Distributionally Fair Stochastic Optimization using Wasserstein Distance

Qing Ye
H. Milton Stewart School of Industrial and Systems Engineering, Georgia Institute of Technology, Atlanta, GA, USA, qye40@gatech.edu

Grani A. Hanasusanto
Department of Industrial & Enterprise Systems Engineering, University of Illinois Urbana-Champaign, Urbana, IL, USA, gah@illinois.edu

Weijun Xie
H. Milton Stewart School of Industrial and Systems Engineering, Georgia Institute of Technology, Atlanta, GA, USA, wxie@gatech.edu

A traditional stochastic program under a finite population typically seeks to optimize efficiency by maximizing the expected profits or minimizing the expected costs, subject to a set of constraints. However, implementing such optimization-based decisions can have varying impacts on individuals, and when assessed using the individuals' utility functions, these impacts may differ substantially across demographic groups delineated by sensitive attributes, such as gender, race, age, and socioeconomic status. As each group comprises multiple individuals, a common remedy is to enforce group fairness, which necessitates the measurement of disparities in the distributions of utilities across different groups. This paper introduces the concept of Distributionally Fair Stochastic Optimization (DFSO) based on the Wasserstein fairness measure. The DFSO aims to minimize distributional disparities among groups, quantified by the Wasserstein distance, while adhering to an acceptable level of inefficiency. Our analysis reveals that: (i) the Wasserstein fairness measure recovers the demographic parity fairness prevalent in binary classification literature; (ii) this measure can approximate the well-known Kolmogorov–Smirnov fairness measure with considerable accuracy; and (iii) despite DFSO's biconvex nature, the epigraph of the Wasserstein fairness measure is generally Mixed-Integer Convex Programming Representable (MICP-R). Additionally, we introduce two distinct lower bounds for the Wasserstein fairness measure: the Jensen bound, applicable to the general Wasserstein fairness measure, and the Gelbrich bound, specific to the type-2 Wasserstein fairness measure. We establish the exactness of the Gelbrich bound and quantify the theoretical difference between the Wasserstein fairness measure and the Gelbrich bound. Lastly, the theoretical underpinnings of the Wasserstein fairness measure enable us to design efficient algorithms to solve DFSO problems. Our numerical studies validate the effectiveness of these algorithms, confirming their practical use in achieving distributional fairness in several societally pertinent real-world stochastic optimization problems.

Key words: Wasserstein Distance, Group Fairness, Stochastic Optimization, Gelbrich Bound, Mixed-Integer Convex Programming

1. Introduction

Optimization empowers decision-making by providing an efficient solution to address complex problems in many domains. Its widespread use has motivated research studies focusing on the societal impact of optimization-based decisions. Since the traditional approach optimizes efficiency relevant to profits or costs, the optimization outcomes can have varying impacts across demographic groups delineated by sensitive attributes, including gender, race, age, and socioeconomic status. As each group comprises multiple individuals, enforcing group fairness necessitates the measurement of disparities of probability distributions of individual utilities between different groups. Traditional fairness measures are often based on summary statistics, such as minimum, mean, or deviation, which can be insufficient to quantify distributional disparities since each notion only characterizes a particular aspect of the probability distributions. On the other hand, statistical distance metrics, such as the Wasserstein distance, can be employed to quantify distributional fairness accurately. However, these metrics introduce significant computational challenges, and hence they remain largely unexplored in the field of fair decision-making. This motivates us to study distributional fairness.

1.1. Setting

The conventional decision-making problem under uncertainty is to optimize the total expected cost efficiency. Such an optimization problem can be formulated as the stochastic program

$$V^* = \min_{\boldsymbol{x}\in\mathcal{X}} \mathbb{E}_{\mathbb{P}}[Q(\boldsymbol{x},\tilde{\boldsymbol{\xi}})], \tag{1}$$

where $\mathcal{X}\subseteq\mathbb{R}^n$ specifies a mixed-integer convex representable decision space (Lubin et al. 2022), $Q(\cdot,\cdot)$ is a recourse function in stochastic programming or a loss function in machine learning, and $\tilde{\boldsymbol{\xi}}\in\mathbb{R}^\kappa$ are the random problem parameters governed by a probability distribution $\mathbb{P}$ with support Ξ. The stochastic program (1) and its variants with risk aversion and distributional robustness have been a prevailing modeling paradigm for numerous decision-making problems (see the survey paper Rahimian and Mehrotra 2019).

Many real-life decision-making problems may often involve a sensitive attribute such as gender, race, or age in the random parameters $\tilde{\boldsymbol{\xi}}$, designated by the component $\tilde{\xi}_\kappa\in A$, where the set A denotes a finite collection of possible outcomes in the sensitive attribute (e.g., $A=\{\text{male, female}\}$). This sensitive attribute partitions the outcome space into groups. Thus, by invoking the law of total expectation, we can rewrite the stochastic program (1) equivalently as

$$V^* = \min_{\boldsymbol{x}\in\mathcal{X}} \sum_{a\in A} \mathbb{P}(\tilde{\xi}_\kappa = a)\mathbb{E}_{\mathbb{P}_a}[Q(\boldsymbol{x},\tilde{\boldsymbol{\xi}}_a)], \tag{2}$$

where $\mathbb{P}_a$ is a shorthand for the conditional distribution of $\tilde{\boldsymbol{\xi}}$ given $\tilde{\xi}_\kappa = a$. Observe that the objective function constitutes a weighted sum of conditional expectations, where the weights correspond to the marginal distribution of the sensitive attribute. From this vantage point, the optimal solution $\boldsymbol{x}$ may treat the minority groups unfairly as it emphasizes groups of higher weight. This observation motivates us to study fair stochastic programming. Since many pertinent decision-making problems with sensitive attributes are concerned with a finite population, we assume that the entire support set Ξ is finite (i.e., $\Xi = \{\boldsymbol{\xi}_i\}_{i\in[m]}$), and we assume that each group $a \in A$ consists of m_a individuals represented by the set C_a, i.e., $\mathbb{P}_a\{\tilde{\boldsymbol{\xi}}_a = \boldsymbol{\xi}_i\} = 1/m_a$ for any $i \in C_a$. Evidently, the following identities hold in view of our assumption: $\Xi = \cup_{a\in A}\{\boldsymbol{\xi}_i\}_{i\in C_a}$ and $C_a \cap C_{\bar{a}} = \emptyset$ for any $a < \bar{a} \in A$. Under this setting, the stochastic program (2) further simplifies to

$$V^* = \min_{\boldsymbol{x}\in\mathcal{X}} \sum_{a\in A} \frac{m_a}{m} \sum_{i\in C_a} \frac{1}{m_a} Q(\boldsymbol{x}, \boldsymbol{\xi}_i). \tag{3}$$

Many deterministic optimization problems involving multiple groups of individuals can be viewed as a special case of (3) by treating each individual as an equiprobable sample.

To measure fairness, given a decision $\boldsymbol{x} \in \mathcal{X}$, for an individual realization $\boldsymbol{\xi}_a$ in each $a \in A$, we suppose that the function $f(\boldsymbol{x}, \boldsymbol{\xi}_a)$ denotes its utility value, which may not be monotonic (see, e.g., Kliegr 2009). Our goal is to match the probability distributions of the random utility values $\{f(\boldsymbol{x}, \tilde{\boldsymbol{\xi}}_a)\}_{a\in A}$ among different groups to attain fairness, where we quantify the utility distributional disparities using a statistical distance metric. Since the random utilities may have different support sets, we employ the Wasserstein distance and propose the following Distributionally Fair Stochastic Optimization (DFSO):

$$v^*(q) = \min_{\boldsymbol{x}\in\mathcal{X}} \left\{ \mathrm{WD}_q^q(\boldsymbol{x}) := \max_{a<\bar{a}\in A} W_q^q\left(\mathbb{P}_{f(\boldsymbol{x},\tilde{\boldsymbol{\xi}}_a)}, \mathbb{P}_{f(\boldsymbol{x},\tilde{\boldsymbol{\xi}}_{\bar{a}})}\right) : \mathbb{E}_{\mathbb{P}}[Q(\boldsymbol{x}, \tilde{\boldsymbol{\xi}})] \le V^* + \epsilon|V^*| \right\}. \tag{DFSO}$$

Here, the objective function represents the qth power of type-q Wasserstein fairness measure. Particularly, the type-q Wasserstein distance $W_q(\cdot,\cdot)$ is defined as

$$W_q(\mathbb{P}_1, \mathbb{P}_2) = \inf_{\mathbb{Q}} \left\{ \sqrt[q]{\int_{\Xi\times\Xi} \|\boldsymbol{\zeta}_1 - \boldsymbol{\zeta}_2\|^q \, \mathbb{Q}(d\boldsymbol{\zeta}_1, d\boldsymbol{\zeta}_2)} : \begin{array}{c} \mathbb{Q} \text{ is a joint distribution of } \tilde{\boldsymbol{\zeta}}_1 \text{ and } \tilde{\boldsymbol{\zeta}}_2 \\ \text{with marginals } \mathbb{P}_1 \text{ and } \mathbb{P}_2, \text{ respectively} \end{array} \right\},$$

where $\|\cdot\|$ is a norm and $q \in [1, \infty]$. In DFSO, the goal is to minimize the maximum distributional disparities of utilities quantified by the Wasserstein distance among all pairs of groups (i.e., the Wasserstein fairness) while maintaining the cost efficiency around a near-optimal region, where $\epsilon \ge 0$ denotes the inefficiency level prescribed by the decision-maker. In practice, the utility function $f(\boldsymbol{x}, \boldsymbol{\xi})$ can be quite general. If it is equal to the recourse function $Q(\boldsymbol{x}, \boldsymbol{\xi})$, then the decision-maker, in this case, tries to achieve the distributional fairness of random cost among different groups.

1.2. Literature Review

Optimization has served as an essential tool in decision-making over the past decades. Throughout the years, the issues of fairness in optimization have been recognized and studied in the fields of resource allocation, facility location, and communication networks (Ogryczak et al. 2014, Karsu and Morton 2015). The commonly adopted definitions of fairness pertain to the utilities of all individuals in the population, e.g., max-min fairness, proportional fairness, and alpha fairness, or to some particular characteristics of the distribution of the utilities, e.g., spread, deviation, Jain's index, and Gini coefficient. Contrary to traditional definitions that consider the entire population, this paper concentrates on fairness among different groups of individuals. These fairness measures at the population level can be simply generalized to the group level by applying them to each group instead of the entire population. For example, Samorani et al. (2022) studied the max-min fairness at the group level. They addressed the racial disparity in medical appointment scheduling by minimizing the maximum waiting time among the racial groups. Cohen et al. (2022) discussed price discrimination against protected groups and attempted to enforce nearly equal prices for different groups. Patel et al. (2020) considered group fairness for the knapsack problem when each item belongs to a particular group. They defined three fair knapsack notions, i.e., to bound the number of items from each group, to bound the total weight of items from each group, and to bound the total value of items from each group. Since the traditional fairness measures are often based on summary statistics, they might be inadequate for quantifying group disparities in a comprehensive way. Our distributional fairness notion overcomes this limitation by using the Wasserstein distance to quantify the distributional disparities among different groups. The Wasserstein distance has also been used in a variety of optimization problems such as Wasserstein distributional robust optimization (Mohajerin Esfahani and Kuhn 2018, Blanchet and Murthy 2019, Gao and Kleywegt 2023, Hanasusanto and Kuhn 2018, Chen et al. 2022, Xie 2021).

Recent studies in the growing field of fair machine learning have proposed various methods for a number of tasks (Caton and Haas 2020). The majority of the literature has focused on group fairness, which seeks to treat different groups equally. Group fairness in binary classification has been extensively studied (Kamishima et al. 2012, Feldman et al. 2015, Barocas and Selbst 2016, Hardt et al. 2016, Zafar et al. 2017, Donini et al. 2018, Aghaei et al. 2019, Kallus et al. 2022, Taskesen et al. 2020, Ye and Xie 2020, Wang et al. 2021, Lowy et al. 2021). However, the number of works on group fairness in regression with continuous outcomes is rather limited. Berk et al. (2017) introduced a family of convex fairness regularizers such that each group should have similar predicted outcomes weighted by the nearness of the true outcomes on average. Agarwal et al. (2019), Chzhen et al. (2020), Rychener et al. (2022) used the Kolmogorov–Smirnov distance to achieve demographic parity. Additionally, Rychener et al. (2022) summarized the common integral probability metrics for quantifying fairness, including the Kolmogorov–Smirnov distance and the Wasserstein distance. To achieve fairness, Agarwal et al. (2019) designed a reduction-based algorithm, while Chzhen et al. (2020) developed a post-processing algorithm for fair regression. Rychener et al. (2022) proposed to solve fair regression via a stochastic gradient descent algorithm. According to the definition in fair machine learning literature, the demographic parity-based fairness notion ensures the probability distribution of outcomes is independent of the sensitive attribute groups. Our distributional fairness notion coincides with demographic parity when applied to machine learning problems. Furthermore, the proposed DFSO formulation is a general stochastic optimization problem where fairness is integrated with efficiency. Thus, it provides flexibility to model various decision-making problems, including classification with binary utilities and regression with continuous utilities. More importantly, different from existing results in the literature, we thoroughly investigate the optimization properties of the Wasserstein fairness measure and exploit them to systematically design efficient solution algorithms with provable guarantees.

1.3. Summary of Contributions

The main contributions of this paper are summarized as follows:

- From a fresh scope, this paper establishes the fundamental result that the Wasserstein fairness measure is essentially equivalent to matching the probability distributions of distinct groups comonotonically and computing the distance of the comonotonic distributions. Using this equivalence, we show that the Wasserstein fairness measure recovers the well-known demographic parity fairness from the binary classification literature, and we reveal that the Wasserstein fairness measure is relatively close to the Kolmogorov–Smirnov one.
- We prove that the DFSO under the Wasserstein fairness measure, in general, is NP-hard. However, different from other biconvex programs, we show that the epigraph of the Wasserstein fairness measure is, in general, Mixed-Integer Convex Programming Representable (MICP-R), and we provide four different representations. These are the first known MICP-R results for Wasserstein distance-based distributional fairness models.
- We derive two different lower bounds for the Wasserstein fairness measure: the Jensen bound for the general Wasserstein fairness measure and the Gelbrich bound for the type-2 Wasserstein fairness measure. We prove a broader condition than the well-known elliptical distributions under which the Gelbrich bound is asymptotically tight, and we provide a theoretical gap between the Wasserstein fairness measure and the Gelbrich bound. We also prove that computing the Gelbrich bound is NP-hard.
- Inspired by the theoretical properties of the Wasserstein fairness measure, we design effective solutions algorithms to solve the DFSO to near-optimality. Our numerical study confirms the effectiveness of the proposed algorithms.

The remainder of the paper is organized as follows. Section 2 presents properties of the Wasserstein fairness measure. Section 3 formalizes definitions and develops two exact mixed-integer convex programming representations of the epigraph of the Wasserstein fairness measure. Section 4 studies two lower bounds of the Wasserstein fairness measure. Section 5 reports the numerical study, and Section 6 concludes the paper. Proofs and additional results are relegated to the appendix.

Notation. Bold lowercase letters (e.g., $\boldsymbol{x}$) denote vectors, bold uppercase letters (e.g., $\boldsymbol{Z}$) denote matrices, and the corresponding regular letters (e.g., x_i, Z_{ij}) denote their components. For any $n \in \mathbb{Z}_+$, we let $[n] := \{1, 2, \dots, n\}$ and use $\mathbb{R}^n_+ := \{\boldsymbol{x} \in \mathbb{R}^n : x_i \geq 0, \forall i \in [n]\}$. For any $n_1 < n_2 \in \mathbb{Z}_+$, we let $[n_1, n_2] := \{n_1, n_1 + 1, \dots, n_2\}$. For a set A, we let $a < \bar{a} \in A$ denote $a, \bar{a} \in A$ such that $a < \bar{a}$. The indicator function $\mathbb{I}(B)$ takes value 1 if B is true and 0 otherwise. Additional notation will be introduced as needed.

2. Properties of the Wasserstein Fairness Measure

This section presents various notable properties of the Wasserstein fairness measure. To begin with, let us define the cumulative distribution functions of the random functions $\{f(\boldsymbol{x}, \tilde{\boldsymbol{\xi}}_a)\}_{a \in A}$ as $F_a(t \mid \boldsymbol{x}) = \mathbb{P}_a\{f(\boldsymbol{x}, \tilde{\boldsymbol{\xi}}_a) \leq t\}$ for all $a \in A$. Correspondingly, we define the inverse distribution functions $F_a^{-1}(y \mid \boldsymbol{x}) = \inf\{t : F_a(t \mid \boldsymbol{x}) \geq y\}$ for all $a \in A$.

2.1. Comonotonicity and Complexity

This subsection investigates the comonotonicity property of the Wasserstein fairness measure and the complexity of DFSO, which motivate us to develop strong mixed-integer convex programming formulations for DFSO.

One property of the Wasserstein fairness measure in DFSO is that it can be simplified as the integral of the difference of inverse cumulative distributions.

Lemma 1 **(Proposition 2.17 in Santambrogio (2015))**. *For any $a < \bar{a} \in A$ and a fixed decision $\boldsymbol{x}$, the Wasserstein distance $W_q\left(\mathbb{P}_{f(\boldsymbol{x}, \tilde{\boldsymbol{\xi}}_a)}, \mathbb{P}_{f(\boldsymbol{x}, \tilde{\boldsymbol{\xi}}_{\bar{a}})}\right)$ can be expressed as*

$$W_q\left(\mathbb{P}_{f(\boldsymbol{x}, \tilde{\boldsymbol{\xi}}_a)}, \mathbb{P}_{f(\boldsymbol{x}, \tilde{\boldsymbol{\xi}}_{\bar{a}})}\right) = \sqrt[q]{\int_0^1 \left|F_a^{-1}(y \mid \boldsymbol{x}) - F_{\bar{a}}^{-1}(y \mid \boldsymbol{x})\right|^q dy}, \tag{4}$$

where F_a^{-1} is the inverse distribution function of the random function $f(\boldsymbol{x}, \tilde{\boldsymbol{\xi}}_a)$ for each $a \in A$. When $q = 1$, the type-1 Wasserstein distance $W_1\left(\mathbb{P}_{f(\boldsymbol{x}, \tilde{\boldsymbol{\xi}}_a)}, \mathbb{P}_{f(\boldsymbol{x}, \tilde{\boldsymbol{\xi}}_{\bar{a}})}\right)$ coincides with the L_1 distance between the cumulative distribution functions

$$W_1\left(\mathbb{P}_{f(\boldsymbol{x}, \tilde{\boldsymbol{\xi}}_a)}, \mathbb{P}_{f(\boldsymbol{x}, \tilde{\boldsymbol{\xi}}_{\bar{a}})}\right) = \int_{\mathbb{R}} \left|F_a(t \mid \boldsymbol{x}) - F_{\bar{a}}(t \mid \boldsymbol{x})\right| dt, \tag{5}$$

where F_a is the cumulative distribution function of the random function $f(\boldsymbol{x}, \tilde{\boldsymbol{\xi}}_a)$ for each $a \in A$.

Lemma 1 shows that the Wasserstein fairness measure $\mathrm{WD}_q(\boldsymbol{x}) := \max_{a<\bar{a}\in A} W_q(\mathbb{P}_{f(\boldsymbol{x},\tilde{\boldsymbol{\xi}}_a)}, \mathbb{P}_{f(\boldsymbol{x},\tilde{\boldsymbol{\xi}}_{\bar{a}})})$ can be viewed as the largest L_q-norm of the difference between inverse distribution functions. Hence, in DFSO, minimizing $\mathrm{WD}_q(\boldsymbol{x})$ implies attempting to match the distributions of utilities between any two groups $a < \bar{a} \in A$. Remarkably, as established in the existing literature, type-1 Wasserstein fairness measure $\mathrm{WD}_1(\boldsymbol{x})$ is equivalent to the maximum L_1 distance between the cumulative distribution functions. For each pair of groups $a < \bar{a} \in A$, under our assumption of discrete distributions, the integrals in (4) and (5) can be simplified to be summations. These properties motivate us to study the exact MICP formulations of the Wasserstein fairness measure.

Another interesting byproduct of Lemma 1 is that when achieving the infimum of the Wasserstein distance, the two distributions must be aligned comonotonically, where the comonotonicity of two random variables is formally defined as follows.

Definition 1. A pair of random variables $X = (X_1, X_2)$ is comonotonic if and only if it can be represented as $(X_1, X_2) \overset{\mathrm{d}}{=} (F_{X_1}^{-1}(U), F_{X_2}^{-1}(U))$, where U is the standard uniform random variable, and $F_{X_1}^{-1}, F_{X_2}^{-1}$ are the inverse distribution functions of X_1, X_2.

This gives rise to an interesting result for the following Wasserstein fairness measure.

Proposition 1. *For a given decision $\boldsymbol{x} \in \mathcal{X}$, when computing the Wasserstein fairness measure in DFSO, the optimal joint distribution is comonotonic for any pair $a < \bar{a} \in A$.*

Proof. See Appendix B.1. □

Proposition 1 shows that the Wasserstein fairness measure, in fact, aligns the two distinct groups' utility function values comonotonically, computes the L_q norm of the difference of their inverse distribution functions, and then takes the maximum value among all the pairs of groups. It helps us study the new exactness conditions of the well-known lower bound (i.e., the Gelbrich bound) of the type-2 Wasserstein fairness measures. This result also motivates us to study its relation with another popular distributional fairness notion: the Kolmogorov–Smirnov fairness measure.

We conclude the subsection by proving the NP-hardness of DFSO via a reduction from the well-known chance-constrained stochastic program (Charnes and Cooper 1959, Ahmed and Xie 2018).

Theorem 1. *Solving DFSO is, in general, strongly NP-hard, even when $\mathcal{X}$ is a polytope, $\epsilon = \infty$, and $f(\boldsymbol{x}, \boldsymbol{\xi})$ is a linear function.*

Proof. See Appendix B.4. □

2.2. Recovering the Demographic Parity Fairness Measure of Binary Outcomes

Demographic parity of binary outcomes, defined as $\mathrm{DP}(\boldsymbol{x})$, requires the probability of beneficial or detrimental outcomes to be independent of the sensitive attribute. In the following, we show that the proposed $\mathrm{WD}_q(\boldsymbol{x})$ recovers $\mathrm{DP}(\boldsymbol{x})$ if the utility function is Bernoulli. Let us first define $\mathrm{DP}(\boldsymbol{x})$.

Definition 2. Suppose that $\mathbb{P}\{f(\boldsymbol{x},\tilde{\boldsymbol{\xi}})\in\{0,1\}\}=1$. The binary demographic parity fairness measure is defined as

$$\mathrm{DP}(\boldsymbol{x})=\max_{a<\bar{a}\in A}\left|\mathbb{P}\{f(\boldsymbol{x},\tilde{\boldsymbol{\xi}}_a)=0\}-\mathbb{P}\{f(\boldsymbol{x},\tilde{\boldsymbol{\xi}}_{\bar{a}})=0\}\right|=\max_{a<\bar{a}\in A}\left|\mathbb{P}_a\{f(\boldsymbol{x},\tilde{\boldsymbol{\xi}}_a)=1\}-\mathbb{P}_{\bar{a}}\{f(\boldsymbol{x},\tilde{\boldsymbol{\xi}}_{\bar{a}})=1\}\right|.$$

We next show that the Wasserstein fairness measure $\mathrm{WD}_q(\boldsymbol{x})$ is equivalent to $\mathrm{DP}(\boldsymbol{x})$ in view of Lemma 1.

Proposition 2. *For a Bernoulli utility function* $f(\boldsymbol{x},\tilde{\boldsymbol{\xi}})\in\{0,1\}$, $\mathrm{WD}_q(\boldsymbol{x})$ *is equivalent to* $\mathrm{DP}(\boldsymbol{x})$.

Proof. See Appendix B.2. □

The result in Proposition 2 reveals that the proposed Wasserstein fairness measure constitutes a generalization of the binary demographic parity fairness measure.

2.3. Comparison with the Kolmogorov–Smirnov Fairness Measure

Instead of the sum of differences, we can use the supremum of differences to measure the demographic parity fairness as defined below.

Definition 3 (Kolmogorov–Smirnov Fairness Measure, Agarwal et al. 2019). The distributional fairness of a decision $\boldsymbol{x}$ can be measured using the Kolmogorov–Smirnov distance:

$$\mathrm{KSD}(\boldsymbol{x})=\max_{a<\bar{a}\in A}\sup_t\left|F_a(t\,|\,\boldsymbol{x})-F_{\bar{a}}(t\,|\,\boldsymbol{x})\right|. \tag{6}$$

The Kolmogorov–Smirnov distance measures the largest difference of cumulative distribution functions between any two distinct groups. To compute $\mathrm{KSD}(\boldsymbol{x})$, one needs to discretize t, which is easily done in view of our assumption of finite populations. Specifically, the assumption implies that the cumulative distributions $\{F_a(y\mid\boldsymbol{x})\}_{a\in A}$ and their inverse counterparts $\{F_a^{-1}(y\mid\boldsymbol{x})\}_{a\in A}$ are of finitely many values, defined formally as follows.

Definition 4 (Breaking Points). For any $\boldsymbol{x}\in\mathcal{X}$ and any pair $a<\bar{a}\in A$, the breaking points of $F_a^{-1}(y\,|\,\boldsymbol{x})-F_{\bar{a}}^{-1}(y\,|\,\boldsymbol{x})$ are denoted by $\boldsymbol{b}_{a\bar{a}}(\boldsymbol{x})=(b_{ja\bar{a}}(\boldsymbol{x}))_{j\in J_{a\bar{a}}}$ with index set $J_{a\bar{a}}:=\{1,2,\cdots,|J_{a\bar{a}}|\}$. We further define the widths $w_{ja\bar{a}}(\boldsymbol{x})=b_{ja\bar{a}}(\boldsymbol{x})-b_{(j-1)a\bar{a}}(\boldsymbol{x})$ for $j\in J_{a\bar{a}}\setminus\{1\}$ and calculate the largest width as $\eta(\boldsymbol{x})=\max_{a<\bar{a}\in A,j\in J_{a\bar{a}}\setminus\{1\}}w_{ja\bar{a}}(\boldsymbol{x})$.

Based on Definition 4, we propose the following lower and upper bounds on $\mathrm{KSD}(\boldsymbol{x})$ in terms of $\mathrm{WD}_q(\boldsymbol{x})$, which shows that the two measures are close to each other within a constant factor. The key idea is to recast the type-q Wasserstein fairness measure as

$$\begin{aligned}\mathrm{WD}_q(\boldsymbol{x}) &= \max_{a<\bar{a}\in A} \sqrt[q]{\int_0^1 \left|F_a^{-1}(y \mid \boldsymbol{x}) - F_{\bar{a}}^{-1}(y \mid \boldsymbol{x})\right|^q dy} \\ &= \max_{a<\bar{a}\in A} \sqrt[q]{\sum_{j\in J_{a\bar{a}}\setminus\{1\}} w_{ja\bar{a}}(\boldsymbol{x}) \left|F_a^{-1}\left(b_{ja\bar{a}}(\boldsymbol{x}) \mid \boldsymbol{x}\right) - F_{\bar{a}}^{-1}\left(b_{ja\bar{a}}(\boldsymbol{x}) \mid \boldsymbol{x}\right)\right|^q},\end{aligned}$$

in the spirit of Definition 4. Then we bound the difference between $\mathrm{WD}_1(\boldsymbol{x})$ and $\mathrm{KSD}(\boldsymbol{x})$. Next, we use the relationship between $\mathrm{WD}_1(\boldsymbol{x})$ and $\mathrm{WD}_q(\boldsymbol{x})$ to finally bound $\mathrm{WD}_q(\boldsymbol{x})$ and $\mathrm{KSD}(\boldsymbol{x})$.

Proposition 3. *For any feasible $\boldsymbol{x} \in \mathcal{X}$ and $q \in [1, \infty]$, the following inequalities hold:*

$$\frac{1}{\max_{a<\bar{a}\in A} \eta(\boldsymbol{x})^{\frac{1-q}{q}} (t_{2a\bar{a}}(\boldsymbol{x}) - t_{1a\bar{a}}(\boldsymbol{x}))} \mathrm{WD}_q(\boldsymbol{x}) \le \mathrm{KSD}(\boldsymbol{x}) \le \frac{1}{\min_{a<\bar{a}\in A} \mu(\Delta_{a\bar{a}}(\boldsymbol{x}))} \mathrm{WD}_q(\boldsymbol{x}).$$

Here, $t_{1a\bar{a}}(\boldsymbol{x}) = \min\{\min_t\{t : F_a(t \mid \boldsymbol{x}) > 0\}, \min_t\{t : F_{\bar{a}}(t \mid \boldsymbol{x}) > 0\}\}$, $t_{2a\bar{a}}(\boldsymbol{x}) = \max\{\sup_t\{t : F_a(t \mid \boldsymbol{x}) < 1\}, \sup_t\{t : F_{\bar{a}}(t \mid \boldsymbol{x}) < 1\}\}$, and $\Delta_{a\bar{a}}(\boldsymbol{x}) = \{\bar{t} : |F_a(\bar{t} \mid \boldsymbol{x}) - F_{\bar{a}}(\bar{t} \mid \boldsymbol{x})| = \sup_t |F_a(t \mid \boldsymbol{x}) - F_{\bar{a}}(t \mid \boldsymbol{x})|\}$ with its Lebesgue measure $\mu(\Delta_{a\bar{a}}(\boldsymbol{x}))$.

Proof. See Appendix B.3. □

Proposition 3 theoretically establishes that the Wasserstein and Kolmogorov–Smirnov fairness measures are rather similar to each other. The bounds can be independent of the decision variables $\boldsymbol{x}$ by finding the least-favorable coefficients. It is worth mentioning that the existing literature (see, e.g., Ross 2011) only bound Kolmogorov–Smirnov fairness measure from above by type-1 Wasserstein fairness measure when the underlying random variables are continuous. In our following derivation, the Wasserstein fairness measure shows amenable optimization properties. Our numerical study demonstrates the advantage of the proposed methods for the Wasserstein fairness measure compared to the existing ones for the Kolmogorov–Smirnov fairness measure.

3. Mixed-Integer Convex Programming Formulations of DFSO

This section focuses on deriving exact Mixed-Integer Convex Programming (MICP) formulations of DFSO. To begin with, we observe that under the discrete-distribution assumption, using epigraphical variable ν, the proposed DFSO can be formulated as the mathematical program

$$v^*(q) = \min_{(\boldsymbol{x},\nu)\in\mathcal{F}_q} \quad \nu, \tag{7a}$$

$$\text{s.t.} \quad \sum_{i\in[m]} \frac{1}{m} Q(\boldsymbol{x}, \boldsymbol{\xi}_i) \le V^* + \epsilon|V^*|, \tag{7b}$$

where we introduce the set $\mathcal{F}_q$ to denote the epigraph of the Wasserstein fairness measure, as follows:

$$\mathcal{F}_q = \left\{(\boldsymbol{x},\nu) \in \mathcal{X} \times \mathbb{R}_+ : W_q^q\left(\mathbb{P}_{f(\boldsymbol{x},\tilde{\boldsymbol{\xi}}_a)}, \mathbb{P}_{f(\boldsymbol{x},\tilde{\boldsymbol{\xi}}_{\bar{a}})}\right) \le \nu, \forall a < \bar{a} \in A\right\}. \tag{8}$$

For the formulations, we will also utilize the following set that corresponds to the graph of the function $f(\cdot,\boldsymbol{\xi}_i)$ for each realization $\boldsymbol{\xi}_i \in \Xi$:

$$X_i = \{(\boldsymbol{x},\bar{w}_i) \in \mathcal{X} \times \mathbb{R} : f(\boldsymbol{x},\boldsymbol{\xi}_i) = \bar{w}_i\}.$$

Section 3.1 discusses the concept of MICP representability and presents MICP formulations for the graphs of utility functions $\{X_i\}_{i\in[m]}$. Section 3.2 and Section 3.3 explore two different ways of representing the epigraph of Wasserstein fairness measure $\mathcal{F}_q$ in (8). The first formulation uses Lemma 1 to represent quantiles using mixed-integer programming formulations. The second formulation is a variation of the first one, using aggregate rather than individual quantiles. We have two additional formulations presented in Appendix A, where the Discretized Formulation (see Appendix A.1) is based on the discretization of the transportation decisions by observing that the inflated transportation decision variables can be restricted to integers, and the Complementary Formulation (see Appendix A.2) is to recast the set $\mathcal{F}_q$ using linear programming with complementary slackness constraints and linearize the complementary slackness constraints. Besides, we derive an equivalent MICP formulation for the Kolmogorov–Smirnov fairness measure KSD($\boldsymbol{x}$), which can be found in Appendix A.3.

