hat{y}|e=0)|$ for classification and χ^2-divergence for regression, and (b) separation (lower is better) using equal opportunity, i.e., $|P(\hat{y}|e=1, y=1) - P(\hat{y}|e=0, y=1)|$ for classification and χ^2-divergence for regression.

Results. We compare with `Baseline` which does not have the additional constraints of `Bootstrap-S`, i.e., it solves for (5) with $\lambda_1, \cdots, \lambda_S$ fixed to 0. For reference, we also compare with `Oracle` that has access to all the true sensitive attributes. Figure 3 show that `Baseline` and `Bootstrap-S` exhibit a concentration of fairness levels near the extreme values of ϵ with `Bootstrap-S` being more noisy than `Baseline`. However, `Bootstrap-S` achieves significantly smaller fairness levels compared to `Baseline`. In fact, in most cases, `Bootstrap-S` achieves fairness levels that are comparable to `Oracle` while maintaining a relatively high predictive power.

References

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

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

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

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

Calmon, F., Wei, D., Vinzamuri, B., Ramamurthy, K. N., and Varshney, K. R. Optimized pre-processing for discrimination prevention. In *Advances in Neural Information Processing Systems*, pp. 3992–4001, 2017.

Castelnovo, A., Crupi, R., Greco, G., Regoli, D., Penco, I. G., and Cosentini, A. C. A clarification of the nuances in the fairness metrics landscape. *Scientific Reports*, 12 (1):1–21, 2022.

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

Celis, L. E., Mehrotra, A., and Vishnoi, N. Fair classification with adversarial perturbations. *Advances in Neural Information Processing Systems*, 34:8158–8171, 2021b.

Chen, J., Kallus, N., Mao, X., Svacha, G., and Udell, M. Fairness under unawareness: Assessing disparity when protected class is unobserved. In *Proceedings of the conference on fairness, accountability, and transparency*, pp. 339–348, 2019.

Diamond, S. and Boyd, S. Cvxpy: A python-embedded modeling language for convex optimization. *The Journal of Machine Learning Research*, 17(1):2909–2913, 2016.

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

Gölz, P., Kahng, A., and Procaccia, A. D. Paradoxes in fair machine learning. In *Advances in Neural Information Processing Systems*, pp. 8340–8350, 2019.

Gupta, M., Cotter, A., Fard, M. M., and Wang, S. Proxy fairness. *arXiv preprint arXiv:1806.11212*, 2018.

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

Hashimoto, T., Srivastava, M., Namkoong, H., and Liang, P. Fairness without demographics in repeated loss minimization. In *International Conference on Machine Learning*, pp. 1929–1938. PMLR, 2018.

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

Jung, S., Chun, S., and Moon, T. Learning fair classifiers with partially annotated group labels. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 10348–10357, 2022.

Kallus, N., Mao, X., and Zhou, A. Assessing algorithmic fairness with unobserved protected class using data combination. *Management Science*, 68(3):1959–1981, 2022.

Kamishima, T., Akaho, S., and Sakuma, J. Fairness-aware learning through regularization approach. In *2011 IEEE 11th International Conference on Data Mining Workshops*, pp. 643–650. IEEE, 2011.

Krumpal, I. Determinants of social desirability bias in sensitive surveys: a literature review. *Quality & quantity*, 47(4):2025–2047, 2013.

Lahoti, P., Beutel, A., Chen, J., Lee, K., Prost, F., Thain, N., Wang, X., and Chi, E. Fairness without demographics through adversarially reweighted learning. *Advances in neural information processing systems*, 33:728–740, 2020.

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

Lantz, B. *Machine learning with R: expert techniques for predictive modeling*. Packt publishing ltd, 2019.

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

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

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

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

Pleiss, G., Raghavan, M., Wu, F., Kleinberg, J., and Weinberger, K. Q. On fairness and calibration. In *Advances in Neural Information Processing Systems*, pp. 5680–5689, 2017.

Redmond, M. and Baveja, A. A data-driven software tool for enabling cooperative information sharing among police departments. *European Journal of Operational Research*, 141(3):660–678, 2002.

Shah, A., Bu, Y., Lee, J. K., Das, S., Panda, R., Sattigeri, P., and Wornell, G. W. Selective regression under fairness criteria. In *International Conference on Machine Learning*, pp. 19598–19615. PMLR, 2022.

Verma, S. and Rubin, J. Fairness definitions explained. In *2018 ieee/acm international workshop on software fairness (fairware)*, pp. 1–7. IEEE, 2018.

Vershynin, R. *High-dimensional probability: An introduction with applications in data science*, volume 47. Cambridge university press, 2018.

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

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

Zafar, M. B., Valera, I., Rogriguez, M. G., and Gummadi, K. P. Fairness constraints: Mechanisms for fair classification. In *Artificial Intelligence and Statistics*, pp. 962–970. PMLR, 2017.

Zemel, R., Wu, Y., Swersky, K., Pitassi, T., and Dwork, C. Learning fair representations. In *International Conference on Machine Learning*, pp. 325–333, 2013.

A. Related Work

Group fairness in machine learning. Various metrics and criteria have been proposed to ensure group fairness in machine learning (Verma & Rubin, 2018; Mehrabi et al., 2021; Castelnovo et al., 2022; Shah et al., 2022), but many of these criteria are mutually exclusive in non-trivial cases (Gölz et al., 2019). For example, the independence and the separation criteria cannot both be satisfied simultaneously. Different approaches exist to enforce these criteria, mainly falling into one of three categories: (a) pre-processing methods (Zemel et al., 2013; Calmon et al., 2017), (b) post-processing methods (Hardt et al., 2016; Pleiss et al., 2017), and (c) in-processing methods (Kamishima et al., 2011; Zafar et al., 2017; Agarwal et al., 2018). In this work, we consider independence and separation, using an in-processing approach where the objective function accounts for both accuracy and fairness.

Fairness without certain sensitive attributes. The growing literature on fairness in the absence of true sensitive attributes can be broadly categorized into the following three groups.

A. Perturbed sensitive attribute. Several approaches have been proposed to handle perturbations in sensitive attributes. For example, in-processing methods for fair classification have been developed by (Lamy et al., 2019) and (Celis et al., 2021a) to deal with noisy group labels, while (Awasthi et al., 2020) investigated the performance of a post-processing algorithm proposed by (Hardt et al., 2016) with noisy sensitive labels. Additionally, (Mozannar et al., 2020) and (Celis et al., 2021b) explored achieving group fairness with adversarially perturbed and differentially private data, respectively.

B. Proxy variables. (Gupta et al., 2018), (Chen et al., 2019), and (Kallus et al., 2022) proposed methods to achieve group fairness when proxy variables are available as substitutes for the sensitive attribute (e.g., zip code as a proxy of race). However, the effectiveness of these methods may be reduced if the correlation between the sensitive attribute and the proxy variables is weak. (Jung et al., 2022) proposed a semi-supervised learning approach to generate proxy pseudo-labels for partially observed sensitive attributes. While these proxy-based methods can be useful, they risk perpetuating biases.

C. No sensitive attribute. (Hashimoto et al., 2018) and (Lahoti et al., 2020) proposed methods for achieving fairness without relying on a labeled sensitive attribute, utilizing distributionally robust optimization to improve the performance of the worst-case risk for all distributions close to the empirical distribution. They aim to achieve Rawlsian max-min fairness, but their notion of fairness is not defined by the population distribution, which sets it apart from our focus on group fairness. Additionally, it may not be straightforward to combine their methods with existing fair-training methods, while our method can be generally applied to any fair-training method.

Our work is most closely related to (Wang et al., 2020), who also focus on achieving a strict group fairness guarantee given uncertain sensitive attributes. However, they focus on classification problems with discrete sensitive attributes. In contrast, our approach is widely applicable, including both regression and classification, as well as both discrete and continuous sensitive attributes.

B. Gaussian setting

For the ease of exposition, we consider zero-mean Gaussian variables, and assume that the marginal distribution $p_{\mathbf{x},y}$ is known or can be learned from $\mathcal{D}^{(p)}$.n We think of u as a representation of the features and let the predictor be $\mathbb{E}[y|u]$. Naturally, the loss function ℓ is chosen to be the mean squared loss as $\mathcal{Y} = \mathbb{R}$. We focus on the independence criterion of fairness and measure the degree of independence between u and e using the notion of $\overline{D}$-divergence, a second-order approximation of Kullback–Leibler divergence, introduced by (Huang et al., 2019).

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

$$\overline{D}(p_{\mathbf{v}} \| p_{\mathbf{w}}) \triangleq \frac{1}{2} \| \boldsymbol{\Sigma}_w^{-1/2} (\boldsymbol{\Sigma}_v - \boldsymbol{\Sigma}_w) \boldsymbol{\Sigma}_w^{-1/2} \|_{\mathrm{F}}^2, \tag{6}$$

For these choices, the optimization in (3) reduces to learning a Gaussian variable u such that

$$u^* \in \arg\min_u \mathbb{E}\big[(y - \mathbb{E}[y|u])^2)\big] \quad \text{s.t.} \quad \overline{D}(p_{e,u} \| p_e p_u) \leq \epsilon. \tag{7}$$

Next, we reformulate (7) into a quadratically constrained quadratic program (QCQP) by utilizing the notion of canonical correlation matrices (CCMs) defined by (Huang et al., 2019).

Definition 2 (Canonical correlation matrix)**.** *The canonical correlation matrix (CCM) between zero-mean jointly Gaussian random vectors* $\mathbf{v} \sim \mathcal{N}(\boldsymbol{0}, \boldsymbol{\Sigma}_v)$ *and* $\mathbf{w} \sim \mathcal{N}(\boldsymbol{0}, \boldsymbol{\Sigma}_w)$ *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* $\mathbf{v}$ *and* $\mathbf{w}$*.*

The $\overline{D}$-divergence is conveniently represented by these CCMs. We now formally state the equivalence between (7) and a QCQP that uses CCMs. We prove this in Appendix C.1 by drawing connections to the information bottleneck principle (Bu et al., 2021).

Theorem B.1 (Gaussian Fair Learning $\iff$ QCQP)**.** *The optimization problem in* (7) *is equivalent to*

$$\max_{\mathbf{a}\in\mathcal{B}(0,1)} \left\langle \mathbf{a}, \mathbf{b}_{yx} \right\rangle^2 \quad \textit{s.t.} \quad \left\langle \mathbf{a}, \mathbf{b}_{ex} \right\rangle^2 \leq \varepsilon, \tag{8}$$

where $\mathcal{B}(0,1)$ *denotes an* 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 (8) has the same dimension as $\mathbf{x}$, i.e., d. The following result, with a proof in Appendix C.2, shows that any d-dimensional QCQP in (8) can be mapped to a 2-dimensional QCQP.

Proposition 1 ($d = 2$ suffices for QCQP)**.** *The optimal solution* $\mathbf{a}^\star$ *of the QCQP in* (8) *lies in the subspace spanned by the vectors* $\mathbf{b}_{yx}$ *and* $\mathbf{b}_{ex}$*.*

In Appendix C.3, we characterize the optimal $\mathbf{a}^\star$ in (8) as a function of $\mathbf{b}_{yx}$, $\mathbf{b}_{ex}$, and ε for $d = 2$. Theorem B.1 demonstrates that considering the uncertainty in the canonical correlation matrix b_{ex} is sufficient to capture the uncertainty in sensitive attributes, which we will now explore in detail.

B.1. Theoretical Results

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 (8) to ensure strict fairness guarantee with high probability, as well as how this robustification affects the optimal objective.

Let $\widehat{\mathbf{b}}_{ex}$ be an estimate of $\mathbf{b}_{ex}$, say obtained from $\mathcal{D}^{(u)}$, such that $\|\mathbf{b}_{ex} - \widehat{\mathbf{b}}_{ex}\|_2 \leq \tau$ (with probability $1-\delta$), for some $\tau \geq 0$,[1] and denoted by $\mathbf{b}_{ex} \in \mathcal{B}(\widehat{\mathbf{b}}_{ex}, \tau)$. To achieve fairness as in (8) with probability $1-\delta$, in the worst case, $\langle \mathbf{a}, \mathbf{b}\rangle^2 \leq \varepsilon$ should hold for all $\mathbf{b} \in \mathcal{B}(\widehat{\mathbf{b}}_{ex}, \tau)$. Then, the following robust optimization maximizes the desired objective while achieving fairness as in (8) (with probability $1-\delta$) without the precise knowledge of $\mathbf{b}_{ex}$:

$$\max_{\mathbf{a}\in\mathcal{B}(0,1)} \langle \mathbf{a}, \mathbf{b}_{yx}\rangle^2 \quad \text{s.t.} \quad \langle \mathbf{a}, \mathbf{b}\rangle^2 \leq \varepsilon,\ \forall \mathbf{b} \in \mathcal{B}(\widehat{\mathbf{b}}_{ex}, \tau). \tag{9}$$

In the following proposition, we show that any d-dimensional robust QCQP in (9) can be mapped to a 2-dimensional QCQP. See Appendix D for a proof.

