DISTRIBUTIONALLY FAVORABLE OPTIMIZATION: A FRAMEWORK FOR DATA-DRIVEN DECISION-MAKING WITH ENDOGENOUS OUTLIERS

NAN JIANG* AND WEIJUN XIE†

Abstract. A typical data-driven stochastic program seeks the best decision that minimizes the sum of a deterministic cost function and an expected recourse function under a given distribution. Recently, much success has been witnessed in the development of Distributionally Robust Optimization (DRO), which considers the worst-case expected recourse function under the least favorable probability distribution from a distributional family. However, in the presence of endogenous outliers such that their corresponding recourse function values are very large or even infinite, the commonly-used DRO framework alone tends to over-emphasize these endogenous outliers and cause undesirable or even infeasible decisions. On the contrary, Distributionally Favorable Optimization (DFO), concerning the best-case expected recourse function under the most favorable distribution from the distributional family, can serve as a proper measure of the stochastic recourse function and mitigate the effect of endogenous outliers. We show that DFO recovers many robust statistics, suggesting that the DFO framework might be appropriate for the stochastic recourse function in the presence of endogenous outliers. A notion of decision outlier robustness is proposed for selecting a DFO framework for data-driven optimization with outliers. We also provide a unified way to integrate DRO with DFO, where DRO addresses the out-of-sample performance, and DFO properly handles the stochastic recourse function under endogenous outliers. We further extend the proposed DFO framework to solve two-stage stochastic programs without relatively complete recourse. The numerical study demonstrates the framework is promising.

Key words. Distributionally Favorable Optimization; Distributionally Robust Optimization; Robust Statistics

1 Introduction.

In many stochastic programs, their underlying probability distribution $\mathbb{P}$ may not be precisely characterized, whereas empirical data or historical information is often available. Therefore, to hedge against distributional uncertainty, instead of committing to a particular probability distribution, the decision-makers can find their best decisions by first figuring out a family of probability distributions, termed "ambiguity set" (denoted as set $\mathcal{P}$), then optimizing the sum of a deterministic function $\boldsymbol{c}^\top \boldsymbol{x}$ and the worst-case expected recourse function $\mathbb{E}_{\mathbb{P}}[Q(\boldsymbol{x}, \tilde{\boldsymbol{\xi}})]$ with respect to the least favorable distribution $\mathbb{P} \in \mathcal{P}$. This type of model is known as Distributionally Robust Optimization (DRO) of the form

$$\min_{\boldsymbol{x}\in\mathcal{X}} \left\{ \boldsymbol{c}^\top \boldsymbol{x} + \sup_{\mathbb{P}\in\mathcal{P}} \mathbb{E}_{\mathbb{P}} \left[Q(\boldsymbol{x}, \tilde{\boldsymbol{\xi}}) \right] \right\}, \tag{1.1}$$

where $\mathcal{X} \subseteq \mathbb{R}^n$ is a deterministic set and $\mathcal{P} \subseteq \{\mathbb{P}\colon \mathbb{P}\{\tilde{\boldsymbol{\xi}} \in \mathcal{U}\} = 1\}$ with support $\mathcal{U} \subseteq \mathbb{R}^m$ (also known as "uncertainty set" throughout this paper). The DRO model (1.1) has successfully addressed many decision-making problems under uncertainty to achieve decision robustness, and better out-of-sample performance guarantees (see the discussions in [20, 47, 62, 68]). The inherent assumption in DRO is that the expectation of the recourse function is finite for any distribution $\mathbb{P}$ from the ambiguity set $\mathcal{P}$. This assumption may not hold when the data used to construct the ambiguity set are contaminated, i.e., in the presence of outliers. We first introduce two notions of outliers, which are formally defined below:

- For a given ball $\mathbb{B}(\widehat{\boldsymbol{\xi}}, \delta)$ around a scenario $\widehat{\boldsymbol{\xi}}$ with radius $\delta > 0$, the scenario $\widehat{\boldsymbol{\xi}}$ is an "exogenous outlier" when $\mathbb{P}_0\{\tilde{\boldsymbol{\xi}} : \tilde{\boldsymbol{\xi}} \in \mathbb{B}(\widehat{\boldsymbol{\xi}}, \delta)\} = 0$ for a given probability distribution $\mathbb{P}_0$;
- For a given large number M_1, a scenario $\widehat{\boldsymbol{\xi}}$ is an "endogenous outlier" when the recourse function value $Q(\boldsymbol{x}, \widehat{\boldsymbol{\xi}}) > M_1$ for some $\boldsymbol{x} \in \mathcal{X}$.

Notice that exogenous outliers are independent from the decision variable $\boldsymbol{x} \in \mathcal{X}$, i.e., exogenous outliers are caused by abnormal data measurement or intentional data distortion. The definition of exogenous outliers dates back to the work [5] and we rephrase the definition based on the statistical properties. The endogenous outliers are from the intrinsic property of the problem itself and are latently dependent on the decision variable $\boldsymbol{x} \in \mathcal{X}$, i.e., the recourse function value may be very large or even unbounded under some extreme scenarios for certain decisions. Since exogenous outliers can be easily detected by preprocessing via a properly-selected robust statistic, in this regard, this work mainly focuses on endogenous outliers. Under such circumstances, the DRO model (1.1) tends to over-emphasize the endogenous outliers and causes undesirable or infeasible decisions. In light of this issue, this paper studies the following Distributionally

*H. Milton Stewart School of Industrial and Systems Engineering, Georgia Institute of Technology, Atlanta, GA 30332 nanjiang@gatech.edu

†H. Milton Stewart School of Industrial and Systems Engineering, Georgia Institute of Technology, Atlanta, GA 30332 wxie@gatech.edu

Favorable Optimization (DFO) by providing a proper measure to mitigate the effect of endogenous outliers

$$v^* = \min_{\boldsymbol{x}\in\mathcal{X}} \left\{ \boldsymbol{c}^\top \boldsymbol{x} + \inf_{\mathbb{P}\in\mathcal{P}} \mathbb{E}_{\mathbb{P}} \left[Q(\boldsymbol{x}, \tilde{\boldsymbol{\xi}}) \right] \right\}, \tag{1.2}$$

which instead seeks the best decision under the most favorable distribution. We formally define a notion of decision outlier robustness for selecting a proper DFO in Section 3. It is worthy of mentioning that since DRO can achieve better out-of-sample performance guarantees, Section 4 studies the worst-case DFO which integrates DRO with DFO.

Note that if there is only support information $\mathcal{U}$ available (i.e., $\mathcal{P} = \{\mathbb{P} \colon \mathbb{P}\{\tilde{\boldsymbol{\xi}} \in \mathcal{U}\} = 1\}$), then the DFO (1.2) degenerates to a regular one (rDFO), i.e.,

$$v^* = \min_{\boldsymbol{x}\in\mathcal{X}} \left\{ \boldsymbol{c}^\top \boldsymbol{x} + \inf_{\boldsymbol{\xi}\in\mathcal{U}} Q(\boldsymbol{x}, \boldsymbol{\xi}) \right\}. \tag{1.3}$$

The special cases of the rDFO (1.3) have been successfully applied in bandit and reinforcement learning literature such as Upper Confidence Bound (UCB) algorithm (see, e.g., [4]), where the DFO framework has been demonstrated to be useful as a tool for uncertainty exploration. However, a thorough study of DFO is missing, in particular, for the decision-making problems under uncertainty. More importantly, our results in Section 2 show that DFO, especially, rDFO, naturally recovers many robust statistics, evidencing that DFO might be desirable for stochastic programming under endogenous outliers. As illustrated in Figure 1, in the presence of endogenous outliers, i.e., $Q(\boldsymbol{x}, \boldsymbol{\xi}) \approx \infty$, DRO may over-emphasize the endogenous outliers, while DFO can mitigate the effect of endogenous outliers.

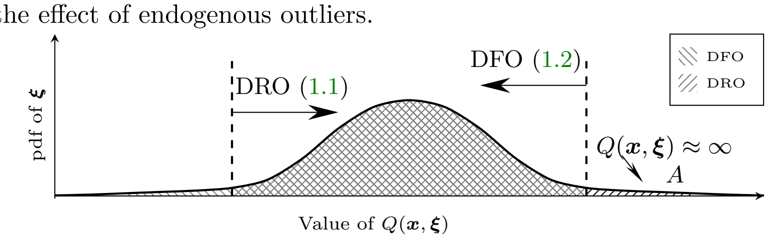

Fig. 1: Illustration of DFO vs. DRO in the Presence of Endogenous Outliers. In region A, due to the effect of endogenous outliers, the recourse function value can be very large or even infinite, where we denote it as "$Q(\boldsymbol{x}, \boldsymbol{\xi}) \approx \infty$."

As mentioned above, the study of DFO is motivated by optimization problems highly affected by endogenous outliers. Throughout the paper, we make the following assumptions for DFO (1.2).

Assumption 1. *(i) Set $\mathcal{X}$ is convex, compact, and has a non-empty interior; and*
(ii) The recourse function $Q(\boldsymbol{x}, \boldsymbol{\xi})$ is bounded below by a constant $-M$ for all $\boldsymbol{x} \in \mathcal{X}$ and $\boldsymbol{\xi} \in \mathcal{U}$.

Both parts in Assumption 1 are standard in literature (see, e.g., section 5 in [7] and chapter 12 in [53]). Part (i) in Assumption 1 is useful to derive big-M coefficients. Part (ii) in Assumption 1 ensures that any expectation of the recourse function is bounded from below, which is particularly useful for the notion of decision outlier robustness in Section 3.

1.1 Motivating Examples.

In this subsection, we provide two examples to illustrate the importance of the DFO framework. The first example uses the DFO framework to explain the connection between chance constrained programming and robust optimization.

Example 1. **Chance Constrained Programming.** Some endogenous outliers can make the problem infeasible in the robust optimization, thus causing the decisions to be practically meaningless (see more discussions in [6]). However, since some extreme scenarios are highly unlikely to occur, to avoid such over-conservatism in robust optimization, the authors in [6] mentioned that "there is no need to care about such highly improbable scenarios" and suggested using the chance constrained programming as a better alternative, which can be well justified through the lens of DFO. In the DFO (1.2), if the objective of the recourse function is 0 with the uncertain inequalities $G(\boldsymbol{x}, \boldsymbol{\xi}) \le 0$, where $G(\cdot,\cdot) : \mathbb{R}^n \times \mathbb{R}^m \to \mathbb{R}$ is a continuous function, i.e., $Q(\boldsymbol{x}, \tilde{\boldsymbol{\xi}}) = \min\{0 \colon G(\boldsymbol{x}, \tilde{\boldsymbol{\xi}}) \le 0\}$ and $\tilde{\boldsymbol{\xi}}$ follows distribution $\mathbb{P}_0$, then the corresponding DFO (1.2) resorts to

$$\min_{\boldsymbol{x}\in\mathcal{X}} \left\{ \boldsymbol{c}^\top \boldsymbol{x} \colon G(\boldsymbol{x}, \boldsymbol{\xi}) \le 0, \forall \boldsymbol{\xi} \in \mathcal{U} \right\} = \min_{\boldsymbol{x}\in\mathcal{X}} \left\{ \boldsymbol{c}^\top \boldsymbol{x} \colon \mathbb{E}_{\mathbb{P}_0} \left[\mathbb{I} \left(G(\boldsymbol{x}, \tilde{\boldsymbol{\xi}}) > 0 \right) \right] \le 0 \right\}. \tag{1.4a}$$

where support $\mathcal{U} := \mathrm{supp}(\mathbb{P}_0)$. This is indeed a conventional robust optimization problem. Applying the following interval ambiguity set, i.e., $\mathcal{P}_I = \{\mathbb{P} : \mathbb{P}(\mathcal{U}) = 1, 0 \preceq \mathbb{P} \preceq \mathbb{P}_0/(1-\varepsilon)\}$ with $\varepsilon \in (0,1)$, the DFO counterpart of the robust optimization (1.4a) can be written as

$$v^* = \min_{\boldsymbol{x}\in\mathcal{X}} \left\{ \boldsymbol{c}^\top \boldsymbol{x} \colon \inf_{\mathbb{P}\in\mathcal{P}_I} \mathbb{E}_{\mathbb{P}} \left[\mathbb{I}\left(G(\boldsymbol{x}, \tilde{\boldsymbol{\xi}}) > 0 \right) \right] \le 0 \right\}, \tag{1.4b}$$

and can be further reduced to a regular chance constrained program. The formal derivations can be found in Proposition A.1 of Appendix A. ⋄

The link between chance constrained programming and robust optimization shows that applying the DFO framework reduces the over-conservatism of robust optimization and explains why a chance constrained program can be less conservative.

The second example focuses on a two-stage stochastic program without relatively complete recourse, where endogenous outliers can cause the underlying problem to be infeasible. The condition of relatively complete recourse states that given a reference distribution $\mathbb{P}_0$, the finiteness of recourse function $Q(\boldsymbol{x}, \tilde{\boldsymbol{\xi}}) < \infty$ holds for every $\boldsymbol{x} \in \mathcal{X}$ and $\mathbb{P}_0$-almost every $\tilde{\boldsymbol{\xi}} \in \mathcal{U}$. This condition guarantees the feasibility of the second-stage problem, and this concept has been elaborated in [56, 65]. However, many problems in practice genuinely fail to have relatively complete recourse, i.e., warehouses may not fulfill the demand due to the disruptions of extreme scenarios. When the second-stage problem can be infeasible, i.e., for the two-stage stochastic program without relatively complete recourse, the optimal objective value of that two-stage problem does not exist. In this case, we adopt the convention that $\mathbb{E}_{\mathbb{P}_0}[Q(\boldsymbol{x}, \tilde{\boldsymbol{\xi}})] = \infty$ for a given reference distribution $\mathbb{P}_0$. We show that DFO serves as a proper measure to address infeasibility, reduces the effect of endogenous outliers, and delivers desirable decisions. It is worth mentioning that our DFO framework does not remove the endogenous outliers, but we change the corresponding probability measures of the endogenous outliers to ensure that the corresponding objective value is finite.

Example 2. **Endogenous Outliers in Two-stage Stochastic Programs without Relatively Complete Recourse.** Consider the following two-stage stochastic program:

$$\min_{x\ge 1} \left\{ x + \mathbb{E}_{\mathbb{P}_0} \left[Q(x, \tilde{\xi}) := \min_{y\in\mathcal{Y}} \left\{ y \colon |\tilde{\xi}| y \ge x \right\} \right] \right\},$$

where the set $\mathcal{Y} = \{y : 0 \le y \le 10\}$ and $\tilde{\xi}$ follows the standard Gaussian distribution $\mathbb{P}_0$, i.e., $\tilde{\xi} \sim \mathcal{N}(0,1)$ (see, e.g., Figure 2). Under this setting, due to the lack of relatively complete recourse, the two-stage stochastic program is infeasible, and so is its DRO counterpart. If the machine learning techniques were employed to preprocess the data ξ to resolve the infeasibility, one may simply relegate the region A or region C or both as outliers since they belong to light-tail parts. However, the problem remains infeasible, and the actual endogenous outliers (i.e., region B) may not be detected unless exploring the optimization problem structure. On the other hand, applying DFO can properly mitigate the effect of the endogenous outliers and address the infeasibility issue using the similar interval ambiguity set in Example 1, i.e., $\mathcal{P}_I = \{\mathbb{P} : \mathbb{P}(\mathcal{U}) = 1, 0 \preceq \mathbb{P} \preceq \mathbb{P}_0/(2 - 2\Phi(0.1))\}$ and $\Phi(\cdot)$ denotes the cumulative distribution function of the standard normal distribution. Thus, let us consider the following DFO:

$$\min_{x\ge 1} \left\{ x + \inf_{\mathbb{P}\in\mathcal{P}_I} \mathbb{E}_{\mathbb{P}} \left[Q(x, \tilde{\xi}) := \min_{y\in\mathcal{Y}} \left\{ y \colon |\tilde{\xi}| y \ge x \right\} \right] \right\} = 1 + \frac{1}{2 - 2\Phi(0.1)} 2 \left[\int_{0.1}^{\infty} \frac{1}{\xi} \frac{1}{\sqrt{2\pi}} e^{-\frac{\xi^2}{2}} d\xi \right] = 3.049.$$

Thus, the resulting favorable two-stage problem is feasible and mitigates the effect of endogenous outliers. We provide more detailed discussions in Section 2.3. ⋄

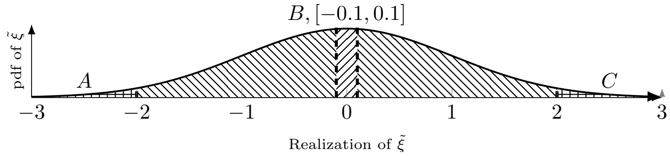

Fig. 2: Illustration of Example 2.

1.2 Literature Review. In literature, in contrast to DRO (see more details in [54]), researchers tend to use optimistic optimization (i.e., special cases of DFO) to tackle learning problems in various areas such as reinforcement learning [1, 67], Bayesian optimization [49–51], classification [10], image reconstruction [26], machine learning [52], etc. For instance, the authors in [67] applied the optimistic DRO approach to the trust-region constrained optimization problem in reinforcement learning and obtained the globally optimal policy in each iteration. The trade-off between exploration and exploitation in reinforcement learning has been discussed using optimistic optimization in [1]. In [50], the authors found that when using the Wasserstein distance, the optimistic likelihood problem can be interpreted as solving a linear program using a greedy heuristic, where the decay pattern is an exponential kernel approximation. They also provided the theoretical guarantees for the variational posterior inference problems under the KL divergence and the Wasserstein distance. The work [51] introduced a novel moment-based divergence ambiguity set and proposed a Bayesian contextual classification model using an optimistic score ratio. The researchers in [49] developed the optimistic likelihood, which can be reduced to a one-dimensional convex optimization problem. In [26], the authors investigated the favorable chance constrained problem, derived the conic reformulation, demonstrated the limits of tractability, and showed its effectiveness in image reconstruction. However, all of these works lack evidence to connect robust statistics and DFO, where a robust statistic aims to yield a good performance when the data are contaminated, as discussed in the literature for decades [34, 45].

There are also a few works focusing on special classes of the rDFO problems (see, e.g., [10, 52]). The work [10] proposed a novel formulation of support vector classification and derived a geometric interpretation of the proposed formulation to handle the uncertainty in classification. In [52], the authors argued that the optimistic assumption could be easier to realize regarding real-world economic resources compared with the pessimistic or worst-case one. However, the literature lacks a framework for DFO or optimistic optimization, and the connection to robust statistics is also missing. This paper fills the gap.

While this paper was prepared to submit, we became aware of the independent works from [12, 21], which discussed the class of distributionally optimistic optimization problems and their applications to contextual bandit problems. The fundamental difference between this work and theirs is that we focus on data-driven optimization with endogenous outliers, connecting to and motivating from robust statistics.

1.3 Summary of Contributions. In this paper, we study DFO (1.2) via various perspectives from statistics, machine learning, and optimization. Each perspective justifies and extends DFO. Particularly, we show the following two fundamental aspects of DFO: framework and unification.

- For the framework aspect, we show that DFO can recover many robust statistics. We also show that in the presence of endogenous outliers, DFO can be a proper framework for decision-making. We introduce a new notion of decision outlier robustness that is easy to check and is useful to characterize whether a DFO model is indeed decision outlier robust.
- For the unification aspect, we integrate DRO with DFO, termed "worst-case DFO," since DRO improves the out-of-sample performance given that the sample size is finite. We show a proper way to integrate both. In particular, we focus on the data-driven ambiguity set for DRO and decision outlier robust ambiguity set for DFO. The convergence analysis shows that the error of the worst-case DFO decreases proportionally to the square root of the sample size. On the other hand, the decision outlier robustness notion also suggests that while the same rate of convergence can be guaranteed, the ambiguity set of DRO should not be too large (i.e., never be overly pessimistic).

The roadmap of contributions in our paper is shown in Figure 3.

Organization. The remainder of the paper is organized as follows. Section 2 shows the equivalence between DFO and many robust statistics and introduces the DFO framework for data-driven optimization with endogenous outliers. Section 3 introduces the notion of decision outlier robustness and Section 4 integrates distributional robustness with DFO to achieve better out-of-sample performance guarantees. Section 5 numerically illustrates the proposed methods. Section 6 concludes the paper.

Notation. The following notation is used throughout the paper. We use bold letters (e.g., $\boldsymbol{x}, \boldsymbol{A}$) to denote vectors and matrices and use corresponding non-bold letters to denote their components. We let $\|\cdot\|_*$ denote the dual norm of a general norm $\|\cdot\|$. We let $\boldsymbol{e}$ be the vector or matrix of all ones, and let $\boldsymbol{e}_i$ be the ith standard basis vector. Given an integer n, we let $[n] := \{1, 2, \ldots, n\}$, and use $\mathbb{R}^n_+ := \{\boldsymbol{x} \in \mathbb{R}^n : x_i \geq 0, \forall i \in [n]\}$. Given a real number t, we let $(t)_+ := \max\{t, 0\}$ and $(t)_- := \min\{t, 0\}$. Given a finite set I, we let $|I|$ denote its cardinality. We let $\tilde{\boldsymbol{\xi}}$ denote a random vector and denote its realizations by $\boldsymbol{\xi}$. Given a vector $\boldsymbol{x} \in \mathbb{R}^n$, let $\text{supp}(\boldsymbol{x})$ be its support, i.e., $\text{supp}(\boldsymbol{x}) := \{i \in [n] : x_i \neq 0\}$. Given a probability distribution $\mathbb{P}$ defined on support $\mathcal{U}$ with sigma-algebra $\mathcal{F}$ and a $\mathbb{P}$-measurable function $g(\boldsymbol{\xi})$, we use $\mathbb{P}\{A\}$ to denote $\mathbb{P}\{\tilde{\boldsymbol{\xi}} : \text{condition } A(\tilde{\boldsymbol{\xi}}) \text{ holds}\}$ when $A(\tilde{\boldsymbol{\xi}})$ is a condition on $\boldsymbol{\xi}$, and to denote $\mathbb{P}\{\tilde{\boldsymbol{\xi}} : \tilde{\boldsymbol{\xi}} \in A\}$ when $A \in \mathcal{F}$ is $\mathbb{P}$-measurable, and we let $\text{ess.sup}_{\mathbb{P}}(g(\tilde{\boldsymbol{\xi}}))$ denote the essential supremum of the deterministic function $g(\tilde{\boldsymbol{\xi}})$. We define a nonnegative measure $\boldsymbol{\mu}$ as $\boldsymbol{\mu} \succeq 0$ when $\boldsymbol{\mu}(A) \geq 0$ for any $A \in \mathcal{F}$, and further define $\boldsymbol{\mu}_2 \succeq \boldsymbol{\mu}_1$

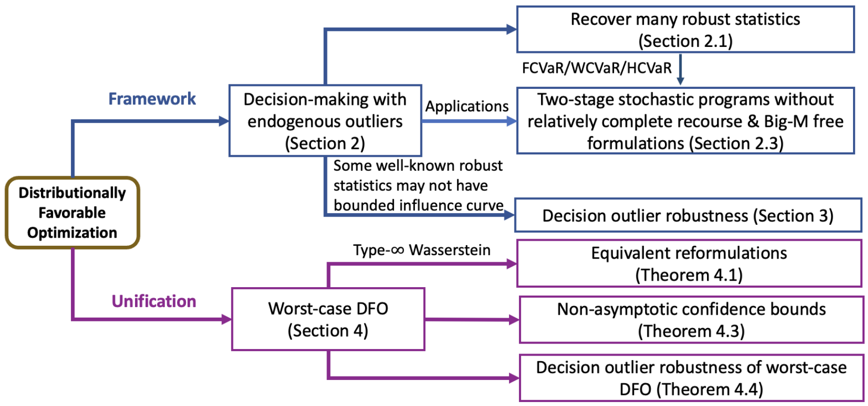

Fig. 3: A Roadmap of the Main Results in This Paper.

if $\boldsymbol{\mu}_2 - \boldsymbol{\mu}_1 \succeq 0$ for any two measures $\boldsymbol{\mu}_1, \boldsymbol{\mu}_2$. We use $\otimes$ to denote the Kronecker product. Given a set R, the characteristic function $\chi_R(\boldsymbol{x}) = 0$ if $\boldsymbol{x} \in R$, and ∞, otherwise; the indicator function $\mathbb{I}(\boldsymbol{x} \in R) = 1$ if $\boldsymbol{x} \in R$, and 0, otherwise. We let δ_ω denote for the Dirac distribution that places unit mass on the realization ω. We use $\lfloor x \rfloor$ to denote the largest integer y satisfying $y \leq x$, for any $x \in \mathbb{R}$. Additional notations will be introduced as needed.