3.1. Mixed-Integer Convex Programming Representability

To begin with, we introduce the notion of MICP representability and develop formulations for various families of utility functions, depending on whether the sets $\{X_i\}_{i\in[m]}$ are MICP representable (MICP-R) or not MICP-R. The MICP-R sets are defined as follows.

Definition 5 (Theorem 4.1 in Lubin et al. 2022). A set $S \subseteq \mathbb{R}^n$ is MICP-R if and only if there exists $d,p \in \mathbb{Z}_+$, a convex set $C \subseteq \mathbb{R}^d$, and a closed convex family $(B_z)_{z\in C} \subseteq \mathbb{R}^{n+p}$ such that $S = \bigcup_{z\in C\cap\mathbb{Z}^d} \mathrm{proj}_x(B_z)$.

In addition, Lubin et al. (2022) also provided the following sufficient condition for not MICP-R.

Lemma 2 **(Lemma 4.1 in Lubin et al. 2022)**. *A set $S \subseteq \mathbb{R}^n$ is not MICP-R if there exists $R \subseteq S, |R| = \infty$ such that $(\boldsymbol{x}+\boldsymbol{x}')/2 \notin S$ for all $\boldsymbol{x},\boldsymbol{x}' \in R, \boldsymbol{x} \neq \boldsymbol{x}'$.*

Our MICP-R results rely on the McCormick representation.

Definition 6 (McCormick Representation, McCormick 1976). Consider a bilinear set $\{(\psi,\kappa,\nu)\in\mathbb{R}\times\{\kappa^L,\kappa^U\}\times[\nu^L,\nu^U]:\psi=\kappa\nu\}$ with given lower bounds κ^L,ν^L and upper bounds κ^U,ν^U. Its McCormick representation is

$$\mathrm{MC}(\kappa^L,\kappa^U,\nu^L,\nu^U)=\left\{(\psi,\kappa,\nu):\begin{array}{c}\psi\in\mathbb{R},\kappa\in\{\kappa^L,\kappa^U\},\nu^L\le\nu\le\nu^U,\\ \psi\ge\kappa^L\nu+\kappa\nu^L-\kappa^L\nu^L,\psi\ge\kappa^U\nu+\kappa\nu^U-\kappa^U\nu^U,\\ \psi\le\kappa^U\nu+\kappa\nu^L-\kappa^U\nu^L,\psi\le\kappa\nu^U+\kappa^L\nu-\kappa^L\nu^U\end{array}\right\}.$$

Next, we discuss three special cases when the sets $\{X_i\}_{i\in[m]}$ are MICP-R.

Proposition 4. *Suppose that* $f(\boldsymbol{x},\boldsymbol{\xi})=\boldsymbol{\xi}^\top\boldsymbol{r}(\boldsymbol{x})+s(\boldsymbol{x})$, *where* $\boldsymbol{r}(\boldsymbol{x})$ *and* $s(\boldsymbol{x})$ *are linear functions. Then the sets* $\{X_i\}_{i\in[m]}$ *are MICP-R.*

Proof. We have

$$X_i=\left\{(\boldsymbol{x},\bar{w}_i)\in\mathcal{X}\times\mathbb{R}:\boldsymbol{\xi}_i^\top\boldsymbol{r}(\boldsymbol{x})+s(\boldsymbol{x})=\bar{w}_i\right\},$$

which is an MICP-R set. □

Proposition 5. *Suppose that* $f(\boldsymbol{x},\boldsymbol{\xi})=\max_{\tau\in T}\left\{\boldsymbol{\xi}^\top\boldsymbol{r}_\tau(\boldsymbol{x})+s_\tau(\boldsymbol{x})\right\}$, *where* $\{\boldsymbol{r}_\tau(\boldsymbol{x})\}_{\tau\in T}$ *and* $\{s_\tau(\boldsymbol{x})\}_{\tau\in T}$ *are linear functions. Then the sets* $\{X_i\}_{i\in[m]}$ *are MICP-R.*

Proof. Suppose that $M_i\ge\max_{\boldsymbol{x}\in\mathcal{X},(7\mathrm{b})}\left|\max_{\tau\in T}\left\{\boldsymbol{\xi}_i^\top\boldsymbol{r}_\tau(\boldsymbol{x})+s_\tau(\boldsymbol{x})\right\}\right|$ for each $i\in[m]$. Then, we have

$$\begin{aligned}X_i&=\left\{(\boldsymbol{x},\bar{w}_i)\in\mathcal{X}\times\mathbb{R}:\max_{\tau\in T}\left\{\boldsymbol{\xi}_i^\top\boldsymbol{r}_\tau(\boldsymbol{x})+s_\tau(\boldsymbol{x})\right\}=\bar{w}_i\right\}\\&=\left\{(\boldsymbol{x},\bar{w}_i)\in\mathcal{X}\times\mathbb{R}:\begin{array}{c}\bar{w}_i\ge\boldsymbol{\xi}_i^\top\boldsymbol{r}_\tau(\boldsymbol{x})+s_\tau(\boldsymbol{x}),\forall\tau\in T,\\ \bar{w}_i\le\boldsymbol{\xi}_i^\top\boldsymbol{r}_\tau(\boldsymbol{x})+s_\tau(\boldsymbol{x})+M_i(1-z_{i\tau}),\forall\tau\in T,\\ \sum_{\tau\in T}z_{i\tau}=1,z_{i\tau}\in\{0,1\},\forall\tau\in T\end{array}\right\},\end{aligned}$$

which is an MICP-R set. □

Proposition 6. *Suppose that* $f(\boldsymbol{x},\boldsymbol{\xi})=\min_{\tau\in T}\left\{\boldsymbol{\xi}^\top\boldsymbol{r}_\tau(\boldsymbol{x})+s_\tau(\boldsymbol{x})\right\}$, *where* $\{\boldsymbol{r}_\tau(\boldsymbol{x})\}_{\tau\in T}$ *and* $\{s_\tau(\boldsymbol{x})\}_{\tau\in T}$ *are linear functions. Then the sets* $\{X_i\}_{i\in[m]}$ *are MICP-R.*

Proof. Recall that $M_i\ge\max_{\boldsymbol{x}\in\mathcal{X},(7\mathrm{b})}\left|\min_{\tau\in T}\left\{\boldsymbol{\xi}_i^\top\boldsymbol{r}_\tau(\boldsymbol{x})+s_\tau(\boldsymbol{x})\right\}\right|$ for each $i\in[m]$. Thus, we have

$$\begin{aligned}X_i&=\left\{(\boldsymbol{x},\bar{w}_i)\in\mathcal{X}\times\mathbb{R}:\min_{\tau\in T}\left\{\boldsymbol{\xi}_i^\top\boldsymbol{r}_\tau(\boldsymbol{x})+s_\tau(\boldsymbol{x})\right\}=\bar{w}_i\right\}\\&=\left\{(\boldsymbol{x},\bar{w}_i)\in\mathcal{X}\times\mathbb{R}:\begin{array}{c}\bar{w}_i\le\boldsymbol{\xi}_i^\top\boldsymbol{r}_\tau(\boldsymbol{x})+s_\tau(\boldsymbol{x}),\forall\tau\in T,\\ \bar{w}_i\ge\boldsymbol{\xi}_i^\top\boldsymbol{r}_\tau(\boldsymbol{x})+s_\tau(\boldsymbol{x})-M_i(1-z_{i\tau}),\forall\tau\in T,\\ \sum_{\tau\in T}z_{i\tau}=1,z_{i\tau}\in\{0,1\},\forall\tau\in T\end{array}\right\},\end{aligned}$$

which is an MICP-R set. □

When the utility functions are exponential or logarithmic, their corresponding $\mathcal{F}_q$ sets are typically not MICP-R according to Lemma 2. Hence we propose to approximate them using piecewise linear functions (see Appendix C).

3.2. Quantile Formulation

In this subsection, we propose a quantile-based formulation to represent the set $\mathcal{F}_q$ motivated by Lemma 1. That is, we first equivalently rewrite set $\mathcal{F}_q$ as

$$\mathcal{F}_q = \left\{ (\boldsymbol{x}, \nu) \in \mathcal{X} \times \mathbb{R}_+ : \int_0^1 \left| F_a^{-1}(y \mid \boldsymbol{x}) - F_{\bar{a}}^{-1}(y \mid \boldsymbol{x}) \right|^q dy \le \nu, \forall a < \bar{a} \in A \right\}. \tag{9}$$

Since all the random parameters have finite support, let us sort the distinct elements of the set

$$\{0\} \cup \left\{ \frac{i}{m_a} \right\}_{i \in [m_a]} \cup \left\{ \frac{i}{m_{\bar{a}}} \right\}_{i \in [m_{\bar{a}}]} := \left\{ \widehat{b}_{ia\bar{a}} \right\}_{i \in [\widehat{m}_{a\bar{a}}]}$$

in the ascending order as $0 := \widehat{b}_{1a\bar{a}} < \cdots < \widehat{b}_{(\widehat{m}_{a\bar{a}})a\bar{a}} := 1$ for each $a < \bar{a} \in A$. Observe that in the equation (4), the value $F_a^{-1}(y \mid \boldsymbol{x}) - F_{\bar{a}}^{-1}(y \mid \boldsymbol{x})$ is a constant whenever $y \in (\widehat{b}_{ia\bar{a}}, \widehat{b}_{(i+1)a\bar{a}}]$ for $i \in [\widehat{m}_{a\bar{a}} - 1]$. Thus, the set $\{\widehat{b}_{ia\bar{a}}\}_{i \in [\widehat{m}_{a\bar{a}}]}$ helps simplify the Wasserstein fairness measure as

$$\begin{aligned} \mathrm{WD}_q^q(\boldsymbol{x}) &= \max_{a < \bar{a} \in A} \sum_{i \in [\widehat{m}_{a\bar{a}} - 1]} \int_{\widehat{b}_{ia\bar{a}}}^{\widehat{b}_{(i+1)a\bar{a}}} \left| F_a^{-1}(y \mid \boldsymbol{x}) - F_{\bar{a}}^{-1}(y \mid \boldsymbol{x}) \right|^q dy, \\ &= \max_{a < \bar{a} \in A} \sum_{i \in [\widehat{m}_{a\bar{a}} - 1]} \left(\widehat{b}_{(i+1)a\bar{a}} - \widehat{b}_{ia\bar{a}} \right) \left| F_a^{-1}(\widehat{b}_{(i+1)a\bar{a}} \mid \boldsymbol{x}) - F_{\bar{a}}^{-1}(\widehat{b}_{(i+1)a\bar{a}} \mid \boldsymbol{x}) \right|^q. \end{aligned} \tag{10}$$

Next, we define the quantile set $\Omega_a(k) = \{(\boldsymbol{x}, t_{ka}) \in \mathcal{X} \times \mathbb{R} : F_a^{-1}(k/m_a \mid \boldsymbol{x}) = t_{ka}\}$ for each $k \in [m_a]$ and $a \in A$. Using the graph representation $(\boldsymbol{x}, \bar{w}_i) \in X_i$ for each $i \in [m]$, we propose the following equivalent formulation of the quantile set $\Omega_a(k)$.

Proposition 7. *Suppose that $M_i \ge \max_{\boldsymbol{x} \in \mathcal{X}, (7\mathrm{b})} |f(\boldsymbol{x}, \boldsymbol{\xi}_i)|$ for each $i \in [m]$. For each $k \in [m_a]$ and $a \in A$, the quantile set $\Omega_a(k)$ is equivalent to*

$$\Omega_a(k) = \left\{ (\boldsymbol{x}, t_{ka}) \in \mathcal{X} \times \mathbb{R} : \begin{array}{l} \pi_{ika} \in \{0,1\}, z_{ika} \in \{0,1\}, \pi_{ika} \le z_{ika}, (\boldsymbol{x}, \bar{w}_i) \in X_i, \forall i \in C_a, \\ \sum_{i \in C_a} z_{ika} = k, \sum_{i \in C_a} \pi_{ika} = 1, t_{ka} = \sum_{i \in C_a} \widehat{t}_{ika}, \\ t_{ka} \ge \bar{w}_i - (M_i + M_{(k)})(1 - z_{ika}), t_{ka} \le \bar{w}_i + (M_i + M_{(k)}) z_{ika}, \\ (\widehat{t}_{ika}, \pi_{ika}, \bar{w}_i) \in \mathrm{MC}(0, 1, -M_i, M_i), \forall i \in C_a \end{array} \right\}, \tag{11}$$

where $M_{(i)}$ is the ith smallest value of the vector $\boldsymbol{M}$.

Proof. See Appendix B.5. □

To reformulate the set $\mathcal{F}_q$ defined in (9), we use the quantile-based representation (10) of the Wasserstein fairness measure by plugging in the MICP-R quantile sets $\{\Omega_a(k)\}_{k \in [m_a], a \in A}$ defined in (11).

Theorem 2. ***(Quantile Formulation)*** *Suppose that the set $X_i = \{(\boldsymbol{x}, \bar{w}_i) \in \mathcal{X} \times \mathbb{R} : f(\boldsymbol{x}, \boldsymbol{\xi}_i) = \bar{w}_i\}$ is MICP-R and $M_i \ge \max_{\boldsymbol{x} \in \mathcal{X}, (7\mathrm{b})} |f(\boldsymbol{x}, \boldsymbol{\xi}_i)|$ for each $i \in [m]$. We further define the quantile set*

$\Omega_a(k) = \{(\boldsymbol{x}, t_{ka}) \in \mathcal{X} \times \mathbb{R} : F_a^{-1}(k/m_a \mid \boldsymbol{x}) = t_{ka}\}$, *which admits a MICP-R form* (11)*. Then* $\mathcal{F}_q$ *can be represented as*

$$\mathcal{F}_q = \left\{ (\boldsymbol{x}, \nu) \in \mathcal{X} \times \mathbb{R}_+ : \begin{array}{c} \displaystyle\sum_{i \in [\widehat{m}_{a\bar{a}} - 1]} \left(\widehat{b}_{(i+1)a\bar{a}} - \widehat{b}_{ia\bar{a}} \right) \eta_{ia\bar{a}}^q \leq \nu, \forall a < \bar{a} \in A, \\ \left| \displaystyle\sum_{j \in [m_a]} \delta_{ija\bar{a}1} t_{ja} - \sum_{j \in [m_{\bar{a}}]} \delta_{ija\bar{a}2} t_{j\bar{a}} \right| \leq \eta_{ia\bar{a}}, \forall i \in [\widehat{m}_{a\bar{a}} - 1], a < \bar{a} \in A, \\ (\boldsymbol{x}, t_{ja}) \in \Omega_a(j), \forall j \in [m_a], a \in A \end{array} \right\}, \quad (12)$$

where

$$\delta_{ija\bar{a}1} = \mathbb{I}\left(\left(\widehat{b}_{ia\bar{a}}, \widehat{b}_{(i+1)a\bar{a}} \right] \subseteq \left(\frac{j-1}{m_a}, \frac{j}{m_a} \right] \right), \forall i \in [\widehat{m}_{a\bar{a}} - 1], j \in [m_a], a < \bar{a} \in A,$$

and

$$\delta_{ija\bar{a}2} = \mathbb{I}\left(\left(\widehat{b}_{ia\bar{a}}, \widehat{b}_{(i+1)a\bar{a}} \right] \subseteq \left(\frac{j-1}{m_{\bar{a}}}, \frac{j}{m_{\bar{a}}} \right] \right), \forall i \in [\widehat{m}_{a\bar{a}} - 1], j \in [m_{\bar{a}}], a < \bar{a} \in A.$$

Proof. See Appendix B.6. □

According to (11), we see that the continuous variable $t_{ja} = F_a^{-1}(j/m_a \mid \boldsymbol{x})$ represents the j/m_ath smallest quantile value, and the binary variable z_{ija} indicates whether up to the ith smallest quantile value is selected or not for each $i \in C_a, j \in [m_a]$, and $a \in A$. Therefore, we obtain the following monotonicity-based valid inequalities.

Proposition 8. *The following inequalities are valid for the Quantile Formulation*

$$t_{ja} \leq t_{(j+1)a}, z_{ija} \geq z_{i(j+1)a}, \forall i \in C_a, j \in [m_a], a \in A. \quad (13)$$

3.3. Aggregate Quantile Formulation

Motivated by the quantile set $\Omega_a(k)$ in (11), we develop another formulation using the aggregate quantiles in this subsection. We also show that this formulation can be quite strong compared to others. To begin with, let us define the aggregate quantile variable $\bar{t}$ and the aggregate quantile set $\bar{\Omega}_a(k) = \{(\boldsymbol{x}, \bar{t}_{ka}) \in \mathcal{X} \times \mathbb{R} : \sum_{i=1}^k F_a^{-1}(i/m_a \mid \boldsymbol{x}) = \bar{t}_{ka}\}$ for each $k \in [m_a]$ and $a \in A$. Letting $(\boldsymbol{x}, \bar{w}_i) \in X_i$ for each $i \in [m]$, we present the following representation of the set $\bar{\Omega}_a(k)$

$$\bar{\Omega}_a(k) = \left\{ (\boldsymbol{x}, \bar{t}_{ka}) \in \mathcal{X} \times \mathbb{R} : \begin{array}{l} (\boldsymbol{x}, \bar{w}_i) \in X_i, \forall i \in C_a, \\ \bar{t}_{ka} \leq \min\limits_{\bar{\boldsymbol{z}}} \left\{ \displaystyle\sum_{i \in C_a} \bar{z}_{ika} \bar{w}_i : \bar{z}_{ika} \in \{0,1\}, \forall i \in C_a, \sum_{i \in C_a} \bar{z}_{ika} = k \right\}, \\ \bar{t}_{ka} \geq \min\limits_{\boldsymbol{z}} \left\{ \displaystyle\sum_{i \in C_a} z_{ika} \bar{w}_i : z_{ika} \in \{0,1\}, \forall i \in C_a, \sum_{i \in C_a} z_{ika} = k \right\} \end{array} \right\},$$

where similar to (11), we let the binary variables $z_{ika}, \bar{z}_{ika}$ indicate whether up to the ith smallest quantile values is selected or not for each $i \in C_a, k \in [m_a]$, and $a \in A$. By dualizing the first minimization problem and linearizing the bilinear terms in the second minimization problem, we arrive at the following MICP-R set.

Proposition 9. *Suppose that $M_i \ge \max_{\boldsymbol{x}\in\mathcal{X},(7\mathrm{b})} |f(\boldsymbol{x},\boldsymbol{\xi}_i)|$ for each $i \in [m]$. For each $k \in [m_a]$ and $a \in A$, the aggregate quantile set $\bar{\Omega}_a(k)$ is equivalent to*

$$\bar{\Omega}_a(k) = \left\{ (\boldsymbol{x}, \bar{t}_{ka}) \in \mathcal{X}\times\mathbb{R} : \begin{array}{c} z_{ika}\in\{0,1\}, (\boldsymbol{x},\bar{w}_i)\in X_i, \forall i\in C_a, \sum_{i\in C_a} z_{ika} = k, \\ \bar{t}_{ka} \le k\pi_{ka} - \sum_{i\in C_a}\rho_{ika}, \pi_{ka}-\rho_{ika}\le \bar{w}_i, \rho_{ika}\ge 0, \forall i\in C_a, \\ \bar{t}_{ka} \ge \sum_{i\in C_a} s_{ika}, (s_{ika}, z_{ika}, \bar{w}_i)\in \mathrm{MC}(0,1,-M_i,M_i), \forall i \in C_a \end{array} \right\}. \tag{14}$$

To represent the set $\mathcal{F}_q$ defined in (9), we simply plug in the representation of the aggregate quantile sets $\{\bar{\Omega}_a(k)\}_{k\in[m_a],a\in A}$ into the representation (9), which motivates the following formulation.

Theorem 3. ***(Aggregate Quantile Formulation)*** *Suppose that the set $X_i = \{(\boldsymbol{x},\bar{w}_i)\in\mathcal{X}\times\mathbb{R} : f(\boldsymbol{x},\boldsymbol{\xi}_i) = \bar{w}_i\}$ is MICP-R and $M_i \ge \max_{\boldsymbol{x}\in\mathcal{X},(7\mathrm{b})} |f(\boldsymbol{x},\boldsymbol{\xi}_i)|$ for each $i\in[m]$. We further define the aggregate quantile set $\bar{\Omega}_a(k) = \{(\boldsymbol{x},\bar{t}_{ka})\in\mathcal{X}\times\mathbb{R} : \sum_{i=1}^{k} F_a^{-1}(i/m_a \mid \boldsymbol{x}) = \bar{t}_{ka}\}$, which admits a MICP-R form (14). Then $\mathcal{F}_q$ can be represented as*

$$\mathcal{F}_q = \left\{ (\boldsymbol{x},\nu)\in\mathcal{X}\times\mathbb{R}_+ : \begin{array}{c} \sum_{i\in[\widehat{m}_{a\bar{a}}-1]} \left(\widehat{b}_{(i+1)a\bar{a}} - \widehat{b}_{ia\bar{a}}\right)\eta^q_{ia\bar{a}} \le \nu, \forall a<\bar{a}\in A, \\ \left| \sum_{j\in[m_a]} \delta_{ija\bar{a}1} t_{ja} - \sum_{j\in[m_{\bar{a}}]} \delta_{ija\bar{a}2} t_{j\bar{a}} \right| \le \eta_{ia\bar{a}}, \forall i\in[\widehat{m}_{a\bar{a}}-1], a<\bar{a}\in A, \\ t_{ja} = \bar{t}_{ja} - \bar{t}_{(j-1)a}, (\boldsymbol{x},\bar{t}_{ja})\in\bar{\Omega}_a(j), \bar{t}_{0a}=0, \forall j\in[m_a], a\in A \end{array} \right\}, \tag{15}$$

where the parameters $\boldsymbol{\delta}$ are defined in Theorem 2.

Proof. The proof follows Theorem 2 with the fact that $t_{ja} = \bar{t}_{ja} - \bar{t}_{(j-1)a}$ for all $j\in[m_a], a\in A$. □

We remark that the inequalities (13) are also valid for the Aggregate Quantile Formulation.

3.4. Summary of the Different Formulations

The different formulations have their own strengths from the derivations according to their developments. Their formulation complexities are summarized in Table 1, where we suppress the term $O(|A|^2)$ for simplicity. In our numerical study, we observe that the Aggregate Quantile Formulation consistently outperforms the others in terms of computational time, which might be because it has the least amount of binary variables and the smallest big-M coefficients.

In the following, we show that the Quantile Formulation and the Aggregate Quantile Formulation can be stronger than the other two under some assumptions.

Proposition 10. *Suppose that the big-M coefficients $\boldsymbol{M}, \widehat{\boldsymbol{M}}$ are large enough as specified in the proof. Then, by relaxing the binary variables,*

Table 1 Formulation Complexity Comparisons

Formulation	# of Constraints	# of Binary Variables	# of Continuous Variables	Largest Big-M Coefficient
Discretized	$O(m^2 \log(m))$	$O(m^2 \log(m))$	$O(m^2 \log(m))$	$\max_{i\in[m]} M_i$
Complementary	$O(m^2)$	$O(m^2)$	$O(m^2)$	$\max_{a<\bar{a}\in A} \sum_{(i,j)\in C_a\times C_{\bar{a}}} (M_i+M_j)^q$
Quantile	$O(m^2)$	$O(m^2)$	$O(m^2)$	$2\max_{i\in[m]} M_i$
Aggregate Quantile	$O(m^2)$	$O(m^2)$	$O(m^2)$	$\max_{i\in[m]} M_i$

(i) the continuous relaxation value of the Discretized Formulation is zero;

(ii) the continuous relaxation value of the Complementary Formulation is zero;

(iii) the continuous relaxation value of the Quantile Formulation is

$$\min_{\boldsymbol{x}\in\mathcal{X}} \max_{a<\bar{a}\in A} \left(\widehat{b}_{(\widehat{m}_{a\bar{a}})a\bar{a}} - \widehat{b}_{(\widehat{m}_{a\bar{a}}-1)a\bar{a}}\right) \left|F_a^{-1}(1\mid\boldsymbol{x}) - F_{\bar{a}}^{-1}(1\mid\boldsymbol{x})\right|^q;$$

(iv) the continuous relaxation value of the Aggregate Quantile Formulation is at least

$$\min_{\boldsymbol{x}\in\mathcal{X}} \max_{a<\bar{a}\in A} \left|\frac{1}{m_a}\sum_{i\in C_a} F_a^{-1}\left(\frac{i}{m_a}\,\middle|\,\boldsymbol{x}\right) - \frac{1}{m_{\bar{a}}}\sum_{i\in C_{\bar{a}}} F_{\bar{a}}^{-1}\left(\frac{i}{m_{\bar{a}}}\,\middle|\,\boldsymbol{x}\right)\right|^q.$$

Proof. See Appendix B.7 □

As a side product of Proposition 10, we see that

Corollary 1. *For any $q \geq 1$, the continuous relaxation value of the Aggregate Quantile Formulation is at least as good as the Jensen bound presented in Section 4.1.*

The continuous relaxations of all the formulations can have nonzero objective values if one optimizes the big-M coefficients or adds valid inequalities. We numerically test each formulation in Section 5.1 and observe that the Aggregate Quantile Formulation performs best overall.

3.5. An Alternating Minimization (AM) Algorithm

When solving large instances with thousands of populations, the exact formulations in the previous subsections may suffer from slow convergence to find an optimal solution. Therefore, in this subsection, motivated by the representation in Lemma 1, we design a fast AM algorithm that can effectively solve DFSO instances to near optimality.

To this end, according to (10), we can recast DFSO as

$$v^*(q) = \min_{\boldsymbol{x}\in\mathcal{X},\nu} \quad \nu, \tag{16a}$$

$$\text{s.t.} \quad \sum_{i\in[\widehat{m}_{a\bar{a}}-1]} \left(\widehat{b}_{(i+1)a\bar{a}} - \widehat{b}_{ia\bar{a}}\right) \left|F_a^{-1}\left(\widehat{b}_{(i+1)a\bar{a}}\mid\boldsymbol{x}\right) - F_{\bar{a}}^{-1}\left(\widehat{b}_{(i+1)a\bar{a}}\mid\boldsymbol{x}\right)\right|^q \leq \nu, \forall a<\bar{a}\in A, \tag{16b}$$

(7b),

which has been used to derive the Quantile Formulation and the Aggregate Quantile Formulation of DFSO. This formulation is also valuable for deriving the AM algorithm. Specifically, we can run the AM algorithm as follows: (i) First, we pick a feasible solution $\boldsymbol{x}_0$ (e.g., an optimal solution that minimizes the total cost (1)); (ii) At iteration $t \geq 0$, we find the inverse distribution functions $\{F_a^{-1}(\cdot \mid \boldsymbol{x}_t)\}_{a \in A}$, which can be done via sorting with time complexity $O(m \log m)$; (iii) For each $i \in [\widehat{m}_{a\bar{a}} - 1]$ and $a < \bar{a} \in A$, let $f(\boldsymbol{x}_t, \boldsymbol{\xi}_{\widehat{s}_a(i)}) := F_a^{-1}(\widehat{b}_{(i+1)a\bar{a}} \mid \boldsymbol{x}_t)$ and $f(\boldsymbol{x}_t, \boldsymbol{\xi}_{\widehat{s}_{\bar{a}}(i)}) := F_{\bar{a}}^{-1}(\widehat{b}_{(i+1)a\bar{a}} \mid \boldsymbol{x}_t)$; (iv) Next, we solve the following program by fixing the inverse distribution functions in the DFSO (16):

$$\begin{aligned} v_{t+1}(q) = \min_{\boldsymbol{x} \in \mathcal{X}, \nu} \quad & \nu, \\ \text{s.t.} \quad & \sum_{i \in [\widehat{m}_{a\bar{a}} - 1]} \left(\widehat{b}_{(i+1)a\bar{a}} - \widehat{b}_{ia\bar{a}} \right) \left| f\left(\boldsymbol{x}, \boldsymbol{\xi}_{\widehat{s}_a(i)}\right) - f\left(\boldsymbol{x}, \boldsymbol{\xi}_{\widehat{s}_{\bar{a}}(i)}\right) \right|^q \leq \nu, \forall a < \bar{a} \in A, \\ & \text{(7b)}, \end{aligned}$$

with an optimal solution $\boldsymbol{x}_{t+1}$; and (v) Let $t := t + 1$ and repeat Step (ii) to Step (iv) until the stopping criterion is invoked (e.g., $|v_t - v_{t+1}| < \bar{\epsilon}$ for some small threshold $\bar{\epsilon}$). The benefit of the proposed AM algorithm is that it completely eliminates the necessity of auxiliary binary variables introduced by the exact MICP-R formulations. In addition to its computational advantage, our numerical study shows that the proposed AM algorithm can successfully find optimal solutions in many instances.

4. Two Lower Bounds for the Wasserstein Fairness Measure

In this section, we study a compact Jensen lower bound for type-q Wasserstein fairness measure (i.e., $\mathrm{WD}_q^q(\boldsymbol{x})$) and the well-known Gelbrich lower bound for type-2 Wasserstein fairness measure (i.e., $\mathrm{WD}_2^2(\boldsymbol{x})$). In particular, we derive new conditions under which the Gelbrich bound is tight. To obtain the equivalent MICP-R formulations, we assume that $f(\boldsymbol{x}, \boldsymbol{\xi}) = \boldsymbol{\xi}^\top \boldsymbol{r}(\boldsymbol{x}) + s(\boldsymbol{x})$, where $\boldsymbol{r}(\boldsymbol{x}) = \boldsymbol{A}\boldsymbol{x} + \widehat{\boldsymbol{a}}_0$ and $s(\boldsymbol{x}) = \widehat{\boldsymbol{a}}_1^\top \boldsymbol{x} + \widehat{a}_2$ are linear functions. For notational convenience, we define the mean and covariance matrix for each group $a \in A$ as $\boldsymbol{\mu}_a = \mathbb{E}_{\mathbb{P}}[\tilde{\boldsymbol{\xi}}_a]$ and $\boldsymbol{\Sigma}_a = \mathrm{Cov}_{\mathbb{P}}[\tilde{\boldsymbol{\xi}}_a]$, respectively.