Proposition 2 ($d = 2$ suffices for robust QCQP)**.** *The optimal solution* $\mathbf{a}^\star$ *of the robust QCQP in* (9) *lies in the subspace spanned by the vectors* $\mathbf{b}_{yx}$ *and* $\widehat{\mathbf{b}}_{ex}$*.*

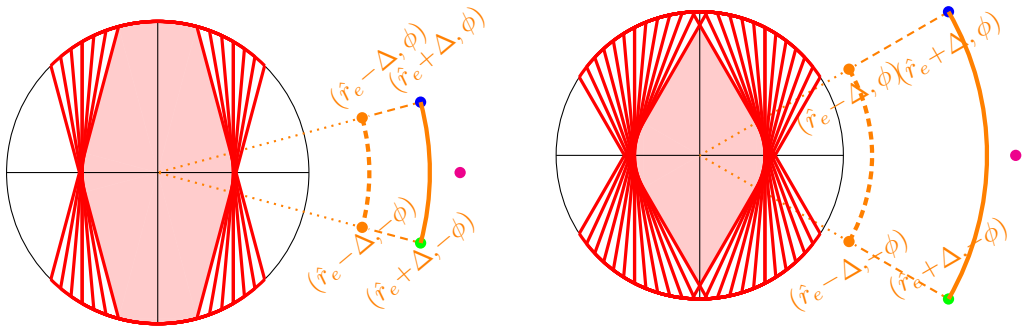
Now, to characterize the solution of the robust QCQP in (9), we focus on $d = 2$ and work with polar coordinates. Further, to analyze the corresponding feasible space, we relax the uncertainty space $\mathcal{B}(\widehat{\mathbf{b}}_{ex}, \tau)$ from a ball to an annular sector. Formally, we let $\widehat{\mathbf{b}}_{ex} \triangleq \widehat{r}_e(\cos\widehat{\theta}_e, \sin\widehat{\theta}_e)$ be the estimate of $\mathbf{b}_{ex}$ such that $|r_e - \widehat{r}_e| \leq \Delta$ and $|\theta_e - \widehat{\theta}_e| \leq \phi$ with probability $1-\delta$ where $\Delta \triangleq \tau \geq 0$ and $\phi \triangleq \sin^{-1}(\tau/\widehat{\mathbf{b}}_{ex}) \in [0, \pi/2]$. In other words, given $\widehat{r}_e, \widehat{\theta}_e, \Delta$, and ϕ, with probability $1-\delta$,

$$\mathbf{b}_{ex} \in \mathcal{A}(\Delta, \phi) \triangleq \{\mathbf{b} = r(\cos\theta, \sin\theta) \colon |r - \widehat{r}_e| \leq \Delta \text{ and } |\theta - \widehat{\theta}_e| \leq \phi\}, \tag{10}$$

i.e., $\mathcal{A}(\Delta, \phi)(\supset \mathcal{B}(\widehat{\mathbf{b}}_{ex}, \tau))$ denotes the smallest annular sector around $\widehat{\mathbf{b}}_{ex}$ capturing our uncertainty in knowing $\mathbf{b}_{ex}$ (see Figure 4 where $\mathcal{A}(\Delta, \phi)$ is the shown in orange). Now, to achieve fairness as in (8) (with probability $1-\delta$), we constrain the robust QCQP in (9) as follows:

$$\max_{\mathbf{a}\in\mathcal{B}(0,1)} \langle \mathbf{a}, \mathbf{b}_{yx}\rangle^2 \quad \text{s.t.} \quad \left\langle \mathbf{a}, \mathbf{b}\right\rangle^2 \leq \varepsilon,\ \forall \mathbf{b} \in \mathcal{A}(\Delta, \phi). \tag{11}$$

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



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

Figure 4. Visualizing the feasible space of the robust QCQP in Theorem B.2, i.e., (12), 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 B.3 are shown in magenta, blue and green, respectively.

As we show below, the constraint in (11) is equivalent to ensuring $\langle \mathbf{a}, \mathbf{b}\rangle^2 \leq \varepsilon$ 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 4). See Appendix E for a proof. Further, in Appendix E.1, we characterize the optimal $\mathbf{a}$ in (11) as a function of $\widehat{\mathbf{b}}_{ex}$, $\mathbf{b}_{yx}$, Δ, ϕ, and ε.

Theorem B.2 (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* $|\theta - \widehat{\theta}_e| \leq \phi\}$ *be the arc on the boundary of* $\mathcal{A}(\Delta, \phi)$ *with maximum radius. Then, the robust QCQP in* (11) *is equivalent to*

$$\max_{\mathbf{a}\in\mathcal{B}(0,1)} \langle \mathbf{a}, \mathbf{b}_{yx}\rangle^2 \quad s.t. \quad \big\langle \mathbf{a}, \mathbf{b}\big\rangle^2 \leq \varepsilon, \ \forall \mathbf{b} \in \overline{\mathcal{A}}(\Delta, \phi). \tag{12}$$

There is a phase transition in the nature of the feasible space of the robust QCQP in (12). Figure 4(a) illustrates the space if $\sqrt{\varepsilon} \geq (\widehat{r}_e + \Delta)\sin\phi$, and Figure 4(b) illustrates the space if $\sqrt{\varepsilon} \leq (\widehat{r}_e + \Delta)\sin\phi$.

While Theorem B.2 simplifies the optimization in (11), the resulting optimization in (12) still has infinite constraints. Below, we provide an approximation to the feasible space in (12) such that it has finitely many constraints. See Appendix F for a proof. We note that alternative approximations are possible.

Theorem B.3 (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* (12)*:*

$$\max_{\mathbf{a}\in\mathcal{B}(0,1)} \big\langle \mathbf{a}, \mathbf{b}_{yx}\big\rangle^2 \quad s.t \quad \big\langle \mathbf{a}, \mathbf{b}_{ex}^{(i)}\big\rangle^2 \leq \varepsilon \ \textit{for all } i \in [3]. \tag{13}$$

See Appendix F.1 for a characterization of the optimal $\mathbf{a}$ in (13) as a function of $\widehat{\mathbf{b}}_{ex}$, $\mathbf{b}_{yx}$, Δ, ϕ, and ε. We visualize $\mathbf{b}_{ex}^{(1)}$, $\mathbf{b}_{ex}^{(2)}$, and $\mathbf{b}_{ex}^{(3)}$ in Figure 4 (in magenta, blue, and green, respectively), and note that $\mathbf{b}_{ex}^{(1)}$ approximates the effect of the points in-between $\mathbf{b}_{ex}^{(2)}$ and $\mathbf{b}_{ex}^{(3)}$ on $\mathcal{A}(\Delta, \phi)$.

Uncertainty due to randomly missing sensitive attributes. Now, we focus on understanding how the optimal objective of the robust QCQP in (13) changes when the uncertainty set $\mathcal{A}(\Delta, \phi)$ changes. For concreteness, we consider the case where uncertainty only stems from sensitive attributes missing at random, and the uncertainty can be improved by collecting more labeled sensitive attributes. Then, we analyze the power of each new labeled sensitive attribute by characterizing the exact difference in the optimal objectives of the QCQP in (8) and the robust QCQP in (13). To enable this, we show (in Appendix H) that the optimal objective in (13) either monotonically increases or coincides with the optimal objective in (8) whenever the uncertainty set monotonically decreases with n.

The response of our algorithm to collecting more labeled sensitive attributes can be classified into 3 categories (see the formal result in Appendix G): (a) *Any uncertainty hurts:* Here, the optimal objective of the robust QCQP in (13) matches

the optimal objective in (8) only when all the uncertainty is removed, i.e., $n \to N$; (b) *Some uncertainty does not hurt:* Here, the optimal objective of the robust QCQP in (13) matches the optimal objective in (8) when some uncertainty is removed, i.e., by collecting few extra labeled sensitive attributes; (c) *Uncertainty does not hurt:* Here, the optimal objective of the robust QCQP in (13) matches the optimal objective in (8) without collecting any extra labeled sensitive attributes. We state this informally below (see Lemma 4 for a formal statement).

Corollary 1 (Free fairness)**.** *There exist problem instances of the robust QCQP in* (13) *where the uncertainty incurs no performance loss while achieving a strict fairness guarantee, without requiring additional labeled sensitive attributes.*

Solving the robust QCQP in (13) for $d = 2$ is straightforward using a standard convex optimization solver. However, as the dimensionality of $\mathbf{x}$ increases, solving the problem becomes more challenging. Furthermore, real datasets may not be well-modeled by Gaussian distributions, and thus, the proposed robust QCQP may not be directly deployable.

B.2. Motivation for `Bootstrap-S`

In Section 3, we leverage our theoretical analysis to propose a generic algorithm that handles high-dimensional features and non-Gaussian data while accounting for uncertainty. The core idea of the robust QCQP (in Theorem B.3) is to construct an uncertainty set around the estimated canonical correlation matrix $\widehat{\mathbf{b}}_{ex}$ by imposing additional constraints. This approach effectively addresses the unknown nature of the true $\mathbf{b}_{ex}$. Another perspective is to view the robust QCQP in (13) as

$$\max_{\mathbf{a}\in\mathcal{B}(0,1)} \left\langle \mathbf{a}, \mathbf{b}_{yx} \right\rangle^2 \quad \text{s.t} \quad \left\langle \mathbf{a}, \widehat{\mathbf{b}}_{ex} \right\rangle^2 \leq \varepsilon \quad \text{and} \quad \left\langle \mathbf{a}, \mathbf{b}_{ex}^{(i)} \right\rangle^2 \leq \varepsilon \;\; \text{for all } i \in [3], \tag{14}$$

where the constraint $\left\langle \mathbf{a}, \widehat{\mathbf{b}}_{ex} \right\rangle^2 \leq \varepsilon$ becomes redundant in the presence of the constraint $\left\langle \mathbf{a}, \mathbf{b}_{ex}^{(1)} \right\rangle^2 \leq \varepsilon$, and $\{\mathbf{b}_{ex}^{(i)}\}_{i\in[3]}$ can be viewed as multiple estimates of $\widehat{\mathbf{b}}_{ex}$. For general non-Gaussian data, we use a similar idea in (4).

B.3. Synthetic data

In this section, we show the efficacy of the robust QCQP in achieving strict fairness on synthetic Gaussian data and demonstrate that `Bootstrap-S` serves as a good approximation for the robust QCQP.

We generate synthetic data using two Gaussian distributions with zero mean, $d = 2$, and covariances matrices (a) Σ_2^{gen} and (b) Σ_2^{fair} defined in Appendix I.1. We present our results when uncertainty is due to randomly missing sensitive attributes. In Appendix I.1, we present results on (i) another choice of covariance, (ii) $d = 3$, and (iii) uncertainty in sensitive attribute due to noise. We estimate $\widehat{\mathbf{b}}_{ex}$ using n samples of $(\mathbf{x}, e)$ for various choices of n. Then, we compare the robust QCQP in (13) and `Bootstrap-S` applied to the QCQP in (8) (for various `S`) against `Baseline`, which solves the QCQP in (8) using $\widehat{\mathbf{b}}_{ex}$. We solve the resulting optimization problems using the CVXPY library (Diamond & Boyd, 2016). The covariances Σ_2^{gen} an Σ_2^{fair} are designed to demonstrate the general behavior (where uncertainty hurts) and the free-fairness behavior (Corollary 1), respectively.

The results, averaged over 1000 random trials, are shown in Figure 5 (the error bars are too small to see). We observe that robust QCQP always ensures no fairness violations. Additionally, the performance (in terms of average MSE) of robust QCQP monotonically improves as n increases, as stated in Appendix H. More importantly, in Figure 5 (b), robust QCQP does not incur any significant loss in the performance, and thus demonstrates the free-fairness phenomenon in Corollary 1, say, $n \approx 350$ onward. We also see that `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 the number of subsamples `S` by forming a more accurate uncertainty set. However, the benefit of larger `S` comes with an increased computation.

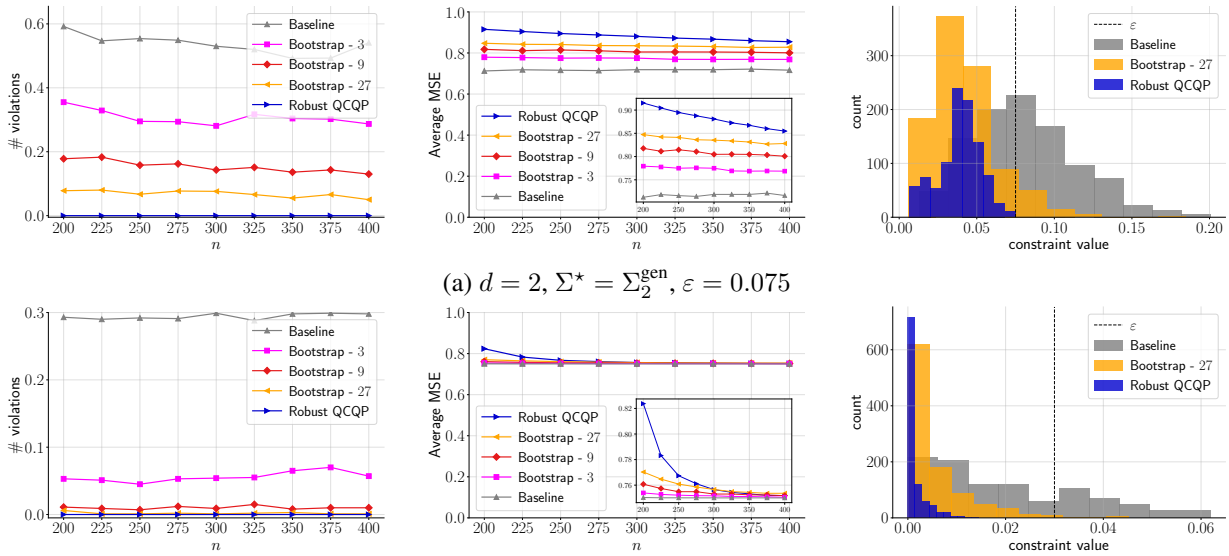
C. Proof of Theorem B.1 and Proposition 1

In this section, we prove Theorem B.1 and Proposition 1 as well as provide a characterization of the optimal $\mathbf{a}^\star$ in (8).

C.1. Proof of Theorem B.1: Gaussian Fair Learning $\iff$ QCQP

First, we show that optimization in (7) is equivalent to the optimization below:

$$\arg\max_{u} \overline{D}(p_{y,u} \| p_y p_u) \quad \text{s.t.} \quad \overline{D}(p_{e,u} \| p_e p_u) \leq \epsilon. \tag{15}$$



(a) $d = 2$, $\Sigma^\star = \Sigma_2^{\text{gen}}$, $\varepsilon = 0.075$

(b) $d = 2$, $\Sigma^\star = \Sigma_2^{\text{free}}$, $\varepsilon = 0.025$

Figure 5. The performance of robust QCQP in (13), `Bootstrap-S` with `S` $\in \{3, 9, 27\}$, and `Baseline` for $d = 2$ and various $\Sigma^\star$. In the left column, we plot the fraction of violations of the true fairness constraint $\langle \mathbf{a}, \mathbf{b}_{ex} \rangle^2 \leq \varepsilon$ 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 $n = 250$.