2 DFO: A Framework to Handle Data-driven Stochastic Programs with Endogenous Outliers. Different from DRO, in this section, we show that DFO can be useful in mitigating the effect of endogenous outliers. We first show that DFO, especially, rDFO, recovers many robust statistics, which can be more desirable for decision-making under uncertainty in the presence of endogenous outliers.

2.1 DFO Recovers Many Robust Statistics. In the literature, robust statistical approaches can effectively provide stable portfolio strategies [19, 74]. For example, the authors in [74] introduced several robust statistical methods to reduce the influence of outliers. Coincidently, DFO can recover many robust statistics, which are detailed in this subsection.

Case I. Least Trimmed Squares. The least trimmed squares (LTS) is a robust regression method that learns from a subset of data not being affected by endogenous outliers (see, e.g., [58]). Given N data points $\{\bar{\boldsymbol{x}}_i, \bar{y}_i\}_{i\in[N]} \subseteq \mathbb{R}^d \times \mathbb{R}$, LTS aims to find an estimator $\boldsymbol{\beta}$ that minimizes the sum of squared residuals over the most favorable size-k subset with an integer $k \in [N]$, i.e., suppose the squared residuals $\boldsymbol{r}^2(\boldsymbol{\beta})$, defined as $r_i^2(\boldsymbol{\beta}) := (\bar{y}_i - \bar{\boldsymbol{x}}_i^\top \boldsymbol{\beta})^2$ for each $i \in [N]$, are sorted in ascending order $r_{(1)}^2(\boldsymbol{\beta}) := (\bar{y}_{(1)} - \bar{\boldsymbol{x}}_{(1)}^\top \boldsymbol{\beta})^2 \leq r_{(2)}^2(\boldsymbol{\beta}) \leq \cdots \leq r_{(N)}^2(\boldsymbol{\beta}) := (\bar{y}_{(N)} - \bar{\boldsymbol{x}}_{(N)}^\top \boldsymbol{\beta})^2$, where $\{(i)\}_{i\in[N]}$ denotes a permutation of set $[N]$. Then the LTS is equivalent to

$$\min_{\boldsymbol{\beta}} \frac{1}{k} \sum_{i\in[k]} r_{(i)}^2(\boldsymbol{\beta}).$$

We can apply the following DFO to recover the LTS, that is,

$$v^* = \min_{\boldsymbol{\beta}} \min_{\boldsymbol{p}\in\mathcal{P}_I} \sum_{i\in[N]} p_i r_i^2(\boldsymbol{\beta}), \tag{2.1}$$

where the interval ambiguity set $\mathcal{P}_I$ is written as $\mathcal{P}_I = \{\boldsymbol{p} \in \mathbb{R}_+^N : \sum_{i\in[N]} p_i = 1, 0 \leq p_i \leq 1/k\}$. A simple calculation shows that the corresponding DFO indeed returns the LTS, that is,

$$v^* = \min_{\boldsymbol{\beta}} \min_{\boldsymbol{p}\in\mathcal{P}_I} \sum_{i\in[N]} p_i r_i^2(\boldsymbol{\beta}) = \min_{\boldsymbol{\beta}} \frac{1}{k} \sum_{i\in[k]} r_{(i)}^2(\boldsymbol{\beta}).$$

We remark that in the above formulation, the DFO recovers LTS by selecting k favorable scenarios and increasing their probability from $1/N$ to $1/k$. Motivated by this case, we show in Section 3 that DFO with interval ambiguity set is equivalent to favorable conditional value-at-risk (FCVaR).

Case II. Winsorized Regression. Winsorized regression (see, e.g., [78]), an effective alternative to the ordinary least-square regression, can reduce the effect of outliers. It involves the calculation of the residual values by replacing the extremal residual values that are beyond an interval with the nearest boundary values. For a fixed $\boldsymbol{\beta}$ and N data points $\{\bar{\boldsymbol{x}}_i, \bar{y}_i\}_{i\in[N]} \subseteq \mathbb{R}^d \times \mathbb{R}$, let the squared residuals $r_i^2(\boldsymbol{\beta}) := (\bar{y}_i - \bar{\boldsymbol{x}}_i^\top \boldsymbol{\beta})^2$ for each $i \in [N]$ and let $r_{(k)}^2(\boldsymbol{\beta})$ be the kth smallest squared residual with an integer number $k \in [N]$. The Winsorized regression can be formulated as

$$\min_{\boldsymbol{\beta}} \frac{1}{N} \sum_{i\in[N]} \min\left\{r_i^2(\boldsymbol{\beta}), r_{(k)}^2(\boldsymbol{\beta})\right\}.$$

The following DFO recovers the Winsorized regression:

$$v^* = \min_{\boldsymbol{\beta}} \min_{\mathbb{P}\in\mathcal{P}(\boldsymbol{\beta})} \mathbb{E}_{\mathbb{P}}[\tilde{\xi}],$$

where the decision-dependent ambiguity set $\mathcal{P}(\boldsymbol{\beta})$ is defined as

$$\mathcal{P}(\boldsymbol{\beta}) = \left\{ \frac{1}{N}\sum_{i\in[N]} \mathbb{P}_i : \begin{array}{l} \mathbb{P}_i\left\{\tilde{\xi} : \tilde{\xi} = r_i^2(\boldsymbol{\beta})\right\} + \mathbb{P}_i\left\{\tilde{\xi} : \tilde{\xi} = r_{(k)}^2(\boldsymbol{\beta})\right\} = 1, \forall i \in [N], \\ \mathbb{P}_i(\mathcal{U}) = 1, \forall i \in [N] \end{array} \right\},$$

with support $\mathcal{U} = \mathbb{R}_+$. The result can also be extended to recover the Ramp loss support vector machine, where the latter was studied in work [33].

Case III. Huber-skip Estimator [34]. Given N data points $\{\bar{\boldsymbol{x}}_i, \bar{y}_i\}_{i\in[N]} \subseteq \mathbb{R}^d \times \mathbb{R}$, suppose the residual $r_i(\boldsymbol{\beta}) = (\bar{y}_i - \bar{\boldsymbol{x}}_i^\top \boldsymbol{\beta})$ for each $i \in [N]$. The Huber-skip estimator truncates the observations with large residuals to mitigate the influence of endogenous outliers, which admits the following formulation

$$\min_{\boldsymbol{\beta}} \frac{1}{N} \sum_{i\in[N]} \min\{r_i^2(\boldsymbol{\beta}), H\},$$

where $H \geq 0$ is the given threshold.

We can apply the following DFO to recover the Huber-skip estimator

$$v^* = \min_{\boldsymbol{\beta}} \min_{\mathbb{P}\in\mathcal{P}(\boldsymbol{\beta})} \mathbb{E}_{\mathbb{P}}[\tilde{\xi}],$$

where the decision-dependent ambiguity set $\mathcal{P}(\boldsymbol{\beta})$ is defined as

$$\mathcal{P}(\boldsymbol{\beta}) = \left\{ \frac{1}{N}\sum_{i\in[N]} \mathbb{P}_i : \begin{array}{l} \mathbb{P}_i\left\{\tilde{\xi} : \tilde{\xi} = r_i^2(\boldsymbol{\beta})\right\} + \mathbb{P}_i\left\{\tilde{\xi} : \tilde{\xi} = H\right\} = 1, \forall i \in [N], \\ \mathbb{P}_i(\mathcal{U}) = 1, \forall i \in [N] \end{array} \right\},$$

with support $\mathcal{U} = \mathbb{R}_+$.

We conclude this section by remarking that DFO can recover many other robust statistics and some machine learning problems. Due to page limit and in agreement with the editor, we relegate additional examples to this extended online technical report version [38], i.e., median in Appendix B.1, Huber estimator and Tukey's bisquare estimator in Appendix B.3, quantile regression in Appendix B.4, and other machine learning examples in Appendix B.5 of [38]. As far as the authors are concerned, there is no prior work on recovering robust statistics using DFO or optimistic optimization. The connections between the DFO framework and robust statistics further show that DFO can be a proper way to handle decision-making under uncertainty in the presence of endogenous outliers, which is illustrated below in detail.

2.2 From Robust Statistics to Decision-making under Uncertainty: DFO Mitigates the Effect of Endogenous Outliers for Stochastic Programming. For a stochastic program with endogenous outliers, motivated by robust statistics, this subsection focuses on a special family of DFO with the interval ambiguity set–the Favorable Conditional Value-at-Risk (FCVaR) as a demonstration and briefly introduces its alternatives. For a given random variable $\tilde{\boldsymbol{X}}$ with probability distribution $\mathbb{P}_0$, cumulative distribution function $F_{\mathbb{P}_0}(\cdot)$, and risk level $\varepsilon \in (0,1)$, the VaR of $\tilde{\boldsymbol{X}}$ is defined as

$$\mathbb{P}_0\text{-VaR}_{1-\varepsilon}(\tilde{\boldsymbol{X}}) := \min_{s}\left\{s : F_{\mathbb{P}_0}(s) \geq 1-\varepsilon\right\},$$

the corresponding FCVaR of $\tilde{\boldsymbol{X}}$ is defined as

$$\mathbb{P}_0\text{-FCVaR}_{1-\varepsilon}(\tilde{\boldsymbol{X}}) := \max_{\beta}\left\{\beta + \frac{1}{1-\varepsilon}\mathbb{E}_{\mathbb{P}_0}\left[\left(\tilde{\boldsymbol{X}} - \beta\right)_-\right]\right\}. \tag{2.2}$$

Roughly speaking, FCVaR (2.2) can be interpreted as the average of the values no larger than $\mathbb{P}_0\text{-VaR}_{1-\varepsilon}(\tilde{\boldsymbol{X}})$.

Proposition 2.1. *(i) Given an interval ambiguity set $\mathcal{P}_I = \{\mathbb{P} : \mathbb{P}(\mathcal{U}) = 1, 0 \preceq \mathbb{P} \preceq \mathbb{P}_0/(1-\varepsilon)\}$ with support $\mathcal{U} = \mathrm{supp}(\mathbb{P}_0)$, we have*

$$\inf_{\mathbb{P}\in\mathcal{P}_I} \mathbb{E}_{\mathbb{P}}\left[\tilde{\boldsymbol{X}}\right] = \max_{\beta}\left\{\beta + \frac{1}{1-\varepsilon}\mathbb{E}_{\mathbb{P}_0}\left[\left(\tilde{\boldsymbol{X}} - \beta\right)_-\right]\right\} = \mathbb{P}_0\text{-FCVaR}_{1-\varepsilon}\left(\tilde{\boldsymbol{X}}\right); \tag{2.3a}$$

(ii) An optimal solution of the right-hand side optimization problem (2.2) is $\beta^* = \mathbb{P}_0\text{-VaR}_{1-\varepsilon}(\tilde{\boldsymbol{X}})$; and

(iii) The $\mathbb{P}_0\text{-FCVaR}_{1-\varepsilon}(\tilde{\boldsymbol{X}})$ can be bounded by two conditional expectations:

$$\mathbb{E}_{\mathbb{P}}\left[\tilde{\boldsymbol{X}}\middle|\tilde{\boldsymbol{X}} < \mathbb{P}_0\text{-VaR}_{1-\varepsilon}\left(\tilde{\boldsymbol{X}}\right)\right] \le \mathbb{P}_0\text{-FCVaR}_{1-\varepsilon}\left(\tilde{\boldsymbol{X}}\right) \le \mathbb{E}_{\mathbb{P}}\left[\tilde{\boldsymbol{X}}\middle|\tilde{\boldsymbol{X}} \le \mathbb{P}_0\text{-VaR}_{1-\varepsilon}\left(\tilde{\boldsymbol{X}}\right)\right]. \tag{2.3b}$$

Proof. See Appendix A.1. ∎

Notice that FCVaR can be viewed as a special case of In-CVaR from work [41] or Range VaR from work [18] (i.e., $\mathbb{P}_0\text{-FCVaR}_{1-\varepsilon}(\tilde{\boldsymbol{X}}) = \text{In-CVaR}_0^{1-\varepsilon}(\tilde{\boldsymbol{X}})$) and a special case of an optimized certainty equivalent from work [8] (i.e., $\mathbb{P}_0\text{-FCVaR}_{1-\varepsilon}(\tilde{\boldsymbol{X}}) = \max_\beta[\beta + \mathbb{E}_{\mathbb{P}_0}[\mu(\tilde{\boldsymbol{X}} - \beta)]]$ with $\mu(t) = -[-t]_+/(1-\varepsilon)$). We can also apply DFO to recover the In-CVaR from [41]. That is, for $0 \le \alpha < \beta \le 1$,

$$\text{In-CVaR}_\alpha^\beta(\tilde{\boldsymbol{X}}) = \inf_{\mathbb{P}\in\mathcal{P}} \mathbb{E}_{\mathbb{P}}\left[\tilde{\boldsymbol{X}}\right],$$

and the ambiguity set $\mathcal{P}$ is defined as

$$\mathcal{P} = \left\{\mathbb{P}\colon \begin{array}{l} \mathbb{P}(\mathcal{U}) = 1, 0 \preceq \mathbb{P} \preceq \mathbb{P}_0/(\beta-\alpha), \\ \mathbb{P}\left\{\tilde{\boldsymbol{X}} \ge \mathbb{P}_0\text{-VaR}_\alpha(\tilde{\boldsymbol{X}})\right\} = 1 \end{array}\right\}.$$

The equivalence (2.3a) shows that FCVaR (2.2) can be a special case of DFO (1.2). That is, letting $\tilde{\boldsymbol{X}} := Q(\boldsymbol{x}, \tilde{\boldsymbol{\xi}})$, $\boldsymbol{c} = \boldsymbol{0}$ and choosing the same interval ambiguity set as Proposition 2.1, DFO (1.2) reduces to the following FCVaR optimization

$$v^* = \min_{\boldsymbol{x}\in\mathcal{X}} \inf_{\mathbb{P}\in\mathcal{P}_I} \mathbb{E}_{\mathbb{P}}\left[Q(\boldsymbol{x}, \tilde{\boldsymbol{\xi}})\right] = \min_{\boldsymbol{x}\in\mathcal{X}} \mathbb{P}_0\text{-FCVaR}_{1-\varepsilon}\left[Q(\boldsymbol{x}, \tilde{\boldsymbol{\xi}})\right]. \tag{2.4}$$

We remark that the LTS introduced in Section 2.1 can be viewed as a special case of FCVaR (2.4). That is, suppose that the random vector $\tilde{\boldsymbol{\xi}}$ has an equiprobable distribution over a finite support $\mathcal{U} = \{\boldsymbol{\xi}^i\}_{i\in[N]} = \{\bar{\boldsymbol{x}}_i, \bar{y}_i\}_{i\in[N]} \subseteq \mathbb{R}^d \times \mathbb{R}$. Let $\varepsilon = (N-k)/N$ with an integer $k \in [N]$ and the recourse function be $Q(\boldsymbol{x}, \boldsymbol{\xi}^i) = (\bar{y}_i - \bar{\boldsymbol{x}}_i^\top \boldsymbol{x})^2$ for each $i \in [N]$. Then the interval ambiguity set in Proposition 2.1 reduces to $\mathcal{P}_I = \{\boldsymbol{p} \in \mathbb{R}_+^N : \sum_{i\in[N]} p_i = 1, 0 \le p_i \le 1/k\}$ and DFO (2.4) reduces to LTS (2.1).

Interestingly, if one replaces the inner infimum operator with the supremum operator on the left-hand side of (2.4), then the left-hand side reduces to the CVaR minimization problem, a well-known DRO model, i.e.,

$$\sup_{\mathbb{P}\in\mathcal{P}_I} \mathbb{E}_{\mathbb{P}}\left[Q(\boldsymbol{x}, \tilde{\boldsymbol{\xi}})\right] = \mathbb{P}_0\text{-CVaR}_{1-\varepsilon}(Q(\boldsymbol{x}, \tilde{\boldsymbol{\xi}}))\colon = \min_\beta\left\{\beta + \frac{1}{\varepsilon}\mathbb{E}_{\mathbb{P}_0}\left[\left(Q(\boldsymbol{x}, \tilde{\boldsymbol{\xi}}) - \beta\right)_+\right]\right\}.$$

Compared with FCVaR, CVaR takes the conditional expectation of unfavorable scenarios. This further demonstrates the non-robustness of DRO models in the existence of outliers. On the other hand, applying the DFO framework can circumvent these outliers. Thus, we remark that FCVaR can be more meaningful and ideal than CVaR in the presence of outliers.

Note that the connection between FCVaR and LTS motivates us to consider the other two alternatives based on the robust statistics in Section 2.1. For example, instead of using LTS, we can use Winsorized approach, e.g., replacing the recourse function values of unfavorable scenarios with the $(1-\varepsilon)$-quantile $\text{VaR}_{1-\varepsilon}(\cdot)$. Similarly, we can also consider the Huber-skip method. That is, we can specify an allowable upper bound for the recourse function value and replace the recourse function value with this bound if going beyond.

Alternative I. Winsorized CVaR. Winsorized CVaR, denoted as WCVaR, is the weighted average between FCVaR and VaR, providing a reasonable estimate of the central tendency of the objective value. Notably, the WCVaR admits the following form:

$$\mathbb{P}_0\text{-WCVaR}_{1-\varepsilon}(\tilde{\boldsymbol{X}})\colon = (1-\varepsilon)\mathbb{P}_0\text{-FCVaR}_{1-\varepsilon}(\tilde{\boldsymbol{X}}) + \varepsilon\mathbb{P}_0\text{-VaR}_{1-\varepsilon}(\tilde{\boldsymbol{X}}), \tag{2.5}$$

for a given random variable $\tilde{\boldsymbol{X}}$. As explained in Section 2, the WCVaR admits a DFO interpretation. An interesting side product is that if we choose a penalty function to be $\mathbb{P}_0\text{-VaR}_{1-\varepsilon}(\tilde{\boldsymbol{X}})$, then WCVaR recovers the two-stage chance constrained program studied in [42].

Alternative II. Huber-skip CVaR. The Huber-skip CVaR, denoted as HCVaR, is to compute the expectation of the minimum of the recourse function value and a given upper bound H, i.e.,

$$\mathbb{P}_0\text{-HCVaR}(\tilde{\boldsymbol{X}}, H) \colon = \mathbb{E}_{\mathbb{P}_0}\left[\min\left\{\tilde{\boldsymbol{X}}, H\right\}\right]. \tag{2.6}$$

As explained in Section 2, the HCVaR admits a DFO interpretation. Notice that a proper choice of the value H decides the quality of Huber-skip CVaR (see, e.g., [29]). We also remark that if we let H be $\mathbb{P}_0\text{-VaR}_{1-\varepsilon}(\cdot)$, then HCVaR (2.6) and WCVaR (2.5) coincide.

The following Example 3 and Example 4 illustrate the differences among VaR, CVaR, FCVaR, WCVaR, HCVaR, and the conventional expectation. We see that compared with CVaR, the proposed methods based on DFO (i.e., FCVaR, WCVaR, and HCVaR) can serve as better alternatives to the expectation, especially when the stochastic recourse function may not be integrable.

Example 3. Let us assume $\tilde{\boldsymbol{X}}$ to be a truncated Cauchy distribution $\mathbb{P}_0$ with a probability density function $f(x) := 2/(\pi(1+x^2)), x \geq 0$. For the demonstration purpose, we let $\varepsilon = 0.1$. Then, we are able to compute the values of $\mathbb{P}_0\text{-FCVaR}_{1-\varepsilon}$, $\mathbb{P}_0\text{-WCVaR}_{1-\varepsilon}$, $\mathbb{P}_0\text{-VaR}_{1-\varepsilon}$, and $\mathbb{P}_0\text{-HCVaR}(\cdot, H)$ with $H = 3$, while the expectation and $\mathbb{P}_0\text{-CVaR}_{1-\varepsilon}$ do not exist. Please see Figure 4 for an illustration. ⋄

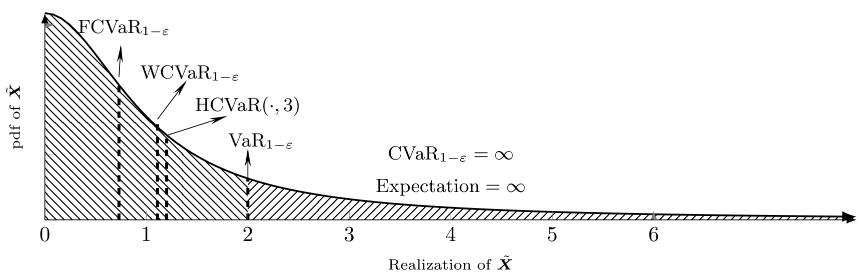

Fig. 4: Illustration of Expectation, FCVaR, WCVaR, HCVaR, VaR, and CVaR with Truncated Cauchy Distribution.