4.1. The Jensen Bound for the qth Power of Type-q Wasserstein Fairness Measure $\mathrm{WD}_q^q(\boldsymbol{x})$

We first introduce the Jensen bound for $\mathrm{WD}_q^q(\boldsymbol{x})$, which enables us to ascertain that the semidefinite relaxation of the Gelbrich bound is relatively weak. The following theorem establishes the relation between the Wasserstein fairness measure and the Jensen bound.

THEOREM 4 **(The Jensen Bound)**. *For any $q \geq 1$, $\mathrm{WD}_q^q(\boldsymbol{x})$ is bounded by*

$$\mathrm{WD}_q^q(\boldsymbol{x}) \geq \max_{a < \bar{a} \in A} \left| \boldsymbol{\mu}_a^\top \boldsymbol{r}(\boldsymbol{x}) - \boldsymbol{\mu}_{\bar{a}}^\top \boldsymbol{r}(\boldsymbol{x}) \right|^q := v_J(q).$$

Proof. For any q, $a < \bar{a} \in A$, and joint distribution $\mathbb{Q}_{a,\bar{a}}$ of $f(\boldsymbol{x}, \tilde{\boldsymbol{\xi}}_a)$ and $f(\boldsymbol{x}, \tilde{\boldsymbol{\xi}}_{\bar{a}})$ with marginals $\mathbb{P}_a, \mathbb{P}_{\bar{a}}$, we have

$$\begin{aligned}\mathbb{E}_{\mathbb{Q}_{a,\bar{a}}}[|f(\boldsymbol{x},\boldsymbol{\xi}_a)-f(\boldsymbol{x},\boldsymbol{\xi}_{\bar{a}})|^q] &\geq \left|\mathbb{E}_{\mathbb{Q}_{a,\bar{a}}}[f(\boldsymbol{x},\boldsymbol{\xi}_a)]-\mathbb{E}_{\mathbb{Q}_{a,\bar{a}}}[f(\boldsymbol{x},\boldsymbol{\xi}_{\bar{a}})]\right|^q, \\ &= |\mathbb{E}_{\mathbb{P}_a}[f(\boldsymbol{x},\boldsymbol{\xi}_a)]-\mathbb{E}_{\mathbb{P}_{\bar{a}}}[f(\boldsymbol{x},\boldsymbol{\xi}_{\bar{a}})]|^q = \left|\boldsymbol{\mu}_a^\top \boldsymbol{r}(\boldsymbol{x})-\boldsymbol{\mu}_{\bar{a}}^\top \boldsymbol{r}(\boldsymbol{x})\right|^q.\end{aligned}$$

Here, the inequality is due to Jensen's inequality, and the first equality is because the random vectors $\boldsymbol{\xi}_a, \boldsymbol{\xi}_{\bar{a}}$ are governed by the marginal distributions $\mathbb{P}_a, \mathbb{P}_{\bar{a}}$, respectively. Then, we obtain

$$\mathrm{WD}_q^q(\boldsymbol{x}) = \max_{a<\bar{a}\in A} W_q^q(\mathbb{P}_a, \mathbb{P}_{\bar{a}}) \geq \max_{a<\bar{a}\in A} \left|\boldsymbol{\mu}_a^\top \boldsymbol{r}(\boldsymbol{x})-\boldsymbol{\mu}_{\bar{a}}^\top \boldsymbol{r}(\boldsymbol{x})\right|^q,$$

which completes the proof. □

This result gives rise to the following model for computing the Jensen bound for $\mathrm{WD}_q^q(\boldsymbol{x})$:

$$v_J(q) = \min_{\boldsymbol{x}\in\mathcal{X},\nu} \left\{\nu : \left|\boldsymbol{\mu}_a^\top \boldsymbol{r}(\boldsymbol{x})-\boldsymbol{\mu}_{\bar{a}}^\top \boldsymbol{r}(\boldsymbol{x})\right|^q \leq \nu, \forall a<\bar{a}\in A, \text{(7b)}\right\}. \tag{18}$$

4.2. The Gelbrich Bound for the Squared Type-2 Wasserstein Fairness Measure $\mathrm{WD}_2^2(\boldsymbol{x})$

When $q=2$, there is a popular Gelbrich bound for $W_2^2(\mathbb{P}_a, \mathbb{P}_{\bar{a}})$ for any $a<\bar{a}\in A$, which has been studied in many optimal transport works (see, e.g., Kuhn et al. 2019). Formally, the Gelbrich bound for $W_2^2(\mathbb{P}_a, \mathbb{P}_{\bar{a}})$ is defined as follows.

Definition 7 (The Gelbrich Bound, Theorem 2.1 in Gelbrich 1990). For any $a<\bar{a}\in A$, the squared type-2 Wasserstein distance $W_2^2(\mathbb{P}_a, \mathbb{P}_{\bar{a}})$ is bounded by

$$W_2^2(\mathbb{P}_a, \mathbb{P}_{\bar{a}}) \geq \left(\boldsymbol{\mu}_a^\top \boldsymbol{r}(\boldsymbol{x})-\boldsymbol{\mu}_{\bar{a}}^\top \boldsymbol{r}(\boldsymbol{x})\right)^2 + \left(\sqrt{\boldsymbol{r}(\boldsymbol{x})^\top \boldsymbol{\Sigma}_a \boldsymbol{r}(\boldsymbol{x})} - \sqrt{\boldsymbol{r}(\boldsymbol{x})^\top \boldsymbol{\Sigma}_{\bar{a}} \boldsymbol{r}(\boldsymbol{x})}\right)^2.$$

According to Definition 7, the Gelbrich bound can be computed via the following nonconvex program:

$$\begin{aligned} v_G = \min_{\boldsymbol{x}\in\mathcal{X},\nu} \quad & \nu, && \text{(19a)}\\ \text{s.t.} \quad & \left(\boldsymbol{\mu}_a^\top \boldsymbol{r}(\boldsymbol{x})-\boldsymbol{\mu}_{\bar{a}}^\top \boldsymbol{r}(\boldsymbol{x})\right)^2 + \left(\sqrt{\boldsymbol{r}(\boldsymbol{x})^\top \boldsymbol{\Sigma}_a \boldsymbol{r}(\boldsymbol{x})} - \sqrt{\boldsymbol{r}(\boldsymbol{x})^\top \boldsymbol{\Sigma}_{\bar{a}} \boldsymbol{r}(\boldsymbol{x})}\right)^2 \leq \nu, \forall a<\bar{a}\in A, && \text{(19b)}\\ & \text{(7b)}. \end{aligned}$$

where $\boldsymbol{\mu}_a$ and $\boldsymbol{\Sigma}_a$ are the mean and covariance of $\tilde{\boldsymbol{\xi}}_a$ for all a.

Using the Cholesky decomposition $\boldsymbol{\Sigma}_a = \boldsymbol{L}_a \boldsymbol{L}_a^\top$ for each $a\in A$, we can recast (19) as

$$\begin{aligned} v_G = \min_{\boldsymbol{x}\in\mathcal{X},\boldsymbol{z},\nu} \quad & \nu, && \text{(20a)}\\ \text{s.t.} \quad & \boldsymbol{z}_a = \boldsymbol{L}_a^\top \boldsymbol{r}(\boldsymbol{x}), \forall a\in A, && \text{(20b)}\\ & \left(\boldsymbol{\mu}_a^\top \boldsymbol{r}(\boldsymbol{x})-\boldsymbol{\mu}_{\bar{a}}^\top \boldsymbol{r}(\boldsymbol{x})\right)^2 + (\|\boldsymbol{z}_a\|_2 - \|\boldsymbol{z}_{\bar{a}}\|_2)^2 \leq \nu, \forall a<\bar{a}\in A, && \text{(20c)} \end{aligned}$$

$$(7\mathrm{b}).$$

Our numerical study shows that the Gelbrich bound (20) can be very close to the true optimal value $v^*(2)$. Unfortunately, computing the Gelbrich bound (20) constitutes an intractable nonconvex program, which we formally prove to be generically NP-hard.

THEOREM 5. *Computing the Gelbrich bound is strongly NP-hard even when $\epsilon = \infty$ and $|A| = 2$.*

Proof. See Appendix B.8. □

We remark that, in practice, one can compute the Gelbrich bound (20) by employing off-the-shelf solvers, which are based on the spatial branch and bound algorithm. To further expedite the solution process, we can tighten the bounds of decision variables and auxiliary variables in formulation (20), which significantly decreases the number of branch and bound nodes and thus accelerates the computation.

AM Algorithm. The complexity result motivates us to solve (20) using a highly effective AM method. We first rewrite the formulation (20) as

$$\begin{aligned} v_G = \min_{\boldsymbol{x}\in\mathcal{X},\boldsymbol{z},\nu} \quad & \nu, \\ \text{s.t.} \quad & \left(\boldsymbol{\mu}_a^\top \boldsymbol{r}(\boldsymbol{x}) - \boldsymbol{\mu}_{\bar{a}}^\top \boldsymbol{r}(\boldsymbol{x})\right)^2 + 2\|\boldsymbol{z}_a\|_2^2 + 2\|\boldsymbol{z}_{\bar{a}}\|_2^2 - \left(\|\boldsymbol{z}_a\|_2 + \|\boldsymbol{z}_{\bar{a}}\|_2\right)^2 \le \nu, \forall a < \bar{a} \in A, \\ & (7\mathrm{b}), (20\mathrm{b}). \end{aligned}$$

Using the convex conjugate representation, we have

$$-\left(\|\boldsymbol{z}_a\|_2 + \|\boldsymbol{z}_{\bar{a}}\|_2\right)^2 = \min_{w_{a\bar{a}}\ge 0,\boldsymbol{\alpha}_a,\boldsymbol{\alpha}_{\bar{a}}} \left\{-2\boldsymbol{\alpha}_a^\top \boldsymbol{z}_a - 2\boldsymbol{\alpha}_a^\top \boldsymbol{z}_a + w_{a\bar{a}}^2 : \|\boldsymbol{\alpha}_a\|_2 \le w_{a\bar{a}}, \|\boldsymbol{\alpha}_{\bar{a}}\|_2 \le w_{a\bar{a}}\right\}.$$

Thus, we can equivalently restate the formulation (20) as

$$\begin{aligned} v_G = \min_{\boldsymbol{x}\in\mathcal{X},\boldsymbol{z},\boldsymbol{w},\boldsymbol{\alpha},\nu} \quad & \nu, & (21\mathrm{a}) \\ \text{s.t.} \quad & \left(\boldsymbol{\mu}_a^\top \boldsymbol{r}(\boldsymbol{x}) - \boldsymbol{\mu}_{\bar{a}}^\top \boldsymbol{r}(\boldsymbol{x})\right)^2 + 2\|\boldsymbol{z}_a\|_2^2 + 2\|\boldsymbol{z}_{\bar{a}}\|_2^2 - 2\boldsymbol{\alpha}_a^\top \boldsymbol{z}_a & \\ & - 2\boldsymbol{\alpha}_a^\top \boldsymbol{z}_a + w_{a\bar{a}}^2 \le \nu, |\boldsymbol{\alpha}_a\|_2 \le w_{a\bar{a}}, \|\boldsymbol{\alpha}_{\bar{a}}\|_2 \le w_{a\bar{a}}, \forall a < \bar{a} \in A, & (21\mathrm{b}) \\ & (7\mathrm{b}), (20\mathrm{b}). & \end{aligned}$$

In the AM method, at each iteration t, given a solution $(\boldsymbol{x}_t, \boldsymbol{z}_t, \nu_t)$, we compute the solution $(\boldsymbol{w}_t, \boldsymbol{\alpha}_t, \bar{\nu}_t)$ in closed-form, as follows:

$$\begin{aligned} & \bar{\nu}_t = \max_{a<\bar{a}\in A} \left\{\left(\boldsymbol{\mu}_a^\top \boldsymbol{r}(\boldsymbol{x}_t) - \boldsymbol{\mu}_{\bar{a}}^\top \boldsymbol{r}(\boldsymbol{x}_t)\right)^2 + \left(\|\boldsymbol{z}_{at}\|_2 - \|\boldsymbol{z}_{\bar{a}t}\|_2\right)^2\right\}, \\ & w_{a\bar{a}t} = \|\boldsymbol{z}_{at}\|_2 + \|\boldsymbol{z}_{\bar{a}t}\|_2, \boldsymbol{\alpha}_{at} = \frac{w_{a\bar{a}t}}{\|\boldsymbol{z}_{at}\|_2} \boldsymbol{z}_{at}, \boldsymbol{\alpha}_{\bar{a}t} = \frac{w_{a\bar{a}t}}{\|\boldsymbol{z}_{\bar{a}t}\|_2} \boldsymbol{z}_{\bar{a}t}. \end{aligned}$$

Then we fix the values of $(\boldsymbol{w}_t, \boldsymbol{\alpha}_t)$ and resolve (21) with respect to the variables $(\boldsymbol{x}, \boldsymbol{z}, \nu)$. The procedure is repeated until we reach a prescribed tolerance. Our numerical study finds that the AM approach works extremely well in quickly finding near-optimal solutions.

Semidefinite Programming Relaxation. Alternatively, in the Gelbrich bound formulation (20), let us introduce a new variable $\sigma_a = \|\boldsymbol{z}_a\|_2$ for each $a \in A$. For each pair $a < \bar{a} \in A$, let us denote $\boldsymbol{s}_{a\bar{a}} = \left[\sigma_a \ \boldsymbol{z}_a \ \sigma_{\bar{a}} \ \boldsymbol{z}_{\bar{a}}\right]^\top$ and $\boldsymbol{Z}_{a\bar{a}} = \boldsymbol{s}_{a\bar{a}} \cdot \boldsymbol{s}_{a\bar{a}}^\top$. Then one can show that the Gelbrich bound (20) can be converted to a semidefinite programming formulation with rank-one constraint, as follows:

$$
\begin{aligned}
v_G = \min_{\boldsymbol{x} \in \mathcal{X}, \boldsymbol{s}, \boldsymbol{Z}, \nu} \quad & \nu, && \text{(22a)}\\
\text{s.t.} \quad & \left(\boldsymbol{\mu}_a^\top \boldsymbol{r}(\boldsymbol{x}) - \boldsymbol{\mu}_{\bar{a}}^\top \boldsymbol{r}(\boldsymbol{x})\right)^2 + \\
& \left(Z_{a\bar{a}11} - 2 Z_{a\bar{a}1(n+2)} + Z_{a\bar{a}(n+2)(n+2)}\right) \leq \nu, \forall a < \bar{a} \in A, && \text{(22b)}\\
& \boldsymbol{z}_a = \boldsymbol{L}_a^\top \boldsymbol{r}(\boldsymbol{x}), \sigma_a \geq 0, \forall a \in A, && \text{(22c)}\\
& Z_{a\bar{a}11} = \sum_{i=2}^{n+1} Z_{a\bar{a}ii}, Z_{a\bar{a}(n+2)(n+2)} = \sum_{i=n+3}^{2n+2} Z_{a\bar{a}ii}, \forall a < \bar{a} \in A, && \text{(22d)}\\
& \boldsymbol{s}_{a\bar{a}} = \left[\sigma_a \ \boldsymbol{z}_a \ \sigma_{\bar{a}} \ \boldsymbol{z}_{\bar{a}}\right]^\top, \forall a < \bar{a} \in A, && \text{(22e)}\\
& \boldsymbol{Z}_{a\bar{a}} = \boldsymbol{s}_{a\bar{a}} \cdot \boldsymbol{s}_{a\bar{a}}^\top, \forall a < \bar{a} \in A, && \text{(22f)}\\
& \text{(7b)}.
\end{aligned}
$$

The rank one constraints in (22f) are difficult to handle in practice. A simple way is to relax (22f) as the semidefinite inequalities

$$
\boldsymbol{Z}_{a\bar{a}} \succeq \boldsymbol{s}_{a\bar{a}} \cdot \boldsymbol{s}_{a\bar{a}}^\top, \forall a < \bar{a} \in A.
$$

Using the Schur complement, we obtain the semidefinite relaxation of the Gelbrich bound (20) as

$$
\begin{aligned}
v_{\underline{G}} = \min_{\boldsymbol{x} \in \mathcal{X}, \boldsymbol{s}, \boldsymbol{Z}, \nu} \quad & \nu, && \text{(23a)}\\
\text{s.t.} \quad & \begin{bmatrix} 1 & \boldsymbol{s}_{a\bar{a}}^\top \\ \boldsymbol{s}_{a\bar{a}} & \boldsymbol{Z}_{a\bar{a}} \end{bmatrix} \succeq 0, \forall a < \bar{a} \in A, && \text{(23b)}\\
& \text{(7b)}, \text{(22b)} - \text{(22e)}.
\end{aligned}
$$

We see that the semidefinite relaxation (23) is stronger than the type-2 Jensen bound $v_J(2)$ in (18) since $\boldsymbol{Z}_{a\bar{a}}$ is positive semidefinite and $\left(Z_{a\bar{a}11} - 2Z_{a\bar{a}1(n+2)} + Z_{a\bar{a}(n+2)(n+2)}\right) \geq 0$ for every pair $a < \bar{a} \in A$. On the other hand, if we allow the relative tolerance of the semidefinite constraints (see MOSEK ApS 2019), then for up to any prescribed tolerance, we can show that $v_{\underline{G}} \leq v_J(2)$. This result is summarized in the following proposition.

Proposition 11. *The semidefinite relaxation* (23) *of the Gelbrich bound model satisfies* $v_{\underline{G}} \geq v_J(2)$*. On the other hand, for any relative tolerance* $\beta > 0$ *of the semidefinite constraints in* (23b) *such that* $\boldsymbol{Z}_{a\bar{a}} - \boldsymbol{s}_{a\bar{a}} \cdot \boldsymbol{s}_{a\bar{a}}^\top \succeq -\beta \lambda_{min}^{+}(\boldsymbol{Z}_{a\bar{a}}) \boldsymbol{I}_{2n+2}$*, where* $\lambda_{min}^{+}(\cdot)$ *denotes the smallest nonzero eigenvalue, we have* $v_{\underline{G}} \leq v_J(2)$*.*

Proof. See Appendix B.9. □

4.3. Tightness of the Gelbrich Bound

The Same Univariate Marginal Distribution Condition: In the literature, it is known that the Gelbrich bound is tight when the random parameters $\{\tilde{\boldsymbol{\xi}}_a\}_{a\in A}$ are asymptotically elliptical as $m_a \to \infty$ for all $a \in A$. We generalize this result by establishing a weaker condition that achieves the tightness of the Gelbrich bound. Our result shows that when the random utility functions of different groups can be linearly transformed to the same univariate random variable, then the Gelbrich bound is asymptotically tight.

THEOREM 6. *Suppose that for any pair $a < \bar{a} \in A$, the optimal comonotonic random variables $(f(\boldsymbol{x}, \tilde{\boldsymbol{\xi}}_a) - \boldsymbol{\mu}_a^\top \boldsymbol{r}(\boldsymbol{x}) - s(\boldsymbol{x}), f(\boldsymbol{x}, \tilde{\boldsymbol{\xi}}_{\bar{a}}) - \boldsymbol{\mu}_{\bar{a}}^\top \boldsymbol{r}(\boldsymbol{x}) - s(\boldsymbol{x})) \xrightarrow{m_a\to\infty, m_{\bar{a}}\to\infty} (\sqrt{\boldsymbol{r}(\boldsymbol{x})^\top \boldsymbol{\Sigma}_a \boldsymbol{r}(\boldsymbol{x})}\tilde{u}, \sqrt{\boldsymbol{r}(\boldsymbol{x})^\top \boldsymbol{\Sigma}_{\bar{a}} \boldsymbol{r}(\boldsymbol{x})}\tilde{u})$ for a univariate random variable $\tilde{u}$ with zero mean and unit variance. Then the Gelbrich bound is asymptotically tight.*

Proof. See Appendix B.10. □

Theorem 6 shows that the tightness of the Gelbrich bound applies to a much broader family of distributions than elliptical. In fact, from the proof, we can see that the Gelbrich bound is derived using the Cauchy-Schwarz inequality, i.e.,

$$\mathbb{E}_{\mathbb{Q}_{a,\bar{a}}}\left[\Big(f(\boldsymbol{x}, \tilde{\boldsymbol{\xi}}_a) - \boldsymbol{\mu}_a^\top \boldsymbol{r}(\boldsymbol{x}) - s(\boldsymbol{x})\Big)\Big(f(\boldsymbol{x}, \tilde{\boldsymbol{\xi}}_{\bar{a}}) - \boldsymbol{\mu}_{\bar{a}}^\top \boldsymbol{r}(\boldsymbol{x}) - s(\boldsymbol{x})\Big)\right] \le \sqrt{\boldsymbol{r}(\boldsymbol{x})^\top \boldsymbol{\Sigma}_a \boldsymbol{r}(\boldsymbol{x})} \cdot \sqrt{\boldsymbol{r}(\boldsymbol{x})^\top \boldsymbol{\Sigma}_{\bar{a}} \boldsymbol{r}(\boldsymbol{x})}.$$

Thus, the tightness result holds whenever there exists a joint distribution such that the Cauchy-Schwarz inequality becomes equality.

More importantly, we can theoretically bound the gap between the optimal Gelbrich bound v_G and the optimal value of DFSO $v^*(2)$ under type $q = 2$ Wasserstein distance.

THEOREM 7. *Suppose that for any group $a \in A$, the individual samples $\{\boldsymbol{\xi}_i\}_{i\in C_a}$ satisfy $f(\boldsymbol{x}, \boldsymbol{\xi}_i) - \boldsymbol{\mu}_a^\top \boldsymbol{r}(\boldsymbol{x}) - s(\boldsymbol{x}) \overset{d}{=} \sqrt{\boldsymbol{r}(\boldsymbol{x})^\top \boldsymbol{\Sigma}_a \boldsymbol{r}(\boldsymbol{x})} u_i$ for each $i \in C_a$, where $\{u_i\}_{i\in C_a}$ are i.i.d. samples of a univariate sub-Gaussian random variable $\tilde{u}_a$ with zero mean and unit variance, and $\{\tilde{u}_a\}_{a\in A}$ obey the same distribution. Then with probability at most $1 - \widehat{\eta}$ such that $\widehat{\eta} > 0$ is small, we have*

$$v^*(2) - \bar{C}_1(\widehat{\eta} \min_{a\in A} \sqrt{m_a})^{-1} \le v_G \le v^*(2)$$

for some positive constant $\bar{C}_1$.

Proof. See Appendix B.11. □

Different Groups with Proportional Covariances: The result in Theorem 6 necessitates the same marginal distributions. We relax this assumption by establishing another tightness condition, such that the marginal distributions of different groups can be distinct.

Theorem 8. *Suppose that for any pair $a < \bar{a} \in A$, the optimal comonotonic random variables $(f(\boldsymbol{x}, \tilde{\boldsymbol{\xi}}_a) - \boldsymbol{\mu}_a^\top \boldsymbol{r}(\boldsymbol{x}) - s(\boldsymbol{x}), f(\boldsymbol{x}, \tilde{\boldsymbol{\xi}}_{\bar{a}}) - \boldsymbol{\mu}_{\bar{a}}^\top \boldsymbol{r}(\boldsymbol{x}) - s(\boldsymbol{x})) \xrightarrow{m_a \to \infty, m_{\bar{a}} \to \infty} (\widehat{\boldsymbol{\xi}}_a^\top \boldsymbol{r}(\boldsymbol{x}), \widehat{\boldsymbol{\xi}}_{\bar{a}}^\top \boldsymbol{r}(\boldsymbol{x}))$, where the random vectors $\widehat{c}_a^{-1} \widehat{\boldsymbol{\xi}}_a, \widehat{c}_{\bar{a}}^{-1} \widehat{\boldsymbol{\xi}}_{\bar{a}}$ obey the same distribution with zero mean and covariance matrix $\boldsymbol{\Sigma}_{a\bar{a}}$ for some positive parameters $\widehat{c}_a, \widehat{c}_{\bar{a}}$. Then the Gelbrich bound is asymptotically tight.*

Proof. See Appendix B.12. □

Similar to Theorem 7, we can theoretically bound the gap between the optimal Gelbrich bound v_G and the optimal value of DFSO $v^*(2)$ under type $q = 2$ Wasserstein distance.

Theorem 9. *Suppose that for any group $a \in A$, the individual samples $\{\boldsymbol{\xi}_i\}_{i \in C_a}$ satisfy $f(\boldsymbol{x}, \boldsymbol{\xi}_i) - \boldsymbol{\mu}_a^\top \boldsymbol{r}(\boldsymbol{x}) - s(\boldsymbol{x}) := \boldsymbol{\xi}_i^\top \boldsymbol{r}(\boldsymbol{x})$ for each $i \in C_a$, where $\{\boldsymbol{\xi}_i\}_{i \in C_a}$ are i.i.d. and sampling from $\widehat{\boldsymbol{\xi}}_a$ and the random vectors $\widehat{c}_a^{-1} \widehat{\boldsymbol{\xi}}_a, \widehat{c}_{\bar{a}}^{-1} \widehat{\boldsymbol{\xi}}_{\bar{a}}$ obey the same sub-Gaussian distribution with zero mean and covariance matrix $\boldsymbol{\Sigma}_{a\bar{a}}$. Then with probability at most $1 - \widehat{\eta}$ such that $\widehat{\eta} > 0$ is small, we have*

$$v^*(2) - \bar{C}_2(\widehat{\eta} \min_{a \in A} \sqrt{m_a})^{-1} \le v_G \le v^*(2)$$

for some positive constant $\bar{C}_2$.

Proof. The proof is similar to that of Theorem 7 and is thus omitted. □

5. Numerical Study

In this section, we apply our framework to several fair optimization problems. We consider the fair regression problem and the fair allocation problem of scarce medical resources. An additional numerical study on the fair knapsack problem can be found in Appendix D.2. All the instances in this section are executed in Python 3.7 with calls to Gurobi 10.0.0 on a PC with an Apple M2 Pro processor and 16GB of memory.

5.1. Fair Regression

Consider the regression problem aiming to predict the response vector $\boldsymbol{y} \in \mathbb{R}^m$ using features $\boldsymbol{\xi} \in \mathbb{R}^{m \times n}$, where the loss function is given by the mean squared error (MSE) $Q(\boldsymbol{x}, \boldsymbol{\xi}_i) = |\boldsymbol{\xi}_i^\top \boldsymbol{x} - y_i|^2$ or the mean absolute error (MAE) $Q(\boldsymbol{x}, \boldsymbol{\xi}_i) = |\boldsymbol{\xi}_i^\top \boldsymbol{x} - y_i|$. In terms of demographic parity fairness, we choose the utility function of the fair regression problem to be $f(\boldsymbol{x}, \boldsymbol{\xi}) = \boldsymbol{\xi}^\top \boldsymbol{x}$. We conduct two experiments to test the proposed methods: (i) using hypothetical data to evaluate the performance of the exact formulations, AM algorithm, and two lower bounds, and (ii) using real data to compare DFSO against two state-of-the-art methods.

5.1.1. Formulation comparisons We compare the proposed methods for solving fair regression with MAE, where we (i) test the exact formulations on small populations and (ii) test the AM algorithm and lower bounds on large populations. In this experiment, we choose $|A|=2$ (i.e., we study the fairness among two groups) and $q=2$ (i.e., we consider type-2 Wasserstein distance), and we set $\epsilon=10\%$ as the inefficiency level. The hypothetical data is generated in the following manner. The response $\tilde{y}$ is generated from $\tilde{y}=\tilde{\boldsymbol{\xi}}^\top(\boldsymbol{x}^0)+\text{noise}$. The first $\lfloor n/2 \rfloor$ components of the vector $\boldsymbol{x}^0$ are randomly sampled i.i.d. from the uniform distribution Unif$(-1,0)$, the next $\lfloor n/2 \rfloor - 1$ components are sampled from Unif$(0,10)$, and the last component is set to zero. The last component $\tilde{\xi}_\kappa \in \{-1,1\}$ of the vector $\tilde{\boldsymbol{\xi}}$ corresponds to the sensitive attribute. In the generated dataset, the first $\lceil m/2 \rceil$ data points are assigned with the sensitive attribute $\xi_\kappa=-1$, where their features $(\tilde{\xi}_j)_{j\in[\kappa-1]}$ are independently drawn from $\{\text{Unif}(0,j)\}_{j\in[\kappa-1]}$. The remaining data points are assigned $\xi_\kappa=1$, where their features $(\tilde{\xi}_j)_{j\in[\kappa-1]}$ are independently drawn from $\{\text{Unif}(0,j+2)\}_{j\in[\kappa-1]}$. The noise follows the uniform distribution Unif$(-0.1,0.1)\times\mathbb{E}[\tilde{\boldsymbol{\xi}}]^\top(\boldsymbol{x}^0)$.

In the first comparison, we generate data sets of a small population with sizes $m \in \{15,20,\ldots,100\}$ and feature dimension $\kappa=10$ to compare the exact formulations against the AM algorithm of DFSO and the two lower bounds. In the second comparison, we generate data sets of a large population with sizes $m\in\{100,200,\ldots,3{,}000\}$ to illustrate the solution quality of the AM algorithm and two lower bounds. We solve the Vanilla Formulation, the four exact MICP-R formulations, the AM algorithm in Section 3.5, the Jensen bound, and the Gelbrich bound in the first comparison. We test the AM algorithm, the Jensen bound, and the Gelbrich bound in the second comparison. Particularly, the AM algorithm of DFSO in Section 3.5 is initialized with the Gelbrich bound solution obtained by executing its corresponding AM algorithm described in Section 4.2.

In the first comparison, we report each instance's objective value, lower bound, optimality gap, and running time. Let "Obj.Val" denote the objective value and "LB" denote the lower bound. We use the dashed line "–" if "Obj.Val" is not available. The optimality gap denoted by "Gap" is computed by (UB-LB)/UB × 100%, where we use the optimal objective value as UB if available. We define the best upper bound as the smallest "Obj.Val" of the exact formulations and the AM algorithm of DFSO, and the best lower bound as the largest "LB" of the exact formulations and "Obj.Val" of Gelbrich bound. For some instances, "Obj.Val" may not be available for the exact formulations. In this case, we use the best upper bound to compute their optimality gaps, use the best lower bound to compute the AM's optimality gap, and use the AM's objective value to compute the Jensen bound's and Gelbrich bound's optimality gaps. The running time in seconds is denoted as "Time". We set the time limit to 3,600 seconds. In the second comparison, we plot the gaps between AM and the two lower bounds over 10 replications, where the gap is computed by

(UB-LB)/UB $\times$ 100%. We report the mean and standard deviation of the gaps and also illustrate the average running time of each method.

The first comparison results are displayed in Tables 2-5. The Vanilla Formulation cannot solve the small population instances within the time limit. The upper bounds of the Vanilla Formulation tend to be close to the optimal value, while the lower bounds are nearly zero. In fact, the gap of the Vanilla Formulation is 100% when the population size of the instance is $m \geq 30$. The Discretized Formulation can solve the instance with $m = 15$. The quality of the incumbent solution at the time limit then deteriorates rapidly as m increases. The Complementary Formulation performs similarly to the Discretized Formulation. Its upper bounds are often worse than other formulations, and the lower bounds are always zero for all the instances. This demonstrates the weakness of the Discretized Formulation and the Complementary Formulation, which is consistent with Proposition 10. The performances of the Quantile Formulation and the Aggregate Quantile Formulation in Table 3 are significantly better. The Quantile Formulation is able to solve instances up to $m \leq 40$ to optimality, and it returns nonzero lower bounds except for the last instance. The optimality gap of the Quantile Formulation becomes larger as m increases. Remarkably, the Aggregate Quantile Formulation can solve instances with $m \leq 60$ and $m = 70$ to optimality. The running time for each instance is less than 10 seconds when the population size is $m \leq 40$. The Aggregate Quantile Formulation cannot be solved optimally for larger instances; however, it still consistently provides high quality lower bounds with small gaps. We observe that the upper bounds of the Quantile Formulation and the Aggregate Quantile Formulation may not be available for instances with large m. This is potentially due to these two MICP-R formulations having many variables and constraints, which causes the solver to have difficulty finding a feasible solution for large instances. Therefore, we instead use the AM algorithm to solve instances for which the Aggregate Quantile Formulation cannot provide an optimal solution within the time limit.