To that end, from the definition of conditional variance, we have

$$\mathbb{E}[(y - \mathbb{E}[y|u])^2] = \mathbb{E}\big[\mathbb{V}\text{ar}[y|u]\big]. \tag{16}$$

Now, for a joint Gaussian vector $(y, u) \in \mathbb{R}^2$ with covariance $\mathbf{\Sigma}$, we have

$$\mathbb{E}\big[\mathbb{V}\text{ar}[y|u]\big] \overset{(a)}{=} \mathbb{E}\big[\mathbf{\Sigma}_{yy} - \mathbf{\Sigma}_{yu}\mathbf{\Sigma}_{uu}^{-1}\mathbf{\Sigma}_{uy}\big] = \mathbf{\Sigma}_{yy} - \mathbf{\Sigma}_{yu}\mathbf{\Sigma}_{uu}^{-1}\mathbf{\Sigma}_{uy}, \tag{17}$$

where (a) follows from Schur's complement. Since $\mathbf{\Sigma}_{yy}$ is a constant w.r.t. u, we can write

$$\arg\min_u \mathbb{E}\big[\mathbb{V}\text{ar}[y|u]\big] = \arg\min_u -\mathbf{\Sigma}_{yu}\mathbf{\Sigma}_{uu}^{-1}\mathbf{\Sigma}_{uy} = \arg\max_u \mathbf{\Sigma}_{yu}\mathbf{\Sigma}_{uu}^{-1}\mathbf{\Sigma}_{uy}. \tag{18}$$

Now, from Huang et al. (2019, Lemma. 68), we have

$$\overline{D}(p_{y,u} \| p_y p_u) = \|\mathbf{\Sigma}_{yy}^{-1/2}\mathbf{\Sigma}_{yu}\mathbf{\Sigma}_{uu}^{-1/2}\|_{\text{F}}^2 \overset{(a)}{=} \mathbf{\Sigma}_{yy}^{-1/2}\mathbf{\Sigma}_{yu}\mathbf{\Sigma}_{uu}^{-1}\mathbf{\Sigma}_{uy}\mathbf{\Sigma}_{yy}^{-1/2}, \tag{19}$$

where $\|\cdot\|_{\text{F}}$ denotes the Frobenius norm, and (a) follows because $\mathbf{\Sigma}_{yy}^{-1/2}\mathbf{\Sigma}_{yu}\mathbf{\Sigma}_{uu}^{-1}\mathbf{\Sigma}_{uy}\mathbf{\Sigma}_{yy}^{-1/2} \in \mathbb{R}$. As before, since $\mathbf{\Sigma}_{yy}$ is a constant w.r.t. u, we can write

$$\arg\max_u \overline{D}(p_{y,u} \| p_y p_u) = \arg\max_u \mathbf{\Sigma}_{yu}\mathbf{\Sigma}_{uu}^{-1}\mathbf{\Sigma}_{uy}. \tag{20}$$

Then, (15) follows by combining (16), (18), and (20).

Next, from Bu et al. (2021, Theorem. 1), the optimization problem in (15) is equivalent to the following semi-definite program (SDP) using the notion of canonical correlation matrices (Definition 2):

$$\max_{\mathbf{A} \in \mathbb{A}^d} \text{Tr}(\mathbf{b}_{yx}^\top \mathbf{b}_{yx}\mathbf{A}) \quad \text{s.t.} \quad \text{Tr}(\mathbf{b}_{ex}^\top \mathbf{b}_{ex}\mathbf{A}) \leq \varepsilon \text{ and } 0 \preceq \mathbf{A} \preceq \mathbf{I}, \tag{21}$$

where $\text{Tr}(\cdot)$ denote the trace of a matrix, $\mathbb{A}^d$ is the space of $d \times d$ symmetric matrices, and $\mathbf{A}$ is of the form $\mathbf{b}_{xu}\mathbf{b}_{xu}^\top$. Finally, we show that the SDP in (21) is equivalent to the QCQP in (8). First, we have

$$\text{Tr}(\mathbf{b}_{yx}^\top \mathbf{b}_{yx}\mathbf{A}) \overset{(a)}{=} \text{Tr}(\mathbf{b}_{yx}^\top \mathbf{b}_{yx}\mathbf{b}_{xu}\mathbf{b}_{xu}^\top) \overset{(b)}{=} \text{Tr}(\mathbf{b}_{xu}^\top \mathbf{b}_{yx}^\top \mathbf{b}_{yx}\mathbf{b}_{xu}) = \text{Tr}((\mathbf{b}_{yx}\mathbf{b}_{xu})^\top \mathbf{b}_{yx}\mathbf{b}_{xu}) \overset{(c)}{=} \langle \mathbf{b}_{yx}, \mathbf{b}_{xu}^\top \rangle^2, \tag{22}$$

where (a) follows because $\mathbf{A} = \mathbf{b}_{xu}\mathbf{b}_{xu}^\top$, (b) follows from the cyclic property of trace, and (c) follows because $\langle \mathbf{b}_{yx}, \mathbf{b}_{xu}^\top \rangle = \mathbf{b}_{yx}\mathbf{b}_{xu} \in \mathbb{R}^{1\times 1}$. Similarly, we have

$$\mathrm{Tr}(\mathbf{b}_{ex}^\top \mathbf{b}_{ex}\mathbf{A}) = \big\langle \mathbf{b}_{ex}, \mathbf{b}_{xu}^\top \big\rangle^2. \tag{23}$$

Finally, noting that $\mathbf{A}$ is a rank 1 matrix, we have

$$0 \preceq \mathbf{A} \preceq \mathbf{I} \iff \mathbf{b}_{xu} \in \mathcal{B}(0,1). \tag{24}$$

Putting together (21) to (24) completes the proof.

C.2. Proof of Proposition 1: $d = 2$ suffices for QCQP

It suffices to show that the projection of any $\mathbf{a} \in \mathcal{B}(0,1)$, satisfying the constraint $\langle \mathbf{a}, \mathbf{b}_{ex}\rangle^2 \le \varepsilon$, onto the subspace spanned by $\mathbf{b}_{yx}$ and $\mathbf{b}_{ex}$ preserves the value of the objective and continues to satisfy the constraint. Fix some $\mathbf{a}$ in the feasible space. Let $\mathrm{proj}(\mathbf{a})$ denote the projection of $\mathbf{a}$ on the subspace spanned by $\mathbf{b}_{yx}$ and $\mathbf{b}_{ex}$.

The value of the objective is preserved. It suffices to show $\langle \mathrm{proj}(\mathbf{a}), \mathbf{b}_{yx}\rangle^2 = \langle \mathbf{a}, \mathbf{b}_{yx}\rangle^2$. Consider an orthonormal basis $\{\mathbf{n}_y, \mathbf{n}_e\} \triangleq \big\{\frac{\mathbf{b}_{yx}}{\|\mathbf{b}_{yx}\|_2}, \mathbf{n}_e\big\}$ spanned by $\mathbf{b}_{yx}$ and $\mathbf{b}_{ex}$ where $\mathbf{n}_y = \frac{\mathbf{b}_{yx}}{\|\mathbf{b}_{yx}\|_2}$ and $\mathbf{n}_e$ is chosen to be orthogonal to $\mathbf{n}_y$. Then,

$$\mathrm{proj}(\mathbf{a}) = \big\langle \mathbf{a}, \mathbf{n}_y \big\rangle \mathbf{n}_y + \big\langle \mathbf{a}, \mathbf{n}_e \big\rangle \mathbf{n}_e. \tag{25}$$

Using the definition of $\mathbf{n}_y$ and the orthogonality between $\mathbf{n}_e$ and $\mathbf{b}_{yx}$, we have

$$\big\langle \mathrm{proj}(\mathbf{a}), \mathbf{b}_{yx} \big\rangle^2 = \big\langle \mathbf{a}, \mathbf{n}_y \big\rangle^2 \|\mathbf{b}_{yx}\|_2^2. \tag{26}$$

Similarly, using the definition of $\mathbf{n}_y$, we have $\big\langle \mathbf{a}, \mathbf{b}_{yx} \big\rangle^2 = \big\langle \mathbf{a}, \mathbf{n}_y \big\rangle^2 \|\mathbf{b}_{yx}\|_2^2$.

The constraint is satisfied. It suffices to show $\mathrm{proj}(\mathbf{a}) \in \mathcal{B}(0,1)$ and $\langle \mathrm{proj}(\mathbf{a}), \mathbf{b}_{ex}\rangle^2 \le \varepsilon$. It is easy to see $\mathrm{proj}(\mathbf{a}) \in \mathcal{B}(0,1)$ because $\|\mathrm{proj}(\mathbf{a})\|_2 \le \|\mathbf{a}\|_2$. Now, consider a different set of orthonormal basis $\{\mathbf{n}_e', \mathbf{n}_y'\} \triangleq \big\{\frac{\mathbf{b}_{ex}}{\|\mathbf{b}_{ex}\|_2}, \mathbf{n}_y'\big\}$ spanned by $\mathbf{b}_{ex}$ and $\mathbf{b}_{yx}$ where $\mathbf{n}_e' = \frac{\mathbf{b}_{ex}}{\|\mathbf{b}_{ex}\|_2}$ and $\mathbf{n}_y'$ is chosen to be orthogonal to $\mathbf{n}_e'$. Then,

$$\mathrm{proj}(\mathbf{a}) = \big\langle \mathbf{a}, \mathbf{n}_e' \big\rangle \mathbf{n}_e' + \big\langle \mathbf{a}, \mathbf{n}_y' \big\rangle \mathbf{n}_y'. \tag{27}$$

Using the definition of $\mathbf{n}_e'$ and the orthogonality between $\mathbf{n}_y'$ and $\mathbf{b}_{ex}$, we have

$$\big\langle \mathrm{proj}(\mathbf{a}), \mathbf{b}_{ex} \big\rangle^2 = \big\langle \mathbf{a}, \mathbf{n}_e' \big\rangle^2 \|\mathbf{b}_{ex}\|_2^2 \overset{(a)}{=} \big\langle \mathbf{a}, \mathbf{b}_{ex} \big\rangle^2 \overset{(8)}{\le} \varepsilon, \tag{28}$$

where (a) follows from the definition of $\mathbf{n}_e'$.

C.3. Characterizing the optimal $\mathbf{a}^\star$ in (8)

Fix any $\varepsilon > 0$. For $d = 2$, it is convenient to work with polar coordinates. Let $\mathbf{b}_{ex} = r_e(\cos\theta_e, \sin\theta_e)$ for some $r_e > 0$ and $\theta_e \in [0, 2\pi]$. Similarly, let $\mathbf{b}_{yx} = r_y(\cos\theta_y, \sin\theta_y)$ for some $r_y > 0$ and $\theta_y \in [0, 2\pi]$. Let $\mathbf{a}^\star$ denote the set of optimal $\mathbf{a}$ in (8), i.e.,

$$\mathbf{a}^\star = \arg\max_{\mathbf{a}\in\mathcal{B}(0,1)} \big\langle \mathbf{a}, \mathbf{b}_{yx} \big\rangle^2 \text{ s.t } \big\langle \mathbf{a}, \mathbf{b}_{ex} \big\rangle^2 \le \varepsilon. \tag{29}$$

Then, the following lemma characterizes $\mathbf{a}^\star$ depending on the values of r_e, θ_e, θ_y, and ε.

Lemma 1. *Let $\alpha \in [0, 2\pi]$ be such that $\cos\alpha = \sqrt{\varepsilon}/r_e$ and $\sin\alpha = \sqrt{1 - \varepsilon/r_e^2}$. Define the function $\mathbf{a} : [0, 2\pi] \to [-1, 1]^2$ such that $\mathbf{a}(\theta) = (\cos\theta, \sin\theta)$. Then,*

Case 1.1. $\mathbf{a}^\star = \big\{\mathbf{a}(\alpha + \theta_e), \mathbf{a}(\pi + \alpha + \theta_e)\big\} \iff \theta_y \in [\theta_e, \alpha + \theta_e] \cup [\pi + \theta_e, \pi + \alpha + \theta_e]$,

Case 1.2. $\mathbf{a}^\star = \big\{\mathbf{a}(\theta_y), \mathbf{a}(\pi + \theta_y)\big\} \iff \theta_y \in [\alpha + \theta_e, \pi - \alpha + \theta_e] \cup [\pi + \alpha + \theta_e, 2\pi - \alpha + \theta_e]$, *and*

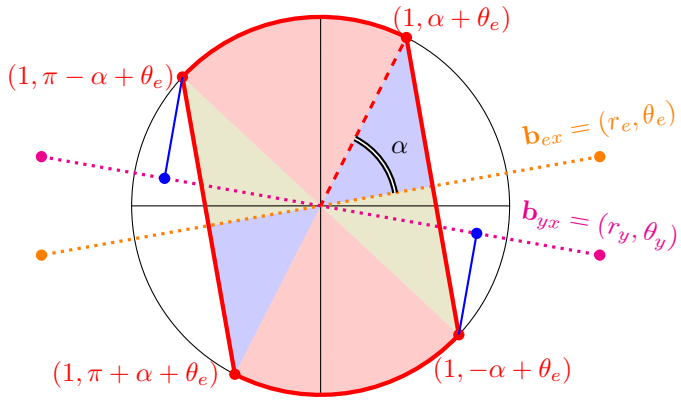


Figure 6. Visualizing QCQP in (8). Each point is shown in polar coordinates, i.e., a point (r, θ) denotes $(r\cos\theta, r\sin\theta)$. The point $\mathbf{b}_{ex}$ is shown in orange. The region enclosed by the solid red lines and red arcs is the feasible space in (8). The feasible space is shaded in blue, red, and olive to represent Case 1.1., Case 1.2., and Case 1.3., respectively. The point $\mathbf{b}_{yx}$ is shown in magenta and falls under Case 1.3. The optimal solution set $\mathbf{a}^\star$ consists of points $(1, \pi - \alpha + \theta_e)$ and $(1, -\alpha, +\theta_e)$.