Example 4. Let us assume $\tilde{\boldsymbol{X}}$ to be a truncated Gaussian distribution $\mathbb{P}_0$ with a probability density function $f(x) := \sqrt{2/\pi}\exp(-x^2/2), x \geq 0$. For the demonstration purpose, we let $\varepsilon = 0.10$. Then, we are able to find the value of expectation, $\mathbb{P}_0\text{-CVaR}_{1-\varepsilon}$, $\mathbb{P}_0\text{-FCVaR}_{1-\varepsilon}$, $\mathbb{P}_0\text{-WCVaR}_{1-\varepsilon}$, $\mathbb{P}_0\text{-VaR}_{1-\varepsilon}$, and $\mathbb{P}_0\text{-HCVaR}(\cdot, H)$ with $H = 2$, which are illustrated in Figure 5. ⋄

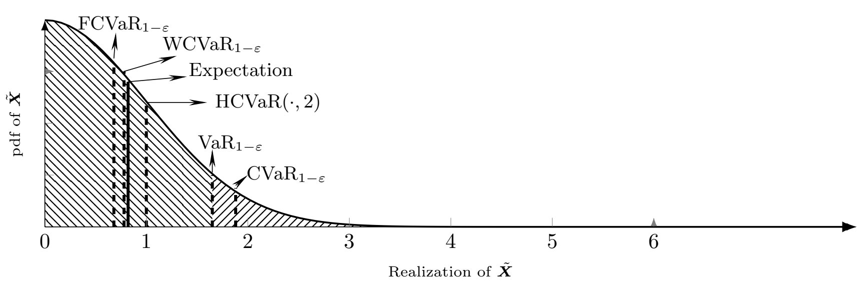

Fig. 5: Illustration of Expectation (solid line), FCVaR, WCVaR, HCVaR, VaR, and CVaR with Truncated Gaussian Distribution.

Next, we apply DFO (i.e., FCVaR, WCVaR, and HCVaR) in the two-stage stochastic programs without relatively complete recourse.

2.3 Two-stage Stochastic Programs without Relatively Complete Recourse. Motivated from the examples in Section 1.1, this subsection focuses on a two-stage stochastic program, which, in general, is defined as

$$\min_{\boldsymbol{x}\in\mathcal{X}} \boldsymbol{c}^\top \boldsymbol{x} + \mathbb{E}_{\mathbb{P}_0}\left[Q(\boldsymbol{x}, \tilde{\boldsymbol{\xi}})\right], \tag{2.7a}$$

where for a realization $\boldsymbol{\xi}$ of $\tilde{\boldsymbol{\xi}}$, the recourse function $Q(\boldsymbol{x}, \boldsymbol{\xi})$ is defined as

$$Q(\boldsymbol{x}, \boldsymbol{\xi}) = \inf_{\boldsymbol{y}\in\mathcal{Y}}\left[(\boldsymbol{Q}\boldsymbol{\xi}_q + \boldsymbol{q})^\top \boldsymbol{y} \colon \boldsymbol{T}(\boldsymbol{x})\boldsymbol{\xi}_T + \boldsymbol{\xi}_W \boldsymbol{y} \geq \boldsymbol{h}(\boldsymbol{x})\right], \tag{2.7b}$$

where $\boldsymbol{y}$ denotes the wait-and-see decisions in the second-stage problem, $\boldsymbol{Q} : \mathbb{R}^{n_2\times m_1}$, $\boldsymbol{T} : \mathbb{R}^n \to \mathbb{R}^{\ell\times m_2}$ and $\boldsymbol{h} : \mathbb{R}^n \to \mathbb{R}^\ell$ represent the technology affine mapping and the right-hand-side affine mapping, separately, and $\boldsymbol{\xi} = (\boldsymbol{\xi}_q, \boldsymbol{\xi}_T, \boldsymbol{\xi}_W) \in \mathbb{R}^{m_1} \times \mathbb{R}^{m_2} \times \mathbb{R}^{\ell\times n_2}$, $\boldsymbol{q} \in \mathbb{R}^{n_2}$. Set $\mathcal{Y} \subseteq \mathbb{R}^{n_2}$ denotes the constraints for $\boldsymbol{y}$, e.g., the boundary constraints of the wait-and-see decisions. In this section, we assume that the set $\mathcal{Y}$ is compact and nonempty, which ensures that $\inf_{\boldsymbol{y}\in\mathcal{Y}}[(\boldsymbol{Q}\tilde{\boldsymbol{\xi}}_q+\boldsymbol{q})^\top \boldsymbol{y}] > -\infty$ almost surely. Following the discussions in Section 2.2, we apply DFO to select favorable scenarios, where the distributionally favorable counterpart of the two-stage programs is defined in (1.2) and $Q(\boldsymbol{x},\tilde{\boldsymbol{\xi}})$ is defined in (2.7b).

Suppose that the empirical distribution $\widehat{\mathbb{P}}$ of the second-stage problem consists of N i.i.d. samples $\{\widehat{\boldsymbol{\xi}}^i\}_{i\in[N]}$ and assume $N\varepsilon$ is an integer, we apply FCVaR to the second-stage problem to focus on some favorable scenarios. This leads to the following favorable two-stage stochastic problem, which can be written as

$$v^* = \min_{\boldsymbol{x}\in\mathcal{X},\boldsymbol{z}} \left\{ \boldsymbol{c}^\top\boldsymbol{x} + \frac{1}{N-N\varepsilon}\sum_{i\in[N]} z_i Q(\boldsymbol{x},\widehat{\boldsymbol{\xi}}^i) \colon \sum_{i\in[N]} z_i = N - N\varepsilon, \boldsymbol{z}\in\{0,1\}^N \right\}, \tag{2.8}$$

where we assume that $\infty\times 0 = 0$. In problem (2.8), for each $i\in[N]$, the product $z_iQ(\boldsymbol{x},\widehat{\boldsymbol{\xi}}^i)$ can be represented as the following MILP

$$z_iQ(\boldsymbol{x},\widehat{\boldsymbol{\xi}}^i) = \min_{\boldsymbol{y}^i\in\mathcal{Y}} \left[(\boldsymbol{Q}\widehat{\boldsymbol{\xi}}^i_q+\boldsymbol{q})^\top\boldsymbol{y}^i - L_i(1-z_i) \colon \boldsymbol{T}(\boldsymbol{x})\widehat{\boldsymbol{\xi}}^i_T + \widehat{\boldsymbol{\xi}}^i_W\boldsymbol{y}^i \geq \boldsymbol{h}(\boldsymbol{x}) - \boldsymbol{M}^i(1-z_i) \right]. \tag{2.9}$$

Above, $\boldsymbol{M}^i$ is a vector of large numbers for each $i\in[N]$, and can be computed as

$$M^i_j \geq \max_{\boldsymbol{x}\in\mathcal{X},\boldsymbol{y}^i\in\mathcal{Y}} h_j(\boldsymbol{x}) - (\boldsymbol{T}(\boldsymbol{x})\widehat{\boldsymbol{\xi}}^i_T + \widehat{\boldsymbol{\xi}}^i_W\boldsymbol{y}^i)_j$$

for each $j\in[\ell]$ and $i\in[N]$, and L_i is the value of the trivial second-stage problem $L_i := \inf_{\boldsymbol{y}^i\in\mathcal{Y}}[(\boldsymbol{Q}\widehat{\boldsymbol{\xi}}^i_q + \boldsymbol{q})^\top\boldsymbol{y}^i] > -\infty$ for each $i\in[N]$.

The purpose of using $\boldsymbol{z}$ variables in the constraints of the second-stage problem (2.8) is to resolve the infeasibility issue and to ensure that the second-stage problem is solvable. For example, when the second-stage problem is infeasible, then $z_i=0$, and the only non-trivial constraint is the boundary constraint, i.e., $\boldsymbol{y}^i\in\mathcal{Y}$. However, the big-M coefficients $\{\boldsymbol{M}^i\}_{i\in[N]}$ are not easy to derive and can be very large. Thus, we further explore the structure of the problem and discuss sufficient conditions under which we can obtain the big-M free formulations. That is, we show that under some conditions, we can represent the bilinear terms $\{z_iQ(\boldsymbol{x},\widehat{\boldsymbol{\xi}}^i)\}_{i\in[N]}$ in problem (2.8) using the big-M free formulations.

Theorem 2.2. *Suppose that the set* $\mathcal{Y} := \{\boldsymbol{y} : \boldsymbol{D}\boldsymbol{y}\geq\boldsymbol{d}, \boldsymbol{y}\geq\boldsymbol{0}\}$ *and* $\boldsymbol{T}(\boldsymbol{x}) = \widehat{\boldsymbol{T}}_1\boldsymbol{x}\otimes\mathbf{e}+\widehat{\boldsymbol{T}}_2, \boldsymbol{h}(\boldsymbol{x}) = \widehat{\boldsymbol{H}}\boldsymbol{x}+\widehat{\boldsymbol{h}}$, $\widehat{\boldsymbol{T}}_1\in\mathbb{R}^{\ell\times n}$, $\widehat{\boldsymbol{T}}_2\in\mathbb{R}^{\ell\times m_2}$, $\widehat{\boldsymbol{H}}\in\mathbb{R}^{\ell\times n}$, $\widehat{\boldsymbol{h}}\in\mathbb{R}^{\ell}$, *vector* $\boldsymbol{0}$ *is contained in the polyhedron* $\{\boldsymbol{y}^i : \widehat{\boldsymbol{T}}_1\boldsymbol{x}\otimes\mathbf{e}\widehat{\boldsymbol{\xi}}^i_T + \widehat{\boldsymbol{\xi}}^i_W\boldsymbol{y}^i - \widehat{\boldsymbol{H}}\boldsymbol{x}\geq\boldsymbol{0}\}$ *for each* $\boldsymbol{x}\in\mathcal{X}$ *and* $i\in[N]$, *and* $\boldsymbol{Q}\widehat{\boldsymbol{\xi}}^i_q+\boldsymbol{q}\geq\boldsymbol{0}$ *for all* $i\in[N]$. *Then, the favorable two-stage stochastic problem* (2.8) *is equivalent to*

$$v^* = \min_{\boldsymbol{x}\in\mathcal{X},\boldsymbol{z}} \left\{ \boldsymbol{c}^\top\boldsymbol{x} + \frac{1}{N-N\varepsilon}\sum_{i\in[N]} \widehat{Q}(\boldsymbol{x},z_i,\widehat{\boldsymbol{\xi}}^i) \colon \sum_{i\in[N]} z_i \geq N - N\varepsilon, \boldsymbol{z}\in\{0,1\}^N \right\}, \tag{2.10}$$

where $\widehat{Q}(\boldsymbol{x},z_i,\widehat{\boldsymbol{\xi}}^i) = z_iQ(\boldsymbol{x},\widehat{\boldsymbol{\xi}}^i)$ *and*

$$\widehat{Q}(\boldsymbol{x},z_i,\widehat{\boldsymbol{\xi}}^i) = \min_{\boldsymbol{y}^i\geq\boldsymbol{0}} \left\{ (\boldsymbol{Q}\widehat{\boldsymbol{\xi}}^i_q+\boldsymbol{q})^\top\boldsymbol{y}^i \colon \widehat{\boldsymbol{T}}_1\boldsymbol{x}\otimes\mathbf{e}\widehat{\boldsymbol{\xi}}^i_T + \widehat{\boldsymbol{\xi}}^i_W\boldsymbol{y}^i - \widehat{\boldsymbol{H}}\boldsymbol{x} \geq \left[\widehat{\boldsymbol{h}} - \widehat{\boldsymbol{T}}_2\widehat{\boldsymbol{\xi}}^i_T\right] z_i, \boldsymbol{D}\boldsymbol{y}^i\geq\boldsymbol{d}z_i \right\}.$$

Proof. In problem (2.10), we first consider $z_i=0$. Since the vector $\boldsymbol{0}$ is contained in the polyhedron $\{\boldsymbol{y}^i : \widehat{\boldsymbol{T}}_1\boldsymbol{x}\otimes\mathbf{e}\widehat{\boldsymbol{\xi}}^i_T + \widehat{\boldsymbol{\xi}}^i_W\boldsymbol{y}^i - \widehat{\boldsymbol{H}}\boldsymbol{x}\geq\boldsymbol{0}\}$ for each $\boldsymbol{x}\in\mathcal{X}$ and $i\in[N]$, then the optimal value of the second-stage problem $\widehat{Q}(\boldsymbol{x},z_i,\widehat{\boldsymbol{\xi}}^i)$ is 0, which is as the same as the value of $z_iQ(\boldsymbol{x},\widehat{\boldsymbol{\xi}}^i)$. If $z_i=1$, then $\widehat{Q}(\boldsymbol{x},z_i,\widehat{\boldsymbol{\xi}}^i)$ is identical to $Q(\boldsymbol{x},\widehat{\boldsymbol{\xi}}^i)$. ◻

Notice that there is no big-M coefficient in the formulation (2.10) and we use the following example to illustrate Theorem 2.2.

Example 5. Let us consider a two-stage resource planning (TRP) problem, which consists of a set of resources (e.g., server types), denoted by $s\in[n]$, that can be used to meet the demand of a set of customer types, denoted by $j\in[n_1]$. Note that similar problems have been studied in many works (see, e.g., [14, 42, 43]). Following the notation, the TRP problem can be formulated as

$$\min_{\boldsymbol{x}\geq\boldsymbol{0},\boldsymbol{z}} \left\{ \boldsymbol{c}^\top\boldsymbol{x} + \frac{1}{N-N\varepsilon}\sum_{i\in[N]} z_iQ(\boldsymbol{x},\widehat{\boldsymbol{\xi}}^i) \colon \sum_{i\in[N]} z_i \geq N - N\varepsilon, \boldsymbol{z}\in\{0,1\}^N \right\}, \tag{2.11a}$$

where for a random $\widehat{\boldsymbol{\xi}}^i = (\boldsymbol{q}^i, \boldsymbol{p}^i, \boldsymbol{u}^i, \boldsymbol{\lambda}^i)$,

$$(2.11b)\qquad Q(\boldsymbol{x}, \widehat{\boldsymbol{\xi}}^i) = \min_{\boldsymbol{y}^i \geq \boldsymbol{0}} \left\{ \sum_{s\in[n]} \sum_{j\in[n_1]} q^i_{sj} y^i_{sj} \colon \sum_{j\in[n_1]} y^i_{sj} \leq p^i_s x_s, \forall s \in [n], \sum_{s\in[n]} u^i_{sj} y^i_{sj} \geq \lambda^i_j, \forall j \in [n_1] \right\}.$$

In this model, c_s represents the unit cost of resource $s \in [n]$. For each $s \in [n]$, variable x_s denotes the amount of resource s to purchase and for $s \in [n]$ and $j \in [n_1]$, variable y_{sj} represents the allocation amount of resource s to customer type j. Parameters $\tilde{\boldsymbol{q}}, \tilde{\boldsymbol{p}}, \tilde{\boldsymbol{u}}, \tilde{\boldsymbol{\lambda}}$ are random, where $\tilde{q}_{sj}$ represents the random cost of allocating resource $s \in [n]$ to customer type $j \in [n_1]$, $\tilde{p}_s$ represents the random utilization rate of resource $s \in [n]$, $\tilde{u}_{sj}$ represents the random service rate of resource $s \in [n]$ for customer type $j \in [n_1]$ and $\tilde{\lambda}_j$ is the random demand of customer type $j \in [n_1]$.

Note that the TRP (2.11a) is a two-stage stochastic program without relatively complete recourse. Besides, when $\lambda^i_j = \lambda^i_j z_i$ with $z_i = 0$ for each $j \in [n_1]$ and $i \in [N]$, for any $\boldsymbol{x} \geq \boldsymbol{0}$, $\boldsymbol{y}^i = \boldsymbol{0}$ is always feasible to (2.11b) for each $i \in [N]$. Hence, we can apply the result in Theorem 2.2. Using the binary variables $\boldsymbol{z}$, we can rewrite the bilinear term as

$$(2.11c)\quad z_i Q(\boldsymbol{x}, \widehat{\boldsymbol{\xi}}^i) = \min_{\boldsymbol{y}^i \geq \boldsymbol{0}} \left\{ \sum_{s\in[n]} \sum_{j\in[n_1]} q^i_{sj} y^i_{sj} \colon p^i_s x_s - \sum_{j\in[n_1]} y^i_{sj} \geq 0, \forall s \in [n], \sum_{s\in[n]} u^i_{sj} y^i_{sj} \geq \lambda^i_j z_i, \forall j \in [n_1] \right\}.$$

Thus, we arrive at a big-M free formulation for (2.11a). ⋄

As a direct corollary of Theorem 2.2, we can provide big-M free formulations for the Winsorized CVaR and the Huber-skip CVaR type of the two-stage problem.

Corollary 2.3. *Under the same assumptions as in Theorem 2.2:*

(i) favorable two-stage stochastic program (2.8) with WCVaR *admits the following formulation*

$$(2.12a)\qquad \min_{\substack{\boldsymbol{x}\in\mathcal{X},\\ \boldsymbol{z}\in\{0,1\}^N}} \left\{ \boldsymbol{c}^\top \boldsymbol{x} + \frac{1}{N} \sum_{i\in[N]} z_i Q(\boldsymbol{x}, \widehat{\boldsymbol{\xi}}^i) + \eta\varepsilon \colon \begin{array}{l} \eta \geq z_i Q(\boldsymbol{x}, \widehat{\boldsymbol{\xi}}^i) + (1 - z_i) L_i, \forall i \in [N], \\ \sum_{i\in[N]} z_i \geq N - N\varepsilon \end{array} \right\},$$

where L_i denotes the value of the trivial second-stage problem $L_i := \inf_{\boldsymbol{y}^i\in\mathcal{Y}}[(\boldsymbol{Q}\widehat{\boldsymbol{\xi}}^i_q + \boldsymbol{q})^\top \boldsymbol{y}^i] > -\infty$ for each $i \in [N]$;

(ii) favorable two-stage stochastic program (2.8) with HCVaR *admits the following formulation*

$$(2.12b)\qquad \min_{\boldsymbol{x}\geq\boldsymbol{0}, \boldsymbol{z}\in\{0,1\}^N} \left\{ \boldsymbol{c}^\top \boldsymbol{x} + \frac{1}{N} \sum_{i\in[N]} \left(z_i Q(\boldsymbol{x}, \widehat{\boldsymbol{\xi}}^i) + (1 - z_i) H \right) \right\},$$

where H denotes the preset upper bound of the second-stage problem.

Notice that the bilinear terms $\{z_i Q(\boldsymbol{x}, \widehat{\boldsymbol{\xi}}^i)\}_{i\in[N]}$ in (2.12a) and (2.12b) can be linearized by applying the result in (2.9) or using Theorem 2.2.

We remark that we show the strength of these big-M free formulations in the numerical study section.

3 Decision Outlier Robustness.

To provide an effective means of evaluating the performance of DFO models, we first review the definition of "outlier robust" in the statistical robustness. In light of its drawbacks, we propose the notion of "decision outlier robust" to address these limitations in evaluating DFO models.

3.1 Counterexamples that Some Well-known Robust Statistics May Not Have Bounded Influence Curve.

In statistical robustness (see the details in [24, 45]), if the influence curve of a statistic estimator is bounded, then that estimator is called "outlier robust." Let $\mathbb{P}_0$ denote the reference probability measure of $\tilde{\boldsymbol{\xi}}$ and $\delta_{\boldsymbol{\xi}^o}$ is the Dirac measure for the perturbation data $\boldsymbol{\xi}^o \in \text{supp}(\mathbb{P}_0)$. For any decision $\boldsymbol{x} \in \mathcal{X}$ with corresponding function values $Q(\boldsymbol{x}, \tilde{\boldsymbol{\xi}})$, the statistic estimator $\mathbb{P}_0$-$T(\cdot)$ is "outlier robust" if the following condition is satisfied:

$$(3.1)\qquad \lim_{\gamma\to 0} \frac{1}{\gamma} \left[[(1-\gamma)\mathbb{P}_0 + \gamma\delta_{\boldsymbol{\xi}^o}]\text{-}T\left(Q(\boldsymbol{x}, \tilde{\boldsymbol{\xi}})\right) - \mathbb{P}_0\text{-}T\left(Q(\boldsymbol{x}, \tilde{\boldsymbol{\xi}})\right) \right] < \infty.$$

Then, based on condition (3.1), we first illustrate that $\mathbb{P}_0$-$\text{VaR}_{1-\varepsilon}\{Q(\boldsymbol{x}, \tilde{\boldsymbol{\xi}})\}$ (i.e., a quantile) may not be outlier robust.

Example 6. Suppose $\mathbb{P}_0\{\tilde{\boldsymbol{\xi}} : \tilde{\boldsymbol{\xi}} = \boldsymbol{\xi}^i\} = 1/N$ for each $i \in [N]$ and the perturbation $Q(\boldsymbol{x}, \boldsymbol{\xi}^o)$, $\mathbb{P}_0$-$\mathrm{VaR}_{1-\varepsilon}\{Q(\boldsymbol{x}, \tilde{\boldsymbol{\xi}})\}$ is "outlier robust" if the condition (3.1) is satisfied. Suppose $\varepsilon = 0.1$, $N = 10\bar{N}$, $\bar{N} = 10$, and $Q(\boldsymbol{x}, \boldsymbol{\xi}^j) = i$ for each $j \in [10(i-1)+1, 10i]$ and $i \in [\bar{N}]$ and $Q(\boldsymbol{x}, \boldsymbol{\xi}^o) = \bar{N}+1$. When $\gamma \to 0$,

$$[(1-\gamma)\mathbb{P}_0 + \gamma\delta_{\boldsymbol{\xi}^o}]\text{-}\mathrm{VaR}_{1-\varepsilon}\{Q(\boldsymbol{x}, \tilde{\boldsymbol{\xi}})\} = \bar{N},$$

and $\mathbb{P}_0$-$\mathrm{VaR}_{1-\varepsilon}\{Q(\boldsymbol{x}, \tilde{\boldsymbol{\xi}})\} = \bar{N} - 1$. Then, condition (3.1) is simplified as

$$\lim_{\gamma \to 0} \frac{1}{\gamma}\left[\bar{N} - (\bar{N} - 1)\right] = \infty,$$

which shows that $\mathbb{P}_0$-$\mathrm{VaR}_{1-\varepsilon}\{Q(\boldsymbol{x}, \tilde{\boldsymbol{\xi}})\}$ may not be outlier robust. ⋄

Under the similar setting of Example 6, we can show that $\mathbb{P}_0$-$\mathrm{FCVaR}_{1-\varepsilon}\{Q(\boldsymbol{x}, \tilde{\boldsymbol{\xi}})\}$ (i.e., LTS) may not be outlier robust.