In fact, as shown in Table 4, the AM algorithm provides very near-optimal solutions to instances of a small population using less than one second. It has a zero gap for most instances when the optimal solution is available, that is, $m \leq 60$ and $m = 70$. Its solution is close to the best lower bound when the optimal solution is unavailable, where the gap is less than 7%. In particular, the AM algorithm has better objective values than the Vanilla Formulation for all instances in this experiment. On the other hand, the Jensen bound has a gap of around 30% for each instance, and its running time is short due to the simplicity of its model formulation. On the contrary, the Gelbrich bound's gap decreases and running time slightly increases when the population size m increases. The gap of the Gelbrich bound is around 15% when the population size is $m \geq 50$. Since the number of features $\kappa = 10$ is small, the Gelbrich bound model can solve all instances to optimality, where each instance's running time is less than 3 seconds. Besides, we also compute the continuous relaxation values of the exact MICP formulations. In Table 5, the continuous relaxation values of the first three formulations are zero for most instances. The Aggregate Quantile Formulation always has a nonzero continuous relaxation value, and it is greater than the objective value of the Jensen bound as shown in Corollary 1. The continuous relaxation gap of the Aggregate Quantile Formulation is around 30% overall, which numerically verifies this formulation's strength.

The second comparison is presented in Figure 1. We see that both gaps stabilize when the population size is large enough. The gap between AM and the Jensen bound decreases from 40% to 21% when the population size m grows from 100 to 1,000. This gap is around 21% when the population size $m \geq 1{,}000$. The gap between AM and the Gelbrich bound drops from 10% to 1% when the population size m grows from 100 to 1,500. This gap decreases to 0.8% after $m = 1{,}500$. The small gap between AM and the Gelbrich bound verifies that the solution of AM is near optimal and demonstrates the strength of the Gelbrich bound. Meanwhile, the running time of these methods grows slowly. Since the Gelbrich bound formulation is nonconvex, it requires a longer time to solve large population instances. Figure 1(c) shows that AM is much faster than the Gelbrich bound, and the Jensen bound is slightly faster than AM. When $m = 3{,}000$, AM, the Jensen bound, and the Gelbrich bound take 15, 11, and 53 seconds on average, respectively. The stable and efficiently solvable lower bound solutions are useful to initialize AM and verify its solution quality.

The two comparisons in this experiment confirm the effectiveness of the proposed methods in solving DFSO. In practice, we suggest choosing the Aggregate Quantile Formulation to solve fair decision-making problems with a small population and switch to the AM method if the population size is large, where we can use the Jensen bound or the Gelbrich bound to initialize and establish the quality of the AM method.

Table 2 Results of Exact MICP Formulations

m	Vanilla Formulation				Discretized Formulation				Complementary Formulation			
	Obj.Val	LB	Gap (%)	Time	Obj.Val	LB	Gap (%)	Time	Obj.Val	LB	Gap (%)	Time
15	342.43	99.92	70.82	3600.00	342.43	342.40	0.01	339.36	342.44	0.00	100.00	3600.00
20	230.62	8.23	96.43	3600.00	230.62	184.95	19.80	3600.00	231.30	0.00	100.00	3600.00
25	136.81	0.62	99.55	3600.00	135.03	84.28	37.58	3600.00	140.37	0.00	100.00	3600.00
30	174.38	0.00	100.00	3600.00	172.21	52.11	69.74	3600.00	199.55	0.00	100.00	3600.00
35	136.94	0.00	100.00	3600.00	133.81	12.34	90.78	3600.00	219.28	0.00	100.00	3600.00
40	256.27	0.00	100.00	3600.00	257.70	53.99	79.05	3600.00	1249.10	0.00	100.00	3600.00
45	226.81	0.01	100.00	3600.00	228.73	20.61	90.99	3600.00	613.29	0.00	100.00	3600.00
50	170.45	0.00	100.00	3600.00	177.92	21.70	87.80	3600.00	708.46	0.00	100.00	3600.00
55	205.50	0.00	100.00	3600.00	230.96	11.95	94.83	3600.00	684.75	0.00	100.00	3600.00
60	134.96	0.00	100.00	3600.00	772.27	1.35	99.82	3600.00	1142.74	0.00	100.00	3600.00
65	150.43	0.00	100.00	3600.00	176.77	0.03	99.98	3600.00	658.52	0.00	100.00	3600.00
70	138.49	0.00	100.00	3600.00	144.42	0.00	100.00	3600.00	596.50	0.00	100.00	3600.00
75	140.58	0.00	100.00	3600.00	242.28	0.00	100.00	3600.00	1021.06	0.00	100.00	3600.00
80	169.10	0.00	100.00	3600.00	253.35	0.00	100.00	3600.00	793.09	0.00	100.00	3600.00
85	148.41	0.00	100.00	3600.00	320.75	0.00	100.00	3600.00	773.68	0.00	100.00	3600.00
90	174.45	0.00	100.00	3600.00	400.22	0.00	100.00	3600.00	835.96	0.00	100.00	3600.00
95	177.90	0.00	100.00	3600.00	587.42	0.00	100.00	3600.00	898.22	0.00	100.00	3600.00
100	161.34	0.00	100.00	3600.00	819.40	0.00	100.00	3600.00	727.02	0.00	100.00	3600.00

Table 3 Results of Exact MICP Formulations

m	Quantile Formulation				Aggregate Quantile Formulation			
	Obj.Val	LB	Gap (%)	Time	Obj.Val	LB	Gap (%)	Time
15	342.43	342.43	0.00	0.51	342.43	342.43	0.00	0.21
20	230.62	230.62	0.00	6.94	230.62	230.62	0.00	0.50
25	135.03	135.03	0.00	40.80	135.03	135.03	0.00	1.50
30	172.21	172.21	0.00	356.06	172.21	172.21	0.00	3.14
35	133.54	133.54	0.00	271.32	133.54	133.54	0.00	5.77
40	252.42	252.42	0.00	2379.63	252.42	252.42	0.00	8.61
45	219.17	177.31	19.10	3600.00	219.17	219.17	0.00	46.05
50	170.16	120.52	29.17	3600.00	169.99	169.99	0.00	169.07
55	204.76	137.83	32.69	3600.00	204.76	204.76	0.00	230.42
60	—	66.85	48.91	3600.00	130.84	130.84	0.00	1022.89
65	—	47.18	66.90	3600.00	142.54	142.27	0.19	3600.00
70	—	30.48	77.57	3600.00	135.92	135.91	0.01	3134.58
75	—	31.39	77.24	3600.00	—	128.64	6.72	3600.00
80	—	25.32	84.03	3600.00	—	154.58	2.51	3600.00
85	—	27.86	80.89	3600.00	157.36	143.52	8.79	3600.00
90	—	26.65	84.47	3600.00	—	165.23	3.71	3600.00
95	—	4.57	97.37	3600.00	—	162.50	6.51	3600.00
100	—	0.00	100.00	3600.00	—	124.59	12.17	3600.00

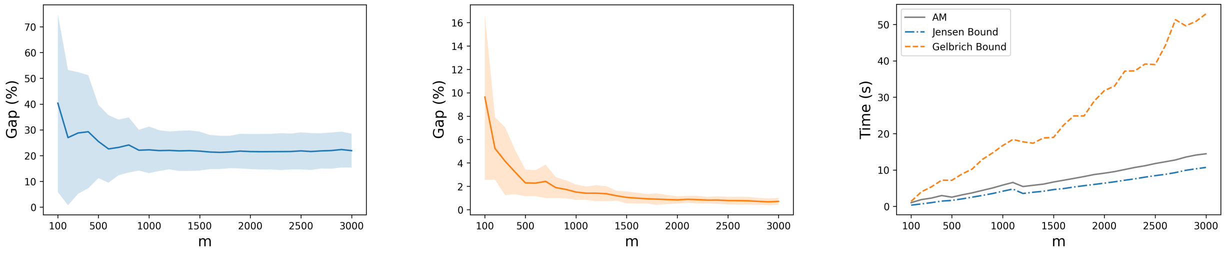
Table 4 Results of AM Algorithm of DFSO, Jensen Bound and Gelbrich Bound

m	AM			Jensen Bound			Gelbrich Bound		
	Obj.Val	Gap (%)	Time	Obj.Val	Gap (%)	Time	Obj.Val	Gap (%)	Time
15	342.43	0.00	0.17	207.12	39.51	0.06	222.14	35.13	0.39
20	230.62	0.00	0.26	89.13	61.35	0.08	105.91	54.07	0.41
25	135.03	0.00	0.37	46.03	65.91	0.09	50.41	62.67	0.48
30	172.21	0.00	0.30	109.56	36.38	0.10	110.85	35.63	0.48
35	133.54	0.00	0.35	92.95	30.39	0.11	93.07	30.31	0.50
40	252.42	0.00	0.23	187.66	25.65	0.13	194.65	22.89	0.54
45	219.17	0.00	0.63	138.17	36.96	0.14	176.41	19.51	0.59
50	170.17	0.10	0.52	116.18	31.65	0.16	143.05	15.85	0.72
55	204.76	0.00	0.62	141.53	30.88	0.16	178.35	12.90	0.65
60	130.84	0.00	0.46	69.87	46.60	0.18	105.48	19.38	1.18
65	142.54	0.00	0.90	83.21	41.63	0.19	118.92	16.57	1.40
70	135.92	0.00	0.51	83.84	38.31	0.20	119.37	12.18	0.81
75	137.91	6.72	0.68	85.22	38.21	0.22	119.79	13.14	0.89
80	158.57	2.51	0.73	112.03	29.35	0.24	145.23	8.41	0.96
85	145.77	1.54	0.81	104.78	28.12	0.24	134.34	7.84	0.99
90	171.59	3.71	0.80	126.73	26.15	0.29	160.95	6.20	1.08
95	173.82	5.36	0.91	132.55	23.74	0.31	164.49	5.36	2.00
100	141.85	6.59	0.92	91.76	35.32	0.32	132.50	6.59	2.27

Table 5 Results of Continuous Relaxation Values of Exact MICP Formulations

m	Discretized Formulation			Complementary Formulation			Quantile Formulation			Aggregate Quantile Formulation		
	Obj.Val	Gap (%)	Time	Obj.Val	Gap (%)	Time	Obj.Val	Gap (%)	Time	Obj.Val	Gap (%)	Time
15	0.20	99.94	0.24	0.00	100.00	0.38	0.34	99.90	0.23	282.42	17.52	0.16
20	0.00	100.00	0.41	0.00	100.00	0.60	0.00	100.00	0.63	147.18	36.18	0.64
25	0.00	100.00	0.50	0.00	100.00	1.00	0.38	99.72	0.60	57.40	57.49	0.63
30	0.00	100.00	0.77	0.00	100.00	1.05	0.00	100.00	0.49	111.36	35.33	1.07
35	0.00	100.00	1.09	0.00	100.00	5.64	0.51	99.62	0.62	98.71	26.08	0.67
40	0.00	100.00	1.36	0.00	100.00	17.50	0.00	100.00	0.72	219.32	13.11	0.82
45	0.00	100.00	1.73	0.00	100.00	37.80	0.96	99.56	1.22	179.74	17.99	1.03
50	0.00	100.00	2.11	0.00	100.00	5.01	0.00	100.00	1.09	134.92	20.63	1.25
55	0.00	100.00	2.63	0.00	100.00	5.76	0.65	99.68	1.81	167.17	18.36	1.51
60	0.00	100.00	3.08	0.00	100.00	6.60	0.00	100.00	1.54	92.30	29.45	1.77
65	0.00	100.00	3.68	0.00	100.00	8.72	0.44	99.69	2.00	101.41	28.85	2.24
70	0.00	100.00	4.76	0.00	100.00	10.27	0.00	100.00	2.08	103.16	24.10	2.43
75	0.00	100.00	5.46	0.00	100.00	13.96	0.32	99.77	2.71	99.28	28.01	3.09
80	0.00	100.00	6.19	0.00	100.00	11.26	0.00	100.00	2.79	120.75	23.85	3.40
85	0.00	100.00	7.25	0.00	100.00	13.57	0.26	99.82	3.49	115.93	20.47	4.08
90	0.00	100.00	8.06	0.00	100.00	14.48	0.00	100.00	4.12	136.94	20.19	4.65
95	0.00	100.00	9.27	0.00	100.00	16.51	0.21	99.88	4.99	140.77	19.01	5.48
100	0.00	100.00	10.59	0.00	100.00	18.72	0.00	100.00	4.35	102.23	27.93	5.66

5.1.2. Comparison with state-of-the-art In the second experiment, we compare our methods against two fair regression methods from the literature using real data. In this experiment, the cost function is set to the mean squared error (MSE). We solve DFSO using its AM algorithm in Section 3.5, solve the Jensen bound by solving (18), and solve the Gelbrich bound using its AM algorithm in Section 4.2. The first approach that we compare is Berk (Berk et al. 2017). Their work proposed a convex optimization method to incorporate group and individual fairness for fair regression. We compare with their group fairness model for demographic parity. The second approach



(a) Gap between AM and the Jensen Bound

(b) Gap between AM and the Gelbrich Bound

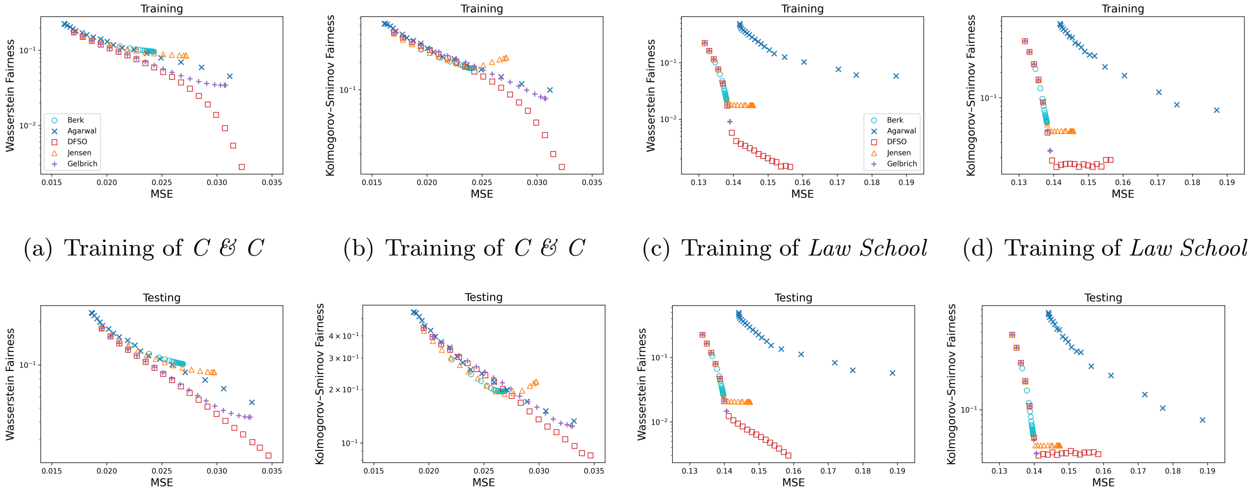
(c) Running time

Figure 1 Gap between AM and the two lower bounds for a large population size. The mean and standard deviation of the gap over 10 replications, as well as the average running time of each method, are illustrated.

is Agarwal (Agarwal et al. 2019), a reduction-based fair regression algorithm that uses the Kolmogorov–Smirnov distance to measure demographic parity. We test the performance of different approaches using criminological and educational datasets. The *Communities and Crime* dataset contains socio-economic, law enforcement, and crime data of different communities in the US. The goal is to predict the number of violent crimes per 100,000 of the population with race (black versus non-black) as the sensitive attribute. It contains 1,994 samples characterized by 127 features. We create the sensitive attribute by thresholding the percentage of the black population following Calders et al. (2013). The *Law School* (Wightman 1998) dataset consists of student records from the Law School Admission Council (LSAC) National Longitudinal Bar Passage Study. The goal is to predict a student's GPA with race (white versus non-white) as the sensitive attribute. The dataset contains 20,649 samples characterized by 12 features.

For both datasets, we split the data into 70% for training and 30% for testing. We repeat this procedure 10 times and report the average performance. We evaluate the different methods based on the trade-off between MSE and fairness scores of Wasserstein fairness and Kolmogorov–Smirnov fairness measures, where we compute $\mathrm{WD}_2(\boldsymbol{x})$ using (10) and $\mathrm{KSD}(\boldsymbol{x})$ using (6) for each method. Note that we only use DFSO to solve the Wasserstein fairness measure and plug in its solution to compute the Kolmogorov–Smirnov fairness measure. The hyperparameters of each method are chosen as follows. For DFSO, the Jensen bound, and the Gelbrich bound, we set the inefficiency level parameter $\epsilon \in \{0.05, 0.1, \ldots, 1\}$ for the *Communities and Crime* dataset and $\epsilon \in \{0.01, 0.02, \ldots, 0.2\}$ for the *Law School* dataset. For Agarwal (Agarwal et al. 2019), we set their unfairness level parameter $\epsilon \in \{0.015, 0.03, \ldots, 0.3\}$ for *Communities and Crime* and $\epsilon \in \{0.035, 0.07, \ldots, 0.7\}$ for *Law School*. For Berk (Berk et al. 2017), we set their unfairness penalty parameter $\lambda \in \{0.5, 1, \ldots, 10\}$ for *Communities and Crime* and $\lambda \in \{0.2, 0.4, \ldots, 4\}$ for *Law School*.

Figure 2 presents the trade-off between fairness scores and MSE on the two datasets described above. We observe that DFSO consistently outperforms other methods in training and testing in view of the Wasserstein fairness measure, where it can reduce the unfairness level to significantly small values (i.e., nearly zero) with a relatively small increase in MSE. DFSO also performs well in terms of the Kolmogorov–Smirnov fairness measure and attains small fairness scores. Two lower bounds (i.e., the Jensen and Gelbrich bounds) also provide good solution quality for both Wasserstein and Kolmogorov–Smirnov fairness measures. Notably, the Jensen bound has similar solutions as Berk (Berk et al. 2017), and the Gelbrich bound is competitive with Agarwal (Agarwal et al. 2019). We observe that the Gelbrich bound tends to be fairer than the Jensen bound. DFSO consistently provides the best Wasserstein fairness score for the *Communities and Crime* and *Law School* datasets. In terms of the Kolmogorov–Smirnov fairness measure, Berk (Berk et al. 2017) and the Jensen bound is effective when MSE is small. Note that they cannot further improve fairness given large inefficiency level parameters or allow large MSE in the experiment. On the contrary, DFSO and the Gelbrich bound have the capacity to improve Kolmogorov–Smirnov fairness with large MSE. It is evident that DFSO is capable of effectively addressing the unfairness issues in fair regression problems.



(a) Training of *C & C* (b) Training of *C & C* (c) Training of *Law School* (d) Training of *Law School*

(e) Testing of *C & C* (f) Testing of *C & C* (g) Testing of *Law School* (h) Testing of *Law School*

Figure 2 Fairness vs MSE for Fair Regression. The Wasserstein fairness versus MSE are shown in (a), (c), (e), (g), and Kolmogorov–Smirnov fairness versus MSE are shown in (b), (d), (f), (h). All the training and testing results are averaged over 10 replications.

5.2. Fair Allocation of Scarce Medical Resources

During public health emergencies such as the influenza pandemic and COVID-19, optimal allocation of scarce medical resources (e.g., therapeutics and vaccines) is a crucial yet challenging task (see, e.g., Sun et al. 2023, Shehadeh and Snyder 2023). DFSO can be adapted to allocate scarce medical resources in a distributionally fair way. In this experiment, we study the fair allocation of COVID-19 vaccine across m counties in Georgia. Given the total amount of vaccines $T \in \mathbb{Z}_+$, counties' population sizes $\boldsymbol{p} \in \mathbb{Z}_+^m$, and thresholds $\boldsymbol{l} \leq \boldsymbol{u} \in (0,1]^m$, we consider a vaccine allocation problem

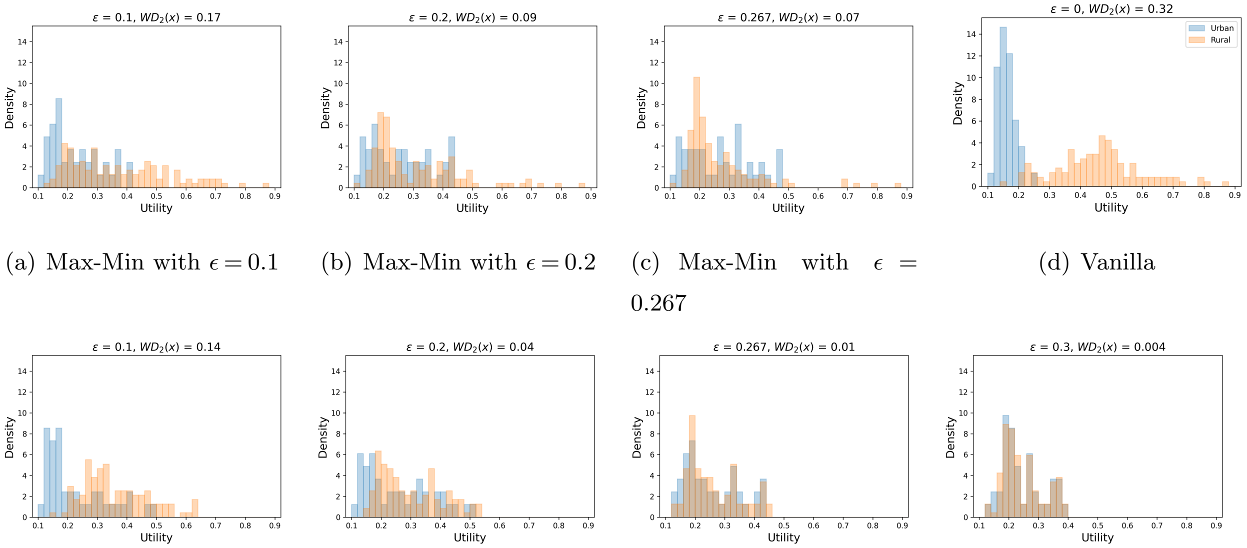
$$V^* = \max_{\boldsymbol{x}} \left\{ \sqrt[m]{\prod_{i\in[m]} x_i} : \sum_{i\in[m]} p_i x_i \leq T, l_i \leq x_i \leq u_i, \forall i \in [m] \right\}, \tag{24}$$

where the coverage rate $x_i \in [0,1]$ is defined as the ratio of the number of allocated vaccines to the population size in each county $i \in [m]$, and the benefit function is $Q(\boldsymbol{x}, \boldsymbol{\xi}_i) = x_i$. We remark that the efficiency function in (24) follows the conventional proportional fairness, which seeks to maximize the product of each individual county's utility. The fair allocation approach that we are comparing is the max-min fairness at the group level, defined as

$$\max_{\boldsymbol{x}} \min_{a\in A} \left\{ \sqrt[m_a]{\prod_{i\in[m_a]} x_i} : \sqrt[m]{\prod_{i\in[m]} x_i} \geq (1-\epsilon)V^*, \sum_{i\in[m]} p_i x_i \leq T, l_i \leq x_i \leq u_i, \forall i \in [m] \right\}. \tag{25}$$

In the experiment, we compare DFSO against the Max-Min formulation (25) using Georgia (GA) population data from the U.S. Census Bureau. We choose the utility function to be $f(\boldsymbol{x}, \boldsymbol{\xi}_i) = x_i$ for each county $i \in [m]$. The dataset includes each county's population size p_i and the size of the population aged 65 years and over, denoted by s_i. We let $A = \{\text{urban, rural}\}$ in order to study the fairness among urban counties (where the population size is at least 50,000) and rural counties (where the population size is less than 50,000). We assume the total amount of vaccine is 20% of the total population. That is, we have $T = 0.2\sum_{i\in[m]} p_i$. Since older people are more vulnerable to COVID-19, we select the minimum and maximum vaccine coverage rates as $l_i = 0.8 s_i T / \sum_{i\in[m]} s_i$ and $u_i = 2 s_i T / \sum_{i\in[m]} s_i$, respectively. We also set the inefficiency level parameter to $\epsilon = \{0.1, 0.2, 0.267, 0.3\}$. We choose type $q = 2$ Wasserstein fairness and solve DFSO using its AM algorithm in Section 3.5. Since the solution of Max-Min (25) remains unchanged when $\epsilon \geq 0.267$, we only display its results for $\epsilon = \{0.1, 0.2, 0.267\}$.

Figure 3 shows the histograms of utility for fair allocation of COVID-19 vaccine in GA. It can be observed that both methods can reduce the disparities of utilities among urban and rural counties, while DFSO always has a smaller Wasserstein fairness score than Max-Min (25) given the same inefficiency level. The solution of Max-Min (25) remains unchanged when $\epsilon \geq 0.267$, thus Figure 3(c) shows the fairest solution that Max-Min (25) can provide. DFSO can achieve a Wasserstein fairness score that is nearly zero, effectively resolving the distributional disparities. We observe that Max-Min (25) is not sufficient to eliminate the disparity between two distributions compared to its counterpart DFSO. This demonstrates that the proposed DFSO can effectively address distributional fairness while achieving relatively high efficiency.



(a) Max-Min with $\epsilon = 0.1$ (b) Max-Min with $\epsilon = 0.2$ (c) Max-Min with $\epsilon = 0.267$ (d) Vanilla

(e) DFSO with $\epsilon = 0.1$ (f) DFSO with $\epsilon = 0.2$ (g) DFSO with $\epsilon = 0.267$ (h) DFSO with $\epsilon = 0.3$

Figure 3 Histograms of Utility for Fair Allocation of COVID-19 Vaccine in GA

6. Conclusion

This paper studies Distributionally Fair Stochastic Optimization (DFSO), where we employ the Wasserstein distance to measure group fairness. We propose exact mixed-integer convex programming formulations for DFSO. By exploring the properties of the Wasserstein fairness measure, we develop an efficient alternating minimization (AM) solution method and two strong lower bounds. Our numerical study shows that the proposed exact methods can solve medium-sized fair learning problems efficiently, while the proposed AM method and lower bounds work efficiently for large-scale fair optimization and learning problems. The convergence rate and solution quality of AM methods are interesting open questions. Stronger lower bounds for the general Wasserstein fairness measure are also interesting to explore in the future. Another future study is properly incorporating distributional robustness into DFSO when the individual data are noise-related.

References

Agarwal A, Dudík M, Wu ZS (2019) Fair regression: Quantitative definitions and reduction-based algorithms. *International Conference on Machine Learning*, 120–129 (PMLR).

Aghaei S, Azizi MJ, Vayanos P (2019) Learning optimal and fair decision trees for non-discriminative decision-making. *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 33, 1418–1426.

Ahmed S, Xie W (2018) Relaxations and approximations of chance constraints under finite distributions. *Mathematical Programming* 170:43–65.

Barocas S, Selbst AD (2016) Big data's disparate impact. *Calif. L. Rev.* 104:671.

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

Blanchet J, Murthy K (2019) Quantifying distributional model risk via optimal transport. *Mathematics of Operations Research* 44(2):565–600.

Calders T, Karim A, Kamiran F, Ali W, Zhang X (2013) Controlling attribute effect in linear regression. *2013 IEEE 13th International Conference on Data Mining*, 71–80 (IEEE).

Caton S, Haas C (2020) Fairness in machine learning: A survey. *ACM Computing Surveys* .

Charnes A, Cooper WW (1959) Chance-constrained programming. *Management science* 6(1):73–79.

Chen Z, Kuhn D, Wiesemann W (2022) Data-driven chance constrained programs over Wasserstein balls. *Operations Research* .

Chzhen E, Denis C, Hebiri M, Oneto L, Pontil M (2020) Fair regression with wasserstein barycenters. *Advances in Neural Information Processing Systems* 33:7321–7331.

Cohen MC, Elmachtoub AN, Lei X (2022) Price discrimination with fairness constraints. *Management Science* 68(12):8536–8552.

Donini M, Oneto L, Ben-David S, Shawe-Taylor JS, Pontil M (2018) Empirical risk minimization under fairness constraints. *Advances in Neural Information Processing Systems*, 2791–2801.

Feldman M, Friedler SA, Moeller J, Scheidegger C, Venkatasubramanian S (2015) Certifying and removing disparate impact. *Proceedings of the 21th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, 259–268.

Fournier N, Guillin A (2015) On the rate of convergence in Wasserstein distance of the empirical measure. *Probability Theory and Related Fields* 162(3-4):707–738.

Gao R, Kleywegt A (2023) Distributionally robust stochastic optimization with Wasserstein distance. *Mathematics of Operations Research* 48(2):603–655.

Gelbrich M (1990) On a formula for the L2 Wasserstein metric between measures on euclidean and hilbert spaces. *Mathematische Nachrichten* 147(1):185–203.

Hanasusanto GA, Kuhn D (2018) Conic programming reformulations of two-stage distributionally robust linear programs over Wasserstein balls. *Operations Research* 66(3):849–869.

Hardt M, Price E, Srebro N (2016) Equality of opportunity in supervised learning. *Advances in Neural Information Processing Systems* 29.

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

Kamishima T, Akaho S, Asoh H, Sakuma J (2012) Fairness-aware classifier with prejudice remover regularizer. *Joint European Conference on Machine Learning and Knowledge Discovery in Databases*, 35–50 (Springer).

Karsu Ö, Morton A (2015) Inequity averse optimization in operational research. *European Journal of Operational Research* 245(2):343–359.

Kliegr T (2009) Uta-nm: Explaining stated preferences with additive non-monotonic utility functions. *Preference Learning* 56.

Kuhn D, Esfahani PM, Nguyen VA, Shafieezadeh-Abadeh S (2019) Wasserstein distributionally robust optimization: Theory and applications in machine learning. *Operations Research & Management Science in the Age of Analytics*, 130–166 (Informs).

Lowy A, Baharlouei S, Pavan R, Razaviyayn M, Beirami A (2021) A stochastic optimization framework for fair risk minimization. *arXiv preprint arXiv:2102.12586* .

Lubin M, Vielma JP, Zadik I (2022) Mixed-integer convex representability. *Mathematics of Operations Research* 47(1):720–749.

McCormick GP (1976) Computability of global solutions to factorable nonconvex programs: Part i—convex underestimating problems. *Mathematical Programming* 10(1):147–175.

Mohajerin Esfahani P, Kuhn D (2018) Data-driven distributionally robust optimization using the Wasserstein metric: performance guarantees and tractable reformulations. *Mathematical Programming* 171(1-2):115–166.

MOSEK ApS (2019) Mosek optimization suite. `https://docs.mosek.com/modeling-cookbook/index.html`, accessed: 01-27-2024.

Ogryczak W, Luss H, Pióro M, Nace D, Tomaszewski A (2014) Fair optimization and networks: A survey. *Journal of Applied Mathematics* 2014.

Patel D, Khan A, Louis A (2020) Group fairness for knapsack problems. *arXiv preprint arXiv:2006.07832* .

Rahimian H, Mehrotra S (2019) Distributionally robust optimization: A review. *arXiv preprint arXiv:1908.05659* .