Case 1.3. $\mathbf{a}^\star = \big\{\mathbf{a}(\pi - \alpha + \theta_e), \mathbf{a}(-\alpha + \theta_e)\big\} \iff \theta_y \in [\pi - \alpha + \theta_e, \pi + \theta_e] \cup [-\alpha + \theta_e, \theta_e]$.

Proof. Consider any $\mathbf{b}_{ex} = r_e(\cos\theta_e, \sin\theta_e)$ (shown in orange in Figure 6) and $\varepsilon > 0$. Let $\Lambda_{\mathbf{a}}(\mathbf{b}_{ex}) \triangleq \{\mathbf{a} : \langle \mathbf{a}, \mathbf{b}_{ex}\rangle^2 \leq \varepsilon$ and $\mathbf{a} \in \mathcal{B}(0,1)\}$ be the feasible space in (8), i.e., the set of all $\mathbf{a}$ satisfying the fairness constraint in (8) w.r.t $\mathbf{b}_{ex}$. Then, $\Lambda_{\mathbf{a}}(\mathbf{b}_{ex})$ is the region enclosed by the solid red lines and red arcs in Figure 6, i.e., the union of the regions shaded in various colors. To obtain this region, consider any $\mathbf{a} \in \mathcal{B}(0,1)$. Then, ensuring $\big\langle \mathbf{a}, \mathbf{b}_{ex}\big\rangle^2 \leq \varepsilon$ is equivalent to ensuring that the projection of $\mathbf{a}$ on the line joining the origin and $\mathbf{b}_{ex}$ (shown in dotted orange in Figure 6) is no more than $\sqrt{\varepsilon}/r_e$. Therefore, we drop perpendiculars to this line at a distance $\sqrt{\varepsilon}/r_e$ on either side of the origin. We obtain the intersection of these perpendiculars and $\mathcal{B}(0,1)$ by standard algebra and trigonometry (these points are shown in Figure 6 in polar coordinates), and the enclosed region is $\Lambda_{\mathbf{a}}(\mathbf{b}_{ex})$.

Now, consider any $\mathbf{b}_{yx} = r_y(\cos\theta_y, \sin\theta_y)$ (shown in magenta in Figure 6). Then, obtaining $\mathbf{a}^\star$ is equivalent to obtaining every $\mathbf{a} \in \Lambda_{\mathbf{a}}(\mathbf{b}_{ex})$ such that the projection of $\mathbf{a}$ on the line joining the origin and $\mathbf{b}_{yx}$ (shown in dotted magenta in Figure 6) is maximized. As a result, every $\mathbf{a} \in \mathbf{a}^\star$ lies on the boundary of this region. To obtain $\mathbf{a}^\star$, we drop perpendiculars to the line joining the origin and $\mathbf{b}_{yx}$ from every $\mathbf{a}$ on the boundary of this region, e.g., we show two such perpendiculars in solid blue in Figure 6. We add $\mathbf{a}$ to $\mathbf{a}^\star$ if the distance between the origin and the point where the corresponding perpendicular intersects the line is maximum. It is straightforward to verify that Case 1.1., Case 1.2., and Case 1.3. correspond to the regions shaded in blue, red, and olive, respectively (see Figure 6). □

D. Proof of Proposition 2: $d = 2$ suffices for robust QCQP

It suffices to show that the projection of any $\mathbf{a} \in \mathcal{B}(0,1)$, satisfying the constraint $\langle \mathbf{a}, \mathbf{b}\rangle^2 \leq \varepsilon$ for all $\mathbf{b} \in \mathcal{B}(\widehat{\mathbf{b}}_{ex}, \tau)$, onto the subspace spanned by $\mathbf{b}_{yx}$ and $\widehat{\mathbf{b}}_{ex}$ preserves the value of the objective and continues to satisfy the constraint. Fix some $\mathbf{a}$ in the feasible space. Let $\mathrm{proj}(\mathbf{a})$ denote the projection of $\mathbf{a}$ on the subspace spanned by $\mathbf{b}_{yx}$ and $\widehat{\mathbf{b}}_{ex}$. The proof that the value of the objective remains preserved is analogous to the proof of Proposition 1 (see Appendix C.2).

The constraint is satisfied. It suffices to show $\mathrm{proj}(\mathbf{a}) \in \mathcal{B}(0,1)$ and $\langle \mathrm{proj}(\mathbf{a}), \mathbf{b}\rangle^2 \leq \varepsilon$ for all $\mathbf{b} \in \mathcal{B}(\widehat{\mathbf{b}}_{ex}, \tau)$. It is easy to see $\mathrm{proj}(\mathbf{a}) \in \mathcal{B}(0,1)$ because $\|\mathrm{proj}(\mathbf{a})\|_2 \leq \|\mathbf{a}\|_2$. Now, consider a set of orthonormal basis $\{\mathbf{n}_e, \mathbf{n}_y\} \triangleq \big\{\frac{\widehat{\mathbf{b}}_{ex}}{\|\widehat{\mathbf{b}}_{ex}\|_2}, \mathbf{n}_y\big\}$

spanned by $\widehat{\mathbf{b}}_{ex}$ and $\mathbf{b}_{yx}$ where $\mathbf{n}_e = \frac{\widehat{\mathbf{b}}_{ex}}{\|\widehat{\mathbf{b}}_{ex}\|_2}$ and $\mathbf{n}_y$ is chosen to be orthogonal to $\mathbf{n}_e$. Then,

$$\mathrm{proj}(\mathbf{a}) = \langle \mathbf{a}, \mathbf{n}_e \rangle \mathbf{n}_e + \langle \mathbf{a}, \mathbf{n}_y \rangle \mathbf{n}_y. \tag{30}$$

Using the definition of $\mathbf{n}_e$ and the orthogonality between $\mathbf{n}_y$ and $\widehat{\mathbf{b}}_{ex}$, we have

$$\langle \mathrm{proj}(\mathbf{a}), \widehat{\mathbf{b}}_{ex} \rangle = \langle \mathbf{a}, \mathbf{n}_e \rangle \|\widehat{\mathbf{b}}_{ex}\|_2 \overset{(a)}{=} \langle \mathbf{a}, \widehat{\mathbf{b}}_{ex} \rangle, \tag{31}$$

where (a) follows from the definition of $\mathbf{n}_e$.

Now, for every $\mathbf{b} \in \mathcal{B}(\widehat{\mathbf{b}}_{ex}, \tau)$, we have $\mathbf{b} = \widehat{\mathbf{b}}_{ex} + \Delta\mathbf{b}$, where $\|\Delta\mathbf{b}\| \leq \tau$. Then,

$$\langle \mathbf{a}, \mathbf{b} \rangle = \langle \mathbf{a}, \widehat{\mathbf{b}}_{ex} + \Delta\mathbf{b} \rangle = \langle \mathrm{proj}(\mathbf{a}), \widehat{\mathbf{b}}_{ex} \rangle + \langle \mathbf{a}, \Delta\mathbf{b} \rangle \overset{(31)}{=} \langle \mathrm{proj}(\mathbf{a}), \widehat{\mathbf{b}}_{ex} \rangle + \langle \mathbf{a}, \Delta\mathbf{b} \rangle \leq \langle \mathrm{proj}(\mathbf{a}), \widehat{\mathbf{b}}_{ex} \rangle + \|\mathbf{a}\|_2 \tau \tag{32}$$

$$\overset{(a)}{\leq} \sqrt{\varepsilon}, \tag{33}$$

where (a) follows because $\langle \mathbf{a}, \mathbf{b} \rangle^2 \leq \varepsilon$. Similarly, we can show $\langle \mathbf{a}, \mathbf{b} \rangle \geq -\sqrt{\varepsilon}$. Now, we can bound $\langle \mathrm{proj}(\mathbf{a}), \mathbf{b} \rangle$ as follows:

$$\langle \mathrm{proj}(\mathbf{a}), \mathbf{b} \rangle = \langle \mathrm{proj}(\mathbf{a}), \widehat{\mathbf{b}}_{ex} + \Delta\mathbf{b} \rangle = \langle \mathrm{proj}(\mathbf{a}), \widehat{\mathbf{b}}_{ex} \rangle + \langle \mathrm{proj}(\mathbf{a}), \Delta\mathbf{b} \rangle \leq \langle \mathrm{proj}(\mathbf{a}), \widehat{\mathbf{b}}_{ex} \rangle + \|\mathrm{proj}(\mathbf{a})\|_2 \tau \tag{34}$$

$$\overset{(a)}{\leq} \langle \mathrm{proj}(\mathbf{a}), \widehat{\mathbf{b}}_{ex} \rangle + \|\mathbf{a}\|_2 r \overset{(33)}{\leq} \sqrt{\varepsilon}, \tag{35}$$

where (a) follows because $\|\mathrm{proj}(\mathbf{a})\|_2 \leq \|\mathbf{a}\|_2$. Similarly, we can show $\langle \mathrm{proj}(\mathbf{a}), \mathbf{b} \rangle \geq -\sqrt{\varepsilon}$ completing the proof.

E. Proof of Theorem B.2: Robust QCQP with infinite constraints

Fix any $\widehat{r}_e \geq 0$, $\widehat{\theta}_e \in [0, 2\pi]$, $\Delta \geq 0$, and $\phi \in [0, \pi/2]$. For any $\mathbf{b}$, let $\Lambda_{\mathbf{a}}(\mathbf{b}) \triangleq \{\mathbf{a} : \langle \mathbf{a}, \mathbf{b} \rangle^2 \leq \varepsilon \text{ and } \mathbf{a} \in \mathcal{B}(0,1)\}$ be the set of all $\mathbf{a}$ satisfying the fairness constraint in (8) w.r.t $\mathbf{b}$. For a given $\mathbf{b}$, $\Lambda_{\mathbf{a}}(\mathbf{b})$ can be constructed as described in the proof of Lemma 1 in Appendix C.3. See Figure 6 for reference.

Now, to show that the optimization problem in (11) is equivalent to the optimization problem in (12), it suffices to show the corresponding constraints are equivalent, i.e.,

$$\bigcap_{\mathbf{b} \in \mathcal{A}(\Delta, \phi)} \Lambda_{\mathbf{a}}(\mathbf{b}) = \bigcap_{\mathbf{b} \in \overline{\mathcal{A}}(\Delta, \phi)} \Lambda_{\mathbf{a}}(\mathbf{b}). \tag{36}$$

Consider any $\mathbf{b}_1, \mathbf{b}_2 \in \mathcal{A}(\Delta, \phi)$ of the form: $\mathbf{b}_1 = r_1(\cos\theta, \sin\theta)$ and $\mathbf{b}_2 = r_2(\cos\theta, \sin\theta)$ for some r_1, r_2, θ such that $|r_1 - \widehat{r}_e| \leq \Delta$, $|r_2 - \widehat{r}_e| \leq \Delta$, and $|\theta - \widehat{\theta}_e| \leq \phi$. Without loss of generality, let $r_1 \geq r_2$. Now, it is straightforward to see that $\Lambda_{\mathbf{a}}(\mathbf{b}_1) \subseteq \Lambda_{\mathbf{a}}(\mathbf{b}_2)$. Therefore, (36) follows from the definition of $\overline{\mathcal{A}}(\Delta, \phi)$.

E.1. Characterizing optimal a in (11)

Fix any $\varepsilon > 0$. Let $\widehat{\mathbf{b}}_{ex} = \widehat{r}_e(\cos\widehat{\theta}_e, \sin\widehat{\theta}_e)$ for some $\widehat{r}_e > 0$ and $\widehat{\theta}_e \in [0, 2\pi]$. Consider some $\Delta \geq 0$ and $\phi \in [0, \pi/2]$ such that $\mathbf{b}_{ex} \in \mathcal{A}(\Delta, \phi)$ with probability at least $1 - \delta$ where $\mathcal{A}(\Delta, \phi)$ is as defined in (10). Let $\mathbf{b}_{yx} = r_y(\cos\theta_y, \sin\theta_y)$ for some $r_y > 0$ and $\theta_y \in [0, 2\pi]$. Let $\overline{\mathbf{a}}$ denote the set of optimal $\mathbf{a}$ in (11), i.e.,

$$\overline{\mathbf{a}} = \arg\max_{\mathbf{a} \in \mathcal{B}(0,1)} \langle \mathbf{a}, \mathbf{b}_{yx} \rangle^2 \text{ s.t } \langle \mathbf{a}, \mathbf{b} \rangle^2 \leq \varepsilon \text{ for all } \mathbf{b} \in \mathcal{A}(\Delta, \phi). \tag{37}$$

Then, the following Lemma characterizes $\overline{\mathbf{a}}$ depending on the values of $\widehat{\mathbf{b}}_{ex}$, $\mathbf{b}_{yx}$, Δ, ϕ, and ε.