Example 7. Suppose $\mathbb{P}_0\{\tilde{\boldsymbol{\xi}} : \tilde{\boldsymbol{\xi}} = \boldsymbol{\xi}^i\} = 1/N$ for each $i \in [N]$ and the perturbation $Q(\boldsymbol{x}, \boldsymbol{\xi}^o)$, $\mathbb{P}_0$-$\mathrm{FCVaR}_{1-\varepsilon}\{Q(\boldsymbol{x}, \tilde{\boldsymbol{\xi}})\}$ is "outlier robust" if the condition (3.1) is satisfied. Suppose $\varepsilon = 0.1$, $N = 10\bar{N}$, $\bar{N} = 10$, and $Q(\boldsymbol{x}, \boldsymbol{\xi}^j) = i$ for each $j \in [10(i-1)+1, 10i]$ and $i \in [\bar{N}]$ and $Q(\boldsymbol{x}, \boldsymbol{\xi}^o) = \bar{N}+1$. Then, when $\gamma \to 0$, condition (3.1) is simplified as

$$\lim_{\gamma \to 0} \frac{1}{\gamma}\frac{1}{1-\varepsilon}\left[\frac{\bar{N}(\bar{N}+1)}{2\bar{N}} - \frac{\bar{N}(\bar{N}-1)}{2\bar{N}}\right] = \infty,$$

which demonstrates that $\mathbb{P}_0$-$\mathrm{FCVaR}_{1-\varepsilon}\{Q(\boldsymbol{x}, \tilde{\boldsymbol{\xi}})\}$ may not be outlier robust. ⋄

The notion of the influence curve has the following two major drawbacks: (i) it focuses on the smoothness of a favorable measure (i.e., a robust statistic), which is quite restrictive; for instance, neither quantiles nor LTS can be well explained due to their nonsmooth nature under a discrete reference distribution (e.g., Example 6). However, in many decision-making problems, the objective function may not be necessarily smooth (e.g., two-stage stochastic integer programming studied in [2]); and (ii) it requires a known reference distribution, which may not be a case in the ambiguity set $\mathcal{P}$ (e.g., a moment ambiguity set). Thus, the influence curve is not appropriate to analyze the outlier robustness of DFO.

3.2 Decision Outlier Robustness. To remedy the issues mentioned in the previous subsection, this subsection proposes a generic way to properly evaluate the decision outlier robustness of a DFO model, motivated by the influence curve from robust statistics. We first define the notion of an unamenable decision.

Definition 3.1. *For a reference distribution $\mathbb{P}_0$, a decision $\boldsymbol{x} \in \mathcal{X}$ is an "unamenable decision" when there exists an outlier $\boldsymbol{\xi}^o \in \mathrm{supp}(\mathbb{P}_0)$ such that the recourse function $Q(\boldsymbol{x}, \boldsymbol{\xi}^o) = +\infty$. The collection of such unamenable decisions is denoted by set $\widehat{\mathcal{X}}$.*

Note that the set of unamenable decisions $\widehat{\mathcal{X}}$ is associated with a reference distribution $\mathbb{P}_0$. Now we are ready to introduce the notion of "*decision outlier robust*," which mainly focuses on unamenable decisions with the reference distribution $\mathbb{P}_0$. In this section, we mainly focus on stochastic programs with unamenable decisions.

Definition 3.2. *The DFO (1.2) is called "decision outlier robust" when the following condition is satisfied:*

$$\inf_{\mathbb{P} \in \mathcal{P}} \left[(1-\gamma)\mathbb{E}_{\mathbb{P}}\left[Q(\boldsymbol{x}, \tilde{\boldsymbol{\xi}})\right] + \gamma Q(\boldsymbol{x}, \boldsymbol{\xi}^o)\mathbb{I}\left(\boldsymbol{\xi}^o \in \mathrm{supp}(\mathbb{P})\right)\right] < \infty, \tag{3.2a}$$

for each unamenable decision $\boldsymbol{x} \in \widehat{\mathcal{X}}$, each outlier $\boldsymbol{\xi}^o \in \mathrm{supp}(\mathbb{P}_0)$, and for any $\gamma \in [0, 1]$. Here, we let $\infty \times 0 = 0$.

Note that condition (3.2a) can also be equivalently written as

$$\inf_{\mathbb{P} \in \mathcal{P}} \left[(1-\gamma)\mathbb{E}_{\mathbb{P}}\left[Q(\boldsymbol{x}, \tilde{\boldsymbol{\xi}})\right] + \gamma \mathrm{ess.sup}_{\mathbb{P}}\left\{Q(\boldsymbol{x}, \tilde{\boldsymbol{\xi}})\right\}\right] < \infty, \tag{3.2b}$$

which implies that by adjusting the probability measure $\mathbb{P}$, a DFO model is decision outlier robust if there exists one probability measure $\mathbb{P}$ such that the left-hand side of condition (3.2b) is bounded. We make the following remarks about Definition 3.2.

(i) In Definition 3.2, for the DFO (1.2) to be decision outlier robust, there exists a probability measure $\mathbb{P} \in \mathcal{P}$ such that an unamenable decision for any mixture distribution of $\mathbb{P}$ and a Dirac measure on an outlier $\boldsymbol{\xi}^o \in \mathrm{supp}(\mathbb{P})$ yields a bounded objective function value. This should hold for any unamenable decision $\boldsymbol{x} \in \widehat{\mathcal{X}}$.

(ii) The purpose of introducing the decision outlier robustness concept is to resolve all issues from the influence curve in the theoretical perspective.

(iii) Although it may require the unamenable decision set beforehand, in practice, one can simply check all the decisions. Besides, the results in Proposition 3.3 can further help simplify the verification process.

PROPOSITION 3.3. *The following statements hold:*

(i) The DFO (1.2) is decision outlier robust if for any unamenable decision $\boldsymbol{x} \in \widehat{\mathcal{X}}$, there exists a probability measure $\mathbb{P} \in \mathcal{P}$ such that $\mathbb{E}_{\mathbb{P}}[Q(\boldsymbol{x}, \tilde{\boldsymbol{\xi}})] < \infty$; and

(ii) The DFO (1.2) is not decision outlier robust if there exists an unamenable decision $\boldsymbol{x} \in \widehat{\mathcal{X}}$ with its outlier $\boldsymbol{\xi}^o$ such that $Q(\boldsymbol{x}, \boldsymbol{\xi}^o) = \infty$ and for any probability measure $\mathbb{P} \in \mathcal{P}$, we have $\boldsymbol{\xi}^o \in \operatorname{supp}(\mathbb{P})$.

The proof of Proposition 3.3 follows directly from Definition 3.2 and thus is omitted.

Using Proposition 3.3, we can immediately demonstrate that the expectation operator with a singleton ambiguity set $\mathcal{P}$ is not decision outlier robust.

COROLLARY 3.4. *Suppose $\mathcal{P}$ is a singleton, and there exists an unamenable decision $\boldsymbol{x} \in \mathcal{X}$. Then, the corresponding DFO, i.e., a regular stochastic program without relatively complete recourse, is not decision outlier robust.*

Proof. Suppose that $\mathcal{P} = \{\mathbb{P}_0\}$. Since there exists an unamenable decision $\boldsymbol{x} \in \mathcal{X}$, according to Definition 3.1, there exists an outlier $\boldsymbol{\xi}^o \in \operatorname{supp}(\mathbb{P}_0)$ with $Q(\boldsymbol{x}, \boldsymbol{\xi}^o) = \infty$. Using part (ii) of Proposition 3.3, we know that the corresponding DFO is not decision outlier robust. ∎

Therefore, without relatively complete recourse, simply taking the expectation with respect to a particular distribution (i.e., sticking to a singleton ambiguity set) may not be ideal (see the discussions in Example 2). A richer and nontrivial ambiguity set is more desirable and is demonstrated in the following subsections.

Moreover, we show that the DFO framework (1.4b) (i.e., the corresponding chance constrained program) is decision outlier robust. In contrast, the robust optimization framework (1.4a) may not be when there are unamenable decisions.

THEOREM 3.5. *Suppose that the unamenable decision set $\widehat{\mathcal{X}}$ is non-empty and for any $\boldsymbol{x} \in \widehat{\mathcal{X}}$, we have $\mathbb{P}_0\{\tilde{\boldsymbol{\xi}} : G(\boldsymbol{x}, \tilde{\boldsymbol{\xi}}) > 0\} \leq \varepsilon$, where $\mathbb{P}_0$ denotes the reference distribution and function $G(\boldsymbol{x}, \tilde{\boldsymbol{\xi}})$ is measurable for any $\boldsymbol{x} \in \mathcal{X}$. Then, the DFO (1.4b) is decision outlier robust, while the robust optimization (1.4a) is not.*

Proof. We split the proof into two parts by checking the DFO (1.4b) and the robust optimization framework (1.4a) separately.

Part I. According to Proposition 3.3, for the DFO framework (1.4b), it is sufficient to show that for any unamenable decision $\boldsymbol{x} \in \widehat{\mathcal{X}}$, there exists a probability measure $\mathbb{P}^* \in \mathcal{P}_I$ such that $\mathbb{E}_{\mathbb{P}^*}[\mathbb{I}(G(\boldsymbol{x}, \tilde{\boldsymbol{\xi}}) > 0)] \leq 0$ and $\mathbb{P}^*\{\tilde{\boldsymbol{\xi}} : G(\boldsymbol{x}, \tilde{\boldsymbol{\xi}}) > 0\} = 0$.

Let us denote set $\mathcal{U}_1 = \{\boldsymbol{\xi} : G(\boldsymbol{x}, \boldsymbol{\xi}) \leq 0\}$, which is measurable (see, e.g., proposition 1 in section 3.1 of [59]). According to our presumption, we know that $\mathbb{P}\{\mathcal{U}_1\} \geq 1 - \varepsilon$. Now let us construct $\mathbb{P}^*(d\boldsymbol{\xi}) = \mathbb{P}_0(d\boldsymbol{\xi})/\mathbb{P}_0\{\mathcal{U}_1\}$ for each $\boldsymbol{\xi} \in \mathcal{U}_1$, 0, otherwise. Note that by our construction, we have $\mathbb{P}^*(\mathcal{U}_1) = 1, 0 \preceq \mathbb{P}^* \preceq \mathbb{P}_0/(1 - \varepsilon)$. Hence, $\mathbb{P}^* \in \mathcal{P}_I$ and $\mathbb{P}^*\{\tilde{\boldsymbol{\xi}} : \tilde{\boldsymbol{\xi}} = \boldsymbol{\xi}^o\} = 0$, where the recourse function can be written as $Q(\boldsymbol{x}, \boldsymbol{\xi}^o) = \min\{0 \colon G(\boldsymbol{x}, \boldsymbol{\xi}^o) > 0\}$. On the other hand, we have

$$\mathbb{E}_{\mathbb{P}^*}\left[\mathbb{I}(G(\boldsymbol{x}, \tilde{\boldsymbol{\xi}}) > 0)\right] = 1 - \mathbb{P}_0\{\mathcal{U}_1\}/\mathbb{P}_0\{\mathcal{U}_1\} = 0, \quad \mathbb{P}^*\left\{\tilde{\boldsymbol{\xi}} : G(\boldsymbol{x}, \tilde{\boldsymbol{\xi}}) > 0\right\} = 0.$$

This proves that $\mathbb{P}^*$ is a desirable probability measure.

Part II. For the robust optimization (1.4a), we have $\mathcal{P} = \{\mathbb{P}_0\}$. According to Proposition 3.3, it is sufficient to show that $G(\boldsymbol{x}, \boldsymbol{\xi}^o) > 0$ for some $\boldsymbol{x} \in \widehat{\mathcal{X}}$ and $\boldsymbol{\xi}^o \in \operatorname{supp}(\mathbb{P}_0)$, which holds due to our preassumption in Proposition 3.3. This proves that the robust optimization framework (1.4a) may not be decision outlier robust. ∎

We make the following remarks on Theorem 3.5:

(i) The result of Theorem 3.5 implies that the value-of-risk (VaR) can also be decision outlier robust. Moreover, letting $\varepsilon = 1/2$ in (A.1) shows that the median is also decision outlier robust;

(ii) For general quantiles, the notion of "outlier robust" based on the influence curve from statistical robustness may not work, as implied in Example 6.

Decision Outlier Robustness of FCVaR and Its Alternatives. Next, we prove the decision outlier robustness of the proposed FCVaR and its alternatives.

THEOREM 3.6. *Suppose that the unamenable decision set $\widehat{\mathcal{X}}$ is non-empty and for any $\boldsymbol{x} \in \widehat{\mathcal{X}}$, there exists an $M \in \mathbb{R}$ such that $\mathbb{P}_0\{\tilde{\boldsymbol{\xi}} : Q(\boldsymbol{x}, \tilde{\boldsymbol{\xi}}) > M\} \le \varepsilon$, where $\mathbb{P}_0$ denotes the reference distribution and $\varepsilon \in (0,1)$ and function $Q(\boldsymbol{x}, \tilde{\boldsymbol{\xi}})$ is measurable for any $\boldsymbol{x} \in \mathcal{X}$. Then $\min_{\boldsymbol{x}\in\mathcal{X}} \mathbb{P}_0\text{-FCVaR}_{1-\varepsilon}\{Q(\boldsymbol{x}, \tilde{\boldsymbol{\xi}})\}$ is decision outlier robust.*

Proof. Based on Proposition 3.3, for $\mathbb{P}_0\text{-FCVaR}_{1-\varepsilon}\{Q(\boldsymbol{x}, \tilde{\boldsymbol{\xi}})\}$ defined in (2.3a), it is sufficient to show that for any unamenable decision $\boldsymbol{x} \in \widehat{\mathcal{X}}$, there exists a probability measure $\mathbb{P}^* \in \mathcal{P}_I$ such that $\mathbb{E}_{\mathbb{P}^*}[Q(\boldsymbol{x}, \tilde{\boldsymbol{\xi}})] < \infty$ and $\mathbb{P}^*\{\tilde{\boldsymbol{\xi}} : Q(\boldsymbol{x}, \tilde{\boldsymbol{\xi}}) = \infty\} = 0$.

Denote set $\mathcal{U}_1 = \{\tilde{\boldsymbol{\xi}} : Q(\boldsymbol{x}, \tilde{\boldsymbol{\xi}}) \le M\}$, which is $\mathbb{P}_0$-measurable (see, e.g., proposition 1 in section 3.1 of [59]). Given the presumption, we have $\mathbb{P}_0\{\mathcal{U}_1\} \ge 1 - \varepsilon$. Let us construct $\mathbb{P}^*(d\boldsymbol{\xi}) = \mathbb{P}_0(d\boldsymbol{\xi})/\mathbb{P}_0\{\mathcal{U}_1\}$ for each $\boldsymbol{\xi} \in \mathcal{U}_1$, 0, otherwise. Note that by our construction, we have $\mathbb{P}^*(\mathcal{U}_1) = 1, 0 \preceq \mathbb{P}^* \preceq \mathbb{P}_0/(1-\varepsilon)$ and hence $\mathbb{P}^* \in \mathcal{P}_I$. On the other hand, we also have

$$\mathbb{E}_{\mathbb{P}^*}\left[Q(\boldsymbol{x}, \tilde{\boldsymbol{\xi}})\right] \le M < \infty, \quad \mathbb{P}^*\left\{\tilde{\boldsymbol{\xi}} : Q(\boldsymbol{x}, \tilde{\boldsymbol{\xi}}) = \infty\right\} = 0.$$

This proves that $\mathbb{P}^*$ is a desirable probability measure. Hence, $\min_{\boldsymbol{x}\in\mathcal{X}} \mathbb{P}_0\text{-FCVaR}_{1-\varepsilon}\{Q(\boldsymbol{x}, \tilde{\boldsymbol{\xi}})\}$ is decision outlier robust. ∎

We make the following remarks about Theorem 3.6:

(i) The assumption that $\mathbb{P}_0\{\tilde{\boldsymbol{\xi}} : Q(\boldsymbol{x}, \tilde{\boldsymbol{\xi}}) > M\} \le \varepsilon$ is crucial to our analysis, which ensures that $\mathbb{E}_{\mathbb{P}^*}[Q(\boldsymbol{x}, \tilde{\boldsymbol{\xi}})] < \infty$ for some $\mathbb{P}^* \in \mathcal{P}_I$.

(ii) Similar to the chance constrained program (A.1), when the reference distribution is discrete, outlier robustness using the influence curve may not work based on the explanation in Example 7.

We conclude this section by remarking that the result in Theorem 3.6 can be extended to Winsorized CVaR and Huber-skip CVaR. The proofs are similar and thus are omitted.

COROLLARY 3.7. *Suppose that the unamenable decision set $\widehat{\mathcal{X}}$ is non-empty. For the reference distribution $\mathbb{P}_0$ and $\varepsilon \in (0,1)$, we have*

(i) the $\min_{\boldsymbol{x}\in\mathcal{X}} \mathbb{P}_0\text{-WCVaR}_{1-\varepsilon}\{Q(\boldsymbol{x}, \tilde{\boldsymbol{\xi}})\}$ is decision outlier robust if for any $\boldsymbol{x} \in \widehat{\mathcal{X}}$, there exists an $M \in \mathbb{R}$ such that $\mathbb{P}_0\{\tilde{\boldsymbol{\xi}} : Q(\boldsymbol{x}, \tilde{\boldsymbol{\xi}}) > M\} \le \varepsilon$; and

(ii) the $\min_{\boldsymbol{x}\in\mathcal{X}} \mathbb{P}_0\text{-HCVaR}\{Q(\boldsymbol{x}, \tilde{\boldsymbol{\xi}}), H\}$ is decision outlier robust.

The detailed comparisons among FCVaR, WCVaR, and HCVaR can be found in the numerical study section.

4 Achieving Out-of-Sample Performance Guarantees: Worst-case DFO. To effectively use i.i.d. samples to approximate the DFO models and achieve better out-of-sample performance guarantees, in this section, we propose applying data-driven distributional robustness (e.g., type$-\infty$ Wasserstein ambiguity set) to the corresponding DFO models. For the first special case of DFO in Section 1.1 (i.e., a chance constrained program), its worst-case counterpart, known as distributionally robust chance constrained programs (DRCCPs), has previously been investigated in the literature, aiming to attain better out-of-sample performance guarantees under conditions of limited available samples (see more discussions in [15, 25, 26, 28, 37, 66, 76]). It is worthy of mentioning that a DRCCP can be viewed as the combination of DFO and DRO, where the underlying chance constrained program aims to reduce the undesirable endogenous outliers and the distributional robustness improves the out-of-sample performances. Hence, to complement the existing results, this section focuses on the other special case of DFO–FCVaR, and studies its worst-case counterpart under the Wasserstein ambiguity set. While, at first glance, the DFO and DRO may seem to behave in opposite directions, in fact, they can be complementary. In an integrated model (the worst-case DFO), DFO and DRO can work together to improve both decision outlier robustness (reduce the effect of endogenous outliers) and out-of-sample performance. By doing so, the integrated model can coordinate the two approaches to achieve better overall performance. Particularly, we study the minimum of the worst-case FCVaR of the form

$$v_W^* = \min_{\boldsymbol{x}\in\mathcal{X}} \left\{\boldsymbol{c}^\top \boldsymbol{x} + \sup_{\mathbb{P}\in\mathcal{P}_\infty^W} \mathbb{P}\text{-FCVaR}_{1-\varepsilon}\left[Q(\boldsymbol{x}, \tilde{\boldsymbol{\xi}})\right]\right\}, \tag{4.1}$$

where we focus on type$-\infty$ Wasserstein ambiguity set

$$\mathcal{P}_\infty^W = \{\mathbb{P} \colon \mathbb{P}\{\tilde{\boldsymbol{\xi}} \in \mathcal{U}\} = 1, W_\infty(\mathbb{P}, \widehat{\mathbb{P}}) \le \theta\}.$$

Recall that $\widehat{\mathbb{P}}$ is a discrete empirical reference distribution of random parameters $\tilde{\boldsymbol{\xi}}$ generated by N i.i.d. samples with support $\mathcal{U}$ such that $\widehat{\mathbb{P}}\{\tilde{\boldsymbol{\xi}} = \widehat{\boldsymbol{\xi}}^i\} = 1/N$, i.e., $\widehat{\mathbb{P}} = 1/N \sum_{i\in[N]} \delta_{\widehat{\boldsymbol{\xi}}^i}$ and $\delta_{\widehat{\boldsymbol{\xi}}^i}$ is the Dirac function that places unit mass on the realization $\tilde{\boldsymbol{\xi}} = \widehat{\boldsymbol{\xi}}^i$ for each $i \in [N]$, $\theta \geq 0$ is the Wasserstein radius, and the $\infty-$Wasserstein distance between two probability distributions $\mathbb{P}_1, \mathbb{P}_2$ with ℓ_p norm is defined as

$$W_\infty(\mathbb{P}_1,\mathbb{P}_2) = \inf\left\{ \text{ess.sup}_{\mathbb{Q}} \left\|\boldsymbol{\xi}^1 - \boldsymbol{\xi}^2\right\|_p : \begin{array}{c} \mathbb{Q} \text{ is a joint distribution of } \tilde{\boldsymbol{\xi}}^1 \text{ and } \tilde{\boldsymbol{\xi}}^2 \\ \text{with marginals } \mathbb{P}_1 \text{ and } \mathbb{P}_2, \text{ respectively} \end{array} \right\}.$$

Le $\mathbb{P}^T$ be the true distribution of random parameters $\tilde{\boldsymbol{\xi}}$ and let $\widehat{\boldsymbol{x}}^*$ denote an optimal solution of the minimum of the worst-case FCVaR (4.1). Motivated by [20], the out-of-sample probability, which is often small, is defined as

$$\mathbb{P}^T\left\{\tilde{\boldsymbol{\xi}} \colon v_W^* < \boldsymbol{c}^\top \widehat{\boldsymbol{x}}^* + \mathbb{P}^T\text{-FCVaR}_{1-\varepsilon}\left[Q(\widehat{\boldsymbol{x}}^*, \tilde{\boldsymbol{\xi}})\right]\right\}. \tag{4.2}$$

That is, it ensures that the probability that the optimal value from the minimum of the worst-case FCVaR (4.1) is smaller than the true objective is small. In the numerical study, we let the probability (4.2) be no larger than 5%.