Rebman KR (1974) Total unimodularity and the transportation problem: a generalization. *Linear Algebra and its Applications* 8(1):11–24.

Ross N (2011) Fundamentals of stein's method. *Probability Surveys* 8:210–293.

Rychener Y, Taskesen B, Kuhn D (2022) Metrizing fairness. *arXiv preprint arXiv:2205.15049* .

Samorani M, Harris SL, Blount LG, Lu H, Santoro MA (2022) Overbooked and overlooked: machine learning and racial bias in medical appointment scheduling. *Manufacturing & Service Operations Management* 24(6):2825–2842.

Santambrogio F (2015) Optimal transport for applied mathematicians. *Birkäuser, NY* 55(58-63):94.

Shehadeh KS, Snyder LV (2023) Equity in stochastic healthcare facility location. *Uncertainty in Facility Location Problems*, 303–334 (Springer).

Sun L, Xie W, Witten T (2023) Distributionally robust fair transit resource allocation during a pandemic. *Transportation Science* 57(4):954–978.

Taskesen B, Nguyen VA, Kuhn D, Blanchet J (2020) A distributionally robust approach to fair classification. *arXiv preprint arXiv:2007.09530* .

Vershynin R (2018) *High-dimensional probability: An introduction with applications in data science*, volume 47 (Cambridge University Press).

Wang Y, Nguyen VA, Hanasusanto GA (2021) Wasserstein robust classification with fairness constraints. *arXiv preprint arXiv:2103.06828* .

Wightman LF (1998) Lsac national longitudinal bar passage study. *LSAC Research Report Series* (ERIC).

Xie W (2021) On distributionally robust chance constrained programs with Wasserstein distance. *Mathematical Programming* 186(1-2):115–155.

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

Zafar MB, Valera I, Rogriguez MG, Gummadi KP (2017) Fairness constraints: Mechanisms for fair classification. *Artificial Intelligence and Statistics*, 962–970 (PMLR).

Appendix A. Two Additional Exact Formulations and An Equivalent MICP-R Formulation for KSD($\boldsymbol{x}$)

A.1 Discretized Formulation: Discretizing the Transportation Decisions

In this subsection, we develop an MICP-R formulation for the Wasserstein fairness measure set $\mathcal{F}_q$ by observing that the balanced transportation polytope can be integral given that the supplies and demands are both integers. To this end, we recast the set as

$$\mathcal{F}_q = \left\{ (\boldsymbol{x}, \nu) \in \mathcal{X} \times \mathbb{R}_+ : \min_{\boldsymbol{\pi}_{a\bar{a}} \in \Pi_{a\bar{a}}} \left\{ \sum_{i \in C_a} \sum_{j \in C_{\bar{a}}} \pi_{ija\bar{a}} \left| f(\boldsymbol{x}, \boldsymbol{\xi}_i) - f(\boldsymbol{x}, \boldsymbol{\xi}_j) \right|^q \right\} \le \nu, \forall a < \bar{a} \in A \right\} \tag{26}$$

where for each $a < \bar{a} \in A$, the transportation feasible set is given by

$$\Pi_{a\bar{a}} = \left\{ \boldsymbol{\pi}_{a\bar{a}} \in \mathbb{R}_+^{m_a \times m_{\bar{a}}} : \sum_{i \in C_a} \pi_{ija\bar{a}} = \frac{1}{m_{\bar{a}}}, \forall j \in C_{\bar{a}}, \sum_{j \in C_{\bar{a}}} \pi_{ija\bar{a}} = \frac{1}{m_a}, \forall i \in C_a \right\}.$$

Observe that, for each $a < \bar{a} \in A$, the constraint in (26) is satisfied if and only if there exists $\boldsymbol{\pi}_{a\bar{a}} \in \Pi_{a\bar{a}}$ such that $\sum_{i \in C_a} \sum_{j \in C_{\bar{a}}} \pi_{ija\bar{a}} \left| f(\boldsymbol{x}, \boldsymbol{\xi}_i) - f(\boldsymbol{x}, \boldsymbol{\xi}_j) \right|^q \le \nu$. Hence, the resulting set has nonconvex terms in $\boldsymbol{\pi}_{a\bar{a}}$ and $\boldsymbol{x}$ that complicate the formulation.

Theorem 10. ***(Discretized Formulation)*** *Suppose that the set $X_i = \{(\boldsymbol{x}, \bar{w}_i) \in \mathcal{X} \times \mathbb{R} : f(\boldsymbol{x}, \boldsymbol{\xi}_i) = \bar{w}_i\}$ is MICP-R for each $i \in [m]$ and $M_i \ge \max_{\boldsymbol{x} \in \mathcal{X}, (7\mathrm{b})} |f(\boldsymbol{x}, \boldsymbol{\xi}_i)|$ for each $i \in [m]$. Then $\mathcal{F}_q$ is equivalent to the MICP-R set*

$$\mathcal{F}_q = \left\{ (\boldsymbol{x}, \nu) \in \mathcal{X} \times \mathbb{R}_+ : \begin{array}{l} \displaystyle\sum_{i \in C_a} \sum_{j \in C_{\bar{a}}} \sum_{k \in [\bar{\Omega}_{a\bar{a}}]} \frac{2^{k-1} \widehat{w}_{ijka\bar{a}}^q}{m_a m_{\bar{a}}} \le \nu, \boldsymbol{z}_{a\bar{a}} \in \Gamma_{a\bar{a}}, \forall a < \bar{a} \in A, \\ (\boldsymbol{x}, \bar{w}_i) \in X_i, \forall i \in [m], |\bar{z}_{ijka\bar{a}1} - \bar{z}_{ijka\bar{a}2}| \le \widehat{w}_{ijka\bar{a}}, \\ (\bar{z}_{ijka\bar{a}1}, z_{ijka\bar{a}}, \bar{w}_i) \in \mathrm{MC}(0, 1, -M_i, M_i), \\ (\bar{z}_{ijka\bar{a}2}, z_{ijka\bar{a}}, \bar{w}_j) \in \mathrm{MC}(0, 1, -M_j, M_j), \\ \forall i \in C_a, j \in C_{\bar{a}}, k \in [\bar{\Omega}_{a\bar{a}}], a < \bar{a} \in A \end{array} \right\}, \tag{27}$$

where for each $a < \bar{a} \in A$, we define $\bar{\Omega}_{a\bar{a}} = \lceil \log_2 (\min\{m_a, m_{\bar{a}}\}) \rceil + 1$ and

$$\Gamma_{a\bar{a}} = \left\{ \boldsymbol{z}_{a\bar{a}} \in \{0,1\}^{m_a \times m_{\bar{a}} \times \bar{\Omega}_{a\bar{a}}} : \begin{array}{l} \displaystyle\sum_{i \in C_a} \sum_{k \in [\bar{\Omega}_{a\bar{a}}]} 2^{k-1} z_{ijka\bar{a}} = m_a, \forall j \in C_{\bar{a}}, \\ \displaystyle\sum_{j \in C_{\bar{a}}} \sum_{k \in [\bar{\Omega}_{a\bar{a}}]} 2^{k-1} z_{ijka\bar{a}} = m_{\bar{a}}, \forall i \in C_a \end{array} \right\}.$$

Proof. Letting $\bar{\pi}_{ija\bar{a}} = \pi_{ija\bar{a}} m_a m_{\bar{a}}$, the set $\mathcal{F}_q$ in (26) can be reformulated as

$$\mathcal{F}_q = \left\{ (\boldsymbol{x}, \nu) \in \mathcal{X} \times \mathbb{R}_+ : \min_{\bar{\boldsymbol{\pi}}_{a\bar{a}} \in \bar{\Pi}_{a\bar{a}}} \left\{ \sum_{i \in C_a} \sum_{j \in C_{\bar{a}}} \frac{\bar{\pi}_{ija\bar{a}}}{m_a m_{\bar{a}}} \left| f(\boldsymbol{x}, \boldsymbol{\xi}_i) - f(\boldsymbol{x}, \boldsymbol{\xi}_j) \right|^q \right\} \le \nu, \forall a < \bar{a} \in A \right\},$$

where for each $a < \bar{a} \in A$, we have

$$\bar{\Pi}_{a\bar{a}} = \left\{ \bar{\boldsymbol{\pi}}_{a\bar{a}} \in \mathbb{R}_+^{m_a \times m_{\bar{a}}} : \sum_{i \in C_a} \bar{\pi}_{ija\bar{a}} = m_a, \forall j \in C_{\bar{a}}, \sum_{j \in C_{\bar{a}}} \bar{\pi}_{ija\bar{a}} = m_{\bar{a}}, \forall i \in C_a \right\}.$$

We see that there exist integer solutions $\bar{\boldsymbol{\pi}}_{a\bar{a}}$ for this transportation problem according to Rebman (1974). Then, we obtain

$$\mathcal{F}_q = \left\{ (\boldsymbol{x}, \nu) \in \mathcal{X} \times \mathbb{R}_+ : \min_{\bar{\boldsymbol{\pi}}_{a\bar{a}} \in \bar{\Pi}_{a\bar{a}} \cap \mathbb{Z}_+^{m_a \times m_{\bar{a}}}} \left\{ \sum_{i \in C_a} \sum_{j \in C_{\bar{a}}} \frac{\bar{\pi}_{ija\bar{a}}}{m_a m_{\bar{a}}} \left| f(\boldsymbol{x}, \boldsymbol{\xi}_i) - f(\boldsymbol{x}, \boldsymbol{\xi}_j) \right|^q \right\} \leq \nu, \forall a < \bar{a} \in A \right\},$$

Since there exists $\bar{\boldsymbol{\pi}}_{a\bar{a}} \in \bar{\Pi}_{a\bar{a}} \cap \mathbb{Z}_+^{m_a \times m_{\bar{a}}}$ for $\mathcal{F}_q$, we have

$$\mathcal{F}_q = \left\{ (\boldsymbol{x}, \nu) \in \mathcal{X} \times \mathbb{R}_+ : \exists \bar{\boldsymbol{\pi}}_{a\bar{a}} \in \bar{\Pi}_{a\bar{a}} \cap \mathbb{Z}_+^{m_a \times m_{\bar{a}}}, \sum_{i \in C_a} \sum_{j \in C_{\bar{a}}} \frac{\bar{\pi}_{ija\bar{a}}}{m_a m_{\bar{a}}} \left| f(\boldsymbol{x}, \boldsymbol{\xi}_i) - f(\boldsymbol{x}, \boldsymbol{\xi}_j) \right|^q \leq \nu, \forall a < \bar{a} \in A \right\},$$

Next, we binarize the integer matrix variables $\bar{\boldsymbol{\pi}}_{a\bar{a}}$ using the expansion

$$\bar{\pi}_{ija\bar{a}} = \sum_{k \in [\bar{\Omega}_{a\bar{a}}]} 2^{k-1} z_{ijka\bar{a}}$$

where $\bar{\Omega}_{a\bar{a}} = \lceil \log_2 (\min\{m_a, m_{\bar{a}}\}) \rceil + 1$ and $z_{ijka\bar{a}} \in \{0,1\}$ for all $i \in C_a$, $j \in C_{\bar{a}}$, and $k \in [\bar{\Omega}_{a\bar{a}}]$.

We thus obtain

$$\mathcal{F}_q = \left\{ (\boldsymbol{x}, \nu) \in \mathcal{X} \times \mathbb{R}_+ : \sum_{i \in C_a} \sum_{j \in C_{\bar{a}}} \sum_{k \in [\bar{\Omega}_{a\bar{a}}]} \frac{2^{k-1} z_{ijka\bar{a}}}{m_a m_{\bar{a}}} \left| f(\boldsymbol{x}, \boldsymbol{\xi}_i) - f(\boldsymbol{x}, \boldsymbol{\xi}_j) \right|^q \leq \nu, \boldsymbol{z}_{a\bar{a}} \in \Gamma_{a\bar{a}}, \forall a < \bar{a} \in A \right\},$$

where for each $a < \bar{a} \in A$, we have

$$\Gamma_{a\bar{a}} = \left\{ \boldsymbol{z}_{a\bar{a}} \in \{0,1\}^{m_a \times m_{\bar{a}} \times \bar{\Omega}_{a\bar{a}}} : \begin{array}{l} \sum_{i \in C_a} \sum_{k \in [\bar{\Omega}_{a\bar{a}}]} 2^{k-1} z_{ijka\bar{a}} = m_a, \forall j \in C_{\bar{a}}, \\ \sum_{j \in C_{\bar{a}}} \sum_{k \in [\bar{\Omega}_{a\bar{a}}]} 2^{k-1} z_{ijka\bar{a}} = m_{\bar{a}}, \forall i \in C_a \end{array} \right\}.$$

Then, letting $\bar{w}_i = f(\boldsymbol{x}, \boldsymbol{\xi}_i)$ and $|z_{ijka\bar{a}} \bar{w}_i - z_{ijka\bar{a}} \bar{w}_j|^q \leq \widehat{w}_{ijka\bar{a}}$ can further linearize the set $\mathcal{F}_q$ as follows

$$\mathcal{F}_q = \left\{ (\boldsymbol{x}, \nu) \in \mathcal{X} \times \mathbb{R}_+ : \begin{array}{l} \sum_{i \in C_a} \sum_{j \in C_{\bar{a}}} \sum_{k \in [\bar{\Omega}_{a\bar{a}}]} \frac{2^{k-1} \widehat{w}_{ijka\bar{a}}^q}{m_a m_{\bar{a}}} \leq \nu, \forall a < \bar{a} \in A, \\ |z_{ijka\bar{a}} \bar{w}_i - z_{ijka\bar{a}} \bar{w}_j| \leq \widehat{w}_{ijka\bar{a}}, \forall i \in C_a, j \in C_{\bar{a}}, k \in [\bar{\Omega}_{a\bar{a}}], a < \bar{a} \in A, \\ \boldsymbol{z}_{a\bar{a}} \in \Gamma_{a\bar{a}}, \forall a < \bar{a} \in A, (\boldsymbol{x}, \bar{w}_i) \in X_i, \forall i \in [m] \end{array} \right\}.$$

The conclusion follows from using McCormick representation of bilinear terms $\{z_{ijka\bar{a}} \bar{w}_i\}_{i \in C_a, j \in C_{\bar{a}}, k \in [\bar{\Omega}_{a\bar{a}}], a < \bar{a} \in A}$ and $\{z_{ijka\bar{a}} \bar{w}_j\}_{i \in C_a, j \in C_{\bar{a}}, k \in [\bar{\Omega}_{a\bar{a}}], a < \bar{a} \in A}$, and invoking the definition of the sets $\{X_i\}_{i \in [m]}$. □

Note that the support size of $\bar{\boldsymbol{\pi}}_{a\bar{a}}$ is $m_a+m_{\bar{a}}$. This motivates us to introduce new binary variables $\widehat{\boldsymbol{z}}_{a\bar{a}}$ such that $z_{ijka\bar{a}} \leq \widehat{z}_{ija\bar{a}}$ for each $i\in C_a, j\in C_{\bar{a}}, k\in[\bar{\Omega}_{a\bar{a}}], a<\bar{a}\in A$ and obtain the following inequalities valid for the set $\mathcal{F}_q$ as

$$\sum_{i\in C_a}\sum_{j\in C_{\bar{a}}} \widehat{z}_{ija\bar{a}} \leq m_a + m_{\bar{a}},$$

for all $a<\bar{a}\in A$.

A.2 Complementary Formulation: Linearizing the Complementary Slackness Constraints

In this subsection, we propose the second formulation of the set $\mathcal{F}_q$ using linear programming complementary slackness. According to the definition of the sets $\{X_i\}_{i\in[m]}$, we can represent the set $\mathcal{F}_q$ in (26) as

$$\mathcal{F}_q = \left\{ (\boldsymbol{x},\nu)\in\mathcal{X}\times\mathbb{R}_+ : \begin{array}{l} \min\limits_{\boldsymbol{\pi}_{a\bar{a}}\in\Pi_{a\bar{a}}} \left\{ \sum\limits_{i\in C_a}\sum\limits_{j\in C_{\bar{a}}} \pi_{ija\bar{a}} w_{ij} \right\} \leq \nu, \forall a<\bar{a}\in A, \\ (\boldsymbol{x},\bar{w}_i)\in X_i, w_{ij}\geq \widehat{w}_{ij}^q, \widehat{w}_{ij}\geq |\bar{w}_i-\bar{w}_j|, \forall i\in[m], j\in[m] \end{array} \right\}, \tag{28}$$

and we have $w_{ij}\leq (M_i+M_j)^q$ for each $(i,j)\in[m]\times[m]$. For each $a<\bar{a}\in A$, the dual of the left-hand side of the first constraint system in (28) is

$$\max_{\boldsymbol{\mu}_{a\bar{a}},\boldsymbol{\lambda}_{a\bar{a}}} \quad \frac{1}{m_{\bar{a}}}\sum_{j\in C_{\bar{a}}}\lambda_{ja\bar{a}} + \frac{1}{m_a}\sum_{i\in C_a}\mu_{ia\bar{a}} \leq \nu, \tag{29a}$$

$$\text{s.t.} \quad \mu_{ia\bar{a}}+\lambda_{ja\bar{a}}\leq w_{ij}, \forall i\in C_a, j\in C_{\bar{a}}. \tag{29b}$$

According to linear programming complementary slackness, the system of linear inequalities in (29) is equivalent to

$$\frac{1}{m_{\bar{a}}}\sum_{j\in C_{\bar{a}}}\lambda_{ja\bar{a}} + \frac{1}{m_a}\sum_{i\in C_a}\mu_{ia\bar{a}} \leq \nu, \tag{30a}$$

$$\sum_{i\in C_a}\pi_{ija\bar{a}} = \frac{1}{m_{\bar{a}}}, \forall j\in C_{\bar{a}}, \tag{30b}$$

$$\sum_{j\in C_{\bar{a}}}\pi_{ija\bar{a}} = \frac{1}{m_a}, \forall i\in C_a, \tag{30c}$$

$$\pi_{ija\bar{a}}\geq 0, \forall i\in C_a, j\in C_{\bar{a}}, \tag{30d}$$

$$w_{ij}-\mu_{ia\bar{a}}-\lambda_{ja\bar{a}}\geq 0, \forall i\in C_a, j\in C_{\bar{a}}, \tag{30e}$$

$$\pi_{ija\bar{a}}\left(w_{ij}-\mu_{ia\bar{a}}-\lambda_{ja\bar{a}}\right)=0, \forall i\in C_a, j\in C_{\bar{a}}. \tag{30f}$$

Then, linearizing the complementary slackness constraints (30f) allows us to derive the second MICP-R.

Theorem 11. ***(Complementary Formulation)*** *Suppose that the set $X_i = \{(\boldsymbol{x}, \bar{w}_i) \in \mathcal{X} \times \mathbb{R} : f(\boldsymbol{x}, \boldsymbol{\xi}_i) = \bar{w}_i\}$ is MICP-R for each $i \in [m]$, $M_i \geq \max_{\boldsymbol{x} \in \mathcal{X}, (7\mathrm{b})} |f(\boldsymbol{x}, \boldsymbol{\xi}_i)|$ for each $i \in [m]$, and $\widehat{M}_{a\bar{a}} = \sum_{(i,j) \in C_a \times C_{\bar{a}}} (M_i + M_j)^q$ for each $a < \bar{a} \in A$. Then the set $\mathcal{F}_q$ is equivalent to*

$$\mathcal{F}_q = \left\{ \boldsymbol{x} : \begin{array}{l} \dfrac{1}{m_{\bar{a}}} \sum\limits_{j \in C_{\bar{a}}} \lambda_{ja\bar{a}} + \dfrac{1}{m_a} \sum\limits_{i \in C_a} \mu_{ia\bar{a}} \leq \nu, \forall a < \bar{a} \in A, \\ \sum\limits_{i \in C_a} \pi_{ija\bar{a}} = \dfrac{1}{m_{\bar{a}}}, \forall j \in C_{\bar{a}}, a < \bar{a} \in A, \sum\limits_{j \in C_{\bar{a}}} \pi_{ija\bar{a}} = \dfrac{1}{m_a}, \forall i \in C_a, a < \bar{a} \in A, \\ (\boldsymbol{x}, \bar{w}_i) \in X_i, \forall i \in [m], w_{ij} \geq \widehat{w}_{ij}^q, \widehat{w}_{ij} \geq |\bar{w}_i - \bar{w}_j|, \forall i \in [m], j \in [m], \\ w_{ij} - \mu_{ia\bar{a}} - \lambda_{ja\bar{a}} \geq 0, w_{ij} - \mu_{ia\bar{a}} - \lambda_{ja\bar{a}} \leq \widehat{M}_{a\bar{a}} \left(1 - z_{ija\bar{a}}\right), \\ \pi_{ija\bar{a}} \leq \min\{m_a^{-1}, m_{\bar{a}}^{-1}\} z_{ija\bar{a}}, \pi_{ija\bar{a}} \geq 0, z_{ija\bar{a}} \in \{0,1\}, \forall i \in C_a, j \in C_{\bar{a}}, a < \bar{a} \in A \end{array} \right\}. \tag{31}$$

Proof. By introducing a large constant $\widehat{M}_{a\bar{a}}$ for each $a < \bar{a} \in A$, (30f) can be linearized as

$$\begin{aligned} &\pi_{ija\bar{a}} \leq \min\{m_a^{-1}, m_{\bar{a}}^{-1}\} z_{ija\bar{a}},\ w_{ij} - \mu_{ia\bar{a}} - \lambda_{ja\bar{a}} \leq \widehat{M}_{a\bar{a}} \left(1 - z_{ija\bar{a}}\right), \\ &z_{ija\bar{a}} \in \{0,1\}, \forall i \in C_a, j \in C_{\bar{a}}. \end{aligned} \tag{32a}$$

It remains to show that for each pair $a < \bar{a} \in A$, the big-M value $\widehat{M}_{a\bar{a}} = \sum_{(i,j) \in C_a \times C_{\bar{a}}} (M_i + M_j)^q$ suffices. That is, any dual feasible solution satisfies $w_{ij} - \mu_{ia\bar{a}} - \lambda_{ja\bar{a}} \leq \widehat{M}_{a\bar{a}}$ for each $i \in C_a, j \in C_{\bar{a}}, a < \bar{a} \in A$. From (30a), we can get

$$0 \leq \frac{1}{m_a m_{\bar{a}}} \sum_{(i,j) \in C_a \times C_{\bar{a}}} (\mu_{ia\bar{a}} + \lambda_{ja\bar{a}}) \leq \nu. \tag{32b}$$

According to (30e), we also know that

$$\mu_{ia\bar{a}} + \lambda_{ja\bar{a}} \leq w_{ij} \leq (M_i + M_j)^q. \tag{32c}$$

Then, we have

$$\mu_{ia\bar{a}} + \lambda_{ja\bar{a}} \geq - \sum_{(i',j') \in C_a \times C_{\bar{a}} \setminus \{i,j\}} (\mu_{i'a\bar{a}} + \lambda_{j'a\bar{a}}) \geq - \sum_{(i',j') \in C_a \times C_{\bar{a}} \setminus \{i,j\}} (M_{i'} + M_{j'})^q,$$

where the first inequality is due to (32b) and the second inequality is because of (32c). Thus, $\widehat{M}_{a\bar{a}}$ can be found by

$$\widehat{M}_{a\bar{a}} := (M_i + M_j)^q + \sum_{(i',j') \in C_a \times C_{\bar{a}} \setminus \{i,j\}} (M_{i'} + M_{j'})^q = \sum_{(i',j') \in C_a \times C_{\bar{a}}} (M_{i'} + M_{j'})^q \geq w_{ij} - \mu_{ia\bar{a}} - \lambda_{ja\bar{a}},$$

which give us $\widehat{M}_{a\bar{a}} = \sum_{(i,j) \in C_a \times C_{\bar{a}}} (M_i + M_j)^q$. □

A.3 A Side Product: An Equivalent MICP-R Formulation for $\mathrm{KSD}(\boldsymbol{x})$

In this subsection, we propose to represent the sublevel set of the Kolmogorov–Smirnov fairness measure $\mathrm{KSD}(\boldsymbol{x})$, denoted as $\mathrm{KSD}_{\widehat{\delta}}(\boldsymbol{x})$, using the sets $\{\Omega_a(k)\}_{a\in A,k\in[m_a]}$ defined in (11). We note that the Kolmogorov–Smirnov fairness problem (similar to DFSO) admits the following form:

$$v^* = \min_{\boldsymbol{x}\in\mathcal{X}} \left\{ \mathrm{KSD}(\boldsymbol{x}) : \mathbb{E}_{\mathbb{P}}[Q(\boldsymbol{x},\tilde{\boldsymbol{\xi}})] \le V^* + \epsilon|V^*| \right\}. \tag{33}$$

Hence, if we can represent the sublevel set of the function $\mathrm{KSD}(\boldsymbol{x})$, then we can simply run a binary search to find the best objective value of problem (33). The formulation of $\mathrm{KSD}_{\widehat{\delta}}(\boldsymbol{x})$ is shown below.

Theorem 12. *Let the quantile set be defined as $\Omega_a(k) = \{(\boldsymbol{x},t_a) \in \mathcal{X}\times\mathbb{R} : F_a^{-1}(k/m_a \mid \boldsymbol{x}) = t_a\}$ for each $a\in A$, which admits a MICP-R form* (11)*. Then for a given $\widehat{\delta} \in \{|i/m_a - j/m_{\bar{a}}|\}_{i\in[0,m_a],j\in[0,m_{\bar{a}}],a<\bar{a}\in A}$, the set $\mathrm{KSD}_{\widehat{\delta}}(\boldsymbol{x})$ can be expressed as*

$$\mathrm{KSD}_{\widehat{\delta}}(\boldsymbol{x}) = \left\{ \boldsymbol{x}\in\mathcal{X} : \begin{array}{c} (\boldsymbol{x},t_{ia}) \in \Omega_a(i), \forall i\in[m_a], a\in A, \\ t_{(\lfloor m_{\bar{a}}(i/m_a-\widehat{\delta})_+\rfloor)\bar{a}} \le t_{ia}, \forall i\in[m_a], a<\bar{a}\in A \\ t_{(i+1)a} \le t_{(\lfloor m_{\bar{a}}\max(i/m_a+\widehat{\delta},1)\rfloor+1)\bar{a}}, \forall i\in[0,m_a-1], a<\bar{a}\in A \end{array} \right\},$$

where we let $t_{0a}=-\infty$ and $t_{(m_a+1)a}=+\infty$ for any $a\in A$.

Proof. Recall that for any $\boldsymbol{x}\in\mathrm{KSD}_{\widehat{\delta}}(\boldsymbol{x})$, we have

$$\max_{a<\bar{a}\in A} \sup_{\tau} |F_a(\tau\mid\boldsymbol{x}) - F_{\bar{a}}(\tau\mid\boldsymbol{x})| \le \widehat{\delta}, \tag{34a}$$

$$\Leftrightarrow \sup_{\tau} |F_a(\tau\mid\boldsymbol{x}) - F_{\bar{a}}(\tau\mid\boldsymbol{x})| \le \nu, \forall a<\bar{a}\in A. \tag{34b}$$

Let us consider the possible values of $F_a(\tau\mid\boldsymbol{x})$ as $\{0,1/m_a,\cdots,m_a/m_a\}$. There are three cases:

Case 1. If $F_a(\tau\mid\boldsymbol{x})=0$, then $-\infty<\tau\le t_{1a}$. For any such τ, we must have $|F_a(\tau\mid\boldsymbol{x}) - F_{\bar{a}}(\tau\mid\boldsymbol{x})| \le \widehat{\delta}$. That is, we must have

$$0 \le F_{\bar{a}}(\tau\mid\boldsymbol{x}) \le \widehat{\delta},$$

or equivalently $-\infty<\tau<t_{(\lfloor m_{\bar{a}}\max(\widehat{\delta},1)\rfloor+1)\bar{a}}$. Therefore, the following inequalities must hold

$$\tau \le t_{1a} \le t_{(\lfloor m_{\bar{a}}\max(\widehat{\delta},1)\rfloor+1)\bar{a}}.$$

Case 2. If $F_a(\tau\mid\boldsymbol{x}) = i/m_a$ for some $i\in[m_a-1]$, then $t_{ia}\le\tau<t_{(i+1)a}$. For any such τ, we must have $|F_a(\tau\mid\boldsymbol{x}) - F_{\bar{a}}(\tau\mid\boldsymbol{x})| \le \widehat{\delta}$. That is, we must have

$$\left(\frac{i}{m_a}-\widehat{\delta}\right)_+ \le F_{\bar{a}}(\tau\mid\boldsymbol{x}) \le \max\left(\frac{i}{m_a}+\widehat{\delta},1\right)$$

or equivalently $t_{(\lfloor m_{\bar{a}}(i/m_a-\widehat{\delta})_+\rfloor)\bar{a}} \le \tau < t_{(\lfloor m_{\bar{a}}\max(i/m_a+\widehat{\delta},1)\rfloor+1)\bar{a}}$. Therefore, the following inequalities must hold

$$t_{(\lfloor m_{\bar{a}}(i/m_a-\widehat{\delta})_+\rfloor)\bar{a}} \le t_{ia} \le t_{(i+1)a} \le t_{(\lfloor m_{\bar{a}}\max(i/m_a+\widehat{\delta},1)\rfloor+1)\bar{a}}.$$

Case 3. If $F_a(\tau \mid \boldsymbol{x}) = 1$, then $t_{(m_a)a} \leq \tau < +\infty$. For any such τ, we must have $|F_a(\tau \mid \boldsymbol{x}) - F_{\bar{a}}(\tau \mid \boldsymbol{x})| \leq \widehat{\delta}$. That is, we must have

$$\left(1 - \widehat{\delta}\right)_+ \leq F_{\bar{a}}(\tau \mid \boldsymbol{x}) \leq 1,$$

or equivalently $t_{(\lfloor m_{\bar{a}}(1-\widehat{\delta})_+ \rfloor)\bar{a}} \leq \tau < +\infty$. Therefore, the following inequalities must hold

$$t_{(\lfloor m_{\bar{a}}(i/m_a - \widehat{\delta})_+ \rfloor)\bar{a}} \leq t_{(m_a)a}.$$

If $F(\tau \mid \boldsymbol{x}) = i/m$, then we must have $i/m - \widehat{\delta} \leq F_a(t_i \mid \boldsymbol{x}) \leq i/m + \widehat{\delta}$ for all $a \in A$ to solve (34b).

Finally, we observe that since all $\{\mathbb{P}_a\}_{a\in A}$ are equiprobable discrete distributions, we must have $\widehat{\delta} \in \{|i/m_a - j/m_{\bar{a}}|\}_{i\in[0,m_a], j\in[0,m_{\bar{a}}], a<\bar{a}\in A}$. □

We remark that to solve (33) to optimality, we can run the binary search to find the optimal $\widehat{\delta} \in \{|i/m_a - j/m_{\bar{a}}|\}_{i\in[0,m_a], j\in[0,m_{\bar{a}}], a<\bar{a}\in A}$. That is, given a current $\widehat{\delta}$ value, we optimize the total cost $\mathbb{E}[Q(\boldsymbol{x}, \tilde{\boldsymbol{\xi}})]$ subject to the set $\mathrm{KSD}_{\widehat{\delta}}(\boldsymbol{x})$. Next, we check whether the optimal value is no larger than $V^* + \epsilon|V^*|$ or not. If yes, we decrease $\widehat{\delta}$; otherwise, we increase it.