Lemma 2. *Let $\overline{\alpha} \in [0, 2\pi]$ be such that $\cos\overline{\alpha} = \sqrt{\varepsilon}/(\widehat{r}_e + \Delta)$ and $\sin\overline{\alpha} = \sqrt{1 - \varepsilon/(\widehat{r}_e + \Delta)^2}$. Define the function $\mathbf{a} : [0, 2\pi] \to [-1, 1]^2$ such that $\mathbf{a}(\theta) = (\cos\theta, \sin\theta)$. Then,*

1. *If $\sqrt{\varepsilon} \geq (\widehat{r}_e + \Delta)\sin\phi$:*

Case 2.1. $\bar{\mathbf{a}} = \frac{\sqrt{\varepsilon}}{\widehat{r}_e+\Delta}\{\mathbf{a}(\theta_y), \mathbf{a}(\pi+\theta_y)\} \iff \theta_y \in [\widehat{\theta}_e - \phi, \widehat{\theta}_e + \phi] \cup [\pi + \widehat{\theta}_e - \phi, \pi + \widehat{\theta}_e + \phi]$,

Case 2.2. $\bar{\mathbf{a}} = \{\mathbf{a}(\widehat{\theta}_e + \phi + \overline{\alpha}), \mathbf{a}(\pi + \widehat{\theta}_e + \phi + \overline{\alpha})\} \iff \theta_y \in [\widehat{\theta}_e + \phi, \widehat{\theta}_e + \phi + \overline{\alpha}] \cup [\pi + \widehat{\theta}_e + \phi, \pi + \widehat{\theta}_e + \phi + \overline{\alpha}]$,

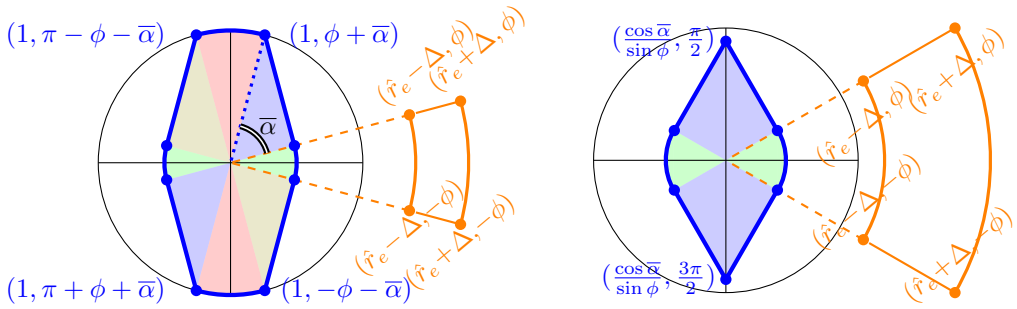
Case 2.3. $\bar{\mathbf{a}} = \{\mathbf{a}(\theta_y), \mathbf{a}(\pi+\theta_y)\} \iff \theta_y \in [\widehat{\theta}_e + \phi + \overline{\alpha}, \pi + \widehat{\theta}_e - \phi - \overline{\alpha}] \cup [\pi + \widehat{\theta}_e + \phi + \overline{\alpha}, 2\pi + \widehat{\theta}_e - \phi - \overline{\alpha}]$,

Case 2.4. $\bar{\mathbf{a}} = \{\mathbf{a}(\pi + \widehat{\theta}_e - \phi - \overline{\alpha}), \mathbf{a}(\widehat{\theta}_e - \phi - \overline{\alpha})\} \iff \theta_y \in [\pi + \widehat{\theta}_e - \phi - \overline{\alpha}, \pi + \widehat{\theta}_e - \phi] \cup [\widehat{\theta}_e - \phi - \overline{\alpha}, \widehat{\theta}_e - \phi]$.

2. *If* $\sqrt{\varepsilon} \leq (\widehat{r}_e + \Delta)\sin\phi$*:*

Case 3.1. $\bar{\mathbf{a}} = \frac{\sqrt{\varepsilon}}{\widehat{r}_e+\Delta}\{\mathbf{a}(\theta_y), \mathbf{a}(\pi+\theta_y)\} \iff \theta_y \in [\widehat{\theta}_e - \phi, \widehat{\theta}_e + \phi] \cup [\pi + \widehat{\theta}_e - \phi, \pi + \widehat{\theta}_e + \phi]$,

Case 3.2. $\bar{\mathbf{a}} = \frac{\sqrt{\varepsilon}}{(\widehat{r}_e+\Delta)\sin\phi}\{\mathbf{a}(\frac{\pi}{2} + \widehat{\theta}_e), \mathbf{a}(\frac{-\pi}{2} + \widehat{\theta}_e)\} \iff \theta_y \in [\widehat{\theta}_e + \phi, \pi + \widehat{\theta}_e - \phi] \cup [\pi + \widehat{\theta}_e + \phi, 2\pi + \widehat{\theta}_e - \phi]$.



(a) QCQP in (11) if $\sqrt{\varepsilon} \geq (\widehat{r}_e + \Delta)\sin\phi$ (b) QCQP in (11) if $\sqrt{\varepsilon} \leq (\widehat{r}_e + \Delta)\sin\phi$

Figure 7. Visualizing QCQP in (11) for $\varepsilon = 0.9$, $\widehat{r}_e = 1.6$, and $\widehat{\theta}_e = 0$. We set $\Delta = 0.2$ and $\phi = 15$ for panel (a), and $\Delta = 0.4$ and $\phi = 30$ 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 shown in orange. The region enclosed by the solid blue lines and blue arcs is the feasible space in (11). In Panel (a), the feasible space is shaded in green, blue, red, and olive to represent Case 2.1., Case 2.2., Case 2.3., and Case 2.4. respectively. In Panel (b), the feasible space is shaded in green and blue to represent Case 3.1. and Case 3.2., respectively.

Proof. For any $\mathbf{b}$, let $\Lambda_{\mathbf{a}}(\mathbf{b}) \triangleq \{\mathbf{a} : \langle \mathbf{a}, \mathbf{b}\rangle^2 \leq \varepsilon \text{ and } \mathbf{a} \in \mathcal{B}(0,1)\}$ be the set of all $\mathbf{a}$ satisfying the fairness constraint in (8) w.r.t $\mathbf{b}$. Then, from (36), the constraint in (37) is equivalent to $\cap_{\mathbf{b}\in\overline{\mathcal{A}}(\Delta,\phi)}\Lambda_{\mathbf{a}}(\mathbf{b})$. Consider any $\mathbf{b}_1 \in \overline{\mathcal{A}}(\Delta, \phi)$. Then, $\mathbf{b}_1 = (\widehat{r}_e + \Delta)(\cos\theta, \sin\theta)$ for some $|\theta - \widehat{\theta}_e| \leq \phi$. Further, $\Lambda_{\mathbf{a}}(\mathbf{b}_1)$ can be constructed as described in the proof of Lemma 1 in Appendix C.3. See Figure 6 for reference. From standard algebra and trigonometry, it is easy to see the two straight lines forming the boundary of $\Lambda_{\mathbf{a}}(\mathbf{b}_1)$ in Figure 6 intersect the y-axis at points $(0, \frac{\sqrt{\varepsilon}}{(\widehat{r}_e+\Delta)\sin\theta})$ and $(0, \frac{-\sqrt{\varepsilon}}{(\widehat{r}_e+\Delta)\sin\theta})$. Then, depending on whether these points lie inside or outside $\mathcal{B}(0,1)$ when $\theta = \phi$, we have two cases.

1. Suppose $\sqrt{\varepsilon} \geq (\widehat{r}_e + \Delta)\sin\phi$. Then, it is straightforward to see that $\cap_{\mathbf{b}\in\overline{\mathcal{A}}(\Delta,\phi)}\Lambda_{\mathbf{a}}(\mathbf{b})$ is the region enclosed by the solid blue lines and blue arcs in Figure 7(a) (see Figure 4(a) for reference), i.e., the union of the regions shaded in various colors. The rest of the proof is similar to the proof of Lemma 1. Lastly, it is straightforward to verify that Case 2.1., Case 2.2., Case 2.3., and Case 2.4. correspond to the regions shaded in green, blue, red, and olive, respectively.

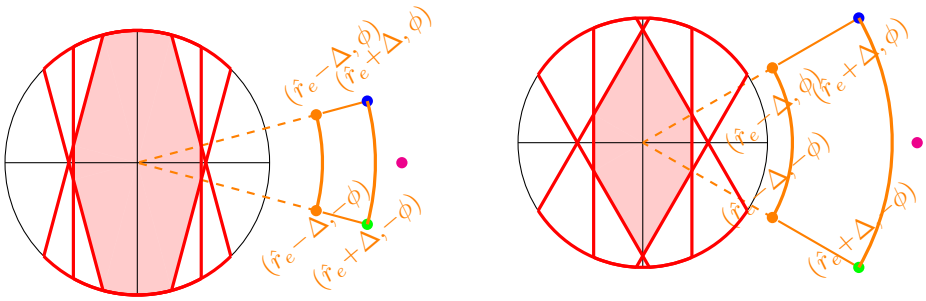
2. Suppose $\sqrt{\varepsilon} \leq (\widehat{r}_e + \Delta)\sin\phi$. Then, it is straightforward to see that $\cap_{\mathbf{b}\in\overline{\mathcal{A}}(\Delta,\phi)}\Lambda_{\mathbf{a}}(\mathbf{b})$ is the region enclosed by the solid blue lines and blue arcs in Figure 7(b) (see Figure 4(b) for reference), i.e., the union of the regions shaded in various colors. The rest of the proof is similar to the proof of Lemma 1. Lastly, it is straightforward to verify that Case 3.1. and Case 3.2. correspond to the regions shaded in green and blue, respectively.

□

F. Proof of Theorem B.3: Robust QCQP with 3 constraints

Fix any $\widehat{r}_e \geq 0$, $\theta_e \in [0, 2\pi]$, $\Delta \geq 0$, and $\phi \in [0, \pi/2]$. For any $\mathbf{b}$, let $\Lambda_{\mathbf{a}}(\mathbf{b}) \triangleq \{\mathbf{a} : \langle \mathbf{a}, \mathbf{b} \rangle^2 \leq \varepsilon \text{ and } \mathbf{a} \in \mathcal{B}(0,1)\}$ be the set of all $\mathbf{a}$ satisfying the fairness constraint in (8) w.r.t $\mathbf{b}$. For a given $\mathbf{b}$, $\Lambda_{\mathbf{a}}(\mathbf{b})$ can be constructed as described in the proof of Lemma 1 in Appendix C.3. See Figure 6 for reference.

First, we note that the straight lines in the feasible space $\cap_{\mathbf{b} \in \overline{\mathcal{A}}(\Delta,\phi)} \Lambda_{\mathbf{a}}(\mathbf{b})$ in Figure 7(a) and Figure 7(b) are generated by the extreme points of the arc $\overline{\mathcal{A}}(\Delta, \phi)$. These extreme points are precisely $\mathbf{b}_{ex}^{(2)}$ and $\mathbf{b}_{ex}^{(3)}$. Therefore, the same straight lines also arise in the feasible space $\Lambda_{\mathbf{a}}(\mathbf{b}_{ex}^{(1)}) \cap \Lambda_{\mathbf{a}}(\mathbf{b}_{ex}^{(2)}) \cap \Lambda_{\mathbf{a}}(\mathbf{b}_{ex}^{(3)})$ in Figure 8(a) and Figure 8(b). Second, we note that the arcs in the feasible space $\cap_{\mathbf{b} \in \overline{\mathcal{A}}(\Delta,\phi)} \Lambda_{\mathbf{a}}(\mathbf{b})$ in Figure 7(a) and Figure 7(b) are at a distance $\sqrt{\varepsilon}/(\widehat{r}_e + \Delta)$ away from the origin. The point $\mathbf{b}_{ex}^{(1)}$ is chosen precisely such that the space $\Lambda_{\mathbf{a}}(\mathbf{b}_{ex}^{(1)})$ intersects each of these arcs at its extreme points (see Figure 8(a) and Figure 8(b)). Therefore, it is easy to see that the boundary of the feasible space $\cap_{\mathbf{b} \in \overline{\mathcal{A}}(\Delta,\phi)} \Lambda_{\mathbf{a}}(\mathbf{b})$ is not closer to the origin than the boundary of the feasible space $\Lambda_{\mathbf{a}}(\mathbf{b}_{ex}^{(1)}) \cap \Lambda_{\mathbf{a}}(\mathbf{b}_{ex}^{(2)}) \cap \Lambda_{\mathbf{a}}(\mathbf{b}_{ex}^{(3)})$. This completes the proof.



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

Figure 8. Visualizing QCQP in (13) for $\varepsilon = 0.9$, $\widehat{r}_e = 1.6$, and $\widehat{\theta}_e = 0$. We set $\Delta = 0.2$ and $\phi = 15$ for panel (a), and $\Delta = 0.4$ and $\phi = 30$ 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 shown in orange. The shaded region is the feasible space in (13). The points $\mathbf{b}_{ex}^{(1)}$, $\mathbf{b}_{ex}^{(2)}$, and $\mathbf{b}_{ex}^{(3)}$ are shown in magenta, blue and green, respectively.

F.1. Characterizing optimal a in (13)

Fix any $\varepsilon > 0$. Let $\widehat{\mathbf{b}}_{ex} = \widehat{r}_e(\cos\widehat{\theta}_e, \sin\widehat{\theta}_e)$ for some $\widehat{r}_e > 0$ and $\widehat{\theta}_e \in [0, 2\pi]$. Consider some $\Delta \geq 0$ and $\phi \in [0, \pi/2]$ such that $\mathbf{b}_{ex} \in \mathcal{A}(\Delta, \phi)$ where $\mathcal{A}(\Delta, \phi)$ is as defined in (10). Let $\mathbf{b}_{yx} = r_y(\cos\theta_y, \sin\theta_y)$ for some $r_y > 0$ and $\theta_y \in [0, 2\pi]$. Let $\overline{\mathbf{a}}$ denote the set of optimal $\mathbf{a}$ in (13), i.e.,

$$\overline{\mathbf{a}} = \arg\max_{\mathbf{a} \in \mathcal{B}(0,1)} \left\langle \mathbf{a}, \mathbf{b}_{yx} \right\rangle^2 \text{ s.t } \left\langle \mathbf{a}, \mathbf{b}_{ex}^{(1)} \right\rangle^2 \leq \varepsilon, \left\langle \mathbf{a}, \mathbf{b}_{ex}^{(2)} \right\rangle^2 \leq \varepsilon, \left\langle \mathbf{a}, \mathbf{b}_{ex}^{(3)} \right\rangle^2 \leq \varepsilon. \tag{38}$$

Then, the following Lemma characterizes $\overline{\mathbf{a}}$ depending on the values of $\widehat{\mathbf{b}}_{ex}$, $\mathbf{b}_{yx}$, Δ, ϕ, and ε. The proof is similar to the proof of Lemma 2 and is omitted for brevity.