4.1 Worst-case FCVaR is Equivalent to DRO with Favorable Sample-selection. We first show that the minimum of the worst-case FCVaR (4.1) admits a neat representation.

Theorem 4.1. *The minimum of the worst-case* FCVaR (4.1) *is equivalent to*

$$v_W^* = \min_{\boldsymbol{x}\in\mathcal{X}} \left\{\boldsymbol{c}^\top \boldsymbol{x} + \widehat{\mathbb{P}}\text{-FCVaR}_{1-\varepsilon}\left[\bar{Q}(\boldsymbol{x}, \widehat{\boldsymbol{\xi}})\right]\right\}, \tag{4.3}$$

where the robustified recourse function is defined as $\bar{Q}(\boldsymbol{x}, \widehat{\boldsymbol{\xi}}) := \max_{\boldsymbol{\xi}}\{Q(\boldsymbol{x},\boldsymbol{\xi}) : \|\boldsymbol{\xi} - \widehat{\boldsymbol{\xi}}\|_p \leq \theta\}$.

Proof. According to the definition of $\text{FCVaR}_{1-\varepsilon}$ (2.2), the minimum of the worst-case FCVaR (4.1) is equivalent to

$$\min_{\boldsymbol{x}\in\mathcal{X}} \left\{\boldsymbol{c}^\top \boldsymbol{x} + \sup_{\mathbb{P}\in\mathcal{P}_\infty^W} \mathbb{P}\text{-FCVaR}_{1-\varepsilon}\left[Q(\boldsymbol{x}, \tilde{\boldsymbol{\xi}})\right]\right\} = \min_{\boldsymbol{x}\in\mathcal{X}} \left\{\boldsymbol{c}^\top \boldsymbol{x} + \sup_{\mathbb{P}\in\mathcal{P}_\infty^W} \max_\beta \left\{\beta + \frac{1}{1-\varepsilon}\mathbb{E}_{\mathbb{P}}\left[\left(Q(\boldsymbol{x}, \tilde{\boldsymbol{\xi}}) - \beta\right)_-\right]\right\}\right\}.$$

Interchanging the supremum operator and the maximum operator, we have

$$\min_{\boldsymbol{x}\in\mathcal{X}} \left\{\boldsymbol{c}^\top \boldsymbol{x} + \sup_{\mathbb{P}\in\mathcal{P}_\infty^W} \mathbb{P}\text{-FCVaR}_{1-\varepsilon}\left[Q(\boldsymbol{x}, \tilde{\boldsymbol{\xi}})\right]\right\} = \min_{\boldsymbol{x}\in\mathcal{X}} \left\{\boldsymbol{c}^\top \boldsymbol{x} + \max_\beta \sup_{\mathbb{P}\in\mathcal{P}_\infty^W} \left\{\beta + \frac{1}{1-\varepsilon}\mathbb{E}_{\mathbb{P}}\left[\left(Q(\boldsymbol{x}, \tilde{\boldsymbol{\xi}}) - \beta\right)_-\right]\right\}\right\}.$$

Recall the following equivalent representation in type$-\infty$ Wasserstein ambiguity set with discrete empirical reference distribution $\widehat{\mathbb{P}}$ and its corresponding random vector $\widehat{\boldsymbol{\xi}}$ (see, e.g., proposition 3 in [9]):

$$\sup_{\mathbb{P}\in\mathcal{P}_\infty^W} \mathbb{E}_{\mathbb{P}}\left[Q(\boldsymbol{x}, \tilde{\boldsymbol{\xi}})\right] = \mathbb{E}_{\widehat{\mathbb{P}}}\left[\max_{\boldsymbol{\xi}}\left\{Q(\boldsymbol{x}, \boldsymbol{\xi}) \colon \|\boldsymbol{\xi} - \widehat{\boldsymbol{\xi}}\|_p \leq \theta\right\}\right] = \mathbb{E}_{\widehat{\mathbb{P}}}\left[\bar{Q}(\boldsymbol{x}, \widehat{\boldsymbol{\xi}})\right],$$

which implies that

$$v_W^* = \min_{\boldsymbol{x}\in\mathcal{X}} \left\{\boldsymbol{c}^\top \boldsymbol{x} + \sup_{\mathbb{P}\in\mathcal{P}_\infty^W} \mathbb{P}\text{-FCVaR}_{1-\varepsilon}\left[Q(\boldsymbol{x}, \tilde{\boldsymbol{\xi}})\right]\right\} = \min_{\boldsymbol{x}\in\mathcal{X}} \left\{\boldsymbol{c}^\top \boldsymbol{x} + \max_\beta \left\{\beta + \frac{1}{1-\varepsilon}\mathbb{E}_{\widehat{\mathbb{P}}}\left[\left(\bar{Q}(\boldsymbol{x}, \widehat{\boldsymbol{\xi}}) - \beta\right)_-\right]\right\}\right\}.$$

Plugging back the definition of $\text{FCVaR}_{1-\varepsilon}$ (2.2), we have the desired formulation. ◻

It turns out that when $N\varepsilon$ is an integer (this can always be done in practice by carefully choosing the sample size or using bootstrapping), the minimum of the worst-case FCVaR (4.1) in fact can be interpreted as the minimum of the a DRO model with sample-selection Wasserstein ambiguity set, i.e., it both selects the most favorable scenarios and guarantees the out-of-sample performance. The key idea of the sample-selection Wasserstein ambiguity set is to optimally select the most favorable $k := N - N\varepsilon$ out of N empirical samples and then construct the corresponding Wasserstein ambiguity set based on selected k empirical samples. For example, given a collection S of k samples, we denote its corresponding type$-\infty$ Wasserstein ambiguity set as $\mathcal{P}_\infty^W(S)$, which is defined as

$$\mathcal{P}_\infty^W(S) = \{\mathbb{P} \colon \mathbb{P}\{\tilde{\boldsymbol{\xi}} \in \mathcal{U}\} = 1, W_\infty(\mathbb{P}, \widehat{\mathbb{P}}(S)) \leq \theta\}.$$

Here, $\widehat{\mathbb{P}}(S)$ denotes an equiprobable discrete probability distribution supported on a size-k subset of samples $\{\widehat{\boldsymbol{\xi}}^i\}_{i\in S\subseteq[N]}$ such that $\widehat{\mathbb{P}}\{\tilde{\boldsymbol{\xi}} = \widehat{\boldsymbol{\xi}}^i\} = 1/k$ for $i \in S$. Intuitively, the DRO with sample-selection Wasserstein ambiguity set can be written as

$$v_R^* = \min_{\substack{\boldsymbol{x}\in\mathcal{X},\\ S\in\mathcal{S}}} \left\{ \boldsymbol{c}^\top \boldsymbol{x} + \sup_{\mathbb{P}\in\mathcal{P}_\infty^W(S)} \mathbb{E}_{\mathbb{P}}\left[Q(\boldsymbol{x},\tilde{\boldsymbol{\xi}})\right]\right\}, \tag{4.4}$$

where $\mathcal{S}$ denotes all the size-k subsets of samples.

Letting the binary variable z_i indicate whether the ith sample is selected or not, according to the result in [9, 75], under type$-\infty$ Wasserstein ambiguity set, problem (4.4) can be represented as

$$v_R^* = \min_{\substack{\boldsymbol{x}\in\mathcal{X},\\ S\in\mathcal{S}}} \left\{ \boldsymbol{c}^\top \boldsymbol{x} + \sup_{\mathbb{P}\in\mathcal{P}_\infty^W(S)} \mathbb{E}_{\mathbb{P}}\left[Q(\boldsymbol{x},\tilde{\boldsymbol{\xi}})\right]\right\} = \min_{\substack{\boldsymbol{x}\in\mathcal{X},\\ \boldsymbol{z}\in\{0,1\}^N}} \left\{ \boldsymbol{c}^\top \boldsymbol{x} + \left\{ \frac{1}{k}\sum_{i\in[N]} z_i \bar{Q}(\boldsymbol{x},\widehat{\boldsymbol{\xi}}^i) \colon \sum_{i\in[N]} z_i = k \right\}\right\}, \tag{4.5}$$

which is exactly the minimum of the worst-case FCVaR. This result is summarized below.

Proposition 4.2. *Given that type$-\infty$ Wasserstein ambiguity set is considered and $N\varepsilon$ is an integer, the minimum of the worst-case* FCVaR (4.1) *is equivalent to the DRO with a favorable sample-selection Wasserstein ambiguity set* (4.4)*, i.e., $v_W^* = v_R^*$.*

This result shows that applying distributional robustness essentially selects favorable samples optimally, consistent with the findings in the previous sections that are beyond the simple preprocessing and are important to eliminate endogenous outliers.

We note that, because of the translation invariance property, we can shift the first-stage objective function $\boldsymbol{c}^\top\boldsymbol{x}$ to the second stage, that is,

$$\boldsymbol{c}^\top \boldsymbol{x} + \widehat{\mathbb{P}}\text{-FCVaR}_{1-\varepsilon}\left[Q(\boldsymbol{x},\widehat{\boldsymbol{\xi}})\right] = \widehat{\mathbb{P}}\text{-FCVaR}_{1-\varepsilon}\left[\boldsymbol{c}^\top \boldsymbol{x} + Q(\boldsymbol{x},\widehat{\boldsymbol{\xi}})\right]. \tag{4.6}$$

For ease of notation in the following discussions within this section, we absorb the linear objective function $\boldsymbol{c}^\top\boldsymbol{x}$ into the recourse function $Q(\boldsymbol{x},\tilde{\boldsymbol{\xi}})$, i.e., we redefine $Q(\boldsymbol{x},\tilde{\boldsymbol{\xi}}) := Q(\boldsymbol{x},\tilde{\boldsymbol{\xi}}) + \boldsymbol{c}^\top\boldsymbol{x}$.

4.2 Confidence Bounds and Decision Outlier Robustness of the Worst-case FCVaR. Given a discrete empirical reference distribution $\widehat{\mathbb{P}}$ generated by N i.i.d. samples of random parameters $\tilde{\boldsymbol{\xi}}$, we proceed in this subsection by comparing the objective value of (4.3) with the optimal value obtained from the true distribution. This analysis further motivates us on how to select the Wasserstein radius θ. Before deriving the confidence bounds, we define the following important quantities. We let v^T denote the minimum FCVaR under the true distribution $\mathbb{P}^T$, that is,

$$v^T = \min_{\boldsymbol{x}\in\mathcal{X}} \max_{\beta} \left\{ \beta + \frac{1}{1-\varepsilon}\mathbb{E}_{\mathbb{P}^T}\left[\left(Q(\boldsymbol{x},\tilde{\boldsymbol{\xi}}) - \beta\right)_-\right]\right\},$$

and for any decision $\boldsymbol{x}\in\mathcal{X}$, we let $\beta^*(\boldsymbol{x})$ denote an optimal solution of inner maximization, i.e., according to Proposition 2.1, we have $\beta^*(\boldsymbol{x}) = \mathbb{P}^T\text{-VaR}_{1-\varepsilon}\{Q(\boldsymbol{x},\tilde{\boldsymbol{\xi}})\}$.

We make the following additional assumptions, which are quite standard in the literature.

Assumption 2. *(i) (**Truncated Concentration Bound**) There exists a positive σ such that $\mathbb{E}_{\mathbb{P}^T}[\exp(([(Q(\boldsymbol{x},\tilde{\boldsymbol{\xi}}) - \beta^*(\boldsymbol{x}))_-] - \mathbb{E}_{\mathbb{P}^T}[(Q(\boldsymbol{x},\tilde{\boldsymbol{\xi}}) - \beta^*(\boldsymbol{x}))_-])^2/\sigma^2)] \le e$ a.s. for each $\boldsymbol{x}\in\mathcal{X}$;*

*(ii) (**Lipschitz Continuity of Recourse Function within a Truncated Support**) There exists a positive parameter $\Delta_1 > 0$ such that within a $\mathbb{P}^T$-measurable set $\widehat{\mathcal{U}}(\Delta_1) := \{\boldsymbol{\xi} : Q(\boldsymbol{x},\boldsymbol{\xi}) \le \beta^*(\boldsymbol{x}) + \Delta_1\}$, the function $Q(\boldsymbol{x},\boldsymbol{\xi})$ is Lipschitz continuous with respect to both $\boldsymbol{x}$ and $\boldsymbol{\xi}$, i.e., $|Q(\boldsymbol{x},\boldsymbol{\xi}^1) - Q(\boldsymbol{y},\boldsymbol{\xi}^2)| \le L\|(\boldsymbol{x},\boldsymbol{\xi}^1) - (\boldsymbol{y},\boldsymbol{\xi}^2)\|_p$ for all $\boldsymbol{x},\boldsymbol{y}\in\mathcal{X}, \boldsymbol{\xi}^1,\boldsymbol{\xi}^2 \in \widehat{\mathcal{U}}(\Delta_1)$; and*

*(iii) (**Local Smoothness of True Cumulative Distribution Function (CDF) around Quantile $\beta^*(\mathrm{x})$**) There exist $\Delta_2 > 0$ and $\ell > 0$ such that $|\mathbb{P}^T\{\tilde{\boldsymbol{\xi}} : Q(\boldsymbol{x},\tilde{\boldsymbol{\xi}}) \le \beta^*(\boldsymbol{x}) + \widehat{\Delta}\} - \mathbb{P}^T\{\tilde{\boldsymbol{\xi}} : Q(\boldsymbol{x},\tilde{\boldsymbol{\xi}}) \le \beta^*(\boldsymbol{x})\}| \ge \ell|\widehat{\Delta}|$ for any $\widehat{\Delta}\in[-\Delta_2,\Delta_2]$ and for all $\boldsymbol{x}\in\mathcal{X}$.*

Note that in Assumption 2, Part (i) is standard in the concentration inequality literature (see, e.g., chapter 2 of [72]). Part (ii) is a common way of addressing the Lipschitz continuity of functions that are smooth within a smaller sub-domain (see more details in [27]). Part (iii) follows from the existing literature on the sample size estimation of the chance constrained programs (see, e.g., [31, 44]), which guarantees that the true underlying distribution has a positive probability density around a neighborhood of the $(1-\varepsilon)$-quantile.

We then develop the non-asymptotic confidence bounds of the minimum of the worst-case FCVaR under type$-\infty$ Wasserstein ambiguity set.

THEOREM 4.3. *(Confidence Bounds) Suppose that Assumption 2 holds. Then for any given* $\widehat{\gamma} \in (0,1)$, *we have: (i)* $\mathbb{P}^T\{v_W^* \leq v^T + 2L\theta\} \geq 1-\widehat{\gamma}$; *and (ii)* $\mathbb{P}^T\{v_W^* \geq v^T - L\theta\} \geq 1-\widehat{\gamma}$, *where* $\theta = \mathcal{O}(1)N^{-1/2}\sqrt{n\log(\widehat{\gamma}^{-1})}$ *for a discrete compact set* $\mathcal{X}$, *and* $\theta = \mathcal{O}(1)N^{-1/2}\sqrt{n\log(nN)\log(\widehat{\gamma}^{-1})}$ *for a general compact set* $\mathcal{X}$.

Proof. The proof of Part (ii) is similar to that of Part (i) and thus is omitted. We split the proof into five steps.

Step I. Let us use v_N^{SAA} to denote the sampling average approximation (SAA) counterpart of the FCVaR with N i.i.d. samples $\{\widehat{\boldsymbol{\xi}}^i\}_{i\in[N]}$, which admits the following form

$$v_N^{SAA} = \min_{\boldsymbol{x}\in\mathcal{X}} \widehat{\mathbb{P}}\text{-FCVaR}_{1-\varepsilon}\left[Q(\boldsymbol{x},\widehat{\boldsymbol{\xi}})\right] = \min_{\boldsymbol{x}\in\mathcal{X}}\max_{\beta_N}\left\{\beta_N + \frac{1}{N-N\varepsilon}\sum_{i\in[N]}\left[\left(Q(\boldsymbol{x},\widehat{\boldsymbol{\xi}}^i) - \beta_N\right)_-\right]\right\}.$$

Under the true distribution $\mathbb{P}^T$, let us define the FCVaR with the decision $\boldsymbol{x}\in\mathcal{X}$ as

$$v^T(\boldsymbol{x}) = \mathbb{P}^T\text{-FCVaR}_{1-\varepsilon}\left[Q(\boldsymbol{x},\tilde{\boldsymbol{\xi}})\right] = \max_{\beta(\boldsymbol{x})}\left\{\beta(\boldsymbol{x}) + \frac{1}{1-\varepsilon}\mathbb{E}_{\mathbb{P}^T}\left[\left(Q(\boldsymbol{x},\tilde{\boldsymbol{\xi}}) - \beta(\boldsymbol{x})\right)_-\right]\right\}.$$

Recall that an optimal $\beta^*(\boldsymbol{x}) = F^{-1}(1-\varepsilon)$, where we let $F(\cdot)$ denote the CDF of random parameter $Q(\boldsymbol{x},\tilde{\boldsymbol{\xi}})$ with respect to true distribution $\mathbb{P}^T$. We also denote the SAA counterpart as

$$v_N^{SAA}(\boldsymbol{x}) = \widehat{\mathbb{P}}\text{-FCVaR}_{1-\varepsilon}\left[Q(\boldsymbol{x},\tilde{\boldsymbol{\xi}})\right] = \max_{\beta_N(\boldsymbol{x})}\left\{\beta_N(\boldsymbol{x}) + \frac{1}{N-N\varepsilon}\sum_{i\in[N]}\left[\left(Q(\boldsymbol{x},\widehat{\boldsymbol{\xi}}^i) - \beta_N(\boldsymbol{x})\right)_-\right]\right\},$$

with an optimal $\beta_N^*(\boldsymbol{x}) = F_N^{-1}(1-\varepsilon)$, where $F_N(\cdot)$ denotes the CDF of random parameter $Q(\boldsymbol{x},\widehat{\boldsymbol{\xi}})$ with respect to empirical distribution $\widehat{\mathbb{P}}$.

According to Hoeffding's inequality (see, e.g., [30]), for a small $\bar{\Delta} > 0$ and $0 < \widehat{\Delta}_N \leq \Delta_2$, we have

$$\mathbb{P}^T\left\{F_N\left(\beta^*(\boldsymbol{x}) + \widehat{\Delta}_N\right) - F\left(\beta^*(\boldsymbol{x}) + \widehat{\Delta}_N\right) \geq -\bar{\Delta}\right\} \geq 1 - \exp\{-2N\bar{\Delta}^2\}. \tag{4.7a}$$

According to Part (iii) of Assumption 2, for some $\ell > 0$, we have

$$F\left(\beta^*(\boldsymbol{x}) + \widehat{\Delta}_N\right) - F(\beta^*(\boldsymbol{x})) \geq \ell\widehat{\Delta}_N.$$

Using this result, inequality (4.7a) implies that

$$\mathbb{P}^T\left\{F_N\left(\beta^*(\boldsymbol{x}) + \widehat{\Delta}_N\right) \geq 1 - \varepsilon + \ell\widehat{\Delta}_N - \bar{\Delta}\right\} \geq 1 - \exp\{-2N\bar{\Delta}^2\}.$$

By letting $\ell\widehat{\Delta}_N = \bar{\Delta}$, we have

$$\mathbb{P}^T\left\{F_N\left(\beta^*(\boldsymbol{x}) + \widehat{\Delta}_N\right) < 1 - \varepsilon\right\} \leq \exp\{-2N(\ell\widehat{\Delta}_N)^2\}.$$

On the other hand, we have $\mathbb{P}^T\{F_N(\beta^*(\boldsymbol{x}) - \widehat{\Delta}_N) > 1-\varepsilon\} \leq \exp\{-2N(\ell\widehat{\Delta}_N)^2\}$. Then, recall the definitions of $\beta_N^*(\boldsymbol{x})$ and $\beta^*(\boldsymbol{x})$, by simple calculations, we have

$$\begin{aligned}
&\mathbb{P}^T\left\{|\beta_N^*(\boldsymbol{x}) - \beta^*(\boldsymbol{x})| \leq \widehat{\Delta}_N\right\} = \mathbb{P}^T\left\{F_N\left(\beta^*(\boldsymbol{x}) + \Delta\right) \geq 1-\varepsilon, F_N\left(\beta^*(\boldsymbol{x}) - \Delta\right) \leq 1-\varepsilon\right\} \\
&\geq 1 - \mathbb{P}^T\left\{F_N\left(\beta^*(\boldsymbol{x}) + \widehat{\Delta}_N\right) < 1-\varepsilon\right\} - \mathbb{P}^T\left\{F_N\left(\beta^*(\boldsymbol{x}) - \widehat{\Delta}_N\right) > 1-\varepsilon\right\} \geq 1 - 2\exp\left\{-2N(\ell\widehat{\Delta}_N)^2\right\}.
\end{aligned} \tag{4.7b}$$

Step II. According to Part (ii) of Assumption 2, we have

$$v_W^* \leq \min_{\boldsymbol{x}\in\mathcal{X}}\max_{\beta_N}\left\{\beta_N + \frac{1}{N-N\varepsilon}\sum_{i\in[N]}\left[\left(Q(\boldsymbol{x},\widehat{\boldsymbol{\xi}}^i) + \max_{\boldsymbol{\xi}}\left\{L\|\boldsymbol{\xi} - \widehat{\boldsymbol{\xi}}^i\| : \|\boldsymbol{\xi} - \widehat{\boldsymbol{\xi}}^i\|_p \leq \theta\right\} - \beta_N\right)_-\right]\right\}.$$

Optimizing over $\boldsymbol{\xi}$ and invoking the definition of v_N^{SAA}, we have

$$v_W^* \leq \min_{\boldsymbol{x}\in\mathcal{X}}\max_{\beta_N}\left\{\beta_N + \frac{1}{N-N\varepsilon}\sum_{i\in[N]}\left[\left(Q(\boldsymbol{x},\widehat{\boldsymbol{\xi}}^i) + L\theta - \beta_N\right)_-\right]\right\} \leq v_N^{SAA} + L\theta.$$

Then, it is sufficient to prove

$$\mathbb{P}^T\left\{v_N^{SAA} \leq v^T + L\theta\right\} \geq 1 - \widehat{\gamma}.$$

Step III. Given that the quantile is close to the true quantile (i.e., the inequalities from Step I hold), we derive the bounds of the difference of the objective functions.

There are two subcases to consider: whether $\beta_N^*(\boldsymbol{x}) - \beta^*(\boldsymbol{x})$ is negative or not.