Alternatively, we can perform difference-of-convex (DC) method at each binary search step, where we solve a continuous relaxation by relaxing the binary variables to be continuous and rewrite the set $\Omega_a(k)$ to

$$\Omega_a(k) = \left\{ (\boldsymbol{x}, t_{ka}) \in \mathcal{X} \times \mathbb{R} : \begin{array}{c} \pi_{ika} \in [0,1], z_{ika} \in [0,1], \pi_{ika} \leq z_{ika}, \forall i \in C_a, \\ \sum_{i\in C_a} z_{ika} = k, \sum_{i\in C_a} \pi_{ika} = 1, t_{ka} = \sum_{i\in C_a} \widehat{t}_{ika}, \\ (\boldsymbol{x}, \bar{w}_i) \in X_i, z_{ika}(t_{ka} - \bar{w}_i) \geq 0, \widehat{t}_{ika} \leq \pi_{ika}\bar{w}_i, \forall i \in C_a \end{array} \right\}.$$

Here, we can rewrite each bilinear term as a difference between two convex functions.

Appendix B. Proofs

B.1 Proof of Proposition 1

Proposition 1. *For a given decision $\boldsymbol{x}\in\mathcal{X}$, when computing the Wasserstein fairness measure in DFSO, the optimal joint distribution is comonotonic for any pair $a<\bar{a}\in A$.*

Proof. Given a standard uniform distribution U, let us define a joint distribution $\mathbb{Q}_{a\bar{a}}$ such that $(f(\boldsymbol{x},\tilde{\boldsymbol{\xi}}_a),f(\boldsymbol{x},\tilde{\boldsymbol{\xi}}_{\bar{a}}))\overset{\mathrm{d}}{=}(F_a^{-1}(U\mid\boldsymbol{x}),F_{\bar{a}}^{-1}(U\mid\boldsymbol{x}))$ for any pair $a<\bar{a}\in A$, and a fixed decision $\boldsymbol{x}\in\mathcal{X}$. Then, we have

$$\sqrt[q]{\int_{\Xi\times\Xi}\|\boldsymbol{\zeta}_1-\boldsymbol{\zeta}_2\|^q\,\mathbb{Q}_{a\bar{a}}(d\boldsymbol{\zeta}_1,d\boldsymbol{\zeta}_2)}=\sqrt[q]{\int_0^1\left|F_a^{-1}(u\mid\boldsymbol{x})-F_{\bar{a}}^{-1}(u\mid\boldsymbol{x})\right|^q du}\geq W_q\left(\mathbb{P}_{f(\boldsymbol{x},\tilde{\boldsymbol{\xi}}_a)},\mathbb{P}_{f(\boldsymbol{x},\tilde{\boldsymbol{\xi}}_{\bar{a}})}\right)$$

where the inequality is because $\mathbb{Q}$ is an admissible joint distribution. According to Lemma 1, $\mathbb{Q}$ is the ideal joint distribution when computing the Wasserstein fairness measure in DFSO. This completes the proof. □

B.2 Proof of Proposition 2

Proposition 2. *For a Bernoulli utility function $f(\boldsymbol{x},\tilde{\boldsymbol{\xi}})\in\{0,1\}$, $\mathrm{WD}_q(\boldsymbol{x})$ is equivalent to $\mathrm{DP}(\boldsymbol{x})$.*

Proof. Since $f(\boldsymbol{x},\tilde{\boldsymbol{\xi}})\in\{0,1\}$, we observe that

$$F_a^{-1}(y\mid\boldsymbol{x})=\begin{cases}0 & y\in(0,F_a(0\mid\boldsymbol{x})],\\ 1 & y\in(F_a(0\mid\boldsymbol{x}),1],\end{cases}$$

for each $a\in A$. According to Lemma 1, the Wasserstein fairness measure $\mathrm{WD}_q(\boldsymbol{x})$ can be simplified as

$$\begin{aligned}\mathrm{WD}_q(\boldsymbol{x})&=\max_{a<\bar{a}\in A}\sqrt[q]{\int_0^1\left|F_a^{-1}(y\mid\boldsymbol{x})-F_{\bar{a}}^{-1}(y\mid\boldsymbol{x})\right|^q dy}=\max_{a<\bar{a}\in A}\sqrt[q]{|F_a(0\mid\boldsymbol{x})-F_{\bar{a}}(0\mid\boldsymbol{x})|^q},\\ &=\max_{a<\bar{a}\in A}|F_a(0\mid\boldsymbol{x})-F_{\bar{a}}(0\mid\boldsymbol{x})|=\max_{a<\bar{a}\in A}\left|\mathbb{P}_a\{f(\boldsymbol{x},\tilde{\boldsymbol{\xi}}_a)=0\}-\mathbb{P}_{\bar{a}}\{f(\boldsymbol{x},\tilde{\boldsymbol{\xi}}_{\bar{a}})=0\}\right|=\mathrm{DP}(\boldsymbol{x}),\end{aligned}$$

where the second and third equalities are due to $\mathbb{P}(f(\boldsymbol{x},\tilde{\boldsymbol{\xi}})\in\{0,1\})=1$ and the observation above, while the fourth one follows from the definition of cumulative distribution functions. □

B.3 Proof of Proposition 3

Proposition 3. *For any feasible $\boldsymbol{x}\in\mathcal{X}$ and $q\in[1,\infty]$, the following inequalities hold:*

$$\frac{1}{\max_{a<\bar{a}\in A}\eta(\boldsymbol{x})^{\frac{1-q}{q}}(t_{2a\bar{a}}(\boldsymbol{x})-t_{1a\bar{a}}(\boldsymbol{x}))}\mathrm{WD}_q(\boldsymbol{x})\leq\mathrm{KSD}(\boldsymbol{x})\leq\frac{1}{\min_{a<\bar{a}\in A}\mu(\Delta_{a\bar{a}}(\boldsymbol{x}))}\mathrm{WD}_q(\boldsymbol{x}).$$

Here, $t_{1a\bar{a}}(\boldsymbol{x})=\min\{\min_t\{t:F_a(t\mid\boldsymbol{x})>0\},\min_t\{t:F_{\bar{a}}(t\mid\boldsymbol{x})>0\}\}$, $t_{2a\bar{a}}(\boldsymbol{x})=\max\{\sup_t\{t:F_a(t\mid\boldsymbol{x})<1\},\sup_t\{t:F_{\bar{a}}(t\mid\boldsymbol{x})<1\}\}$, and $\Delta_{a\bar{a}}(\boldsymbol{x})=\{\bar{t}:|F_a(\bar{t}\mid\boldsymbol{x})-F_{\bar{a}}(\bar{t}\mid\boldsymbol{x})|=\sup_t|F_a(t\mid\boldsymbol{x})-F_{\bar{a}}(t\mid\boldsymbol{x})|\}$ with its Lebesgue measure $\mu(\Delta_{a\bar{a}}(\boldsymbol{x}))$.

Proof. We split the proof into two steps.

Step 1. We first derive the relationship between $\mathrm{WD}_1(\boldsymbol{x})$ and $\mathrm{KSD}(\boldsymbol{x})$.

We let $t_{1a\bar{a}}(\boldsymbol{x}) = \min\{\min_t\{t : F_a(t \mid \boldsymbol{x}) > 0\}, \min_t\{t : F_{\bar{a}}(t \mid \boldsymbol{x}) > 0\}\}$ and $t_{2a\bar{a}}(\boldsymbol{x}) = \max\{\max_t\{t : F_a(t \mid \boldsymbol{x}) < 1\}, \max_t\{t : F_{\bar{a}}(t \mid \boldsymbol{x}) < 1\}\}$. Then, according to Lemma 1, we have

$$\begin{aligned}
\mathrm{WD}_1(\boldsymbol{x}) &= \max_{a<\bar{a}\in A} \int_t |F_a(t \mid \boldsymbol{x}) - F_{\bar{a}}(t \mid \boldsymbol{x})|\, dt, \\
&= \max_{a<\bar{a}\in A} \int_{t_{1a\bar{a}}(\boldsymbol{x})}^{t_{2a\bar{a}}(\boldsymbol{x})} |F_a(t \mid \boldsymbol{x}) - F_{\bar{a}}(t \mid \boldsymbol{x})|\, dt, \\
&\le \max_{a<\bar{a}\in A} (t_{2a\bar{a}}(\boldsymbol{x}) - t_{1a\bar{a}}(\boldsymbol{x})) \max_{a<\bar{a}\in A} \sup_t |F_a(t \mid \boldsymbol{x}) - F_{\bar{a}}(t \mid \boldsymbol{x})| = \max_{a<\bar{a}\in A} (t_{2a\bar{a}}(\boldsymbol{x}) - t_{1a\bar{a}}(\boldsymbol{x}))\mathrm{KSD}(\boldsymbol{x}),
\end{aligned}$$

which yields the lower bound on $\mathrm{KSD}(\boldsymbol{x})$.

To establish the upper bound on $\mathrm{KSD}(\boldsymbol{x})$, we let $\Delta_{a\bar{a}}(\boldsymbol{x}) = \{\bar{t} : |F_a(\bar{t} \mid \boldsymbol{x}) - F_{\bar{a}}(\bar{t} \mid \boldsymbol{x})| = \sup_t |F_a(t \mid \boldsymbol{x}) - F_{\bar{a}}(t \mid \boldsymbol{x})|\}$ and use $\mu(\Delta_{a\bar{a}}(\boldsymbol{x}))$ to denote the Lebesgue measure of the set $\Delta_{a\bar{a}}(\boldsymbol{x})$, which is positive since all the groups are finitely distributed. Then, we have

$$\begin{aligned}
\mathrm{WD}_1(\boldsymbol{x}) &= \max_{a<\bar{a}\in A} \int_t |F_a(t \mid \boldsymbol{x}) - F_{\bar{a}}(t \mid \boldsymbol{x})|\, dt, \\
&\ge \max_{a<\bar{a}\in A} \int_{\Delta_{a\bar{a}}(\boldsymbol{x})} |F_a(t \mid \boldsymbol{x}) - F_{\bar{a}}(t \mid \boldsymbol{x})|\, dt, \\
&\ge \min_{a<\bar{a}\in A} \mu(\Delta_{a\bar{a}}(\boldsymbol{x})) \max_{a<\bar{a}\in A} \sup_t |F_a(t \mid \boldsymbol{x}) - F_{\bar{a}}(t \mid \boldsymbol{x})| = \min_{a<\bar{a}\in A} \mu(\Delta_{a\bar{a}}(\boldsymbol{x}))\mathrm{KSD}(\boldsymbol{x}).
\end{aligned}$$

Combining both lower and upper bounds, we obtain the desired inequalities

$$\frac{1}{\max_{a<\bar{a}\in A}(t_{2a\bar{a}}(\boldsymbol{x}) - t_{1a\bar{a}}(\boldsymbol{x}))}\mathrm{WD}_1(\boldsymbol{x}) \le \mathrm{KSD}(\boldsymbol{x}) \le \frac{1}{\min_{a<\bar{a}\in A} \mu(\Delta_{a\bar{a}}(\boldsymbol{x}))}\mathrm{WD}_1(\boldsymbol{x}).$$

Step 2. Next, we derive the bounds between $\mathrm{WD}_1(\boldsymbol{x})$ and $\mathrm{WD}_q(\boldsymbol{x})$ for any $q \in [1, \infty]$. According to Lemma 1, we know that

$$\begin{aligned}
\mathrm{WD}_q(\boldsymbol{x}) &= \max_{a<\bar{a}\in A} \sqrt[q]{\int_0^1 \left|F_a^{-1}(y \mid \boldsymbol{x}) - F_{\bar{a}}^{-1}(y \mid \boldsymbol{x})\right|^q dy}, \\
&= \max_{a<\bar{a}\in A} \sqrt[q]{\sum_{j\in J_{a\bar{a}}\setminus\{1\}} w_{ja\bar{a}}(\boldsymbol{x}) \left|F_a^{-1}(b_{ja\bar{a}}(\boldsymbol{x}) \mid \boldsymbol{x}) - F_{\bar{a}}^{-1}(b_{ja\bar{a}}(\boldsymbol{x}) \mid \boldsymbol{x})\right|^q}, \\
&\le \max_{a<\bar{a}\in A} \sqrt[q]{\sum_{j\in J_{a\bar{a}}\setminus\{1\}} \frac{w_{ja\bar{a}}(\boldsymbol{x})^q}{\eta(\boldsymbol{x})^{q-1}} \left|F_a^{-1}(b_{ja\bar{a}}(\boldsymbol{x}) \mid \boldsymbol{x}) - F_{\bar{a}}^{-1}(b_{ja\bar{a}}(\boldsymbol{x}) \mid \boldsymbol{x})\right|^q}, \\
&\le \eta(\boldsymbol{x})^{\frac{1-q}{q}} \max_{a<\bar{a}\in A} \sum_{j\in J_{a\bar{a}}\setminus\{1\}} w_{ja\bar{a}}(\boldsymbol{x}) \left|F_a^{-1}(b_{ja\bar{a}}(\boldsymbol{x}) \mid \boldsymbol{x}) - F_{\bar{a}}^{-1}(b_{ja\bar{a}}(\boldsymbol{x}) \mid \boldsymbol{x})\right| = \eta(\boldsymbol{x})^{\frac{1-q}{q}} \mathrm{WD}_1(\boldsymbol{x}),
\end{aligned}$$

where the second equality is due to Definition 4, the first inequality is due to $\eta(\boldsymbol{x}) = \max_{a<\bar{a}\in A, j\in J_a\setminus\{1\}} w_{ja\bar{a}}(\boldsymbol{x})$, and the second one is because $\|\cdot\|_q \le \|\cdot\|_1$.

Meanwhile, by using Hölder's inequality, we have

$$\sqrt[1]{\int_0^1 |F_a^{-1}(y\mid \boldsymbol{x}) - F^{-1}(y\mid \boldsymbol{x})|^1\, dy} \le \sqrt[q]{\int_0^1 |F_a^{-1}(y\mid \boldsymbol{x}) - F^{-1}(y\mid \boldsymbol{x})|^q\, dy}\sqrt[p]{\int_0^1 dy}$$

for any $p,q \in [1,\infty)$ with $1/p+1/q=1$, Thus, the following inequality holds

$$\int_0^1 \left|F_a^{-1}(y\mid \boldsymbol{x}) - F_{\bar{a}}^{-1}(y\mid \boldsymbol{x})\right|^1 dy \le \sqrt[q]{\int_0^1 \left|F_a^{-1}(y\mid \boldsymbol{x}) - F_{\bar{a}}^{-1}(y\mid \boldsymbol{x})\right|^q dy},$$

i.e., $\mathrm{WD}_1(\boldsymbol{x}) \le \mathrm{WD}_q(\boldsymbol{x})$. This concludes the proof. □

B.4 Proof of Theorem 1

THEOREM 1. *Solving DFSO is, in general, strongly NP-hard, even when $\mathcal{X}$ is a polytope, $\epsilon=\infty$, and $f(\boldsymbol{x},\boldsymbol{\xi})$ is a linear function.*

Proof. We derive a reduction from the chance constrained optimization problem, which is strongly NP-hard (Ahmed and Xie 2018). Let us consider the following feasibility problem of the generic chance-constrained stochastic program:

Does there exist a feasible solution to the following chance-constrained set

$$\left\{(\boldsymbol{x},\boldsymbol{z}) \in \mathcal{X} : \boldsymbol{z} \in \{0,1\}^{m'}, \sum_{i\in[m']} z_i = m'-k,\right\}$$

where $\mathcal{X} = \{(\boldsymbol{x},\boldsymbol{z}) \in \mathbb{R}^n \times [0,1]^{m'} : \widehat{\boldsymbol{A}}_i\boldsymbol{x} \ge \boldsymbol{b}_i - \boldsymbol{M}(1-z_i), \forall i \in [m']\}$ with large coefficients $\boldsymbol{M}$ for each $i\in[m']$?

Let us consider a special case of DFSO with $|A|=2$, $m_a = m_{\bar{a}} = m'$, $C_1 = [m']$, $C_2 = [m'+1, 2m']$, and $\boldsymbol{\xi} \in \mathbb{R}^{n+m'+1}$ with

$$f((\boldsymbol{x},\boldsymbol{z}),\boldsymbol{\xi}) = \boldsymbol{\xi}_{1:n}^\top \boldsymbol{x} + \boldsymbol{\xi}_{[n+1:n+m']}^\top \boldsymbol{z} + \xi_{n+m'+1}.$$

Specifically, for each individual $i\in[2m']$, we design their corresponding scenario $\boldsymbol{\xi}_i$ such that

$$\begin{aligned} &f((\boldsymbol{x},\boldsymbol{z}),\boldsymbol{\xi}_i) = z_i, \forall i \in [m'],\\ &f((\boldsymbol{x},\boldsymbol{z}),\boldsymbol{\xi}_j) = 0, \forall j \in [m'+1:m'+k],\\ &f((\boldsymbol{x},\boldsymbol{z}),\boldsymbol{\xi}_j) = 1, \forall j \in [m'+k+1:2m']. \end{aligned}$$

Assuming that $\epsilon=\infty$, this particular DFSO becomes

$$v^*(q) = \min_{(\boldsymbol{x},\boldsymbol{z})\in\mathcal{X}} \quad \mathrm{WD}_q^q((\boldsymbol{x},\boldsymbol{z})).$$

Hence, we see that $v^*(q)=0$ if and only if there exists a point $(\boldsymbol{x},\boldsymbol{z}) \in \mathcal{X}$ such that $\boldsymbol{z} \in \{0,1\}^{m'}, \sum_{i\in[m']} z_i = m'-k$. In other words, $v^*(q)=0$ if and only if $(\boldsymbol{x},\boldsymbol{z})$ is feasible to the chance-constrained stochastic program. This completes the proof. □

B.5 Proof of Proposition 7

PROPOSITION 7. *Suppose that $M_i \geq \max_{\boldsymbol{x}\in\mathcal{X},(7\mathrm{b})} |f(\boldsymbol{x},\boldsymbol{\xi}_i)|$ for each $i \in [m]$. For each $k \in [m_a]$ and $a \in A$, the quantile set $\Omega_a(k)$ is equivalent to*

$$\Omega_a(k) = \left\{ (\boldsymbol{x}, t_{ka}) \in \mathcal{X}\times\mathbb{R} : \begin{array}{l} \pi_{ika} \in \{0,1\}, z_{ika} \in \{0,1\}, \pi_{ika} \leq z_{ika}, (\boldsymbol{x}, \bar{w}_i) \in X_i, \forall i \in C_a, \\ \sum\limits_{i\in C_a} z_{ika} = k, \sum\limits_{i\in C_a} \pi_{ika} = 1, t_{ka} = \sum\limits_{i\in C_a} \widehat{t}_{ika}, \\ t_{ka} \geq \bar{w}_i - (M_i + M_{(k)})(1 - z_{ika}), t_{ka} \leq \bar{w}_i + (M_i + M_{(k)}) z_{ika}, \\ (\widehat{t}_{ika}, \pi_{ika}, \bar{w}_i) \in \mathrm{MC}(0, 1, -M_i, M_i), \forall i \in C_a \end{array} \right\}, \quad (11)$$

where $M_{(i)}$ is the ith smallest value of the vector $\boldsymbol{M}$.

Proof. By definition of the inverse distribution function, we have

$$\begin{aligned} \Omega_a(k) &= \left\{ (\boldsymbol{x}, t_{ka}) \in \mathcal{X}\times\mathbb{R} : F_a^{-1}(k/m_a \,|\, \boldsymbol{x}) = t_{ka} \right\} \\ &= \left\{ (\boldsymbol{x}, t_{ka}) \in \mathcal{X}\times\mathbb{R} : \begin{array}{l} \pi_{ika} \in \{0,1\}, z_{ika} \in \{0,1\}, \pi_{ika} \leq z_{ika}, (\boldsymbol{x}, \bar{w}_i) \in X_i, \forall i \in C_a, \\ \sum\limits_{i\in C_a} z_{ika} = k, \sum\limits_{i\in C_a} \pi_{ika} = 1, t_{ka} = \sum\limits_{i\in C_a} \bar{w}_i \pi_{ika}, \\ t_{ka} \geq \bar{w}_i - (M_i + M_{(k)})(1 - z_{ika}), t_{ka} \leq \bar{w}_i + (M_i + M_{(k)}) z_{ika}, \end{array} \right\}, \end{aligned}$$

for each $k \in [m_a]$ and $a \in A$. Next, we arrive at the conclusion by linearizing the bilinear terms with the McCormick representation. □

B.6 Proof of Theorem 2

THEOREM 2. ***(Quantile Formulation)*** *Suppose that the set $X_i = \{(\boldsymbol{x}, \bar{w}_i) \in \mathcal{X}\times\mathbb{R} : f(\boldsymbol{x}, \boldsymbol{\xi}_i) = \bar{w}_i\}$ is MICP-R and $M_i \geq \max_{\boldsymbol{x}\in\mathcal{X},(7\mathrm{b})} |f(\boldsymbol{x},\boldsymbol{\xi}_i)|$ for each $i \in [m]$. We further define the quantile set $\Omega_a(k) = \{(\boldsymbol{x}, t_{ka}) \in \mathcal{X}\times\mathbb{R} : F_a^{-1}(k/m_a \,|\, \boldsymbol{x}) = t_{ka}\}$, which admits a MICP-R form* (11)*. Then $\mathcal{F}_q$ can be represented as*

$$\mathcal{F}_q = \left\{ (\boldsymbol{x}, \nu) \in \mathcal{X}\times\mathbb{R}_+ : \begin{array}{l} \sum\limits_{i\in[\widehat{m}_{a\bar{a}}-1]} \left(\widehat{b}_{(i+1)a\bar{a}} - \widehat{b}_{ia\bar{a}} \right) \eta^q_{ia\bar{a}} \leq \nu, \forall a < \bar{a} \in A, \\ \left| \sum\limits_{j\in[m_a]} \delta_{ija\bar{a}1} t_{ja} - \sum\limits_{j\in[m_{\bar{a}}]} \delta_{ija\bar{a}2} t_{j\bar{a}} \right| \leq \eta_{ia\bar{a}}, \forall i \in [\widehat{m}_{a\bar{a}} - 1], a < \bar{a} \in A, \\ (\boldsymbol{x}, t_{ja}) \in \Omega_a(j), \forall j \in [m_a], a \in A \end{array} \right\}, \quad (12)$$

where

$$\delta_{ija\bar{a}1} = \mathbb{I}\left(\left(\widehat{b}_{ia\bar{a}}, \widehat{b}_{(i+1)a\bar{a}} \right] \subseteq \left(\frac{j-1}{m_a}, \frac{j}{m_a} \right] \right), \forall i \in [\widehat{m}_{a\bar{a}} - 1], j \in [m_a], a < \bar{a} \in A,$$

and

$$\delta_{ija\bar{a}2} = \mathbb{I}\left(\left(\widehat{b}_{ia\bar{a}}, \widehat{b}_{(i+1)a\bar{a}} \right] \subseteq \left(\frac{j-1}{m_{\bar{a}}}, \frac{j}{m_{\bar{a}}} \right] \right), \forall i \in [\widehat{m}_{a\bar{a}} - 1], j \in [m_{\bar{a}}], a < \bar{a} \in A.$$

Proof. We split the proof into two steps.

Step 1. First, we will reformulate $\mathrm{WD}_q^q(\boldsymbol{x})$. According to (10), we can represent $\mathrm{WD}_q^q(\boldsymbol{x}) \le \nu$ as

$$\sum_{i\in[\widehat{m}_{a\bar{a}}-1]} (\widehat{b}_{(i+1)a\bar{a}} - \widehat{b}_{ia\bar{a}}) \left| F_a^{-1}(\widehat{b}_{(i+1)a\bar{a}} \mid \boldsymbol{x}) - F_{\bar{a}}^{-1}(\widehat{b}_{(i+1)a\bar{a}} \mid \boldsymbol{x}) \right|^q \le \nu, \forall a < \bar{a} \in A. \tag{35}$$

Suppose $F_a^{-1}(y \mid \boldsymbol{x}) = t_{ja}$ for $j \in [m_a], a \in A$. Let $\delta_{ija\bar{a}1} = \mathbb{I}((\widehat{b}_{ia\bar{a}}, \widehat{b}_{(i+1)a\bar{a}}] \subseteq ((j-1)/m_a, j/m_a])$ for each $i \in [\widehat{m}_{a\bar{a}} - 1], j \in [m_a], a < \bar{a} \in A$ and $\delta_{ija\bar{a}2} = \mathbb{I}((\widehat{b}_{ia\bar{a}}, \widehat{b}_{(i+1)a\bar{a}}] \subseteq ((j-1)/m_{\bar{a}}, j/m_{\bar{a}}])$ for each $i \in [\widehat{m}_{a\bar{a}} - 1], j \in [m_{\bar{a}}], a < \bar{a} \in A$. Then, the constraints (35) are equivalent to

$$\sum_{i\in[\widehat{m}_{a\bar{a}}-1]} (\widehat{b}_{(i+1)a\bar{a}} - \widehat{b}_{ia\bar{a}}) \eta_{ia\bar{a}}^q \le \nu, \forall a < \bar{a} \in A, \tag{36}$$

where

$$\left| \sum_{j\in[m_a]} \delta_{ija\bar{a}1} t_{ja} - \sum_{j\in[m_{\bar{a}}]} \delta_{ija\bar{a}2} t_{j\bar{a}} \right| \le \eta_{ia\bar{a}}, \forall i \in [\widehat{m}_{a\bar{a}} - 1], a < \bar{a} \in A. \tag{37}$$

Step 2. By choosing $(\boldsymbol{x}, t_{ja}) \in \Omega_a(j)$ for all $j \in [m_a], a \in A$, we have the formulation (12) for the set $\mathcal{F}_q$. □

B.7 Proof of Proposition 10

Proposition 10. *Suppose that the big-M coefficients $\boldsymbol{M}, \widehat{\boldsymbol{M}}$ are large enough as specified in the proof. Then, by relaxing the binary variables,*

(i) the continuous relaxation value of the Discretized Formulation is zero;

(ii) the continuous relaxation value of the Complementary Formulation is zero;

(iii) the continuous relaxation value of the Quantile Formulation is

$$\min_{\boldsymbol{x}\in\mathcal{X}} \max_{a<\bar{a}\in A} \left(\widehat{b}_{(\widehat{m}_{a\bar{a}})a\bar{a}} - \widehat{b}_{(\widehat{m}_{a\bar{a}}-1)a\bar{a}} \right) \left| F_a^{-1}(1 \mid \boldsymbol{x}) - F_{\bar{a}}^{-1}(1 \mid \boldsymbol{x}) \right|^q;$$

(iv) the continuous relaxation value of the Aggregate Quantile Formulation is at least

$$\min_{\boldsymbol{x}\in\mathcal{X}} \max_{a<\bar{a}\in A} \left| \frac{1}{m_a} \sum_{i\in C_a} F_a^{-1}\left(\frac{i}{m_a} \,\middle|\, \boldsymbol{x}\right) - \frac{1}{m_{\bar{a}}} \sum_{i\in C_{\bar{a}}} F_{\bar{a}}^{-1}\left(\frac{i}{m_{\bar{a}}} \,\middle|\, \boldsymbol{x}\right) \right|^q.$$

Proof.

(i) Since the optimal value of the continuous relaxation of the Discretized Formulation is at least zero, it suffices to show that there exists a feasible solution for the continuous relaxation of the Discretized Formulation such that its objective value is zero.

In the Discretized Formulation, we choose any $\boldsymbol{x} \in \mathcal{X}$ and $\bar{w}_i = f(\boldsymbol{x}, \boldsymbol{\xi}_i)$ for any $i \in [m]$. We also let $\nu = 0, \widehat{w}_{ijka\bar{a}} = \bar{z}_{ijka\bar{a}1} = \bar{z}_{ijka\bar{a}2} = 0$ and $z_{ijka\bar{a}} = 2^{1-k}/\bar{\Omega}_{a\bar{a}}$, for all $i \in C_a, j \in C_{\bar{a}}$, and $k \in [\bar{\Omega}_{a\bar{a}}]$. Then we have

$$\sum_{i\in C_a} \sum_{k\in[\bar{\Omega}_{a\bar{a}}]} 2^{k-1} z_{ijka\bar{a}} = m_a, \forall j \in C_{\bar{a}}, \sum_{j\in C_{\bar{a}}} \sum_{k\in[\bar{\Omega}_{a\bar{a}}]} 2^{k-1} z_{ijka\bar{a}} = m_{\bar{a}}, \forall i \in C_a,$$

for all $a < \bar{a} \in A$. It remains to show that

$$(\bar{z}_{ijka\bar{a}1}, z_{ijka\bar{a}}, \bar{w}_i) \in \mathrm{MC}(0,1,-M_i,M_i), (\bar{z}_{ijka\bar{a}2}, z_{ijka\bar{a}}, \bar{w}_j) \in \mathrm{MC}(0,1,-M_j,M_j),$$

for all $i \in C_a, j \in C_{\bar{a}}, k \in [\bar{\Omega}_{a\bar{a}}], a < \bar{a} \in A$. This is true by choosing

$$M_i \geq \max_{a<\bar{a}\in A} \bar{\Omega}_{a\bar{a}}/(\bar{\Omega}_{a\bar{a}}-1) \max_{\boldsymbol{x}\in\mathcal{X},(7\mathrm{b})} |f(\boldsymbol{x},\boldsymbol{\xi}_i)|$$

for each $i \in [m]$. Hence, we see that $(\boldsymbol{x}, \bar{\boldsymbol{w}}, \nu, \boldsymbol{z}, \bar{\boldsymbol{z}}, \widehat{\boldsymbol{w}})$ satisfies the constraints in (27), which yields an objective value zero.

(ii) Similarly, in the Complementary Formulation, we choose any $\boldsymbol{x} \in \mathcal{X}$ and $\bar{w}_i = f(\boldsymbol{x},\boldsymbol{\xi}_i)$ for any $i \in [m]$. We let $z_{ija\bar{a}} = \min\{m_a, m_{\bar{a}}\}/(m_a m_{\bar{a}}), w_{ij} = \widehat{w}_{ij}^q, \widehat{w}_{ij} = |\bar{w}_i - \bar{w}_j|$, and $\pi_{ija\bar{a}} = 1/(m_a m_{\bar{a}})$ for all $i \in C_a, j \in C_{\bar{a}}, a < \bar{a} \in A$. We also let $\mu_{ia\bar{a}} = 0$ and $\lambda_{ja\bar{a}} = 0$ for all $i \in C_a, j \in C_{\bar{a}}, a < \bar{a} \in A$. We note that

$$\sum_{i\in C_a} \pi_{ija\bar{a}} = \frac{1}{m_{\bar{a}}}, \sum_{j\in C_{\bar{a}}} \pi_{ija\bar{a}} = \frac{1}{m_a}, \pi_{ija\bar{a}} \leq \min\{m_a^{-1}, m_{\bar{a}}^{-1}\} z_{ija\bar{a}}, \pi_{ija\bar{a}} \geq 0,$$

for all $i \in C_a, j \in C_{\bar{a}}, a < \bar{a} \in A$.

It remains to show that

$$w_{ij} \leq \widehat{M}_{a\bar{a}} \left(1 - z_{ija\bar{a}}\right),$$

for all $i \in C_a, j \in C_{\bar{a}}, a < \bar{a} \in A$. This is true by choosing $\widehat{M}_{a\bar{a}} \geq (m_a m_{\bar{a}})/(m_a m_{\bar{a}} - \min\{m_a, m_{\bar{a}}\}) \sum_{(i,j)\in C_a\times C_{\bar{a}}} (M_i + M_j)^q$ for each $a < \bar{a} \in A$. Hence, we see that $(\boldsymbol{x}, \bar{\boldsymbol{w}}, \widehat{\boldsymbol{w}}, \boldsymbol{w}, \boldsymbol{\mu}, \boldsymbol{\lambda}, \nu, \boldsymbol{z}, \bar{\boldsymbol{z}}, \boldsymbol{\pi})$ satisfies the constraints in (31), which yields an objective value zero.