Lemma 3. *Let $\overline{\alpha} \in [0, 2\pi]$ be such that $\cos\overline{\alpha} = \sqrt{\varepsilon}/(\widehat{r}_e + \Delta)$ and $\sin\overline{\alpha} = \sqrt{1 - \varepsilon/(\widehat{r}_e + \Delta)^2}$. Define the function $\mathbf{a} : [0, 2\pi] \to [-1, 1]^2$ such that $\mathbf{a}(\theta) = (\cos\theta, \sin\theta)$. Then,*

1. *If $\sqrt{\varepsilon} \geq (\widehat{r}_e + \Delta)\sin\phi$:*

   Case 4.1. $\overline{\mathbf{a}} = \frac{\sqrt{\varepsilon}}{\widehat{r}_e + \Delta}\left\{\mathbf{a}(\widehat{\theta}_e + \phi), \mathbf{a}(\pi + \widehat{\theta}_e + \phi)\right\} \iff \theta_y \in [\widehat{\theta}_e, \widehat{\theta}_e + \phi] \cup [\pi + \widehat{\theta}_e, \pi + \widehat{\theta}_e + \phi]$,

   Case 4.2. $\overline{\mathbf{a}} = \left\{\mathbf{a}(\widehat{\theta}_e + \phi + \overline{\alpha}), \mathbf{a}(\pi + \widehat{\theta}_e + \phi + \overline{\alpha})\right\} \iff \theta_y \in [\widehat{\theta}_e + \phi, \widehat{\theta}_e + \phi + \overline{\alpha}] \cup [\pi + \widehat{\theta}_e + \phi, \pi + \widehat{\theta}_e + \phi + \overline{\alpha}]$,

   Case 4.3. $\overline{\mathbf{a}} = \left\{\mathbf{a}(\theta_y), \mathbf{a}(\pi + \theta_y)\right\} \iff \theta_y \in [\widehat{\theta}_e + \phi + \overline{\alpha}, \pi + \widehat{\theta}_e - \phi - \overline{\alpha}] \cup [\pi + \widehat{\theta}_e + \phi + \overline{\alpha}, 2\pi + \widehat{\theta}_e - \phi - \overline{\alpha}]$,

   Case 4.4. $\overline{\mathbf{a}} = \left\{\mathbf{a}(\pi + \widehat{\theta}_e - \phi - \overline{\alpha}), \mathbf{a}(\widehat{\theta}_e - \phi - \overline{\alpha})\right\} \iff \theta_y \in [\pi + \widehat{\theta}_e - \phi - \overline{\alpha}, \pi + \widehat{\theta}_e - \phi] \cup [\widehat{\theta}_e - \phi - \overline{\alpha}, \widehat{\theta}_e - \phi]$,

   Case 4.5. $\overline{\mathbf{a}} = \frac{\sqrt{\varepsilon}}{\widehat{r}_e + \Delta}\left\{\mathbf{a}(\widehat{\theta}_e - \phi), \mathbf{a}(\pi + \widehat{\theta}_e - \phi)\right\} \iff \theta_y \in [\pi + \widehat{\theta}_e - \phi, \pi + \widehat{\theta}_e] \cup [\widehat{\theta}_e - \phi, \widehat{\theta}_e]$.

2. *If* $\sqrt{\varepsilon} \leq (\widehat{r}_e + \Delta)\sin\phi$*:*

Case 5.1. $\overline{\mathbf{a}} = \frac{\sqrt{\varepsilon}}{\widehat{r}_e+\Delta}\left\{\mathbf{a}(\widehat{\theta}_e+\phi), \mathbf{a}(\pi+\widehat{\theta}_e+\phi)\right\} \iff \theta_y \in [\widehat{\theta}_e, \widehat{\theta}_e+\phi] \cup [\pi+\widehat{\theta}_e, \pi+\widehat{\theta}_e+\phi]$,

Case 5.2. $\overline{\mathbf{a}} = \frac{\sqrt{\varepsilon}}{(\widehat{r}_e+\Delta)\sin\phi}\left\{\mathbf{a}(\frac{\pi}{2}+\widehat{\theta}_e), \mathbf{a}(\frac{-\pi}{2}+\widehat{\theta}_e)\right\} \iff \theta_y \in [\widehat{\theta}_e+\phi, \pi+\widehat{\theta}_e-\phi] \cup [\pi+\widehat{\theta}_e+\phi, 2\pi+\widehat{\theta}_e-\phi]$,

Case 5.3. $\overline{\mathbf{a}} = \frac{\sqrt{\varepsilon}}{\widehat{r}_e+\Delta}\left\{\mathbf{a}(\widehat{\theta}_e-\phi), \mathbf{a}(\pi+\widehat{\theta}_e-\phi)\right\} \iff \theta_y \in [\pi+\widehat{\theta}_e-\phi, \pi+\widehat{\theta}_e] \cup [\widehat{\theta}_e-\phi, \widehat{\theta}_e]$.

G. Analyzing the power of labeled sensitive attributes

In this section, we consider the case where uncertainty only stems from sensitive attributes missing at random, and the uncertainty can be improved by collecting more labeled sensitive attributes. We characterize the difference in the optimal objectives in (8) and (13) to obtain the power of each new labeled sensitive attribute.

Fix any $\varepsilon > 0$. For any n, let $\widehat{\mathbf{b}}_{ex}(n) \triangleq \widehat{r}_e(n)(\cos\widehat{\theta}_e(n), \sin\widehat{\theta}_e(n))$ denote the estimate of $\mathbf{b}_{ex}$ as a function of n and let $\Delta(n) \geq 0$ and $\phi(n) \in [0, \pi/2]$ denote the corresponding uncertainty parameters as a function of n, i.e., $\mathbf{b}_{ex} = r_e(\cos\theta_e, \sin\theta_e) \in \mathcal{A}(\Delta(n), \phi(n))$, with probability $1-\delta$. Let $\Psi(n)$ denote the difference in the optimal objectives in (8) and (13), i.e., $\Psi(n) \triangleq \Psi - \Psi_2(n)$ where

$$\Psi \triangleq \max_{\mathbf{a}\in\mathcal{B}(0,1)} \left\langle \mathbf{a}, \mathbf{b}_{yx}\right\rangle^2 \quad \text{s.t} \quad \left\langle \mathbf{a}, \mathbf{b}_{ex}\right\rangle^2 \leq \varepsilon \tag{39}$$

and

$$\Psi_2(n) \triangleq \max_{\mathbf{a}\in\mathcal{B}(0,1)} \left\langle \mathbf{a}, \mathbf{b}_{yx}\right\rangle^2 \quad \text{s.t} \quad \left\langle \mathbf{a}, \mathbf{b}_{ex}^{(1)}(n)\right\rangle^2 \leq \varepsilon, \left\langle \mathbf{a}, \mathbf{b}_{ex}^{(2)}(n)\right\rangle^2 \leq \varepsilon, \left\langle \mathbf{a}, \mathbf{b}_{ex}^{(3)}(n)\right\rangle^2 \leq \varepsilon. \tag{40}$$

Then, the following lemma characterizes $\Psi(n)$ as a function of n when $\theta_e \geq \widehat{\theta}_e(n)$. The characterization when $\theta_e \leq \widehat{\theta}_e(n)$ can be obtained analogously. The proof follows from Lemma 1 and Lemma 3, and is omitted for brevity.

Lemma 4. *Let* $\alpha, \overline{\alpha}(n) \in [0, 2\pi]$ *be such that* (a) $\cos\alpha = \sqrt{\varepsilon}/r_e$ *and* $\sin\alpha = \sqrt{1-\varepsilon/r_e^2}$ *and* (b) $\cos\overline{\alpha}(n) = \sqrt{\varepsilon}/(\widehat{r}_e(n)+\Delta(n))$ *and* $\sin\overline{\alpha}(n) = \sqrt{1-\varepsilon/(\widehat{r}_e(n)+\Delta(n))^2}$*. Then,*

1. *If* $\sqrt{\varepsilon} \geq (\widehat{r}_e(n)+\Delta(n))\sin\phi(n)$*:*

Case A. $\Psi(n) = \cos(\alpha+\theta_e-\theta_y) - \cos\overline{\alpha}(n)\cos(\widehat{\theta}_e(n)+\phi(n)-\theta_y)$ *when* $\theta_y \in [\theta_e, \widehat{\theta}_e(n)+\phi(n)] \cup [\pi+\theta_e, \pi+\widehat{\theta}_e(n)+\phi(n)]$,

Case B. $\Psi(n) = \cos(\alpha+\theta_e-\theta_y) - \cos(\widehat{\theta}_e(n)+\overline{\alpha}(n)+\phi(n)-\theta_y)$ *when* $\theta_y \in [\widehat{\theta}_e(n)+\phi(n), \alpha+\theta_e] \cup [\pi+\widehat{\theta}_e(n)+\phi(n), \pi+\alpha+\theta_e]$,

Case C. $\Psi(n) = 1 - \cos(\widehat{\theta}_e(n)+\overline{\alpha}(n)+\phi(n)-\theta_y)$ *when* $\theta_y \in [\alpha+\theta_e, \widehat{\theta}_e(n)+\phi(n)+\overline{\alpha}(n)] \cup [\pi+\alpha+\theta_e, \pi+\widehat{\theta}_e(n)+\phi(n)+\overline{\alpha}(n)]$,

Case D. $\Psi(n) = 0$ *when* $\theta_y \in [\widehat{\theta}_e(n)+\phi(n)+\overline{\alpha}(n), \pi+\widehat{\theta}_e(n)-\phi(n)-\overline{\alpha}(n)] \cup [\pi+\widehat{\theta}_e(n)+\phi(n)+\overline{\alpha}(n), 2\pi+\widehat{\theta}_e(n)-\phi(n)-\overline{\alpha}(n)]$,

Case E. $\Psi(n) = 1 - \cos(\theta_y+\overline{\alpha}(n)+\phi(n)-\widehat{\theta}_e(n))$ *when* $\theta_y \in [\pi+\widehat{\theta}_e(n)-\phi(n)-\overline{\alpha}(n), \pi-\alpha+\theta_e] \cup [\widehat{\theta}_e(n)-\phi(n)-\overline{\alpha}(n), -\alpha+\theta_e]$,

Case F. $\Psi(n) = \cos(\theta_y+\alpha-\theta_e) - \cos(\theta_y+\overline{\alpha}(n)+\phi(n)-\widehat{\theta}_e(n))$*when* $\theta_y \in [\pi-\alpha+\theta_e, \pi+\widehat{\theta}_e(n)-\phi(n)] \cup [-\alpha+\theta_e, \widehat{\theta}_e(n)-\phi(n)]$,

Case G. $\Psi(n) = \cos(\theta_y+\alpha-\theta_e) - \cos\overline{\alpha}(n)\cos(\theta_y+\phi(n)-\widehat{\theta}_e(n))$ *when* $\theta_y \in [\pi+\widehat{\theta}_e(n)-\phi(n), \pi+\widehat{\theta}_e(n)] \cup [\widehat{\theta}_e(n)-\phi(n), \widehat{\theta}_e(n)]$,

Case H. $\Psi(n) = \cos(\theta_y+\alpha-\theta_e) - \cos\overline{\alpha}(n)\cos(\theta_y-\phi(n)-\widehat{\theta}_e(n))$ *when* $\theta_y \in [\pi+\widehat{\theta}_e(n), \pi+\theta_e] \cup [\widehat{\theta}_e(n), \theta_e]$.

2. *If* $\sqrt{\varepsilon}s \leq (\widehat{r}_e(n)+\Delta(n))\sin\phi(n)$*:*

Case A. $\Psi(n) = \cos(\alpha+\theta_e-\theta_y) - \cos\overline{\alpha}(n)\cos(\widehat{\theta}_e(n)+\phi(n)-\theta_y)$ *when* $\theta_y \in [\theta_e, \widehat{\theta}_e(n)+\phi(n)] \cup [\pi+\theta_e, \pi+\widehat{\theta}_e(n)+\phi(n)]$,

Case B. $\Psi(n) = \cos(\alpha + \theta_e - \theta_y) - \cos\overline{\alpha}(n)\frac{\sin(\theta_y - \widehat{\theta}_e(n))}{\sin\phi(n)}$ *when* $\theta_y \in [\widehat{\theta}_e(n) + \phi(n), \alpha + \theta_e] \cup [\pi + \widehat{\theta}_e(n) + \phi(n), \pi + \alpha + \theta_e]$,

Case C. $\Psi(n) = 1 - \cos\overline{\alpha}(n)\frac{\sin(\theta_y - \widehat{\theta}_e(n))}{\sin\phi(n)}$ *when* $\theta_y \in [\alpha + \theta_e, \pi - \alpha + \theta_e] \cup [\pi + \alpha + \theta_e, 2\pi - \alpha + \theta_e]$,

Case D. $\Psi(n) = \cos(\theta_y + \alpha - \theta_e) - \cos\overline{\alpha}(n)\frac{\sin(\theta_y - \widehat{\theta}_e(n))}{\sin\phi(n)}$ *when* $\theta_y \in [\pi - \alpha + \theta_e, \pi + \widehat{\theta}_e(n) - \phi(n)] \cup [-\alpha + \theta_e, \widehat{\theta}_e(n) - \phi(n)]$,

Case E. $\Psi(n) = \cos(\theta_y + \alpha - \theta_e) - \cos\overline{\alpha}(n)\cos(\theta_y + \phi(n) - \widehat{\theta}_e(n))$ *when* $\theta_y \in [\pi + \widehat{\theta}_e(n) - \phi(n), \pi + \widehat{\theta}_e(n)] \cup [\widehat{\theta}_e(n) - \phi(n), \widehat{\theta}_e(n)]$,

Case F. $\Psi(n) = \cos(\theta_y + \alpha - \theta_e) - \cos\overline{\alpha}(n)\cos(\theta_y - \phi(n) - \widehat{\theta}_e(n))$ *when* $\theta_y \in [\pi + \widehat{\theta}_e(n), \pi + \theta_e] \cup [\widehat{\theta}_e(n), \theta_e]$.