Case (a). When $0 \leq \beta_N^*(\boldsymbol{x}) - \beta^*(\boldsymbol{x}) \leq \widehat{\Delta}_N$, we have

$$\begin{aligned}&\beta_N^*(\boldsymbol{x}) - \beta^*(\boldsymbol{x}) + \frac{1}{N - N\varepsilon} \sum_{i\in[N]} \left[\left[\left(Q(\boldsymbol{x}, \widehat{\boldsymbol{\xi}}^i) - \beta_N^*(\boldsymbol{x}) \right)_- \right] - \mathbb{E}_{\mathbb{P}^T} \left[\left(Q(\boldsymbol{x}, \tilde{\boldsymbol{\xi}}) - \beta^*(\boldsymbol{x}) \right)_- \right] \right] \\ \leq & \frac{1}{N - N\varepsilon} \sum_{i\in[N]} \left[\left[\left(Q(\boldsymbol{x}, \widehat{\boldsymbol{\xi}}^i) - \beta^*(\boldsymbol{x}) \right)_- \right] - \mathbb{E}_{\mathbb{P}^T} \left[\left(Q(\boldsymbol{x}, \tilde{\boldsymbol{\xi}}) - \beta^*(\boldsymbol{x}) \right)_- \right] \right] + \frac{\widehat{\Delta}_N}{1-\varepsilon}.\end{aligned}$$

where the inequality is due to the conditions $\beta_N^*(\boldsymbol{x}) - \beta^*(\boldsymbol{x}) \leq \widehat{\Delta}_N$ and $\varepsilon \in (0, 1)$.

Case (b). When $-\widehat{\Delta}_N \leq \beta_N^*(\boldsymbol{x}) - \beta^*(\boldsymbol{x}) < 0$, we have

$$\begin{aligned}&\beta_N^*(\boldsymbol{x}) - \beta^*(\boldsymbol{x}) + \frac{1}{N - N\varepsilon} \sum_{i\in[N]} \left[\left[\left(Q(\boldsymbol{x}, \widehat{\boldsymbol{\xi}}^i) - \beta_N^*(\boldsymbol{x}) \right)_- \right] - \mathbb{E}_{\mathbb{P}^T} \left[\left(Q(\boldsymbol{x}, \tilde{\boldsymbol{\xi}}) - \beta^*(\boldsymbol{x}) \right)_- \right] \right] \\ \leq & \frac{1}{N - N\varepsilon} \sum_{i\in[N]} \left[\left[\left(Q(\boldsymbol{x}, \widehat{\boldsymbol{\xi}}^i) - \beta^*(\boldsymbol{x}) \right)_- \right] - \mathbb{E}_{\mathbb{P}^T} \left[\left(Q(\boldsymbol{x}, \tilde{\boldsymbol{\xi}}) - \beta^*(\boldsymbol{x}) \right)_- \right] \right] + \frac{\widehat{\Delta}_N}{1-\varepsilon}.\end{aligned}$$

where the inequality is due to the conditions $\beta_N^*(\boldsymbol{x}) - \beta^*(\boldsymbol{x}) < 0$, $\widehat{\Delta}_N/(1-\varepsilon) > 0$, and $\beta_N^*(\boldsymbol{x}) \geq \beta^*(\boldsymbol{x}) - \widehat{\Delta}_N$.

Therefore, when $|\beta_N^*(\boldsymbol{x}) - \beta^*(\boldsymbol{x})| \leq \widehat{\Delta}_N$, we have

$$\begin{aligned}&\beta_N^*(\boldsymbol{x}) - \beta^*(\boldsymbol{x}) + \frac{1}{N - N\varepsilon} \sum_{i\in[N]} \left[\left[\left(Q(\boldsymbol{x}, \widehat{\boldsymbol{\xi}}^i) - \beta_N^*(\boldsymbol{x}) \right)_- \right] - \mathbb{E}_{\mathbb{P}^T} \left[\left(Q(\boldsymbol{x}, \tilde{\boldsymbol{\xi}}) - \beta^*(\boldsymbol{x}) \right)_- \right] \right] \\ \leq & \frac{1}{N - N\varepsilon} \sum_{i\in[N]} \left[\left[\left(Q(\boldsymbol{x}, \widehat{\boldsymbol{\xi}}^i) - \beta^*(\boldsymbol{x}) \right)_- \right] - \mathbb{E}_{\mathbb{P}^T} \left[\left(Q(\boldsymbol{x}, \tilde{\boldsymbol{\xi}}) - \beta^*(\boldsymbol{x}) \right)_- \right] \right] + \frac{\widehat{\Delta}_N}{1-\varepsilon}. \quad (4.7c)\end{aligned}$$

Now, we are going to apply lemma A.1 in [22] to provide the probability bound for $\mathbb{P}^T\{v_N(\boldsymbol{x}) - \lambda_2\sigma/\sqrt{N} \leq v^T(\boldsymbol{x})\}$ for any $\lambda_2 > 0$. Given a positive parameter $\lambda_1 > 0$, let us define $\lambda_2 = 2\lambda_1/(1-\varepsilon)$ and $\widehat{\Delta}_N = \lambda_1\sigma/\sqrt{N} \leq \min\{\Delta_1, \Delta_2\}$, that is,

$$\frac{\widehat{\Delta}_N}{1-\varepsilon} = \frac{\lambda_1\sigma}{(1-\varepsilon)\sqrt{N}} = \frac{\lambda_2\sigma}{2\sqrt{N}}. \quad (4.7d)$$

According to equation (4.7d), we have

$$\mathbb{P}^T \left\{ v_N(\boldsymbol{x}) - \frac{\lambda_2\sigma}{\sqrt{N}} \leq v^T(\boldsymbol{x}) \right\} = \mathbb{P}^T \left\{ v_N(\boldsymbol{x}) - \frac{\lambda_1\sigma}{(1-\varepsilon)\sqrt{N}} - \frac{\widehat{\Delta}_N}{1-\varepsilon} \leq v^T(\boldsymbol{x}) \right\}. \quad (4.7e)$$

Invoking the definition of $v^T(\boldsymbol{x})$ and $v_N(\boldsymbol{x})$, we can rewrite (4.7e) as

$$\begin{aligned}&\mathbb{P}^T \left\{ v_N(\boldsymbol{x}) - \frac{\lambda_2\sigma}{\sqrt{N}} \leq v^T(\boldsymbol{x}) \right\} \\ =&\mathbb{P}^T \left\{ \beta_N^*(\boldsymbol{x}) - \beta^*(\boldsymbol{x}) + \frac{1}{N - N\varepsilon} \sum_{i\in[N]} \left[\left[\left(Q(\boldsymbol{x}, \widehat{\boldsymbol{\xi}}^i) - \beta_N^*(\boldsymbol{x}) \right)_- \right] - \mathbb{E}_{\mathbb{P}^T} \left[\left(Q(\boldsymbol{x}, \tilde{\boldsymbol{\xi}}) - \beta^*(\boldsymbol{x}) \right)_- \right] \right] - \frac{\widehat{\Delta}_N}{1-\varepsilon} \right. \\ &\left. \leq \frac{\lambda_1\sigma}{(1-\varepsilon)\sqrt{N}} \right\}.\end{aligned}$$

By the law of total probability (see, e.g., appendix A of [70]), we have

$$\begin{aligned}&\mathbb{P}^T \left\{ v_N(\boldsymbol{x}) - \frac{\lambda_2\sigma}{\sqrt{N}} \leq v^T(\boldsymbol{x}) \right\} \\ \geq&\mathbb{P}^T \left\{ \beta_N^*(\boldsymbol{x}) - \beta^*(\boldsymbol{x}) + \frac{1}{N - N\varepsilon} \sum_{i\in[N]} \left[\left[\left(Q(\boldsymbol{x}, \widehat{\boldsymbol{\xi}}^i) - \beta_N^*(\boldsymbol{x}) \right)_- \right] - \mathbb{E}_{\mathbb{P}^T} \left[\left(Q(\boldsymbol{x}, \tilde{\boldsymbol{\xi}}) - \beta^*(\boldsymbol{x}) \right)_- \right] \right] - \frac{\widehat{\Delta}_N}{1-\varepsilon} \right.\end{aligned}$$

$$\leq \frac{\lambda_1\sigma}{(1-\varepsilon)\sqrt{N}}, |\beta_N^*(\boldsymbol{x}) - \beta^*(\boldsymbol{x})| \leq \widehat{\Delta}_N \Bigg\}.$$

According to inequality (4.7c), we have

$$\begin{aligned}
&\mathbb{P}^T \left\{ v_N(\boldsymbol{x}) - \frac{\lambda_2\sigma}{\sqrt{N}} \leq v^T(\boldsymbol{x}) \right\} \\
\geq &\mathbb{P}^T \Bigg\{ \frac{1}{N-N\varepsilon} \sum_{i\in[N]} \left[\left[\left(Q(\boldsymbol{x}, \widehat{\boldsymbol{\xi}}^i) - \beta^*(\boldsymbol{x}) \right)_- \right] - \mathbb{E}_{\mathbb{P}^T} \left[\left(Q(\boldsymbol{x}, \tilde{\boldsymbol{\xi}}) - \beta^*(\boldsymbol{x}) \right)_- \right] \right] \\
&\qquad \leq \frac{\lambda_1\sigma}{(1-\varepsilon)\sqrt{N}}, |\beta_N^*(\boldsymbol{x}) - \beta^*(\boldsymbol{x})| \leq \widehat{\Delta}_N \Bigg\} \\
\geq &\mathbb{P}^T \left\{ \frac{1}{N} \sum_{i\in[N]} \left[\left[\left(Q(\boldsymbol{x}, \widehat{\boldsymbol{\xi}}^i) - \beta^*(\boldsymbol{x}) \right)_- \right] - \mathbb{E}_{\mathbb{P}^T} \left[\left(Q(\boldsymbol{x}, \tilde{\boldsymbol{\xi}}) - \beta^*(\boldsymbol{x}) \right)_- \right] \right] \leq \frac{\lambda_1\sigma}{\sqrt{N}} \right\} \\
&\qquad + \mathbb{P}^T \left\{ |\beta_N^*(\boldsymbol{x}) - \beta^*(\boldsymbol{x})| \leq \widehat{\Delta}_N \right\} - 1,
\end{aligned}$$

where the second equality is due to the union bound (see, e.g., [11]).

Defining $c^i = [Q(\boldsymbol{x}, \widehat{\boldsymbol{\xi}}^i) - \beta^*(\boldsymbol{x})]_-$ and $c^T = \mathbb{E}_{\mathbb{P}^T}[Q(\boldsymbol{x}, \tilde{\boldsymbol{\xi}}) - \beta^*(\boldsymbol{x})]_-$ and applying lemma A.1 in [22] with $d_i = c^i - c^T$ for each $i \in [N]$, together with inequalities (4.7b), for any $\boldsymbol{x} \in \mathcal{X}$, we have

$$\begin{aligned}
&\mathbb{P}^T \left\{ v_N(\boldsymbol{x}) - \frac{\lambda_2\sigma}{\sqrt{N}} \leq v^T(\boldsymbol{x}) \right\} \geq \left[1 - \exp\{\lambda_1^2/3\}\right] + \left[1 - 2\exp\{-2N(\ell\widehat{\Delta}_N)^2\}\right] - 1 \\
\geq &1 - \exp\{-\lambda_1^2/3\} - 2\exp\{-\ell^2(1-\varepsilon)^2\lambda_1^2\sigma^2/2\}.
\end{aligned}$$

Step IV. When set $\mathcal{X}$ is discrete, then applying the union bound, we have

$$\mathbb{P}^T \left\{ v_N^{SAA} - \frac{\lambda_2\sigma}{\sqrt{N}} \leq v^T \right\} \geq 1 - |\mathcal{X}|\exp\{-\lambda_1^2/3\} - 2|\mathcal{X}|\exp\{-\ell^2(1-\varepsilon)^2\lambda_1^2\sigma^2/2\},$$

with sample size N at least to be $\log(2/\widehat{\gamma})/(2(\ell\Delta_N)^2)$.

Assume that $|\mathcal{X}| \leq r^n$ and let $\widehat{\gamma}/3 = r^n \max\left\{\exp\{-\lambda_1^2/3\}, \exp\{-l^2(1-\varepsilon)^2\lambda_1^2\sigma^2/2\}\right\}$, which implies that

$$\frac{\widehat{\gamma}}{3} \geq r^n \exp\{-\lambda_1^2/3\}, \quad \frac{\widehat{\gamma}}{3} \geq r^n \exp\{-\ell^2(1-\varepsilon)^2\lambda_1^2\sigma^2/2\}.$$

By simple calculation, we have

$$\lambda_1 = \max\left\{ \sqrt{3n\log(r) - 3\log(\widehat{\gamma}/3)}, \sqrt{\frac{2n\log(r) - 2\log(\widehat{\gamma}/3)}{\ell^2(1-\varepsilon)^2\sigma^2}} \right\}.$$

We can choose $\theta := 2\lambda_1\sigma L^{-1}N^{-1/2}(1-\varepsilon)^{-1} = \mathcal{O}(1)N^{-1/2}\sqrt{n\log(\widehat{\gamma}^{-1})}$ and we have the conclusion.

Step V. We are going to analyze the more general setting, i.e., when set $\mathcal{X}$ is not discrete. Suppose $\mathcal{X} \subseteq [-M, M]^n$, by discretization, where for any $\widehat{\boldsymbol{x}} \in \mathcal{X}$, there exists $\widehat{\boldsymbol{y}} \in \mathcal{X}^\nu$, such that $\|\widehat{\boldsymbol{x}} - \widehat{\boldsymbol{y}}\|_\infty \leq \nu$ and $|\mathcal{X}^\nu| \leq |2M/\nu|^n$. For notational convenience, we let

$$v_N^{SAA}(\nu) = \min_{\boldsymbol{x}\in\mathcal{X}^\nu} \widehat{\mathbb{P}}\text{-FCVaR}_{1-\varepsilon}\left[Q(\boldsymbol{x}, \widehat{\boldsymbol{\xi}})\right], \quad v^T(\nu) = \min_{\boldsymbol{x}\in\mathcal{X}^\nu} \mathbb{P}^T\text{-FCVaR}_{1-\varepsilon}\left[Q(\boldsymbol{x}, \tilde{\boldsymbol{\xi}})\right].$$

According to Part (iii) of Assumption 2, when $L\nu\sqrt[q]{n} \leq \min\{\Delta_1, \Delta_2\}$, we have

$$|\beta^*(\widehat{\boldsymbol{x}}) - \beta^*(\widehat{\boldsymbol{y}})| \leq L\nu\sqrt[q]{n}.$$

We then bound the difference between objective functions. There are two subcases to consider: whether $\beta^*(\widehat{\boldsymbol{y}}) - \beta^*(\widehat{\boldsymbol{x}})$ is negative or not.

Case (a). When $-L\nu\sqrt[q]{n} \leq \beta^*(\widehat{\boldsymbol{y}}) - \beta^*(\widehat{\boldsymbol{x}}) \leq 0$, we have

$$\begin{aligned}
&\beta^*(\widehat{\boldsymbol{y}}) - \beta^*(\widehat{\boldsymbol{x}}) + \frac{1}{1-\varepsilon}\left[\mathbb{E}_{\mathbb{P}^T}\left[\left(Q(\widehat{\boldsymbol{y}}, \tilde{\boldsymbol{\xi}}) - \beta^*(\widehat{\boldsymbol{y}})\right)_-\right] - \mathbb{E}_{\mathbb{P}^T}\left[\left(Q(\widehat{\boldsymbol{x}}, \tilde{\boldsymbol{\xi}}) - \beta^*(\widehat{\boldsymbol{x}})\right)_-\right]\right] \\
\leq &\beta^*(\widehat{\boldsymbol{y}}) - \beta^*(\widehat{\boldsymbol{x}}) + \frac{1}{1-\varepsilon}\left[\mathbb{E}_{\mathbb{P}^T}\left[\left(Q(\widehat{\boldsymbol{x}}, \tilde{\boldsymbol{\xi}}) + L\|\widehat{\boldsymbol{y}} - \widehat{\boldsymbol{x}}\|_\infty - \beta^*(\widehat{\boldsymbol{y}})\right)_-\right] - \mathbb{E}_{\mathbb{P}^T}\left[\left(Q(\widehat{\boldsymbol{x}}, \tilde{\boldsymbol{\xi}}) - \beta^*(\widehat{\boldsymbol{x}})\right)_-\right]\right] \\
\leq &\beta^*(\widehat{\boldsymbol{y}}) - \beta^*(\widehat{\boldsymbol{x}}) + \frac{1}{1-\varepsilon}\left[\mathbb{E}_{\mathbb{P}^T}\left[\left(Q(\widehat{\boldsymbol{x}}, \tilde{\boldsymbol{\xi}}) + L\nu - \beta^*(\widehat{\boldsymbol{y}})\right)_-\right] - \mathbb{E}_{\mathbb{P}^T}\left[\left(Q(\widehat{\boldsymbol{x}}, \tilde{\boldsymbol{\xi}}) - \beta^*(\widehat{\boldsymbol{x}})\right)_-\right]\right]
\end{aligned}$$

$$\leq \frac{1}{1-\varepsilon}\left[\mathbb{E}_{\mathbb{P}^T}\left[\left(Q(\widehat{\boldsymbol{x}},\tilde{\boldsymbol{\xi}}) + L\nu(1+\sqrt[q]{n}) - \beta^*(\widehat{\boldsymbol{x}})\right)_-\right] - \mathbb{E}_{\mathbb{P}^T}\left[\left(Q(\widehat{\boldsymbol{x}},\tilde{\boldsymbol{\xi}}) - \beta^*(\widehat{\boldsymbol{x}})\right)_-\right]\right]$$

$$\leq \frac{1}{1-\varepsilon}\left[L\nu(1+\sqrt[q]{n})\right],$$

where the first inequality is due to Part (ii) of Assumption 2, the second one is based on the discretization, the third one is due to the presumption in this case, the last one is due to subadditivity of the concave function $h(t) = \min\{t, 0\}$.

Case (b). When $0 < \beta^*(\widehat{\boldsymbol{y}}) - \beta^*(\widehat{\boldsymbol{x}}) \leq L\nu\sqrt[q]{n}$, we have

$$\beta^*(\widehat{\boldsymbol{y}}) - \beta^*(\widehat{\boldsymbol{x}}) + \frac{1}{1-\varepsilon}\left[\mathbb{E}_{\mathbb{P}^T}\left[\left(Q(\widehat{\boldsymbol{y}},\tilde{\boldsymbol{\xi}}) - \beta^*(\widehat{\boldsymbol{y}})\right)_-\right] - \mathbb{E}_{\mathbb{P}^T}\left[\left(Q(\widehat{\boldsymbol{x}},\tilde{\boldsymbol{\xi}}) - \beta^*(\widehat{\boldsymbol{x}})\right)_-\right]\right]$$

$$\leq \beta^*(\widehat{\boldsymbol{y}}) - \beta^*(\widehat{\boldsymbol{x}}) + \frac{1}{1-\varepsilon}\left[\mathbb{E}_{\mathbb{P}^T}\left[\left(Q(\widehat{\boldsymbol{x}},\tilde{\boldsymbol{\xi}}) + L\nu - \beta^*(\widehat{\boldsymbol{y}})\right)_-\right] - \mathbb{E}_{\mathbb{P}^T}\left[\left(Q(\widehat{\boldsymbol{x}},\tilde{\boldsymbol{\xi}}) - \beta^*(\widehat{\boldsymbol{y}})\right)_-\right]\right]$$

$$\leq L\nu(\sqrt[q]{n}) + \frac{1}{1-\varepsilon}L\nu \leq \frac{1}{1-\varepsilon}\left[L\nu(1+\sqrt[q]{n})\right],$$

where the first inequality is due to Part (ii) of Assumption 2, discretization, and $\beta^*(\widehat{\boldsymbol{x}}) < \beta^*(\widehat{\boldsymbol{y}})$, the second one is due to subadditivity of concave function $h(t) = \min\{t, 0\}$, and the last one is due to $\varepsilon \in (0, 1)$.

Therefore, when $|\beta^*(\widehat{\boldsymbol{x}}) - \beta^*(\widehat{\boldsymbol{y}})| \leq L\nu\sqrt[q]{n}$, we have

$$\beta^*(\widehat{\boldsymbol{y}}) - \beta^*(\widehat{\boldsymbol{x}}) + \frac{1}{1-\varepsilon}\left[\mathbb{E}_{\mathbb{P}^T}\left[\left(Q(\widehat{\boldsymbol{y}},\tilde{\boldsymbol{\xi}}) - \beta^*(\widehat{\boldsymbol{y}})\right)_-\right] - \mathbb{E}_{\mathbb{P}^T}\left[\left(Q(\widehat{\boldsymbol{x}},\tilde{\boldsymbol{\xi}}) - \beta^*(\widehat{\boldsymbol{x}})\right)_-\right]\right] \leq \frac{1}{1-\varepsilon}\left[L\nu(1+\sqrt[q]{n})\right],$$

which implies that $v^T(\nu) \leq v^T + [L\nu(1+\sqrt[q]{n})]/(1-\varepsilon)$ holds a.s..