(iii) We observe that for any feasible solution of the Quantile Formulation, we must have $z_{i(m_a)a} = 1$ for each $i \in C_a$ and $a \in A$. Therefore, $t_{(m_a)a} = F_a^{-1}(1 \mid \boldsymbol{x}), t_{(m_{\bar{a}})\bar{a}} = F_{\bar{a}}^{-1}(1 \mid \boldsymbol{x})$ and $\eta_{\widehat{m}_{a\bar{a}}-1} \geq |t_{(m_a)a} - t_{(m_{\bar{a}})\bar{a}}|$. That is,

$$\nu \geq \max_{a<\bar{a}\in A} \left(\widehat{b}_{(\widehat{m}_{a\bar{a}})a\bar{a}} - \widehat{b}_{(\widehat{m}_{a\bar{a}}-1)a\bar{a}}\right) |F_a^{-1}(1 \mid \boldsymbol{x}) - F_{\bar{a}}^{-1}(1 \mid \boldsymbol{x})|^q.$$

We show that this bound is tight by constructing a solution such that $t_{ka} = 0$ for any $k \in [m_a - 1]$ and $a \in A$. In fact, for any $\boldsymbol{x} \in \chi$, let $\bar{w}_i = f(\boldsymbol{x},\boldsymbol{\xi}_i)$ for each $i \in [m]$. Then for any $i \in C_a, k \in [m_a - 1]$, and $a \in A$, we let $z_{ika} = k/m_a$, $\pi_{ika} = 1/m_a$, $\widehat{t}_{ika} = 0$. It remains to show that for any $i \in C_a, k \in [m_a - 1]$, and $a \in A$,

$$(\widehat{t}_{ika}, \pi_{ika}, \bar{w}_i) \in \mathrm{MC}(0,1,-M_i,M_i)$$

which must hold when $M_i \geq \max_{a\in A} m_a \max_{\boldsymbol{x}\in\mathcal{X},(7\mathrm{b})} |f(\boldsymbol{x},\boldsymbol{\xi}_i)|$.

(iv) We observe that for any feasible solution of the Aggregate Quantile Formulation, we must have $z_{i(m_a)a} = 1$ for each $i \in C_a$ and $a \in A$. Therefore, $\bar{t}_{(m_a)a} = \sum_{i \in C_a} F_a^{-1}(i/m_a \mid \boldsymbol{x}), \bar{t}_{(m_{\bar{a}})\bar{a}} = \sum_{i \in C_{\bar{a}}} F_{\bar{a}}^{-1}(i/m_{\bar{a}} \mid \boldsymbol{x})$. According to the first two constraints in (15), we have

$$\begin{aligned} \nu &\geq \max_{a<\bar{a}\in A} \sum_{i\in[\widehat{m}_{a\bar{a}}-1]} \left(\widehat{b}_{(i+1)a\bar{a}} - \widehat{b}_{ia\bar{a}}\right) \left| \sum_{j\in[m_a]} \delta_{ija\bar{a}1} t_{ja} - \sum_{j\in[m_{\bar{a}}]} \delta_{ija\bar{a}2} t_{j\bar{a}} \right|^q \\ &\geq \max_{a<\bar{a}\in A} \left| \sum_{i\in[\widehat{m}_{a\bar{a}}-1]} \left(\widehat{b}_{(i+1)a\bar{a}} - \widehat{b}_{ia\bar{a}}\right) \left[\sum_{j\in[m_a]} \delta_{ija\bar{a}1} t_{ja} - \sum_{j\in[m_{\bar{a}}]} \delta_{ija\bar{a}2} t_{j\bar{a}} \right] \right|^q \\ &= \max_{a<\bar{a}\in A} \left| m_a^{-1} \sum_{i\in C_a} F_a^{-1}(i/m_a \mid \boldsymbol{x}) - m_{\bar{a}}^{-1} \sum_{i\in C_{\bar{a}}} F_{\bar{a}}^{-1}(i/m_{\bar{a}} \mid \boldsymbol{x}) \right|^q \end{aligned}$$

where the second inequality is due to Jensen's inequality. □

B.8 Proof of Theorem 5

Theorem 5. *Computing the Gelbrich bound is strongly NP-hard even when $\epsilon = \infty$ and $|A| = 2$.*

Proof. We show a reduction from the integer programming feasibility problem, which is known to be strongly NP-complete. Consider the following feasibility problem of a binary integer program:

Does there exist a feasible solution to the binary program $X = \{\boldsymbol{x} \in \{-1,1\}^{n-1} : \boldsymbol{A}\boldsymbol{x} \geq \boldsymbol{b}\}$?

We consider a special case of the Gelbrich bound (20) by letting $\epsilon = \infty$, $|A| = 2$, $\boldsymbol{r}(\boldsymbol{x}, y) = (\boldsymbol{x}^\top, y)^\top$, $\boldsymbol{\mu}_a = \boldsymbol{\mu}_{\bar{a}} = \boldsymbol{0}$, $\boldsymbol{L}_a = \begin{bmatrix} I_{m_{\bar{a}}} & \boldsymbol{0} \\ \boldsymbol{0} & 0 \end{bmatrix}$, $\boldsymbol{L}_{\bar{a}} = \begin{bmatrix} \boldsymbol{0} & \boldsymbol{0} \\ \boldsymbol{0} & 1 \end{bmatrix}$, and defining

$$\mathcal{X} := \left\{(\boldsymbol{x}, y) \in [-1,1]^{n-1} \times \{\sqrt{n-1}\} : \boldsymbol{A}\boldsymbol{x} \geq \boldsymbol{b}\right\}.$$

Under this setting, the Gelbrich bound (20) simplifies to

$$v_G = \min_{\boldsymbol{x}} \quad \left(\sqrt{\boldsymbol{x}^\top \boldsymbol{x}} - \sqrt{n-1}\right)^2, \tag{38a}$$

$$\text{s.t.} \quad \boldsymbol{x} \in [-1,1]^{n-1}, \boldsymbol{A}\boldsymbol{x} \geq \boldsymbol{b}. \tag{38b}$$

We see that $v_G = 0$ in the formulation (38) if and only if there exists a binary feasible solution to the set X. Thus, the claim follows. □

B.9 Proof of Proposition 11

Proposition 11. *The semidefinite relaxation* (23) *of the Gelbrich bound model satisfies $v_{\underline{G}} \geq v_J(2)$. On the other hand, for any relative tolerance $\beta > 0$ of the semidefinite constraints in* (23b) *such that $\boldsymbol{Z}_{a\bar{a}} - \boldsymbol{s}_{a\bar{a}} \cdot \boldsymbol{s}_{a\bar{a}}^\top \succeq -\beta \lambda_{min}^{+}(\boldsymbol{Z}_{a\bar{a}}) \boldsymbol{I}_{2n+2}$, where $\lambda_{min}^{+}(\cdot)$ denotes the smallest nonzero eigenvalue, we have $v_{\underline{G}} \leq v_J(2)$.*

Proof. We split the proof into two steps.

Step 1. We know that $\boldsymbol{Z}_{a\bar{a}}$ is positive semidefinite and $\left(Z_{a\bar{a}11} - 2Z_{a\bar{a}1(n+2)} + Z_{a\bar{a}(n+2)(n+2)}\right) \geq 0$. Therefore, by removing the second term in the left-hand side of (22b), we have

$$\begin{aligned}
v_{\underline{G}} \geq \min_{\boldsymbol{x}\in\mathcal{X},\boldsymbol{s},\boldsymbol{Z},\nu} \quad & \nu, && \text{(39a)}\\
\text{s.t.} \quad & \left|\boldsymbol{\mu}_a^\top \boldsymbol{r}(\boldsymbol{x}) - \boldsymbol{\mu}_{\bar{a}}^\top \boldsymbol{r}(\boldsymbol{x})\right|^2 \leq \nu, \forall a < \bar{a} \in A, && \text{(39b)}\\
& \text{(7b)}, \text{(22c)} - \text{(22e)}, \text{(23b)}
\end{aligned}$$

where the optimal value of the minimization problem is equal to $v_J(2)$. Thus, we have $v_{\underline{G}} \geq v_J(2)$.

Step 2. Let $(\boldsymbol{x}^*, \nu^*)$ denote the optimal solution of (18) for $q = 2$. We compute $\boldsymbol{z}_a^* = \boldsymbol{L}_a^\top \boldsymbol{r}(\boldsymbol{x}^*)$, $\sigma_a^* = \|\boldsymbol{z}_a^*\|_2$, $\boldsymbol{s}_{a\bar{a}}^* = \left[\sigma_a^* \ \boldsymbol{z}_a^* \ \sigma_{\bar{a}}^* \ \boldsymbol{z}_{\bar{a}}^*\right]^\top$, and $\boldsymbol{Z}_{a\bar{a}}^* = \boldsymbol{s}_{a\bar{a}}^* \cdot \boldsymbol{s}_{a\bar{a}}^{*\top}$ for all $a < \bar{a} \in A$. Suppose $\sigma_a^* \geq \sigma_{\bar{a}}^*$ and $\gamma = \sigma_a^*/\sigma_{\bar{a}}^* \geq 1$. Let $\widehat{\boldsymbol{s}}_{a\bar{a}} = \left[\widehat{\sigma}_a \ \widehat{\boldsymbol{z}}_a \ \widehat{\sigma}_{\bar{a}} \ \widehat{\boldsymbol{z}}_{\bar{a}}\right]^\top = \left[\sigma_a^* \ \boldsymbol{z}_a^* \ \gamma\sigma_{\bar{a}}^* \ \gamma\boldsymbol{z}_{\bar{a}}^*\right]^\top$. For any given $\alpha \geq \bar{\alpha}$ (we will specify $\bar{\alpha}$ later), we construct $\widehat{\boldsymbol{Z}}_{a\bar{a}}$ as

$$\widehat{\boldsymbol{Z}}_{a\bar{a}} = \begin{bmatrix} \widehat{\sigma}_a^2 & \boldsymbol{0} & \widehat{\sigma}_a\widehat{\sigma}_{\bar{a}} & \boldsymbol{0} \\ \boldsymbol{0} & \widehat{\boldsymbol{z}}_a^2 & \boldsymbol{0} & \boldsymbol{0} \\ \widehat{\sigma}_a\widehat{\sigma}_{\bar{a}} & \boldsymbol{0} & \widehat{\sigma}_{\bar{a}}^2 & \boldsymbol{0} \\ \boldsymbol{0} & \boldsymbol{0} & \boldsymbol{0} & \widehat{\boldsymbol{z}}_{\bar{a}}^2 \end{bmatrix}.$$

It is evident that $(\boldsymbol{x}^*, \boldsymbol{s}^*, \alpha\widehat{\boldsymbol{Z}}, \nu^*)$ satisfies constraints (7b), (22b)-(22c), (22e).

We next compute $\alpha\widehat{\boldsymbol{Z}}_{a\bar{a}} - \boldsymbol{s}_{a\bar{a}}^* \cdot {\boldsymbol{s}_{a\bar{a}}^*}^\top$. Without loss of generality, we can assume that $z_{ia}^* \neq 0$ for all $i \in [n]$ and $a \in A$. We first observe that by dropping the first row and column, the resulting principal submatrix has rank $2n+1$. Thus, the rank of $\widehat{\boldsymbol{Z}}_{a\bar{a}}$ is at least $2n+1$. On the other hand, let $\boldsymbol{u} = [1, \boldsymbol{0}, -1, \boldsymbol{0}]^\top$. Then we have $\boldsymbol{u}^\top \widehat{\boldsymbol{Z}}_{a\bar{a}} \boldsymbol{u} = 0$. Thus, the rank of $\widehat{\boldsymbol{Z}}_{a\bar{a}}$ is $2n+1$. Let $\{\boldsymbol{v}_i\}_{i\in[2n+1]}$ denote the nonzero eigenvectors of $\widehat{\boldsymbol{Z}}_{a\bar{a}}$ with the corresponding positive eigenvalues $\{\lambda_i\}_{i\in[2n+1]} \subseteq \mathbb{R}_{++}$. Then we can rewrite the matrix as

$$\widehat{\boldsymbol{Z}}_{a\bar{a}} = \boldsymbol{V}\boldsymbol{\Lambda}\boldsymbol{V}^\top,$$

where $\boldsymbol{V} = [\boldsymbol{v}_1, \ldots, \boldsymbol{v}_{2n+1}]$ and $\boldsymbol{\Lambda} = \operatorname{diag}(\boldsymbol{\lambda})$. On the other hand, since $\boldsymbol{s}^* \in \operatorname{span}(\{\boldsymbol{v}_i\}_{i\in[2n+1]} \cup \{\boldsymbol{u}\})$, we have

$$\boldsymbol{s}^* = t_0\boldsymbol{u} + \sum_{i\in[2n+1]} q_i\boldsymbol{v}_i.$$

Thus, we have

$$\alpha\widehat{\boldsymbol{Z}}_{a\bar{a}} - \boldsymbol{s}_{a\bar{a}}^* \cdot {\boldsymbol{s}_{a\bar{a}}^*}^\top = \boldsymbol{V}(\alpha\boldsymbol{\Lambda} - \boldsymbol{q}\boldsymbol{q}^\top)\boldsymbol{V}^\top - t_0^2\boldsymbol{u}\boldsymbol{u}^\top,$$

where $t_0 = (\boldsymbol{s}_{a\bar{a}}^*)^\top\boldsymbol{u}/2$. Since $\boldsymbol{\Lambda} \succ \boldsymbol{0}$, we can choose $\bar{\alpha} = \max\{\boldsymbol{q}^\top\boldsymbol{q}/\min_{i\in[2n+1]}\lambda_i, 2t_0^2/(\beta\min_{i\in[2n+1]}\lambda_i)\}$ such that for any $\alpha \geq \bar{\alpha}$, we have

$$\alpha\widehat{\boldsymbol{Z}}_{a\bar{a}} - \boldsymbol{s}_{a\bar{a}}^* \cdot {\boldsymbol{s}_{a\bar{a}}^*}^\top = \boldsymbol{V}(\alpha\boldsymbol{\Lambda} - \boldsymbol{q}\boldsymbol{q}^\top)\boldsymbol{V}^\top - t_0^2\boldsymbol{u}\boldsymbol{u}^\top \succeq -t_0^2\boldsymbol{u}\boldsymbol{u}^\top \succeq -\beta\lambda_{\min}^+(\alpha\widehat{\boldsymbol{Z}}_{a\bar{a}})\boldsymbol{I}_{2n+2}.$$

This completes the proof. □

B.10 Proof of Theorem 6

Theorem 6. *Suppose that for any pair $a < \bar{a} \in A$, the optimal comonotonic random variables* $(f(\boldsymbol{x},\tilde{\boldsymbol{\xi}}_a) - \boldsymbol{\mu}_a^\top \boldsymbol{r}(\boldsymbol{x}) - s(\boldsymbol{x}), f(\boldsymbol{x},\tilde{\boldsymbol{\xi}}_{\bar{a}}) - \boldsymbol{\mu}_{\bar{a}}^\top \boldsymbol{r}(\boldsymbol{x}) - s(\boldsymbol{x})) \xrightarrow{m_a\to\infty, m_{\bar{a}}\to\infty} (\sqrt{\boldsymbol{r}(\boldsymbol{x})^\top \boldsymbol{\Sigma}_a \boldsymbol{r}(\boldsymbol{x})}\tilde{u}, \sqrt{\boldsymbol{r}(\boldsymbol{x})^\top \boldsymbol{\Sigma}_{\bar{a}} \boldsymbol{r}(\boldsymbol{x})}\tilde{u})$ *for a univariate random variable $\tilde{u}$ with zero mean and unit variance. Then the Gelbrich bound is asymptotically tight.*

Proof. Note that for any $a < \bar{a} \in A$, and an optimal comonotonic joint distribution $\mathbb{Q}_{a,\bar{a}}$ of $f(\boldsymbol{x},\tilde{\boldsymbol{\xi}}_a)$ and $f(\boldsymbol{x},\tilde{\boldsymbol{\xi}}_{\bar{a}})$ with marginals $\mathbb{P}_a, \mathbb{P}_{\bar{a}}$ such that $(f(\boldsymbol{x},\tilde{\boldsymbol{\xi}}_a) - \boldsymbol{\mu}_a^\top \boldsymbol{r}(\boldsymbol{x}) - s(\boldsymbol{x}), f(\boldsymbol{x},\tilde{\boldsymbol{\xi}}_{\bar{a}}) - \boldsymbol{\mu}_{\bar{a}}^\top \boldsymbol{r}(\boldsymbol{x}) - s(\boldsymbol{x})) \xrightarrow{m_a\to\infty, m_{\bar{a}}\to\infty} (\sqrt{\boldsymbol{r}(\boldsymbol{x})^\top \boldsymbol{\Sigma}_a \boldsymbol{r}(\boldsymbol{x})}\tilde{u}, \sqrt{\boldsymbol{r}(\boldsymbol{x})^\top \boldsymbol{\Sigma}_{\bar{a}} \boldsymbol{r}(\boldsymbol{x})}\tilde{u})$, we have

$$
\begin{aligned}
&\mathbb{E}_{\mathbb{Q}_{a,\bar{a}}}[|f(\boldsymbol{x},\tilde{\boldsymbol{\xi}}_a) - f(\boldsymbol{x},\tilde{\boldsymbol{\xi}}_{\bar{a}})|^2] = \mathbb{E}_{\mathbb{P}_a}\left[\left(f(\boldsymbol{x},\tilde{\boldsymbol{\xi}}_a) - \boldsymbol{\mu}_a^\top \boldsymbol{r}(\boldsymbol{x}) - s(\boldsymbol{x})\right)^2\right] + \mathbb{E}_{\mathbb{P}_{\bar{a}}}\left[\left(f(\boldsymbol{x},\tilde{\boldsymbol{\xi}}_{\bar{a}}) - \boldsymbol{\mu}_{\bar{a}}^\top \boldsymbol{r}(\boldsymbol{x}) - s(\boldsymbol{x})\right)^2\right] \\
&- 2\mathbb{E}_{\mathbb{P}_a}\left[\left(f(\boldsymbol{x},\tilde{\boldsymbol{\xi}}_a) - \boldsymbol{\mu}_a^\top \boldsymbol{r}(\boldsymbol{x}) - s(\boldsymbol{x})\right)\left(\boldsymbol{\mu}_a^\top \boldsymbol{r}(\boldsymbol{x}) - \boldsymbol{\mu}_{\bar{a}}^\top \boldsymbol{r}(\boldsymbol{x})\right)\right] \\
&- 2\mathbb{E}_{\mathbb{P}_{\bar{a}}}\left[\left(f(\boldsymbol{x},\tilde{\boldsymbol{\xi}}_{\bar{a}}) - \boldsymbol{\mu}_{\bar{a}}^\top \boldsymbol{r}(\boldsymbol{x}) - s(\boldsymbol{x})\right)\left(\boldsymbol{\mu}_a^\top \boldsymbol{r}(\boldsymbol{x}) - \boldsymbol{\mu}_{\bar{a}}^\top \boldsymbol{r}(\boldsymbol{x})\right)\right] \\
&+ \left(\boldsymbol{\mu}_a^\top \boldsymbol{r}(\boldsymbol{x}) - \boldsymbol{\mu}_{\bar{a}}^\top \boldsymbol{r}(\boldsymbol{x})\right)^2 - 2\mathbb{E}_{\mathbb{Q}_{a,\bar{a}}}\left[\left(f(\boldsymbol{x},\tilde{\boldsymbol{\xi}}_a) - \boldsymbol{\mu}_a^\top \boldsymbol{r}(\boldsymbol{x}) - s(\boldsymbol{x})\right)\left(f(\boldsymbol{x},\tilde{\boldsymbol{\xi}}_{\bar{a}}) - \boldsymbol{\mu}_{\bar{a}}^\top \boldsymbol{r}(\boldsymbol{x}) - s(\boldsymbol{x})\right)\right] \\
&\xrightarrow{m_a\to\infty, m_{\bar{a}}\to\infty} \left(\boldsymbol{\mu}_a^\top \boldsymbol{r}(\boldsymbol{x}) - \boldsymbol{\mu}_{\bar{a}}^\top \boldsymbol{r}(\boldsymbol{x})\right)^2 + \left(\sqrt{\boldsymbol{r}(\boldsymbol{x})^\top \boldsymbol{\Sigma}_a \boldsymbol{r}(\boldsymbol{x})} - \sqrt{\boldsymbol{r}(\boldsymbol{x})^\top \boldsymbol{\Sigma}_{\bar{a}} \boldsymbol{r}(\boldsymbol{x})}\right)^2.
\end{aligned}
$$

Thus, the claim follows. □

B.11 Proof of Theorem 7

Theorem 7. *Suppose that for any group $a \in A$, the individual samples $\{\boldsymbol{\xi}_i\}_{i\in C_a}$ satisfy $f(\boldsymbol{x},\boldsymbol{\xi}_i) - \boldsymbol{\mu}_a^\top \boldsymbol{r}(\boldsymbol{x}) - s(\boldsymbol{x}) \overset{d}{=} \sqrt{\boldsymbol{r}(\boldsymbol{x})^\top \boldsymbol{\Sigma}_a \boldsymbol{r}(\boldsymbol{x})}u_i$ for each $i \in C_a$, where $\{u_i\}_{i\in C_a}$ are i.i.d. samples of a univariate sub-Gaussian random variable $\tilde{u}_a$ with zero mean and unit variance, and $\{\tilde{u}_a\}_{a\in A}$ obey the same distribution. Then with probability at most $1-\widehat{\eta}$ such that $\widehat{\eta} > 0$ is small, we have*

$$
v^*(2) - \bar{C}_1(\widehat{\eta}\min_{a\in A}\sqrt{m_a})^{-1} \le v_G \le v^*(2)
$$

for some positive constant $\bar{C}_1$.

Proof. Note that the inequality $v_G \le v^*$ is due to the derivation of the Gelbrich bound. It remains to establish the other direction. For notational convenience, we define $\widehat{\mathbb{P}}_a$ as the true distribution of random variable $s(\boldsymbol{x}) + \boldsymbol{\mu}_a^\top \boldsymbol{r}(\boldsymbol{x}) + \sqrt{\boldsymbol{r}(\boldsymbol{x})^\top \boldsymbol{\Sigma}_a \boldsymbol{r}(\boldsymbol{x})}\tilde{u}_a$ for each $a \in A$, and let $\bar{u}_a$ be the discrete random variable with a uniform distribution on the samples $\{u_i\}_{i\in C_a}$, where $\bar{\mu}_a$ and $\mathrm{var}(\bar{u}_a)$ are the corresponding sample mean and variance, respectively. We also let $\mathbb{P}_a'$ be the empirical distribution of $s(\boldsymbol{x}) + \boldsymbol{\mu}_a^\top \boldsymbol{r}(\boldsymbol{x}) + \sqrt{\boldsymbol{r}(\boldsymbol{x})^\top \boldsymbol{\Sigma}_a \boldsymbol{r}(\boldsymbol{x})}\bar{u}_a$ for each $a \in A$, and $\bar{\mathbb{P}}$ be the joint distribution of all the random variables $\{\tilde{u}_a\}_{a\in A}$. In this case, for any $a < \bar{a} \in A$, the Gelbrich bound reduces to

$$
\left(\boldsymbol{\mu}_a^\top \boldsymbol{r}(\boldsymbol{x}) - \boldsymbol{\mu}_{\bar{a}}^\top \boldsymbol{r}(\boldsymbol{x}) + \sqrt{\boldsymbol{r}(\boldsymbol{x})^\top \boldsymbol{\Sigma}_a \boldsymbol{r}}\bar{\mu}_a - \sqrt{\boldsymbol{r}(\boldsymbol{x})^\top \boldsymbol{\Sigma}_{\bar{a}} \boldsymbol{r}(\boldsymbol{x})}\bar{\mu}_{\bar{a}}\right)^2 +
$$

$$\left(\sqrt{\boldsymbol{r}(\boldsymbol{x})^\top \boldsymbol{\Sigma}_a \boldsymbol{r}(\boldsymbol{x})\mathrm{var}(\bar{u}_a)} - \sqrt{\boldsymbol{r}(\boldsymbol{x})^\top \boldsymbol{\Sigma}_{\bar{a}} \boldsymbol{r}(\boldsymbol{x})\mathrm{var}(\bar{u}_{\bar{a}})}\right)^2.$$

We split the proof into five steps.

Step 1. Following the proof in Theorem 6, for any $a < \bar{a} \in A$, we have

$$W_2^2\left(\widehat{\mathbb{P}}_a, \widehat{\mathbb{P}}_{\bar{a}}\right) = \left(\boldsymbol{\mu}_a^\top \boldsymbol{r}(\boldsymbol{x}) - \boldsymbol{\mu}_{\bar{a}}^\top \boldsymbol{r}(\boldsymbol{x})\right)^2 + \left(\sqrt{\boldsymbol{r}(\boldsymbol{x})^\top \boldsymbol{\Sigma}_a \boldsymbol{r}(\boldsymbol{x})} - \sqrt{\boldsymbol{r}(\boldsymbol{x})^\top \boldsymbol{\Sigma}_{\bar{a}} \boldsymbol{r}(\boldsymbol{x})}\right)^2.$$

According to the triangle inequality, for any $a < \bar{a} \in A$, we have

$$W_2\left(\mathbb{P}'_a, \mathbb{P}'_{\bar{a}}\right) \le W_2\left(\mathbb{P}'_a, \widehat{\mathbb{P}}_a\right) + W_2\left(\widehat{\mathbb{P}}_a, \widehat{\mathbb{P}}_{\bar{a}}\right) + W_2\left(\widehat{\mathbb{P}}_{\bar{a}}, \mathbb{P}'_{\bar{a}}\right).$$

Step 2. In view of Theorem 2 in Fournier and Guillin (2015), for each $a \in A$, there exist two constants $c_{1a} > 0, c_{2a} > 0$ such that for any $\widehat{\eta}_a > 0$, we have

$$\bar{\mathbb{P}}\left\{W_2^2\left(\mathbb{P}'_a, \widehat{\mathbb{P}}_a\right) > \widehat{\eta}_a\right\} \le c_{1a}\exp(-c_{2a} m_a \widehat{\eta}_a).$$

By letting the right-hand side probability be no larger than $\widehat{\delta} \in (0, 0.1)$, we obtain $\widehat{\eta}_a := -\log(c_{1a}/\widehat{\delta})/(c_{2a} m_a)$.

Step 3. We have

$$\begin{aligned}
\left(\boldsymbol{\mu}_a^\top \boldsymbol{r}(\boldsymbol{x}) - \boldsymbol{\mu}_{\bar{a}}^\top \boldsymbol{r}(\boldsymbol{x})\right)^2 \le{} & \left(\boldsymbol{\mu}_a^\top \boldsymbol{r}(\boldsymbol{x}) - \boldsymbol{\mu}_{\bar{a}}^\top \boldsymbol{r}(\boldsymbol{x}) + \sqrt{\boldsymbol{r}(\boldsymbol{x})^\top \boldsymbol{\Sigma}_a \boldsymbol{r}}\bar{\mu}_a - \sqrt{\boldsymbol{r}(\boldsymbol{x})^\top \boldsymbol{\Sigma}_{\bar{a}} \boldsymbol{r}(\boldsymbol{x})}\bar{\mu}_{\bar{a}}\right)^2 \\
& + \left|\boldsymbol{\mu}_a^\top \boldsymbol{r}(\boldsymbol{x}) - \boldsymbol{\mu}_{\bar{a}}^\top \boldsymbol{r}(\boldsymbol{x})\right| \left|\sqrt{\boldsymbol{r}(\boldsymbol{x})^\top \boldsymbol{\Sigma}_a \boldsymbol{r}}\bar{\mu}_a - \sqrt{\boldsymbol{r}(\boldsymbol{x})^\top \boldsymbol{\Sigma}_{\bar{a}} \boldsymbol{r}(\boldsymbol{x})}\bar{\mu}_{\bar{a}}\right| \\
\le{} & \left(\boldsymbol{\mu}_a^\top \boldsymbol{r}(\boldsymbol{x}) - \boldsymbol{\mu}_{\bar{a}}^\top \boldsymbol{r}(\boldsymbol{x}) + \sqrt{\boldsymbol{r}(\boldsymbol{x})^\top \boldsymbol{\Sigma}_a \boldsymbol{r}}\bar{\mu}_a - \sqrt{\boldsymbol{r}(\boldsymbol{x})^\top \boldsymbol{\Sigma}_{\bar{a}} \boldsymbol{r}(\boldsymbol{x})}\bar{\mu}_{\bar{a}}\right)^2 + 4M^2\left|\bar{\mu}_a - \bar{\mu}_{\bar{a}}\right|,
\end{aligned}$$

where $M := \max_{a \in A}\max_{\boldsymbol{x} \in \mathcal{X}}\left\{\max\{\sqrt{\boldsymbol{r}(\boldsymbol{x})^\top \boldsymbol{\Sigma}_a \boldsymbol{r}(\boldsymbol{x})}, |\boldsymbol{\mu}_a^\top \boldsymbol{r}(\boldsymbol{x})|\} : (7\mathrm{b})\right\}$. According to the Chebyshev inequality, for each $a \in A$, the following probabilistic bound holds:

$$\bar{\mathbb{P}}\left\{|\bar{\mu}_a| > \frac{1}{\sqrt{\widehat{\eta} m_a}}\right\} \le \widehat{\eta}.$$

Thus, using the union bound, we obtain

$$\begin{aligned}
\left(\boldsymbol{\mu}_a^\top \boldsymbol{r}(\boldsymbol{x}) - \boldsymbol{\mu}_{\bar{a}}^\top \boldsymbol{r}(\boldsymbol{x})\right)^2 \le{} & \left(\boldsymbol{\mu}_a^\top \boldsymbol{r}(\boldsymbol{x}) - \boldsymbol{\mu}_{\bar{a}}^\top \boldsymbol{r}(\boldsymbol{x}) + \sqrt{\boldsymbol{r}(\boldsymbol{x})^\top \boldsymbol{\Sigma}_a \boldsymbol{r}}\bar{\mu}_a - \sqrt{\boldsymbol{r}(\boldsymbol{x})^\top \boldsymbol{\Sigma}_{\bar{a}} \boldsymbol{r}(\boldsymbol{x})}\bar{\mu}_{\bar{a}}\right)^2 \\
& + 4M^2\left(\sqrt{\frac{1}{\sqrt{\widehat{\eta} m_a}}} + \sqrt{\frac{1}{\sqrt{\widehat{\eta} m_{\bar{a}}}}}\right)^2,
\end{aligned}$$

that is,

$$\begin{aligned}
& \left|\boldsymbol{\mu}_a^\top \boldsymbol{r}(\boldsymbol{x}) - \boldsymbol{\mu}_{\bar{a}}^\top \boldsymbol{r}(\boldsymbol{x})\right| \le \left|\boldsymbol{\mu}_a^\top \boldsymbol{r}(\boldsymbol{x}) - \boldsymbol{\mu}_{\bar{a}}^\top \boldsymbol{r}(\boldsymbol{x}) + \sqrt{\boldsymbol{r}(\boldsymbol{x})^\top \boldsymbol{\Sigma}_a \boldsymbol{r}}\bar{\mu}_a - \sqrt{\boldsymbol{r}(\boldsymbol{x})^\top \boldsymbol{\Sigma}_{\bar{a}} \boldsymbol{r}(\boldsymbol{x})}\bar{\mu}_{\bar{a}}\right| \\
& + 2M\left(\sqrt{\frac{1}{\sqrt{\widehat{\eta} m_a}}} + \sqrt{\frac{1}{\sqrt{\widehat{\eta} m_{\bar{a}}}}}\right),
\end{aligned}$$

with probability at least $1-2\widehat{\eta}$.