Remark 1. *The cases above can be classified into one of the following three categories:*

Category 1. ***Any uncertainty hurts:*** *Here, the optimal performance of the robust QCQP in (11) matches the optimal performance in (8) only when all the uncertainty is removed, i.e., when $n \to N$. When the uncertainty parameters are not too large, i.e., $(\widehat{r}_e(n) + \Delta(n))\sin\phi(n) \leq \sqrt{\varepsilon}$, Case A., Case B., Case F., Case G., and Case H. fall into this category. When the uncertainty parameters are large, i.e., $(\widehat{r}_e(n) + \Delta(n))\sin\phi(n) \geq \sqrt{\varepsilon}$, Case A., Case B., Case D., Case E., and Case F. fall into this category.*

Category 2. ***Some uncertainty does not hurt:*** *Here, the optimal performance of the robust QCQP in (11) matches the optimal performance in (8) when some uncertainty is removed, i.e., by collecting some additional labeled sensitive attributes. When the uncertainty parameters are not too large, i.e., $(\widehat{r}_e(n) + \Delta(n))\sin\phi(n) \leq \sqrt{\varepsilon}$, Case C. and Case E. fall into this category. When the uncertainty parameters are large, i.e., $(\widehat{r}_e(n) + \Delta(n))\sin\phi(n) \geq \sqrt{\varepsilon}$, Case C. fall into this category.*

Category 3. ***Uncertainty does not hurt:*** *Here, the optimal performance of the robust QCQP in (11) matches the optimal performance in (8) without removing any uncertainty, i.e., without collecting any additional labeled sensitive attributes. This only happens in Case D. where $\Psi(n) = 0$ when the uncertainty parameters are not too large, i.e., $(\widehat{r}_e(n) + \Delta(n))\sin\phi(n) \leq \sqrt{\varepsilon}$. Such a situation arises when $\mathbf{b}_{yx}$ and $\widehat{\mathbf{b}}_{ex}$ are very close to being perpendicular to each other. In other words, when $\mathbf{b}_{yx}$ and $\widehat{\mathbf{b}}_{ex}$ are close to being independent, our proposed robust QCQP achieves optimal performance while ensuring a strict fairness guarantee. We call such a phenomenon as "free fairness" (see Corollary 1).*

H. Monotonic performance with new labeled sensitive attributes

As in the previous section, we consider the case where uncertainty only stems from sensitive attributes missing at random. Then, we show that the optimal objective in (13) either monotonically increases or coincides with the optimal objective in (8) whenever the uncertainty set monotonically decreases with n.

To that end, below, we express the uncertainty set $\mathcal{A}(\Delta, \phi)$ and the associated uncertainty parameters Δ and ϕ as a function of n.

Proposition 3 (Uncertainty as a function of n)**.** *Fix any $\delta > 0$. Given n samples $\{(e^{(i)}, \boldsymbol{x}^{(i)})\}_{i\in[n]}$ of $(\mathsf{e}, \mathbf{x})$, let*

$$\widehat{\mathbf{b}}_{ex} \triangleq \boldsymbol{\Sigma}_{ee}^{-1/2}\widehat{\boldsymbol{\Sigma}}_{ex}\boldsymbol{\Sigma}_{xx}^{-1/2} \text{ with } \widehat{\boldsymbol{\Sigma}}_{ex} \triangleq \frac{1}{n}\sum_{i\in[n]} e^{(i)}\boldsymbol{x}^{(i)}. \tag{41}$$

Then, $\mathbf{b}_{ex} \in \mathcal{B}(\widehat{\mathbf{b}}_{ex}, \tau(n))$, with probability $1 - \delta$, where

$$\tau(n) \triangleq \|\boldsymbol{\Sigma}_{xx}^{-1/2}\|_2 \frac{c\sqrt{\sigma_e}\max_{i\in[d]}\{\sigma_i\}}{n\sqrt{d}} \log\frac{4}{\delta}, \tag{42}$$

with $\sigma_1^2, \sigma_2^2, \sigma_e^2$ denoting variances of $\mathbf{x}_1$, $\mathbf{x}_2$, e, respectively, and c is a universal constant. Further, for $d = 2$, $\mathbf{b}_{ex} \in \mathcal{A}(\Delta(n), \phi(n))$, with probability at least $1 - \delta$, where

$$\phi(n) \triangleq \sin^{-1}\left(\frac{\tau(n)}{\|\widehat{\mathbf{b}}_{ex}\|_2}\right) \text{ and } \Delta(n) \triangleq \tau(n). \tag{43}$$

Proof. For any sub-Gaussian (sub-exponential) random variable, we denote its sub-Gaussian (sub-exponential) norm by $\|\cdot\|_{\Psi_2}$ ($\|\cdot\|_{\Psi_1}$). For every $i \in [d]$, we note that x_i is a sub-Gaussian random variable with sub-Gaussian norm $\|x_i\|_{\Psi_2} = c\sigma_i$ where $\sigma_i^2 = \mathbf{\Sigma}_{x_i x_i}$. Similarly, e is a sub-Gaussian random variable with sub-Gaussian norm $\|e\|_{\Psi_2} = c\sigma_e$ where $\sigma_e^2 = \mathbf{\Sigma}_{ee}$. Then, from Vershynin (2018, Lemma. 2.7.7), for every $i \in [d]$, $x_i e$ is a sub-exponential random variable with sub-exponential norm $\|x_i e\|_{\Psi_1} = c\sigma_i\sigma_e$. Therefore, from Bernstein's inequality (Vershynin, 2018, Corollary 2.8.3), with probability $1-\delta$, we have $\|\widehat{\mathbf{\Sigma}}_{ex} - \mathbf{\Sigma}_{ex}\|_\infty \leq t$ whenever $t \geq \frac{c\sigma_e \max_{i\in[d]} \sigma_i}{n} \log \frac{2d}{\delta}$. As a result, we have $\|\widehat{\mathbf{\Sigma}}_{ex} - \mathbf{\Sigma}_{ex}\|_2 \leq t$ with probability $1-\delta$ whenever $t \geq \frac{c\sigma_e \max_{i\in[d]} \sigma_i}{n\sqrt{d}} \log \frac{2d}{\delta}$. Conditioning on this event, we have, with probability $1-\delta$,

$$\|\widehat{\mathbf{b}}_{ex} - \mathbf{b}_{ex}\|_2 \overset{(a)}{=} \|\mathbf{\Sigma}_{ee}^{-1/2}(\widehat{\mathbf{\Sigma}}_{ex} - \mathbf{\Sigma}_{ex})\mathbf{\Sigma}_{xx}^{-1/2}\|_2 \overset{(b)}{\leq} \frac{1}{\sqrt{\sigma_e}}\|\widehat{\mathbf{\Sigma}}_{ex} - \mathbf{\Sigma}_{ex}\|_2\|\mathbf{\Sigma}_{xx}^{-1/2}\|_2 \leq \frac{1}{\sqrt{\sigma_e}}\|\mathbf{\Sigma}_{xx}^{-1/2}\|_2 t, \tag{44}$$

where (a) follows from Definition 2 and (b) follows because induced matrix norms are sub-multiplicative. Then, letting $d = 2$ and $\tau(n) = \mathbf{\Sigma}_{ee}^{-1/2}\|\mathbf{\Sigma}_{xx}^{-1/2}\|_\infty t$, it is easy to verify $\mathbf{b}_{ex} \in \mathcal{A}(\Delta(n), \phi(n))$ whenever

$$\Delta(n) = \tau(n) \quad \text{and} \quad \phi(n) = \sin^{-1}\left(\frac{\tau(n)}{\|\widehat{\mathbf{b}}_{ex}\|_2}\right). \tag{45}$$

□

H.1. Monotonic performance with decrease in uncertainty set

Now, in the following theorem, we show that the optimal objective of the robust QCQP with 3 constraints in (13) either monotonically increases or coincides with the optimal objective of the QCQP in (8) whenever the uncertainty set $\mathcal{A}(\Delta(n), \phi(n))$ monotonically decreases with n.

Theorem H.1 (Monotonic performance)**.** *For any number of labeled sensitive attributes n, let $\mathcal{A}(\Delta(n), \phi(n))$ denote the uncertainty set containing $\mathbf{b}_{ex}$. Let $\overline{\mathbf{a}}(n)$ denote the optimal solution $\mathbf{a}$ in (13) as a function of n. If $\mathcal{A}(\Delta(n+1), \phi(n+1)) \subset \mathcal{A}(\Delta(n), \phi(n))$, then*

$$\left\langle \overline{\mathbf{a}}(n+1), \mathbf{b}_{yx} \right\rangle^2 > \left\langle \overline{\mathbf{a}}(n), \mathbf{b}_{yx} \right\rangle^2 \quad \textit{or} \tag{46}$$

$$\left\langle \overline{\mathbf{a}}(n+1), \mathbf{b}_{yx} \right\rangle^2 = \left\langle \overline{\mathbf{a}}(n), \mathbf{b}_{yx} \right\rangle^2 = \left\langle \mathbf{a}^\star, \mathbf{b}_{yx} \right\rangle^2, \tag{47}$$

where $\mathbf{a}^\star$ is the optimal solution of the QCQP in (8).

Fix any $\varepsilon > 0$. For any $\mathbf{b}$, let $\Lambda_{\mathbf{a}}(\mathbf{b}) \triangleq \{\mathbf{a} : \langle \mathbf{a}, \mathbf{b}\rangle^2 \leq \varepsilon \text{ and } \mathbf{a} \in \mathcal{B}(0,1)\}$ be the set of all $\mathbf{a}$ satisfying the fairness constraint in (8) w.r.t $\mathbf{b}$. For a given $\mathbf{b}$, $\Lambda_{\mathbf{a}}(\mathbf{b})$ can be constructed as described in the proof of Lemma 1 in Appendix C.3. See Figure 6 for reference.

Now, using $\mathcal{A}(\Delta(n+1), \phi(n+1)) \subset \mathcal{A}(\Delta(n), \phi(n))$, it is easy to see that $\widehat{r}_e(n+1) < \widehat{r}_e(n)$, $\left|\widehat{\theta}_e(n+1) + \phi(n+1)\right| < \left|\widehat{\theta}_e(n) + \phi(n)\right|$, $\left|\widehat{\theta}_e(n+1) - \phi(n+1)\right| < \left|\widehat{\theta}_e(n) - \phi(n)\right|$, and $\left|\cos\phi(n+1)\right| < \left|\cos\phi(n)\right|$. Then, it follows that the feasible space increases with n, i.e., $\Lambda_{\mathbf{a}}(\mathbf{b}_{ex}^{(1)}(n+1)) \cap \Lambda_{\mathbf{a}}(\mathbf{b}_{ex}^{(2)}(n+1)) \cap \Lambda_{\mathbf{a}}(\mathbf{b}_{ex}^{(3)}(n+1)) \supset \Lambda_{\mathbf{a}}(\mathbf{b}_{ex}^{(1)}(n)) \cap \Lambda_{\mathbf{a}}(\mathbf{b}_{ex}^{(2)}(n)) \cap \Lambda_{\mathbf{a}}(\mathbf{b}_{ex}^{(3)}(n))$. As a result, the optimal objective in (13) at $n+1$ is either less than the optimal objective in (13) at n or equal to the optimal objective in (13) at n. It remains to show that if the optimal objective in (13) at $n+1$ equals the optimal objective in (13) at n, then it also equals the optimal objective of the QCQP in (8). This follows directly from the expressions for $\Psi(n)$ in various cases of Lemma 4.

I. Additional experimental results

In this section, we provide additional empirical results as well as more implementation details.

I.1. Additional results for Gaussian data

Covariance matrices. First, we provide the covariance matrices of $(\mathbf{x}, y, e)$ used in the experiments in Appendix B.3:

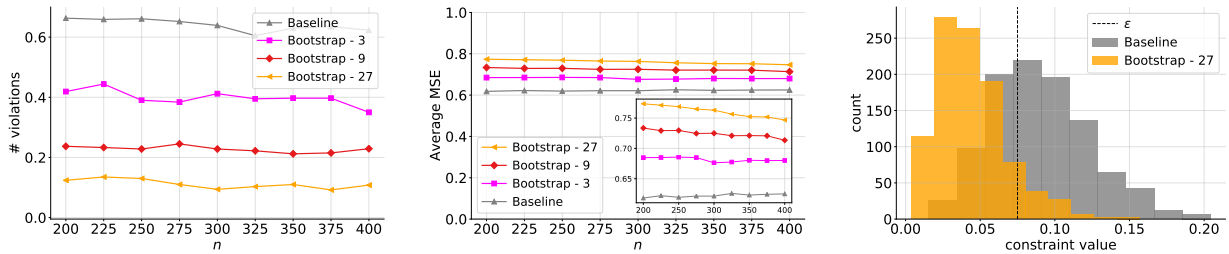
$$\Sigma_2^{\text{gen}} = \left[\begin{array}{cc|c|c} 1 & 0.1 & 0.5 & 0.4 \\ 0.1 & 1 & 0.5 & 0.25 \\ \hline 0.5 & 0.5 & 1 & 0.75 \\ \hline 0.4 & 0.25 & 0.75 & 1 \end{array}\right] \qquad \Sigma_2^{\text{fair}} = \left[\begin{array}{cc|c|c} 1 & 0.1 & 0.5 & 0.05 \\ 0.1 & 1 & 0.05 & 0.25 \\ \hline 0.5 & 0.05 & 1 & 0.75 \\ \hline 0.05 & 0.25 & 0.75 & 1 \end{array}\right]$$

The choice of Σ_2^{fair} is deliberate to demonstrate the free-fairness behavior. As pointed out in Category 3 of Remark 1, the free-fairness behavior may happen when $\mathbf{b}_{yx}$ and $\widehat{\mathbf{b}}_{ex}$ are almost perpendicular. It is easy to check that this can be achieved by the following steps, for example: (1) Strongly correlate e with x_1, and y with x_2. (2) Set the covariance of $\mathbf{x}$ to be close to the identity matrix.