Together with the fact that the inequality $v_N^{SAA} \leq v_N^{SAA}(\nu)$ holds a.s. and the inequality $v_N^{SAA}(\nu) \leq v^T(\nu) + \lambda_2\sigma/\sqrt{N}$ with probability $1 - \exp\{-\lambda_1^2/3\} - 2\exp\{-\ell^2(1-\varepsilon)^2\lambda_1^2\sigma^2/2\}$ from Step III, we have

$$\mathbb{P}^T\left\{v_N^{SAA}(\nu) - \frac{\lambda_2\sigma}{\sqrt{N}} - \frac{1}{1-\varepsilon}\left[L\nu(1+\sqrt[q]{n})\right] \leq v^T(\nu)\right\} \geq 1 - \left[\exp\{-\lambda_1^2/3\} + 2\exp\{-\ell^2(1-\varepsilon)^2\lambda_1^2\sigma^2/2\}\right].$$

Then, the confidence bound can be written as

$$\mathbb{P}^T\left\{v_N^{SAA} - \frac{\lambda_2\sigma}{\sqrt{N}} - \frac{1}{1-\varepsilon}\left[L\nu(1+\sqrt[q]{n})\right] \leq v^T\right\}$$

$$\geq 1 - (2M/\nu)^n\left[\exp\{-\lambda_1^2/3\} + 2\exp\{-\ell^2(1-\varepsilon)^2\lambda_1^2\sigma^2/2\}\right].$$

Letting $\widehat{\gamma}/3 = |2M/\nu|^n \max\left\{\exp\{-\lambda_1^2/3\}, \exp\{-l^2(1-\varepsilon)^2\lambda_1^2\sigma^2/2\}\right\}$, which implies that

$$\frac{\widehat{\gamma}}{3} \geq |2M/\nu|^n \exp\{-\lambda_1^2/3\}, \quad \frac{\widehat{\gamma}}{3} \geq |2M/\nu|^n \exp\{-\ell^2(1-\varepsilon)^2\lambda_1^2\sigma^2/2\},$$

and we have

$$\lambda_1 = \max\left\{\sqrt{3n\log(2M/\nu) - 3\log(\widehat{\gamma}/3)}, \sqrt{\frac{2n\log(2M/\nu) - 2\log(\widehat{\gamma}/3)}{\ell^2(1-\varepsilon)^2\sigma^2}}\right\}.$$

Letting $\lambda_2\sigma/\sqrt{N} = L\nu(1+\sqrt[q]{n})(1-\varepsilon)$ and setting

$$\theta := 4\lambda_1\sigma L^{-1}N^{-1/2}(1-\varepsilon)^{-1} = \mathcal{O}(1)N^{-1/2}\sqrt{n\log(nN)\log(\widehat{\gamma}^{-1})},$$

we arrive at the conclusion. □

We make the following remarks on Theorem 4.3:

(i) Parts (i) and (ii) together show that with high probability, the value of the minimum of the worst-case FCVaR is at most $L\theta$ less than the true value v^T and $2L\theta$ larger than v^T, implying that the Wasserstein radius θ in $\mathcal{O}(N^{-1/2}\sqrt{\log(N)})$ or $\mathcal{O}(N^{-1/2})$ suffices;

(ii) Due to the discretization error, the non-asymptotic Wasserstein radius for the general compact support is in the order of $\mathcal{O}(N^{-1/2}\sqrt{\log(N)})$, which is slightly larger than the one with the discrete compact support one (i.e., $\mathcal{O}(N^{-1/2})$);

(iii) In our numerical study, we numerically verify the order magnitude of the proposed confidence bound. We observe that the appropriate Wasserstein radius θ is nearly proportional to $1/\sqrt{N}$, where N denotes the sample size.

We then demonstrate that the worst-case FCVaR can also be decision outlier robust when Part (ii) of Assumption 2 holds. To begin with, let us define the following two constants. For a given $\alpha_1 \in (0, \varepsilon)$ and a set $\widehat{\mathcal{U}}(\Delta_1)$ defined in Part (ii) of Assumption 2, we define

$$\Delta_1^L = \inf\left\{\Delta_1 \colon \mathbb{P}^T\left\{\widehat{\mathcal{U}}(\Delta_1) \geq 1-\varepsilon+\alpha_1\right\}\right\}, \quad \Delta_1^U = \sup\left\{\Delta_1 \colon \mathbb{P}^T\left\{\widehat{\mathcal{U}}(\Delta_1) \geq 1-\varepsilon+\alpha_1\right\}\right\},$$

which represent the smallest and largest perturbations, respectively, that preserve the Lipschitz continuity property in Part (ii) of Assumption 2.

THEOREM 4.4. *(Decision Outlier Robustness) Suppose that for any unamenable decision $\boldsymbol{x} \in \widehat{\mathcal{X}}$, there exists a $\Delta_1 \in (\Delta_1^L, \Delta_1^U)$ such that Part (ii) of Assumption 2 holds and $\mathbb{P}^T\{\widehat{\mathcal{U}}(\Delta_1)\} \geq 1-\varepsilon+\alpha_1$ for some $\alpha_1 \in (0, \varepsilon]$. Then, if $\Delta_1 + L\theta < \Delta_1^U$ and sample size $N \geq \log(\widehat{\gamma}^{-1})/(2\alpha_1^2)$, then with probability $1-\widehat{\gamma}$, the worst-case* FCVaR *is decision outlier robust.*

Proof. We split the proof into two steps.
Step I. First of all, we need to ensure that with probability at least $1-\widehat{\gamma}$, the number of N i.i.d. empirical samples $\{\widehat{\boldsymbol{\xi}}^i\}_{i\in[N]}$ is large enough, such that the number of the samples which fall outside the set $\widehat{\mathcal{U}}(\Delta_1)$ is at most $\lfloor N\varepsilon \rfloor$. Since $\alpha_1 \in (0, \varepsilon]$, by applying Hoeffding's inequality (see, e.g., [30]), we have

$$\mathbb{P}^T\left\{\sum_{i\in[N]} \mathbb{I}\left(\widehat{\boldsymbol{\xi}}^i \notin \widehat{\mathcal{U}}(\Delta_1)\right) \leq \lfloor N\varepsilon \rfloor\right\} \leq \exp\left\{-2N\left(\alpha_1 + \frac{\lfloor N\varepsilon \rfloor}{N} - \varepsilon\right)^2\right\} \approx \exp\left\{-2N\alpha_1^2\right\}.$$

Letting $\exp\left\{-2N\alpha_1^2\right\} \leq \widehat{\gamma}$, the sample size is at least $N \geq \log(\widehat{\gamma}^{-1})/(2\alpha_1^2)$.
Step II. Note that $\Delta_1^L < \Delta_1 + L\theta < \Delta_1^U$ and the function $\bar{Q}(\boldsymbol{x}, \widehat{\boldsymbol{\xi}})$ is defined as

$$\bar{Q}(\boldsymbol{x}, \widehat{\boldsymbol{\xi}}) = \max_{\boldsymbol{\xi}}\left\{Q(\boldsymbol{x}, \boldsymbol{\xi}) \colon \|\boldsymbol{\xi} - \widehat{\boldsymbol{\xi}}\|_p \leq \theta\right\}.$$

According to the definition of set $\widehat{\mathcal{U}}(\Delta_1)$, we conclude that if $Q(\boldsymbol{x}, \widehat{\boldsymbol{\xi}})$ is finite and $\widehat{\boldsymbol{\xi}} \in \widehat{\mathcal{U}}(\Delta_1)$, then $\bar{Q}(\boldsymbol{x}, \widehat{\boldsymbol{\xi}})$ must also be finite by the Lipschitz continuity and is bounded by $Q(\boldsymbol{x}, \widehat{\boldsymbol{\xi}}) + L\theta$. According to the definition of set $\widehat{\mathcal{U}}(\Delta_1^U)$, $\Delta_1 + L\theta < \Delta_1^U$, and the result in Step I, with probability at least $1-\widehat{\gamma}$, we have

$$\eta = \widehat{\mathbb{P}}\left\{\bar{Q}(\boldsymbol{x}, \widehat{\boldsymbol{\xi}}) < \infty\right\} \geq \widehat{\mathbb{P}}\left\{\bar{Q}(\boldsymbol{x}, \tilde{\boldsymbol{\xi}}) \leq \beta^*(\boldsymbol{x}) + \Delta_1 + L\theta\right\} \geq 1-\varepsilon.$$

Step III. For the worst-case distribution $\bar{\mathbb{P}} \in \mathcal{P}_\infty^W$, according to [9], it can be represented as

$$\bar{\mathbb{P}} = \sum_{i\in[N]} \delta_{(\tilde{\boldsymbol{\xi}} = \bar{\boldsymbol{\xi}}^i)}/N$$

with $\bar{\boldsymbol{\xi}}^i \in \operatorname{argmax}_{\boldsymbol{\xi}}\{Q(\boldsymbol{x}, \boldsymbol{\xi}) \colon \|\boldsymbol{\xi} - \widehat{\boldsymbol{\xi}}^i\|_p \leq \theta\}$ for each $i \in [N]$.

Next, we construct the favorable distribution $\mathbb{P}^*$ such that $\mathbb{P}^*\{\tilde{\boldsymbol{\xi}} = \bar{\boldsymbol{\xi}}^i\} = \mathbb{I}\{\bar{Q}(\boldsymbol{x}, \widehat{\boldsymbol{\xi}}^i) < \infty\}/(N\eta)$ for each $i \in [N]$. By our construction, we have $\mathbb{P}^*\{\mathcal{U}\} = 1, 0 \preceq \mathbb{P}^* \preceq \bar{\mathbb{P}}/(1-\varepsilon)$. On the other hand, we have

$$\mathbb{E}_{\mathbb{P}^*}\left[\bar{Q}(\boldsymbol{x}, \tilde{\boldsymbol{\xi}})\right] < \infty, \quad \mathbb{P}^*\left\{\tilde{\boldsymbol{\xi}} : \bar{Q}(\boldsymbol{x}, \tilde{\boldsymbol{\xi}}) = \infty\right\} = 0.$$

This proves that $\mathbb{P}^*$ is a desirable probability measure, such that the condition in Proposition 3.3 is satisfied. Hence, we conclude that with probability $1-\widehat{\gamma}$, the worst-case FCVaR is decision outlier robust. ◻

According to Theorem 4.4, to preserve the decision outlier robustness, we need to guarantee that the radius of type$-\infty$ Wasserstein ambiguity set θ is small, i.e., $0 \leq \theta < (\Delta_1^U - \Delta_1^L)/L$. In fact, to simultaneously achieve out-of-sample performance guarantees and decision outlier robustness, since $\theta \propto 1/\sqrt{N}$ according to Theorem 4.3, it is expected that the sample size should not be too small.

We conclude this section by remarking that the results in Theorem 4.3 and Theorem 4.4 can be extended to Winsorized CVaR and Huber-skip CVaR. The proofs are similar and thus are omitted.

4.3 Achieving Out-of-Sample Performance Guarantees in Favorable Two-stage Stochastic Programs. In this subsection, to achieve the out-of-sample performance, we provide one robustified favorable two-stage stochastic program by applying type$-\infty$ Wasserstein ambiguity set. First of all, if we apply the worst-case FCVaR to a two-stage stochastic program, we have

$$\min_{\substack{\boldsymbol{x}\in\mathcal{X},\\ S\in\mathcal{S}}}\left\{\boldsymbol{c}^\top\boldsymbol{x} + \sup_{\mathbb{P}\in\mathcal{P}_\infty^W(S)}\left\{\mathbb{E}_{\mathbb{P}}\left[Q(\boldsymbol{x}, \tilde{\boldsymbol{\xi}})\right]\right\}\right\},$$

which can be written as

$$\text{(4.8)} \qquad \min_{\substack{\boldsymbol{x}\in\mathcal{X},\\ \boldsymbol{z}\in\{0,1\}^N}} \boldsymbol{c}^\top \boldsymbol{x} + \left\{ \frac{1}{N-N\varepsilon} \sum_{i\in[N]} z_i \max_{\boldsymbol{\xi}} \left\{ Q(\boldsymbol{x},\boldsymbol{\xi}) : \|\boldsymbol{\xi}-\widehat{\boldsymbol{\xi}}^i\|_p \le \theta \right\} : \sum_{i\in[N]} z_i = N - N\varepsilon \right\},$$

Notice that in general, for a given $\boldsymbol{z}$, the optimization problem above is NP-hard (see the details in [75]). Therefore, instead of focusing on (4.8), by exploring the structure of the problem, we consider the following special case of the worst-case favorable two-stage stochastic program. For example, if the recourse function $Q(\boldsymbol{x},\boldsymbol{\xi})$ is monotone in $\boldsymbol{\xi}$ for any $\boldsymbol{x}\in\mathcal{X}$ and the norm is L_∞, then (4.8) is equivalent to

$$\text{(4.9)} \qquad \min_{\substack{\boldsymbol{x}\in\mathcal{X},\\ \boldsymbol{z}\in\{0,1\}^N}} \boldsymbol{c}^\top \boldsymbol{x} + \left\{ \frac{1}{N-N\varepsilon} \sum_{i\in[N]} z_i Q(\boldsymbol{x}, \widehat{\boldsymbol{\xi}}^i \pm \theta \mathbf{e}) \colon \sum_{i\in[N]} z_i = N - N\varepsilon \right\},$$

where we choose $-\theta$ if the recourse function is monotone non-decreasing over a particular parameter, and $+\theta$ if the recourse function is monotone non-increasing over a parameter. Then, we can apply the result in Theorem 2.2 or the MILP (2.9) to derive a proper formulation. Notice that this monotonicity structure has been studied in several recent works (see, e.g., [16, 75, 77]). In order to illustrate the formulation (4.8), we use the two-stage recourse planning problem in Example 5 and apply the worst-case DFO under type-∞ Wasserstein ambiguity set.

Example 8. Consider Example 5 under type$-\infty$ Wasserstein ambiguity set equipped with weighted L_∞ norm (i.e., $\|\boldsymbol{\xi}\|_\infty := \max\{w_q\|\boldsymbol{q}\|_\infty, w_u\|\boldsymbol{u}\|_\infty, w_p\|\boldsymbol{p}\|_\infty, w_\lambda\|\boldsymbol{\lambda}\|_\infty\}$ with positive weights w_q, w_u, w_p, w_λ) constructed based on N i.i.d. samples $\{\widehat{\boldsymbol{\xi}}^i\}_{i\in[N]}$ on the nonnegative support $\mathcal{U}$. Then, the minimum of the worst-case FCVaR (4.9) is equivalent to

$$\text{(4.10a)} \qquad \min_{\boldsymbol{x}\ge \mathbf{0}, \boldsymbol{z}} \left\{ \boldsymbol{c}^\top \boldsymbol{x} + \frac{1}{N-N\varepsilon} \sum_{i\in[N]} z_i Q(\boldsymbol{x}, \widehat{\boldsymbol{\xi}}^i \pm \theta\mathbf{e}) \colon \sum_{i\in[N]} z_i \ge N - N\varepsilon, \boldsymbol{z}\in\{0,1\}^N \right\},$$

where for each $i\in[N]$, we have

(4.10b)

$$z_i Q(\boldsymbol{x}, \widehat{\boldsymbol{\xi}}^i \pm \theta\mathbf{e}) = \min_{\boldsymbol{y}^i \ge \mathbf{0}} \left\{ \sum_{s\in[n]} \sum_{j\in[n_1]} (q^i_{sj} + \frac{\theta}{w_q}) y^i_{sj} \colon \begin{array}{l} \sum_{j\in[n_1]} y^i_{sj} \le (p^i_s - \theta/w_p)_+ x_s, \forall s\in[n], \\ \sum_{s\in[n]} (u^i_{sj} - \theta/w_u)_+ y^i_{sj} \ge (\lambda^i_j + \theta/w_\lambda) z_i, \forall j\in[n_1] \end{array} \right\}.$$

Similarly, the minimum of the worst-case WCVaR in this example can be formulated as follows:

$$\text{(4.10c)} \qquad \min_{\substack{\boldsymbol{x}\in\mathcal{X},\\ \boldsymbol{z}\in\{0,1\}^N}} \left\{ \boldsymbol{c}^\top \boldsymbol{x} + \frac{1}{N} \sum_{i\in[N]} z_i Q(\boldsymbol{x}, \widehat{\boldsymbol{\xi}}^i \pm \theta\mathbf{e}) + \eta\varepsilon \colon \begin{array}{l} \eta \ge z_i Q(\boldsymbol{x}, \widehat{\boldsymbol{\xi}}^i \pm\theta\mathbf{e}) + (1-z_i)L_i, \forall i\in[N], \\ \sum_{i\in[N]} z_i \ge N - N\varepsilon \end{array} \right\},$$

where, for each $i\in[N]$, the scalar L_i is defined in Corollary 2.3 and the product $z_i Q(\boldsymbol{x}, \widehat{\boldsymbol{\xi}}^i \pm \theta\mathbf{e})$ is defined in (4.10b). ⋄

The comprehensive process for selecting θ in Example 8 can be found in the numerical study section. We remark that interested readers are referred to [75] for many reformulation results in the two-stage stochastic program with type$-\infty$ Wasserstein ambiguity set, which can be useful to derive the reformulation of the worst-case DFO.

5 Numerical Study.

This section presents the numerical results to compare the strengths of FCVaR and its alternatives based on Example 5 in Section 2.3, where the relatively complete recourse assumption may not be satisfied.

We generate random instances with varying sample sizes N for the numerical experiments. All the random variables (i.e., the customer demands $\tilde{\boldsymbol{\lambda}}$, random costs $\tilde{\boldsymbol{q}}$, random utilization rates $\tilde{\boldsymbol{p}}$, and random service rates $\tilde{\boldsymbol{u}}$) are truncated to be nonnegative. Particularly, for each instance, we suppose that the components of the cost vector $\boldsymbol{c}$ are i.i.d. truncated Gaussian ones with means 1 and variances 0.2, the components of random utilization rate $\tilde{\boldsymbol{p}}$ are independent truncated Gaussian ones with means uniformly distributed in $(0.9, 1)$ and variance being 0.05, and we let $q^i_{sj} = p^i_s$ for all $s \in [n]$, $j \in [n_1]$, and $i \in [N]$ to let the reliable servers are more expensive in the second-stage cost. The components of the nominal customer demand $\tilde{\boldsymbol{\lambda}}$ are i.i.d. truncated Gaussian ones with means 10 and variances 0.2 and the random service rates $\tilde{\boldsymbol{u}}$ are i.i.d. truncated Gaussian ones with means 1 and variances 0.2. We also assume that there exist some outliers in the customer demand information and service rate information, denoted by $\tilde{\boldsymbol{\lambda}}^o$ and $\tilde{\boldsymbol{u}}^o$, respectively. We assume the components of random vector $\tilde{\boldsymbol{\lambda}}^o$ are i.i.d. truncated Gaussian distributed with mean 30 and variance 5 and the components of random vector $\tilde{\boldsymbol{u}}^o$ are i.i.d. truncated Gaussian distributed with means 0.02 and variances 0.01, which may cause the underlying two-stage problem infeasible. The observed demand vector follows the following distribution $0.85\tilde{\boldsymbol{\lambda}} + 0.15\tilde{\boldsymbol{\lambda}}^o$, and the observed service rate vector follows $0.95\tilde{\boldsymbol{u}} + 0.05\tilde{\boldsymbol{u}}^o$. We let the number of resources $n = 20$ and the number of customers $n_1 = 20$.

In the numerical implementation, since the original SAA problem (2.7a) may be infeasible, we resolve the infeasibility issue from the original SAA by removing the infeasible scenarios until the remaining problem is solvable. This procedure is known as "*Trimmed SAA*" (see more discussions in chapter 7 of [17] and chapter 2.3 of [23]). After solving the corresponding Trimmed SAA, FCVaR, WCVaR, and HCVaR models, we generate additional 50 random testing cases to evaluate the solution performances, i.e., to assess the performance of the first-stage decision in each method. For the worst-case models, we follow Example 8 and focus on type$-\infty$ Wasserstein ambiguity set equipped with weighted infinity norm. All the instances in this section are coded in Python 3.9 with calls to solver Gurobi (version 9.1.1 with default settings) on a personal PC with an Apple M1 Pro processor and 16G of memory. We set the time limit of each instance to be 3600s.

Experiment 1. Model Comparisons When the Testing Distribution is the Same as Training. For each method (i.e., Trimmed SAA, FCVaR, WCVaR, HCVaR, and In-CVaR models), when evaluating the first-stage decision using 50 random generated test instances, i.e., the components of the random utilization rate vector $\tilde{\boldsymbol{p}}$ are i.i.d. truncated Gaussian ones with means sampled uniformly from $(0.9, 1)$ and variances all being 0.05. we record all the $50\%, 60\%, 70\%, 80\%, 90\%$ quantiles of the second-stage values, respectively. We then report the 95% confidence interval (C.I.) of each quantile among these 50 testing instances. We set $\varepsilon = 0.10$ in both FCVaR (2.11a) and WCVaR (2.12a) and consider the sample size with $N \in \{100, 200\}$. To avoid any trivial solution in HCVaR (i.e., $\boldsymbol{x} = \boldsymbol{0}, \boldsymbol{z} = \boldsymbol{0}$ may be a trivial optimal solution in (2.12b) when H is relatively small), we solve the trimmed SAA model first and then select its $(1-\varepsilon)$-quantile as the value of H. We use In-CVaR$^\beta_\alpha$ from [41] with $\alpha = 0.1, \beta = 0.9$ for comparisons. Notice that based on Example 5 in Section 2.3, we may not provide a big-M free formulation for In-CVaR model and therefore, we may not be able to solve all the instances of In-CVaR model to optimality within the time limit. We use "GAP" to denote its optimality gap as GAP(%) $= (|\text{UB} - \text{LB}|)/|\text{LB}| \times 100$, where "UB" and "LB" denote the best upper bound and the best lower bound found by the In-CVaR model, respectively. For each testing instance, we assume the components of customer demand $\tilde{\boldsymbol{\lambda}}$ are i.i.d. truncated Gaussian ones with means 10 and variance 0.2, the components of service rate $\tilde{\boldsymbol{u}}$ are i.i.d. truncated Gaussian ones with means 1 and variances 0.2, and the remaining parameters follow the same assumptions described in the training procedure. The result is shown in Table 1. It is seen that, in a reasonable time, FCVaR, WCVaR, and In-CVaR can consistently provide a favorable solution with a lower cost than the trimmed SAA. However, In-CVaR takes much longer than the other methods and HCVaR performs worst among the four models. Additionally, it is worth noting that when we set the parameter H in the HCVaR to be the $(1-\varepsilon)$-quantile of the trimmed SAA model, we observe that the performances of HCVaR and trimmed SAA are quite similar. We continue to discuss the performance of HCVaR in the next experiment.

Table 1: Quantile Comparisons among Trimmed SAA, FCVaR, WCVaR, HCVaR, and In-CVaR in Experiment 1.