Step 4. In view of Theorem 4.7.1 in Vershynin (2018) and the Markov inequality, for each $a \in A$, there exists a constant $c_{3a} > 0$ such that

$$\bar{\mathbb{P}}\left\{|\mathrm{var}(\bar{u}_a)-1| > \frac{2c_{3a}}{\widehat{\eta}\sqrt{m_a}}\right\} \le \widehat{\eta}.$$

Thus, according to the union bound, with probability at least $1-2\widehat{\eta}$, we have

$$\left(\sqrt{\boldsymbol{r}(\boldsymbol{x})^\top\boldsymbol{\Sigma}_a\boldsymbol{r}(\boldsymbol{x})}-\sqrt{\boldsymbol{r}(\boldsymbol{x})^\top\boldsymbol{\Sigma}_{\bar{a}}\boldsymbol{r}(\boldsymbol{x})}\right)^2 \le \left(\sqrt{\boldsymbol{r}(\boldsymbol{x})^\top\boldsymbol{\Sigma}_a\boldsymbol{r}(\boldsymbol{x})\mathrm{var}(\bar{u}_a)}-\sqrt{\boldsymbol{r}(\boldsymbol{x})^\top\boldsymbol{\Sigma}_{\bar{a}}\boldsymbol{r}(\boldsymbol{x})\mathrm{var}(\bar{u}_{\bar{a}})}\right)^2$$
$$+4M^2\left(\sqrt{\frac{2c_{3a}}{\widehat{\eta}\sqrt{m_a}}}+\sqrt{\frac{2c_{3\bar{a}}}{\widehat{\eta}\sqrt{m_{\bar{a}}}}}\right)^2.$$

That is,

$$\left|\sqrt{\boldsymbol{r}(\boldsymbol{x})^\top\boldsymbol{\Sigma}_a\boldsymbol{r}(\boldsymbol{x})}-\sqrt{\boldsymbol{r}(\boldsymbol{x})^\top\boldsymbol{\Sigma}_{\bar{a}}\boldsymbol{r}(\boldsymbol{x})}\right| \le \left|\sqrt{\boldsymbol{r}(\boldsymbol{x})^\top\boldsymbol{\Sigma}_a\boldsymbol{r}(\boldsymbol{x})\mathrm{var}(\bar{u}_a)}-\sqrt{\boldsymbol{r}(\boldsymbol{x})^\top\boldsymbol{\Sigma}_{\bar{a}}\boldsymbol{r}(\boldsymbol{x})\mathrm{var}(\bar{u}_{\bar{a}})}\right|$$
$$+2M\left(\sqrt{\frac{2c_{3a}}{\widehat{\eta}\sqrt{m_a}}}+\sqrt{\frac{2c_{3\bar{a}}}{\widehat{\eta}\sqrt{m_{\bar{a}}}}}\right).$$

Step 5. Combining all the steps together and using the union bound again, we have with probability at most $1-6\widehat{\eta}$,

$$W_2^2(\mathbb{P}_a,\mathbb{P}_{\bar{a}}) \le \left(\boldsymbol{\mu}_a^\top\boldsymbol{r}(\boldsymbol{x})-\boldsymbol{\mu}_{\bar{a}}^\top\boldsymbol{r}(\boldsymbol{x})+\sqrt{\boldsymbol{r}(\boldsymbol{x})^\top\boldsymbol{\Sigma}_a\boldsymbol{r}}\bar{\mu}_a-\sqrt{\boldsymbol{r}(\boldsymbol{x})^\top\boldsymbol{\Sigma}_{\bar{a}}\boldsymbol{r}(\boldsymbol{x})}\bar{\mu}_{\bar{a}}\right)^2+$$
$$+\left(\sqrt{\boldsymbol{r}(\boldsymbol{x})^\top\boldsymbol{\Sigma}_a\boldsymbol{r}(\boldsymbol{x})\mathrm{var}(\bar{u}_a)}-\sqrt{\boldsymbol{r}(\boldsymbol{x})^\top\boldsymbol{\Sigma}_{\bar{a}}\boldsymbol{r}(\boldsymbol{x})\mathrm{var}(\bar{u}_{\bar{a}})}\right)^2+\bar{C}_1(\widehat{\eta}\min_{a\in A}\sqrt{m_a})^{-1},$$

where $\bar{C}_1$ is a positive constant depending on $\{c_{1a},c_{2a},c_{3a}\}_{a\in A}$ and M. Letting $\boldsymbol{x}$ be an optimal solution of DFSO and redefining $\widehat{\eta} := 6\widehat{\eta}$, the conclusion follows. □

B.12 Proof of Theorem 8

Theorem 8. *Suppose that for any pair $a < \bar{a} \in A$, the optimal comonotonic random variables $(f(\boldsymbol{x},\tilde{\boldsymbol{\xi}}_a)-\boldsymbol{\mu}_a^\top\boldsymbol{r}(\boldsymbol{x})-s(\boldsymbol{x}), f(\boldsymbol{x},\tilde{\boldsymbol{\xi}}_{\bar{a}})-\boldsymbol{\mu}_{\bar{a}}^\top\boldsymbol{r}(\boldsymbol{x})-s(\boldsymbol{x})) \xrightarrow{m_a\to\infty, m_{\bar{a}}\to\infty} (\widehat{\boldsymbol{\xi}}_a^\top\boldsymbol{r}(\boldsymbol{x}), \widehat{\boldsymbol{\xi}}_{\bar{a}}^\top\boldsymbol{r}(\boldsymbol{x}))$, where the random vectors $\widehat{c}_a^{-1}\widehat{\boldsymbol{\xi}}_a, \widehat{c}_{\bar{a}}^{-1}\widehat{\boldsymbol{\xi}}_{\bar{a}}$ obey the same distribution with zero mean and covariance matrix $\boldsymbol{\Sigma}_{a\bar{a}}$ for some positive parameters $\widehat{c}_a, \widehat{c}_{\bar{a}}$. Then the Gelbrich bound is asymptotically tight.*

Proof. Note that for any $a < \bar{a} \in A$, and an optimal comonotonic joint distribution $\mathbb{Q}_{a,\bar{a}}$ of $f(\boldsymbol{x},\tilde{\boldsymbol{\xi}}_a)$ and $f(\boldsymbol{x},\tilde{\boldsymbol{\xi}}_{\bar{a}})$ with marginals $\mathbb{P}_a, \mathbb{P}_{\bar{a}}$ such that $(f(\boldsymbol{x},\tilde{\boldsymbol{\xi}}_a)-\boldsymbol{\mu}_a^\top\boldsymbol{r}(\boldsymbol{x})-s(\boldsymbol{x}), f(\boldsymbol{x},\tilde{\boldsymbol{\xi}}_{\bar{a}})-\boldsymbol{\mu}_{\bar{a}}^\top\boldsymbol{r}(\boldsymbol{x})-s(\boldsymbol{x})) \xrightarrow{m_a\to\infty, m_{\bar{a}}\to\infty} (\widehat{\boldsymbol{\xi}}_a^\top\boldsymbol{r}(\boldsymbol{x}), \widehat{\boldsymbol{\xi}}_{\bar{a}}^\top\boldsymbol{r}(\boldsymbol{x}))$ and the random vectors $\widehat{c}_a^{-1}\widehat{\boldsymbol{\xi}}_a, \widehat{c}_{\bar{a}}^{-1}\widehat{\boldsymbol{\xi}}_{\bar{a}}$ obey the same distribution with zero mean and covariance matrix $\boldsymbol{\Sigma}_{a\bar{a}}$, we have

$$\mathbb{E}_{\mathbb{Q}_{a,\bar{a}}}[|f(\boldsymbol{x},\tilde{\boldsymbol{\xi}}_a)-f(\boldsymbol{x},\tilde{\boldsymbol{\xi}}_{\bar{a}})|^2] = \mathbb{E}_{\mathbb{P}_a}\left[\left(f(\boldsymbol{x},\tilde{\boldsymbol{\xi}}_a)-\boldsymbol{\mu}_a^\top\boldsymbol{r}(\boldsymbol{x})-s(\boldsymbol{x})\right)^2\right]+\mathbb{E}_{\mathbb{P}_{\bar{a}}}\left[\left(f(\boldsymbol{x},\tilde{\boldsymbol{\xi}}_{\bar{a}})-\boldsymbol{\mu}_{\bar{a}}^\top\boldsymbol{r}(\boldsymbol{x})-s(\boldsymbol{x})\right)^2\right]$$

$$
\begin{aligned}
&-2\mathbb{E}_{\mathbb{P}_a}\left[\left(f(\boldsymbol{x},\tilde{\boldsymbol{\xi}}_a)-\boldsymbol{\mu}_a^\top\boldsymbol{r}(\boldsymbol{x})-s(\boldsymbol{x})\right)\left(\boldsymbol{\mu}_a^\top\boldsymbol{r}(\boldsymbol{x})-\boldsymbol{\mu}_{\bar{a}}^\top\boldsymbol{r}(\boldsymbol{x})\right)\right]\\
&-2\mathbb{E}_{\mathbb{P}_{\bar{a}}}\left[\left(f(\boldsymbol{x},\tilde{\boldsymbol{\xi}}_{\bar{a}})-\boldsymbol{\mu}_{\bar{a}}^\top\boldsymbol{r}(\boldsymbol{x})-s(\boldsymbol{x})\right)\left(\boldsymbol{\mu}_a^\top\boldsymbol{r}(\boldsymbol{x})-\boldsymbol{\mu}_{\bar{a}}^\top\boldsymbol{r}(\boldsymbol{x})\right)\right]\\
&+\left(\boldsymbol{\mu}_a^\top\boldsymbol{r}(\boldsymbol{x})-\boldsymbol{\mu}_{\bar{a}}^\top\boldsymbol{r}(\boldsymbol{x})\right)^2-2\mathbb{E}_{\mathbb{Q}_{a,\bar{a}}}\left[\left(f(\boldsymbol{x},\tilde{\boldsymbol{\xi}}_a)-\boldsymbol{\mu}_a^\top\boldsymbol{r}(\boldsymbol{x})-s(\boldsymbol{x})\right)\left(f(\boldsymbol{x},\tilde{\boldsymbol{\xi}}_{\bar{a}})-\boldsymbol{\mu}_{\bar{a}}^\top\boldsymbol{r}(\boldsymbol{x})-s(\boldsymbol{x})\right)\right]\\
&\xrightarrow{m_a\to\infty,m_{\bar{a}}\to\infty}\left(\boldsymbol{\mu}_a^\top\boldsymbol{r}(\boldsymbol{x})-\boldsymbol{\mu}_{\bar{a}}^\top\boldsymbol{r}(\boldsymbol{x})\right)^2+\left(\widehat{c}_a\sqrt{\boldsymbol{r}(\boldsymbol{x})^\top\boldsymbol{\Sigma}_{a\bar{a}}\boldsymbol{r}(\boldsymbol{x})}-\widehat{c}_{\bar{a}}\sqrt{\boldsymbol{r}(\boldsymbol{x})^\top\boldsymbol{\Sigma}_{a\bar{a}}\boldsymbol{r}(\boldsymbol{x})}\right)^2.
\end{aligned}
$$

Thus, the claim follows. □

Appendix C. Not MICP-R Functions and Their Piecewise Linear Approximations

When the utility functions are exponential or logarithmic, their corresponding $\mathcal{F}_q$ sets are typically not MICP-R. Hence we propose to approximate them using piecewise linear functions.

PROPOSITION 12. *If $f(\boldsymbol{x},\boldsymbol{\xi})=\exp\left(\boldsymbol{\xi}^\top \boldsymbol{r}(\boldsymbol{x})+s(\boldsymbol{x})\right)$, where $\boldsymbol{r}(\boldsymbol{x})$ and $s(\boldsymbol{x})$ are linear functions, then set $\mathcal{F}_q$, in general, is not MICP-R even when $m=2$ and $|A|=2$.*

Proof. Let us consider a special case of DFSO with $n=2$, $\mathcal{X}=[0,1]^2$, $m=2$, $|A|=2$, $m_a=m_{\bar{a}}=1$, $\epsilon=0.1$ and $f(\boldsymbol{x},\boldsymbol{\xi}_1)=\exp(x_1), f(\boldsymbol{x},\boldsymbol{\xi}_2)=\exp(x_2)$. Under this setting, we have

$$\mathcal{F}_q=\left\{(\boldsymbol{x},\nu)\in[0,1]^2\times\mathbb{R}_+ : |\exp(x_1)-\exp(x_2)|^q\le\nu\right\}.$$

It suffices to show that the sublevel set of the function $|\exp(x_1)-\exp(x_2)|^q$ is not MICP-R. Specifically, letting $\nu=0.1$ in $\mathcal{F}_q$, we consider the set

$$\widehat{\mathcal{F}}_q=\left\{(\boldsymbol{x},\nu)\in[0,1]^2\times\mathbb{R}_+ : |\exp(x_1)-\exp(x_2)|^q\le 0.1\right\}.$$

Then for any two distinct points $\boldsymbol{x}_1,\boldsymbol{x}_2$ satisfying $x_{11}=\log(\exp(x_{21})+\sqrt[q]{0.1}), x_{12}=\log(\exp(x_{22})+\sqrt[q]{0.1})$, and $x_{21},x_{22}\in(0,\log(e-\sqrt[q]{0.1}))$, one can show that their midpoint $(\boldsymbol{x}_1+\boldsymbol{x}_2)/2\notin\widehat{\mathcal{F}}_q$. Indeed, we have

$$\begin{aligned}
&\exp\left(\frac{1}{2}\left(\log(\exp(x_{21})+\sqrt[q]{0.1})+\log(\exp(x_{22})+\sqrt[q]{0.1})\right)\right)-\exp\left(\frac{1}{2}(x_{21}+x_{22})\right)-\sqrt[q]{0.1}>0\\
(\Leftrightarrow)&\sqrt{\exp(x_{21})+\sqrt[q]{0.1}}\sqrt{\exp(x_{22})+\sqrt[q]{0.1}}-\sqrt{\exp(x_{21})}\sqrt{\exp(x_{22})}-\sqrt[q]{0.1}>0\\
(\Leftrightarrow)&\left(\exp(x_{21})+\sqrt[q]{0.1}\right)\left(\exp(x_{22})+\sqrt[q]{0.1}\right)-\left(\sqrt{\exp(x_{21})}\sqrt{\exp(x_{22})}+\sqrt[q]{0.1}\right)^2>0\\
(\Leftrightarrow)&\sqrt[q]{0.1}\exp(x_{21})+\sqrt[q]{0.1}\exp(x_{22})-2\sqrt[q]{0.1}\sqrt{\exp(x_{21})\exp(x_{22})}>0\\
(\Leftrightarrow)&\left(\sqrt{\exp(x_{21})}-\sqrt{\exp(x_{22})}\right)^2>0
\end{aligned}$$

where the last inequality holds due to $x_{21}\neq x_{22}$.

Since there is an infinite number of these points, according to Lemma 2, the set $\widehat{\mathcal{F}}_q$ is not MICP-R. Hence, the set $\mathcal{F}_q$ is not MICP-R. □

Therefore, when $f(\boldsymbol{x},\boldsymbol{\xi})=\exp\left(\boldsymbol{\xi}^\top\boldsymbol{r}(\boldsymbol{x})+s(\boldsymbol{x})\right)$, we propose to use an iterative discretization method to approximate it. Suppose that there are T candidate points $\{\widehat{\boldsymbol{x}}_\tau\}_{\tau\in[T]}$ and their corresponding $g_\tau=\boldsymbol{\xi}^\top\boldsymbol{r}(\widehat{\boldsymbol{x}}_\tau)+s(\widehat{\boldsymbol{x}}_\tau)$, then we approximate $f(\boldsymbol{x},\boldsymbol{\xi})\approx\max_{\tau\in[T]}\exp(g_\tau)+\exp(g_\tau)(\boldsymbol{\xi}^\top\boldsymbol{r}(\boldsymbol{x})+s(\boldsymbol{x})-g_\tau)$. Thus, we obtain the following approximate MICP set of X_i for each $i\in[m]$ as

$$X_i\approx\widehat{X}_i=\left\{(\boldsymbol{x},\bar{w}_i)\in\mathcal{X}\times\mathbb{R}: \begin{array}{c}\bar{w}_i\ge\exp(g_\tau)+\exp(g_\tau)(\boldsymbol{\xi}^\top\boldsymbol{r}(\boldsymbol{x})+s(\boldsymbol{x})-g_\tau),\forall\tau\in[T],\\ \bar{w}_i\le\exp(g_\tau)+\exp(g_\tau)(\boldsymbol{\xi}^\top\boldsymbol{r}(\boldsymbol{x})+s(\boldsymbol{x})-g_{\tau i})+M_{i\tau}(1-z_{i\tau}),\forall\tau\in[T],\\ \sum_{\tau\in[T]}z_{i\tau}=1,z_{i\tau}\in\{0,1\},\forall\tau\in[T]\end{array}\right\},$$

where $M_{i\tau} = \max_{\boldsymbol{x}\in\mathcal{X}}\{\exp(\boldsymbol{\xi}_i^\top \boldsymbol{r}(\boldsymbol{x}) + s(\boldsymbol{x})) - \exp(g_\tau) - \exp(g_\tau)(\boldsymbol{\xi}^\top \boldsymbol{r}(\boldsymbol{x}) + s(\boldsymbol{x}) - g_{\tau i})\}$ for each $\tau \in [T]$.

A similar negative result holds when $f(\boldsymbol{x},\boldsymbol{\xi}) = \log(\boldsymbol{\xi}^\top \boldsymbol{r}(\boldsymbol{x}) + s(\boldsymbol{x}))$.

Proposition 13. *If $f(\boldsymbol{x},\boldsymbol{\xi}) = \log(\boldsymbol{\xi}^\top \boldsymbol{r}(\boldsymbol{x}) + s(\boldsymbol{x}))$, where $\boldsymbol{r}(\boldsymbol{x})$ and $s(\boldsymbol{x})$ are linear functions, and $\boldsymbol{\xi}^\top \boldsymbol{r}(\boldsymbol{x}) + s(\boldsymbol{x}) > 0$ for all $\boldsymbol{x} \in \mathcal{X}$, then set $\mathcal{F}_q$ is not MICP-R even when $m = 2$ and $|A| = 2$.*

Proof. Let us consider a special case of DFSO with $n = 2$, $\mathcal{X} = [1,+\infty)^2$, $m = 2$, $A = |2|$, $m_a = m_{\bar{a}} = 1$ and $f(\boldsymbol{x},\boldsymbol{\xi}_1) = \log(x_1)$, $f(\boldsymbol{x},\boldsymbol{\xi}_2) = \log(x_2)$. Under this setting, we have

$$\mathcal{F}_q = \left\{(\boldsymbol{x},\nu) \in [1,+\infty)^2 \times \mathbb{R}_+ : |\log(x_1) - \log(x_2)|^q \le \nu\right\}.$$

Consider any two distinct points $(\boldsymbol{x}_1,\nu_1),(\boldsymbol{x}_2,\nu_2) \in [1,+\infty)^2 \times \mathbb{R}_+$ such that $x_{11} = x_{12} = 1$, $x_{21} = \exp(\sqrt[q]{\nu_1}) > \exp(q-1)$, $x_{22} = \exp(\sqrt[q]{\nu_2}) > \exp(q-1)$. Note that when $\nu \ge (q-1)^q$, the function $\exp(\sqrt[q]{\nu})$ is convex in ν. It remains to show that their midpoint $((\boldsymbol{x}_1,\nu_1) + (\boldsymbol{x}_2,\nu_2))/2 \notin \mathcal{F}_q$. Indeed, we have

$$\frac{1}{2}\left(\exp(\sqrt[q]{\nu_1}) + \exp(\sqrt[q]{\nu_2})\right) - \exp\left(\sqrt[q]{\left(\frac{1}{2}(\nu_1+\nu_2)\right)}\right) > 0$$

due to the convexity of the function $\exp(\sqrt[q]{\nu})$.

Since there are infinitely many of these points, according to Lemma 2, the set $\mathcal{F}_q$ is not MICP-R. □

Therefore, when $f(\boldsymbol{x},\boldsymbol{\xi}) = \log(\boldsymbol{\xi}^\top \boldsymbol{r}(\boldsymbol{x}) + s(\boldsymbol{x}))$, we also propose to use an iterative discretization method to approximate it. Suppose that there are T candidate points $\{\widehat{\boldsymbol{x}}_\tau\}_{\tau\in[T]}$ and their corresponding $g_\tau = \widehat{\boldsymbol{x}}_\tau^\top \boldsymbol{r}(\widehat{\boldsymbol{x}}_\tau) + s(\widehat{\boldsymbol{x}}_\tau)$, then we approximate $f(\boldsymbol{x},\boldsymbol{\xi}) \approx \min_{\tau\in[T]} \log(g_\tau) + g_\tau^{-1}(\boldsymbol{\xi}^\top \boldsymbol{r}(\boldsymbol{x}) + s(\boldsymbol{x}) - g_\tau)$. Thus, we obtain the following approximate MICP set of X_i for each $i \in [m]$ as

$$X_i \approx \widehat{X}_i = \left\{ (\boldsymbol{x},\bar{w}_i) \in \mathcal{X}\times\mathbb{R} : \begin{array}{c} \bar{w}_i \le \log(g_\tau) + g_\tau^{-1}(\boldsymbol{\xi}^\top \boldsymbol{r}(\boldsymbol{x}) + s(\boldsymbol{x}) - g_\tau), \forall \tau \in [T], \\ \bar{w}_i \ge \log(g_\tau) + g_\tau^{-1}(\boldsymbol{\xi}^\top \boldsymbol{r}(\boldsymbol{x}) + s(\boldsymbol{x}) - g_{\tau i}) - M_{i\tau}(1 - z_{i\tau}), \forall \tau \in [T], \\ \sum_{\tau\in[T]} z_{i\tau} = 1, z_{i\tau} \in \{0,1\}, \forall \tau \in [T] \end{array} \right\},$$

where $M_{i\tau} = \max_{\boldsymbol{x}\in\mathcal{X}}\{-\log(\boldsymbol{\xi}_i^\top \boldsymbol{r}(\boldsymbol{x}) + s(\boldsymbol{x})) + \log(g_\tau) + g_\tau^{-1}(\boldsymbol{\xi}^\top \boldsymbol{r}(\boldsymbol{x}) + s(\boldsymbol{x}) - g_{\tau i})\}$ for each $\tau \in [T]$.

Appendix D. Additional Numerical Results

D.1 Results of Exact MICP Formulations With Jensen Inequality

Table 6 and Table 7 display the numerical results for the Vanilla Formulation, Discretized Formulation, Complementary Formulation, Quantile Formulation, Aggregate Quantile Formulation by adding Jensen bound as a valid inequality. We see that adding the Jensen bound improves the lower bounds for most instances. However, it often yields worse upper bounds and does not help solve the hard instances to optimality. Therefore, the main paper only reports the exact formulation comparison results without having Jensen bound.

Table 6 Results of Exact MICP Formulations (With Jensen Inequality)

m	Vanilla Formulation				Discretized Formulation				Complementary Formulation			
	Obj.Val	LB	Gap (%)	Time	Obj.Val	LB	Gap (%)	Time	Obj.Val	LB	Gap (%)	Time
15	342.44	207.12	39.51	3600.00	342.43	342.43	0.00	57.36	345.38	207.12	40.03	3600.00
20	232.67	89.13	61.69	3600.00	236.27	192.01	18.73	3600.00	245.28	89.13	63.66	3600.00
25	139.71	46.03	67.05	3600.00	158.40	75.62	52.26	3600.00	248.53	46.03	81.48	3600.00
30	177.52	109.56	38.28	3600.00	218.84	109.56	49.94	3600.00	217.74	109.56	49.68	3600.00
35	140.54	92.95	33.86	3600.00	326.31	92.95	71.51	3600.00	963.45	92.95	90.35	3600.00
40	256.82	187.66	26.93	3600.00	583.48	187.66	67.84	3600.00	294.49	187.66	36.27	3600.00
45	223.49	138.17	38.18	3600.00	480.48	138.17	71.24	3600.00	693.37	138.17	80.07	3600.00
50	170.04	116.18	31.67	3600.00	933.76	116.18	87.56	3600.00	915.49	116.18	87.31	3600.00
55	209.61	141.53	32.48	3600.00	576.83	141.53	75.46	3600.00	636.08	141.53	77.75	3600.00
60	131.32	69.87	46.80	3600.00	546.72	69.87	87.22	3600.00	857.62	69.87	91.85	3600.00
65	153.22	83.21	45.70	3600.00	548.73	83.21	84.84	3600.00	1125.53	83.21	92.61	3600.00
70	136.70	83.84	38.67	3600.00	754.04	83.84	88.88	3600.00	962.53	83.84	91.29	3600.00
75	173.96	85.22	51.01	3600.00	703.20	85.22	87.88	3600.00	881.76	85.22	90.34	3600.00
80	159.68	112.03	29.84	3600.00	567.92	112.03	80.27	3600.00	746.74	112.03	85.00	3600.00
85	194.68	104.78	46.18	3600.00	599.51	104.78	82.52	3600.00	563.55	104.78	81.41	3600.00
90	176.64	126.73	28.26	3600.00	807.08	126.73	84.30	3600.00	1713.39	126.73	92.60	3600.00
95	231.74	132.55	42.80	3600.00	749.70	132.55	82.32	3600.00	1144.84	132.55	88.42	3600.00
100	147.47	91.76	37.78	3600.00	683.09	91.76	86.57	3600.00	1427.90	91.76	93.57	3600.00

Table 7 Results of Exact MICP Formulations (With Jensen Inequality)

m	Quantile Formulation				Aggregate Quantile Formulation			
	Obj.Val	LB	Gap (%)	Time	Obj.Val	LB	Gap (%)	Time
15	342.43	342.43	0.00	0.56	342.43	342.40	0.01	0.54
20	230.62	230.62	0.00	5.26	230.62	230.61	0.01	0.49
25	135.03	135.03	0.00	18.58	135.03	135.03	0.00	1.65
30	172.21	172.21	0.00	53.27	172.21	172.20	0.00	5.92
35	133.54	133.54	0.00	911.07	133.54	133.54	0.00	8.81
40	252.42	252.42	0.00	3068.64	252.42	252.42	0.00	15.39
45	219.63	192.58	12.32	3600.00	219.17	219.17	0.00	21.90
50	170.01	120.79	28.95	3600.00	169.99	169.99	0.00	41.36
55	—	143.47	29.93	3600.00	204.77	204.76	0.00	75.79
60	—	79.65	39.12	3600.00	130.84	130.84	0.00	199.56
65	—	83.25	41.59	3600.00	142.55	142.54	0.00	327.37
70	—	83.94	38.24	3600.00	136.33	134.50	1.34	3600.00
75	—	85.22	38.21	3600.00	—	135.69	1.61	3600.00
80	—	112.03	29.35	3600.00	—	155.35	2.03	3600.00
85	—	104.78	28.12	3600.00	—	144.59	0.81	3600.00
90	—	126.73	26.15	3600.00	—	169.96	0.95	3600.00
95	—	132.55	23.74	3600.00	—	171.93	1.09	3600.00
100	—	91.76	35.32	3600.00	—	139.46	1.68	3600.00

D.2 Fair Knapsack

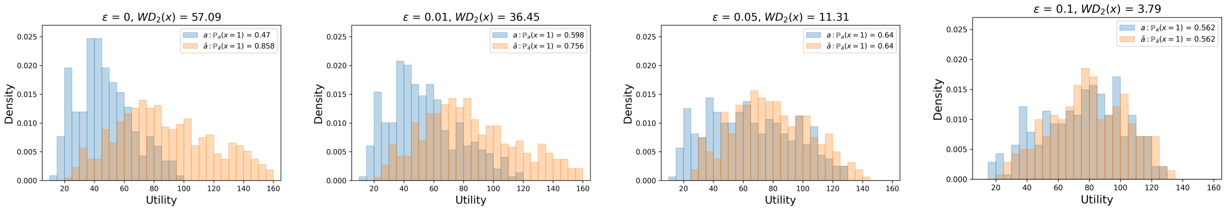
This subsection extends DFSO to the classic knapsack problem. Given weights $\boldsymbol{w} \in \mathbb{Z}_+^m$ and values $\boldsymbol{\xi} \in \mathbb{Z}_+^m$ of a set of m items, the objective of the knapsack problem is to select a subset of items to maximize the total value given that the total weight does not exceed the capacity C. That is, the knapsack problem can be formulated as

$$V^* = \max_{\boldsymbol{x}} \left\{ \sum_{i\in[m]} \xi_i x_i : \sum_{i\in[m]} w_i x_i \leq C, x_i \in \{0,1\}, \forall i \in [m] \right\}, \tag{40}$$

where the binary variable x_i indicates whether item $i \in [m]$ is selected or not. We define the distributional fairness for the knapsack problem, where we choose the utility function of the fair knapsack problem to be the value of item $f(\boldsymbol{x}, \xi_i) = \xi_i x_i$ if being selected for each $i \in [m]$.

In this experiment, we use $A = \{a, \bar{a}\}$ and generate the hypothetical data in the following manner. The weights w_i are drawn from Unif$\{1, 100\}$. The first $\lceil m/2 \rceil$ data points are assigned with the sensitive attribute a, where their values $\{\xi_i\}_{i\in[\lceil m/2\rceil]}$ are drawn independently from $\{\text{Unif}\{w_i + 10, w_i + 30\}\}_{i\in[\lceil m/2\rceil]}$. The remaining data points are assigned with $\bar{a}$, where their values $\{\xi_i\}_{i\in[\lceil m/2\rceil+1,m]}$ are drawn independently from $\{\text{Unif}\{w_i + 20, w_i + 60\}\}_{i\in[\lceil m/2\rceil+1,m]}$. We generate a dataset of $m =$ 1,000 items and choose the capacity $C = 0.5\sum_{i\in[m]} w_i$. We set the inefficiency level parameter to $\epsilon \in \{0.01, 0.05, 0.1\}$. We choose type $q = 2$ Wasserstein fairness and solve DFSO using its AM algorithm in Section 3.5. For ease of illustration, we only display the histograms of each group's utility for selected items $(x = 1)$. The probabilities of selection $\mathbb{P}_a(x = 1)$ and $\mathbb{P}_{\bar{a}}(x = 1)$ are reported.

Figure 4 presents the histograms of utility for fair knapsack. The vanilla knapsack problem (40) has a Wasserstein fairness score of 57.09, where it selects 47% and 85.8% of items from groups a and $\bar{a}$, respectively. In Figures 4(b)-4(d), DFSO improves the Wasserstein fairness score significantly between groups a and $\bar{a}$ by slightly reducing the efficiency. The two groups have the same probability of selected items when $\epsilon \in \{0.05, 0.1\}$. This shows that the proposed approach can achieve distributional fairness for the knapsack problem while maintaining high efficiency.



(a) Vanilla (b) DFSO with $\epsilon = 0.01$ (c) DFSO with $\epsilon = 0.05$ (d) DFSO with $\epsilon = 0.1$

Figure 4 Histograms of Utility for Fair Knapsack