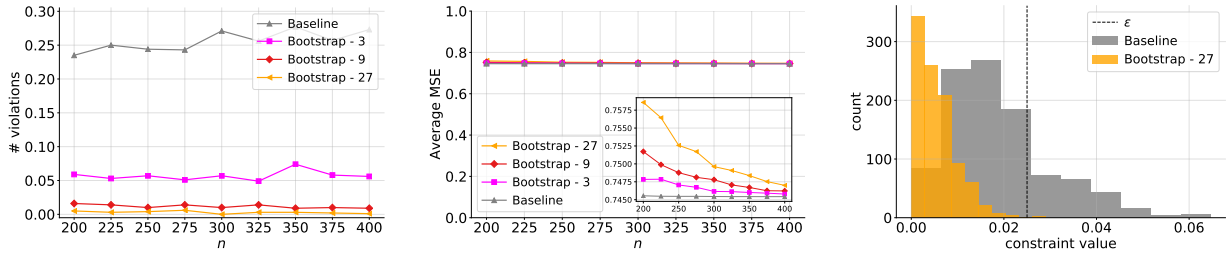
Results for $d = 3$. Next, we provide results for $d = 3$ where we use the following covariance matrices:

$$\Sigma_3^{\text{gen}} = \left[\begin{array}{ccc|c|c} 1 & 0.1 & 0.5 & 0.5 & 0.4 \\ 0.1 & 1 & 0.5 & 0.5 & 0.25 \\ 0.5 & 0.5 & 1 & 0.2 & 0.2 \\ \hline 0.5 & 0.5 & 0.2 & 1 & 0.75 \\ \hline 0.4 & 0.25 & 0.2 & 0.75 & 1 \end{array}\right] \qquad \Sigma_3^{\text{fair}} = \left[\begin{array}{ccc|c|c} 1 & 0.1 & 0.5 & 0.5 & 0.05 \\ 0.1 & 1 & 0.5 & 0.05 & 0.25 \\ 0.5 & 0.5 & 1 & 0.2 & 0.2 \\ \hline 0.5 & 0.05 & 0.2 & 1 & 0.75 \\ \hline 0.05 & 0.25 & 0.2 & 0.75 & 1 \end{array}\right]$$

The reasoning behind the choice of Σ_3^{fair} is the same as above. As we see in Figure 9 (a) and Figure 9 (b), `Bootstrap-S` has significantly less fairness violations compared to `Baseline` with a similar MSE even for $d = 3$.



(a) $d = 3$, $\Sigma^\star = \Sigma_3^{\text{gen}}$, $\varepsilon = 0.075$



(b) $d = 3$, $\Sigma^\star = \Sigma_3^{\text{fair}}$, $\varepsilon = 0.025$

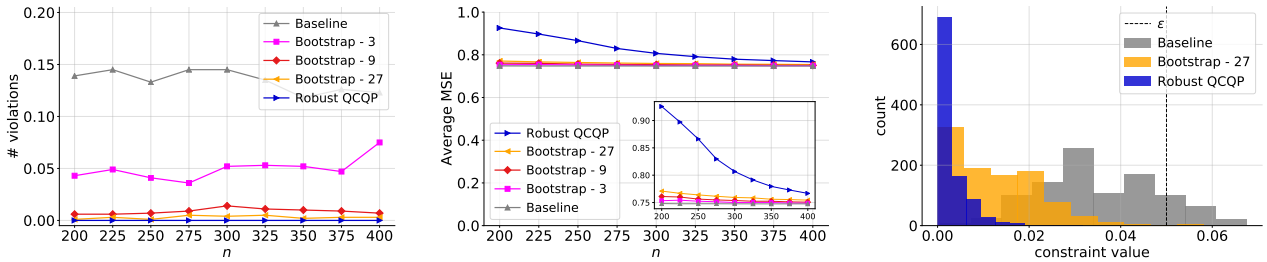
Figure 9. The performance of `Bootstrap-S` with `S` $\in \{3, 9, 27\}$ and `Baseline` for $d = 3$ and various $\Sigma^\star$. In the left column, we plot the fraction of violations of the fairness constraint $\langle \mathbf{a}, \mathbf{b}_{ex} \rangle^2 \leq \varepsilon$ vs. n; in the middle column, 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 $n = 250$.

A different choice of covariance matrices. Next, we provide results with a different choice of covariance matrices:

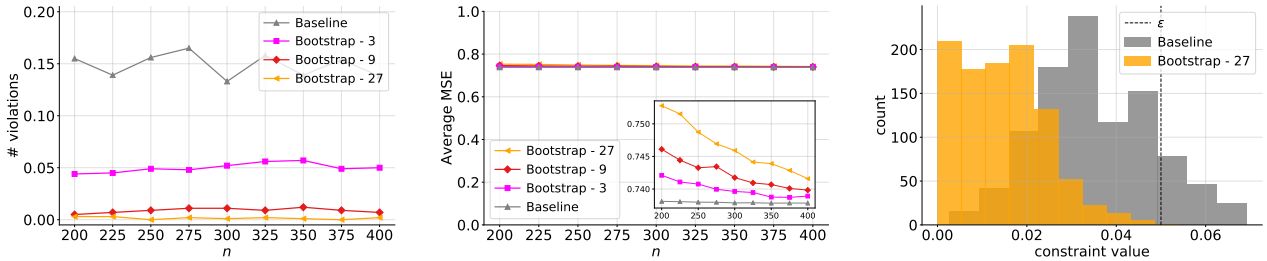
$$\Sigma_2^{\text{a}} = \left[\begin{array}{cc|c|c} 1 & 0.1 & 0.5 & 0.01 \\ 0.1 & 1 & 0.01 & 0.25 \\ \hline 0.5 & 0.01 & 1 & 0.75 \\ \hline 0.01 & 0.25 & 0.75 & 1 \end{array}\right] \qquad \Sigma_3^{\text{a}} = \left[\begin{array}{ccc|c|c} 1 & 0.1 & 0.5 & 0.5 & 0.01 \\ 0.1 & 1 & 0.5 & 0.01 & 0.25 \\ 0.5 & 0.5 & 1 & 0.2 & 0.2 \\ \hline 0.5 & 0.01 & 0.2 & 1 & 0.75 \\ \hline 0.01 & 0.25 & 0.2 & 0.75 & 1 \end{array}\right]$$

As we see in Figure 10, the behavior is similar as in other scenarios.

Results for uncertainty due to noise. In Appendix B.3, we considered uncertainty arising due to limited number of sensitive attributes. Here, we consider uncertainty arising due to noise. Specifically, we generate synthetic data using the covariance matrix Σ_2^{fair} and conduct two sets of experiments as follows. In the first experiment, we add independent zero-mean Gaussian noise, i.e., $\mathcal{N}(0, \sigma^2)$, with varying noise level, i.e., $\sigma = 1, \cdots, 10$, to the sensitive attributes. Such kind of noise is common when the sensitive attributes need to be privatized. In the second experiment, we add independent zero-mean Gaussian noise, i.e., $\mathcal{N}(0, \sigma^2)$, with varying noise level, i.e., $\sigma = 1, \cdots, 10$, to the sensitive attributes only if



(a) $d = 2$, $\Sigma^\star = \Sigma_2^{\mathrm{a}}$, $\varepsilon = 0.05$



(b) $d = 3$, $\Sigma^\star = \Sigma_3^{\mathrm{a}}$, $\varepsilon = 0.05$

Figure 10. The performance of robust QCQP in (13) (for $d = 2$), `Bootstrap-S` with `S` $\in \{3, 9, 27\}$, and `Baseline` for various d and $\Sigma^\star$. In the left column, we plot the fraction of violations of the fairness constraint $\langle \mathbf{a}, \mathbf{b}_{ex} \rangle^2 \leq \varepsilon$ 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 1000 trials for $n = 250$.

the sensitive attribute magnitude is larger than a certain threshold, i.e., $|e| \geq \mathbf{\Sigma}_{ee}$. Such data-dependent noise simulates noisy responses when sensitive attributes are collected via a survey, e.g., minority groups pretend to be the majority due to fear of discrimination. In our experiment, samples with sensitive attribute magnitudes larger than the threshold have noisy responses. We provide the results for both experiments in Figure 11 where we observe that `Bootstrap-S` achieves significantly fewer fairness violations than `Baseline` while maintaining comparable MSE.

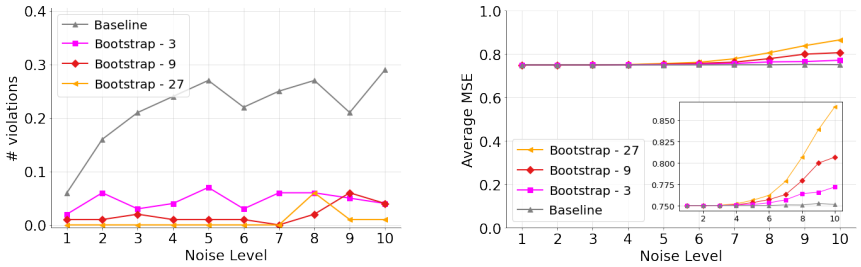
I.2. Additional results for real-world data

Dataset description. The Adult data (Lantz, 2019) (`https://archive.ics.uci.edu/ml/datasets/adult`) considers predicting whether an individual's income is more than \$50,000 from the 1994 Census database using 14 demographic features such as age, education, marital status, and country of origin. The sensitive attribute is sex. During pre-processing, we remove the sensitive attribute from the set of input features and discard rows with any missing data. The data has 48,842 samples, with 30,527 males and 14,695 females. We use a train-test split ratio of 0.72:0.28.

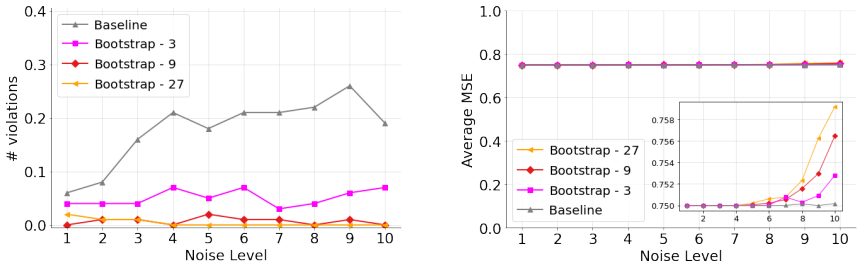
The Insurance data (Lantz, 2019) (`https://www.kaggle.com/datasets/teertha/ushealthinsurancedataset`) considers predicting the total annual medical expenses of individuals using 5 demographic features from the U.S. Census Bureau, such as BMI, number of children, and age. The sensitive attribute is sex. During pre-processing, we remove the sensitive attribute from the set of input features and perform normalization. The data has 1,338 samples with 676 males and 662 females. We use a train-test split ratio 0.8:0.2.

The Crime dataset (Redmond & Baveja, 2002) (`https://archive.ics.uci.edu/ml/datasets/communities+and+crime`) considers predicting the number of violent crimes per 100K population using socio-economic information of communities in the U.S. The sensitive attribute is the percentage of people belonging to a particular race in the community. During pre-processing, we drop all the samples with the value of sensitive attribute less than 5% to remove any outliers. We also remove the non-predictive attributes and the sensitive attribute from the set of input features, and normalize all attributes to the standardized range of $[0, 1]$. The resulting data has 1,112 samples, and we use a train-test split ratio 0.8:0.2.

Implementation details. We use two-layer fully connected networks in all our experiments. For all hidden layers, we use the selu activation function. For the output layer, we use softmax activation for classification and no activation for regression. We use 80 (50) units in the hidden layer and train the network for 30 (200) epochs for classification (regression).



(a) Independence Noise, $d = 2$, $\Sigma^\star = \Sigma_2^{\text{fair}}$, $\varepsilon = 0.025$

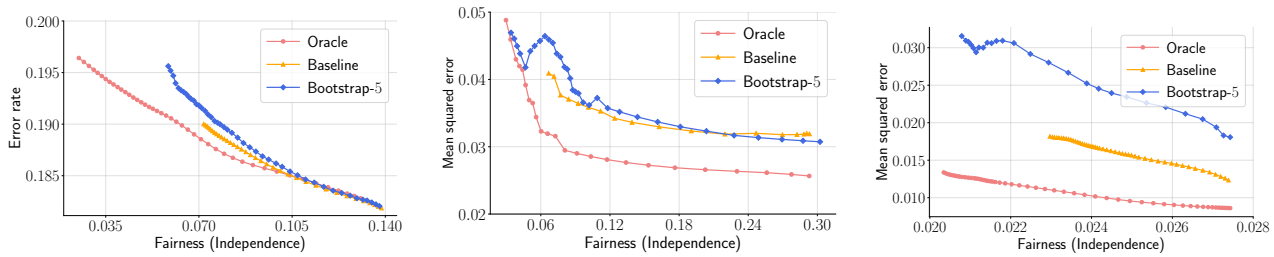


(b) Dependent Noise, $d = 2$, $\Sigma^\star = \Sigma_2^{\text{fair}}$, $\varepsilon = 0.025$

Figure 11. The performance of `Bootstrap-S` with $\mathtt{S} \in \{3, 9, 27\}$, and `Baseline`. In the left column, we plot the fraction of violations of the fairness constraint $\langle \mathbf{a}, \mathbf{b}_{ex} \rangle^2 \leq \varepsilon$ vs. the level of noise added to the sensitive attributes; in the right column, we plot average MSE vs. the level of noise added to the sensitive attributes.

Empirically, we do not observe any substantial improvement in performance for $\mathtt{S} \geq 5$, and thus report for a single $\mathtt{S}$ for each dataset. The hyperparameters $\lambda, \lambda_1 \cdots, \lambda_{\mathtt{S}}$ are initialized to 10 (5) for classification (regression). We optimize the model parameters and the hyperparameters $\lambda, \lambda_1 \cdots, \lambda_{\mathtt{S}}$ using different optimizers. For model optimization, we use Adam optimizer with batch size 128 (100), initial learning rate 10^{-3} (10^{-4}), and weight decay 0 (0.01) for classification (regression). For optimizing the hyperparameters, we use SGD optimizer with learning rate 10^{-2}. The batch sizes for Adult, Crime, and Insurance datasets are set to be 128, 100, and 128, respectively. All experiments are implemented in PyTorch using Tesla V100 GPUs with 32 GB memory.

Results of different levels of uncertainty. In Figure 12, we provide results for independence notion of fairness with $n = 200$ for Adult dataset, $\sigma = 0.25$ for the Crime dataset, and $n = 20$ for Insurance dataset, respectively, where n denotes the number of sensitive attributes kept out of N and σ decides the amount of noise added to the sensitive attributes. The observations are consistent to that in Figure 3(a), (b), and (c).



(a) Adult dataset (b) Crime dataset (c) Insurance dataset

Figure 12. Performance of `Bootstrap-S` and `Baseline` for independence for various datasets. `Bootstrap-S` achieves much better fairness levels compared to `Baseline` throughout.