N	Model	Time (s)	GAP	Quantile				
				50% C.I.	60% C.I.	70% C.I.	80% C.I.	90% C.I.
100	Trimmed SAA	5.58	0.00%	[532.04,535.40]	[535.60,538.94]	[539.16,542.54]	[543.21,546.72]	[549.31,552.89]
	FCVaR (2.11a)	8.05	0.00%	[473.75,477.56]	[478.41,482.40]	[483.97,487.93]	[490.13,494.00]	[498.34,502.26]
	WCVaR (2.12a)	11.05	0.00%	[474.33,477.99]	[478.79,482.69]	[484.10,487.95]	[489.84,493.60]	[497.46,501.31]
	HCVaR (2.12b)	2.44	0.00%	[532.05,535.40]	[535.61,538.94]	[539.14,542.53]	[543.20,546.71]	[549.28,552.86]
	In-CVaR [41]	1740.39	0.00%	[473.97,477.68]	[478.52,482.45]	[483.88,487.77]	[489.75,493.59]	[497.66,501.52]
200	Trimmed SAA	16.93	0.00%	[575.99,579.47]	[579.40,582.74]	[583.24,586.55]	[587.25,590.59]	[593.10,596.41]
	FCVaR (2.11a)	41.36	0.00%	[492.34,495.64]	[495.90,499.15]	[499.92,503.21]	[504.47,507.74]	[510.37,513.74]
	WCVaR (2.12a)	47.10	0.00%	[492.78,496.11]	[496.21,499.55]	[500.31,503.62]	[504.95,508.28]	[511.03,514.46]
	HCVaR (2.12b)	5.06	0.00%	[575.99,579.29]	[579.40,582.68]	[583.24,586.51]	[587.25,590.58]	[593.10,596.41]
	In-CVaR [41]	3600	0.91%	[492.42,495.71]	[495.96,499.24]	[500.03,503.34]	[504.49,507.79]	[510.59,514.02]

Experiment 2. Model Comparisons When the Testing Distribution is Different From the Training one. We follow the same procedure described in Experiment 1, i.e., we record all the 50%, 60%, 70%, 80%, 90% quantiles in the second-stage scenarios for each method (e.g., Trimmed SAA, FCVaR, WCVaR, and HCVaR) in each testing instance, respectively, and report the average of each quantile among these 50 random generated testing instances. The testing setting is the same as that of Experiment 1, except that we assume that the utilization rates have been perturbed, i.e., the components of the random utilization rate vector $\tilde{\boldsymbol{p}}$ are i.i.d. truncated Gaussian ones with means being 0.6 and variances being 0.3. The result is shown in Table 2. As expected, both FCVaR and WCVaR can consistently provide a favorable solution with a lower cost than the trimmed SAA. On the other hand, HCVaR surprisingly performs worse than FCVaR, WCVaR, and In-CVaR. This may be because that HCVaR is very sensitive to its trimmed parameter H. In this experiment, we let the parameter H in HCVaR be $(1-\varepsilon)$-quantile of trimmed SAA model to avoid any trivial solution; that is, when H is small, e.g., H is less than the first-stage cost, it provides a trivial solution is $\boldsymbol{x} = \boldsymbol{0}, \boldsymbol{z} = \boldsymbol{0}$ in (2.12b) . In the following discussions, we focus on FCVaR and WCVaR that have small differences and may not be comparable. Therefore, to measure their relative performances, we report the running time of FCVaR and WCVaR in the following discussions.

Table 2: Quantile Comparisons among Trimmed SAA, FCVaR, WCVaR, HCVaR, and In-CVaR in Experiment 2.

N	Model	Time (s)	GAP	Quantile				
				50% C.I.	60% C.I.	70% C.I.	80% C.I.	90% C.I.
100	Trimmed SAA	5.58	0.00%	[578.05,582.25]	[582.72,586.87]	[587.70,591.98]	[593.75,598.15]	[601.53,605.97]
	FCVaR (2.11a)	8.05	0.00%	[540.41,545.57]	[547.32,552.56]	[554.86,560.01]	[563.72,569.06]	[577.85,583.22]
	WCVaR (2.12a)	11.05	0.00%	[537.08,542.16]	[543.62,548.53]	[550.62,555.61]	[558.62,563.52]	[571.84,576.99]
	HCVaR (2.12b)	2.44	0.00%	[577.96,582.16]	[582.61,586.76]	[587.56,591.84]	[593.55,597.95]	[601.37,605.82]
	In-CVaR [41]	1740.39	0.00%	[538.27,543.37]	[544.88,550.01]	[552.17,557.15]	[560.47,565.57]	[574.09,579.40]
200	Trimmed SAA	16.93	0.00%	[621.98,626.08]	[626.28,630.41]	[631.40,635.46]	[637.22,641.33]	[645.12,649.30]
	FCVaR (2.11a)	41.36	0.00%	[543.94,548.07]	[549.06,553.12]	[554.58,558.74]	[560.62,564.77]	[569.90,574.13]
	WCVaR (2.12a)	47.10	0.00%	[544.62,548.82]	[549.41,553.53]	[554.76,558.82]	[561.22,565.40]	[570.29,574.54]
	HCVaR (2.12b)	5.06	0.00%	[621.88,625.95]	[626.24,630.36]	[631.33,635.37]	[637.17,641.28]	[644.95,649.15]
	InCVaR [41]	3600	0.91%	[544.29,548.45]	[549.24,553.33]	[554.73,558.84]	[560.93,565,13]	[570.30,574.55]

Experiment 3. Comparisons in the Worst-case FCVaR and WCVaR and Finding a Proper Wasserstein Radius. Since HCVaR is quite sensitive to the parameter H and does not work well in general, we focus on FCVaR and WCVaR for the remaining experiments. We follow the same setting and derivation of Example 8 in Section 4.3 for both worst-case FCVaR and worst-case WCVaR models and adopt the same training parameter setting as that in Experiment 1 for training and testing in this experiment. We also let the risk parameter $\varepsilon = 0.10$ and sample size $N = 200$. To choose a proper Wasserstein radius θ, based on out-of-sample probability (4.2), we suggest selecting the smallest θ such that its corresponding training costs of FCVaR and WCVaR are beyond the 95% one-sided testing confidence interval (similar procedure for the out-of-sample performances can be found in section 7.3 of [68]). In the numerical study, we choose the weight of each random vector used in the weighted L_∞ norm to be proportional to the inverse of the average of all the samples of the corresponding random vector, i.e., we let w_q in Example 8 as $\theta/\bar{q}$, where $\bar{q}$ is the average of $\boldsymbol{q}$ in that particular instance. Then, following the same procedure as described in Experiment 2, the result is shown in Table 3. The optimal Wasserstein radius is $\theta = 0.10$ for FCVaR and $\theta = 0.01$ for WCVaR, and we observe that the running time of FCVaR is slightly less than that of WCVaR.

Experiment 4. Value of Confidence Bound. In this experiment, we test the order magnitude of the proposed confidence bound presented in Section 4.2. Since Example 8 lacks a fixed recourse structure, the computation of the required Lipschitz coefficient for Assumption 2 (ii) of Theorem 4.3 is not possible. Instead, we present the asymptotic trend of the optimal θ. In this experiment, we follow the same setting as that in Experiment 3. Then, we follow the same procedure described in Experiment 3 to choose a proper θ for each sample size. We repeat this process 10 times and the result is shown in Figure 6, where we observe that the optimal Wasserstein radius θ decreases when sample size N increases. The curve can well fit the results in the order of $1/\sqrt{N}$, which validates our discussions in Section 4.2.

Experiment 5. Value of Big-M Free Formulations. In this experiment, we follow the same setting as Experiment 1 and compare the Big-M and Big-M free formulations between FCVaR and WCVaR with different choices of θ. The big-M free formulations can be found in Section 2.3. We let the risk parameter $\varepsilon = 0.10$ and generate instances with the varying sample sizes $N \in \{200, 300, 400, 500\}$. The proposed big-M

Table 3: Comparisons in the Worst-case of FCVaR (2.11a) and WCVaR (2.12a) and θ Selection in Experiment 3.

θ	FCVaR (2.11a)		WCVaR (2.12a)		Testing	
	Opt. Val.	Time (s)	Opt. Val.	Time (s)	FCVaR (2.11a) C.I.	WCVaR (2.12a) C.I.
0.00	508.78	41.36	545.81	47.11	[540.49.543.33]	[544.01,546.85]
0.01	519.43	44.68	559.43	52.91	[546.81,550.95]	[550.28,554.42]
0.02	530.39	49.69	569.32	54.82	[553.71,557.86]	[557.18,561.34]
0.03	541.55	52.87	579.48	55.28	[560.88,565.04]	[564.31,568.48]
0.04	552.96	56.18	589.94	58.75	[576.07,580.65]	[574.38,578.60]
0.05	564.63	57.76	600.71	60.88	[583.33,587.92]	[582.05,586.28]
0.06	576.62	63.25	611.78	64.29	[590.34,594.93]	[589.91,594.15]
0.07	588.93	66.21	623.16	69.38	[597.82,602.40]	[598.05,602.31]
0.08	601.59	68.48	634.89	80.39	[605.69,610.27]	[606.50,610.76]
0.09	614.59	71.36	646.96	81.72	[613.83,618.41]	[615.15,619.43]
0.10	627.97	73.33	659.41	83.68	[622.07,626.64]	[624.05,628.33]
0.11	641.73	74.71	672.24	86.25	[630.71,635.25]	[633.32,637.60]
0.12	655.90	77.86	685.49	92.26	[639.90,644.45]	[642.99,647.28]

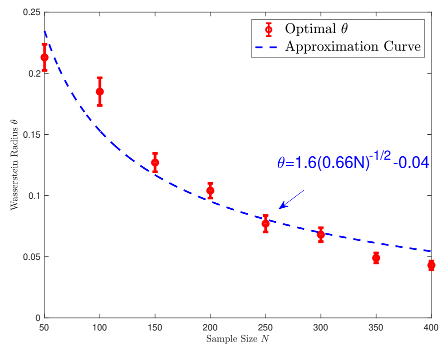

Fig. 6: Optimal θ vs. Sample Size N in Experiment 4.

free formulations can effectively identify better feasible solutions than the exact Big-M model with a much shorter solution time. Recall that we let "UB" and "LB" denote the best upper bound and the best lower bound found by the Big-M model. Since we cannot solve the Big-M model to optimality within the time limit, we use "GAP" to denote its optimality gap as $\text{GAP}(\%) = (|\text{UB} - \text{LB}|)/|\text{LB}| \times 100$. In the corresponding big-M formulations, to select a proper value of the big-M coefficient, we first run the trimmed SAA model and then let the value of the big-M coefficient be the feasible scenario with the largest recourse value. We repeat this process for 10 times, and the average performance can be found in Table 4. Notably, we show that big-M free formulation can improve the running time. We anticipate that the differences will be more striking for larger-scale instances.

6 Conclusion. This paper studied distributionally favorable optimization (DFO) for data-driven optimization with endogenous outliers, where the conventional data-driven stochastic programs may fail. Notably, we showed its connection to robust statistics, established decision outlier robustness concept, and integrated distributional robustness to achieve out-of-sample performance guarantees. Exploring the contextual information in DFO or studying the worst-case regret bound of the FCVaR can be promising future research directions.

Acknowledgment. This research has been supported in part by the National Science Foundation grants 2246414 and 2246417. Valuable comments from the associate editor and two anonymous reviewers are gratefully acknowledged.

References

[1] R. Agarwal, D. Schuurmans, and M. Norouzi, *An optimistic perspective on offline reinforcement learning*, in International Conference on Machine Learning (ICML), 2020.

[2] S. Ahmed, *Two-Stage Stochastic Integer Programming: A Brief Introduction*, American Cancer Society, 2011.

[3] B. Ari and H. A. Güvenir, *Clustered linear regression*, Knowledge-Based Systems, 15 (2002), pp. 169–175.

Table 4: Comparisons Between Big-M and Big-M Free Formulations of FCVaR and WCVaR in Experiment 5

N	θ	Trimmed SAA	FCVaR			WCVaR		
			Big-M (2.11a) & (2.11b)		Big-M Free (2.11a) & (2.11c)	Big-M (2.12a) & (2.11b)		Big-M Free (2.12a) & (2.11c)
		Time (s)	Time (s)	GAP	Time (s)	Time (s)	GAP	Time (s)
200	0.00	17.12	90.23	0.00%	42.17	121.54	0.00%	48.15
	0.01	17.58	105.48	0.00%	49.32	135.67	0.00%	53.28
	0.02	18.01	116.12	0.00%	52.89	143.89	0.00%	58.02
	0.03	18.43	126.78	0.00%	57.04	147.23	0.00%	59.91
	0.04	18.76	135.95	0.00%	59.21	166.42	0.00%	61.34
	0.05	19.23	147.31	0.00%	61.56	178.56	0.00%	64.89
	0.06	19.47	156.54	0.00%	66.73	187.91	0.00%	69.34
	0.07	20.01	165.89	0.00%	72.02	206.78	0.00%	74.12
	0.08	20.34	175.02	0.00%	76.58	218.02	0.00%	78.18
	0.09	20.87	182.76	0.00%	79.91	224.98	0.00%	82.46
	0.10	20.89	191.43	0.00%	82.47	237.12	0.00%	87.01
	0.11	21.15	196.87	0.00%	85.69	242.19	0.00%	91.78
	0.12	21.42	198.56	0.00%	86.84	248.67	0.00%	92.82
300	0.00	34.17	383.27	0.00%	167.23	563.23	0.00%	241.23
	0.01	34.42	397.58	0.00%	176.58	577.45	0.00%	259.48
	0.02	34.76	412.94	0.00%	182.94	589.78	0.00%	273.10
	0.03	35.02	427.12	0.00%	189.12	602.89	0.00%	291.67
	0.04	35.29	435.87	0.00%	197.87	615.12	0.00%	297.89
	0.05	35.67	447.29	0.00%	204.29	629.34	0.00%	312.04
	0.06	36.01	456.66	0.00%	211.66	675.56	0.00%	324.88
	0.07	36.23	468.05	0.00%	219.05	686.23	0.00%	339.56
	0.08	36.47	479.21	0.00%	223.21	702.49	0.00%	355.92
	0.09	37.05	492.37	0.00%	227.37	722.67	0.00%	365.76
	0.10	37.29	505.92	0.00%	228.92	783.64	0.00%	381.34
	0.11	37.58	514.76	0.00%	229.76	789.01	0.00%	396.29
	0.12	37.94	545.03	0.00%	231.93	794.53	0.00%	402.58
400	0.00	57.12	3600	0.04%	985.34	3600	0.09%	1602.45
	0.01	57.45	3600	0.05%	1004.58	3600	0.09%	1632.79
	0.02	57.82	3600	0.05%	1023.94	3600	0.09%	1675.89
	0.03	58.03	3600	0.06%	1046.22	3600	0.09%	1708.11
	0.04	58.27	3600	0.06%	1089.87	3600	0.12%	1765.68
	0.05	58.92	3600	0.08%	1114.26	3600	0.15%	1799.44
	0.06	59.02	3600	0.11%	1136.72	3600	0.17%	1891.67
	0.07	59.18	3600	0.11%	1165.05	3600	0.18%	1910.26
	0.08	59.46	3600	0.12%	1198.21	3600	0.18%	2016.91
	0.09	59.75	3600	0.12%	1211.37	3600	0.20%	2078.57
	0.10	60.09	3600	0.12%	1234.92	3600	0.22%	2111.23
	0.11	60.35	3600	0.15%	1267.76	3600	0.22%	2187.89
	0.12	60.72	3600	0.15%	1299.35	3600	0.23%	2236.42
500	0.00	83.14	3600	0.26%	1175.23	3600	1.92%	3029.23
	0.01	83.36	3600	0.29%	1193.58	3600	1.95%	3071.49
	0.02	83.65	3600	0.34%	1269.94	3600	2.03%	3123.89
	0.03	83.89	3600	0.34%	1283.12	3600	2.24%	3190.68
	0.04	84.28	3600	0.42%	1290.87	3600	2.33%	3234.58
	0.05	84.48	3600	0.57%	1323.29	3600	2.41%	3301.20
	0.06	84.67	3600	0.62%	1378.72	3600	2.52%	3355.97
	0.07	85.03	3600	0.72%	1436.05	3600	2.63%	3422.71
	0.08	85.29	3600	0.75%	1479.21	3600	2.74%	3458.02
	0.09	85.56	3600	0.79%	1527.33	3600	3.18%	3479.03
	0.10	85.82	3600	0.81%	1554.93	3600	3.30%	3510.48
	0.11	86.07	3600	0.84%	1580.76	3600	4.02%	3525.06
	0.12	86.36	3600	0.88%	1595.45	3600	4.49%	3536.44

[4] P. Auer, N. Cesa-Bianchi, and P. Fischer, *Finite-time analysis of the multiarmed bandit problem*, Machine learning, 47 (2002), pp. 235–256.

[5] V. Barnett and T. Lewis, *Outliers in statistical data*, Wiley Series in Probability and Mathematical Statistics. Applied Probability and Statistics, (1984).

[6] A. Ben-Tal, L. El Ghaoui, and A. Nemirovski, *Robust optimization*, Princeton university press, 2009.

[7] A. Ben-Tal and A. Nemirovski, *Lectures on modern convex optimization: analysis, algorithms, and engineering applications*, SIAM, Philadelphia, PA, 2001.

[8] A. Ben-Tal and M. Teboulle, *An old-new concept of convex risk measures: the optimized certainty equivalent*, Mathematical Finance, 17 (2007), pp. 449–476.

[9] D. Bertsimas, S. Shtern, and B. Sturt, *A data-driven approach to multistage stochastic linear optimization*, Management Science, 69 (2023), pp. 51–74.

[10] J. Bi and T. Zhang, *Support vector classification with input data uncertainty*, in Advances in neural information processing systems, 2005, pp. 161–168.

[11] G. Boole, *The mathematical analysis of logic*, Philosophical Library, 1847.

[12] J. Cao and R. Gao, *Contextual decision-making under parametric uncertainty and data-driven optimistic optimization*. Available at Optimization Online, 2021.

[13] N. Cesa-Bianchi and G. Lugosi, *Prediction, learning, and games*, Cambridge university press, 2006.

[14] R. Chen and J. Luedtke, *On sample average approximation for two-stage stochastic programs without relatively complete recourse*, Mathematical Programming, 196 (2022), pp. 719–754.

[15] Z. Chen, D. Kuhn, and W. Wiesemann, *Data-driven chance constrained programs over wasserstein balls*, Operations Research, (2022).

[16] Z. Chen and W. Xie, *Regret in the newsvendor model with demand and yield randomness*, Production and Operations Management, 30 (2021), pp. 4176–4197.

[17] J. W. Chinneck, *Feasibility and Infeasibility in Optimization: Algorithms and Computational Methods*, vol. 118, Springer Science & Business Media, 2007.

[18] R. Cont, R. Deguest, and G. Scandolo, *Robustness and sensitivity analysis of risk measurement procedures*, Quantitative finance, 10 (2010), pp. 593–606.

[19] V. DeMiguel and F. J. Nogales, *Portfolio selection with robust estimation*, Operations research, 57 (2009), pp. 560–577.

[20] P. M. Esfahani and D. Kuhn, *Data-driven distributionally robust optimization using the Wasserstein metric: Performance guarantees and tractable reformulations*, Mathematical Programming, 171 (2018), pp. 115–166.

[21] J.-y. Gotoh, M. J. Kim, and A. E. Lim, *A data-driven approach to beating saa out of sample*, Operations Research, (2023).

[22] V. Guigues, A. Juditsky, and A. Nemirovski, *Non-asymptotic confidence bounds for the optimal value of a stochastic program*, Optimization Methods and Software, 32 (2017), pp. 1033–1058.

[23] L. Gurobi Optimization, *Gurobi optimizer reference manual*, 2022.

[24] F. R. Hampel, *The influence curve and its role in robust estimation*, Journal of the american statistical association, 69 (1974), pp. 383–393.

[25] G. A. Hanasusanto, V. Roitch, D. Kuhn, and W. Wiesemann, *A distributionally robust perspective on uncertainty quantification and chance constrained programming*, Mathematical Programming, 151 (2015), pp. 35–62.

[26] G. A. Hanasusanto, V. Roitch, D. Kuhn, and W. Wiesemann, *Ambiguous joint chance constraints under mean and dispersion information*, Operations Research, 65 (2017), pp. 751–767.

[27] J. Heinonen, *Lectures on Lipschitz analysis*, University of Jyväskylä, 2005.

[28] N. Ho-Nguyen, F. Kilinç-Karzan, S. Küçükyavuz, and D. Lee, *Distributionally robust chance-constrained programs with right-hand side uncertainty under Wasserstein ambiguity*, Mathematical Programming, 196 (2022), p. 641–672.

[29] V. Hodge and J. Austin, *A survey of outlier detection methodologies*, Artificial intelligence review, 22 (2004), pp. 85–126.

[30] W. Hoeffding, *Probability inequalities for sums of bounded random variables*, in The Collected Works of Wassily Hoeffding, Springer, 1994, pp. 409–426.

[31] L. J. Hong, Z. Hu, and G. Liu, *Monte carlo methods for value-at-risk and conditional value-at-risk: a review*, ACM Transactions on Modeling and Computer Simulation (TOMACS), 24 (2014), pp. 1–37.

[32] D. C. Howell, *Median Absolute Deviation*, American Cancer Society, 2014.

[33] X. Huang, L. Shi, and J. A. Suykens, *Ramp loss linear programming support vector machine*, The Journal of Machine Learning Research, 15 (2014), pp. 2185–2211.

[34] P. J. Huber, *Robust estimation of a location parameter*, in Breakthroughs in statistics, Springer, 1992, pp. 492–518.

[35] P. J. Huber, *Robust statistics*, John Wiley & Sons, 2004.

[36] K. Jaganathan, Y. Eldar, and B. Hassibi, *Phase retrieval: an overview of recent developments*. arXiv preprint arXiv:1510.07713, 2015.

[37] R. Ji and M. A. Lejeune, *Data-driven distributionally robust chance-constrained optimization with Wasserstein metric*, Journal of Global Optimization, 79 (2021), pp. 779–811.

[38] N. Jiang and W. Xie, *Distributionally favorable optimization: A framework for data-driven decision-*

making with endogenous outliers, Optimization Online, (2023).
[39] R. Koenker and K. F. Hallock, *Quantile regression*, Journal of economic perspectives, 15 (2001), pp. 143–156.
[40] G. Li, *Robust regression*, Exploring data tables, trends, and shapes, 281 (1985), p. U340.
[41] J. Liu and J.-S. Pang, *Risk-based robust statistical learning by stochastic difference-of-convex value-function optimization*, Operations Research, 71 (2023), pp. 397–414.
[42] X. Liu, S. Küçükyavuz, and J. Luedtke, *Decomposition algorithms for two-stage chance-constrained programs*, Mathematical Programming, 157 (2016), pp. 219–243.
[43] J. Luedtke, *A branch-and-cut decomposition algorithm for solving chance-constrained mathematical programs with finite support*, Mathematical Programming, 146 (2014), pp. 219–244.
[44] J. Luedtke and S. Ahmed, *A sample approximation approach for optimization with probabilistic constraints*, SIAM Journal on Optimization, 19 (2008), pp. 674–699.
[45] R. A. Maronna, R. D. Martin, V. J. Yohai, and M. Salibián-Barrera, *Robust statistics: theory and methods (with R)*, John Wiley & Sons, 2019.
[46] D. L. Massart, L. Kaufman, P. J. Rousseeuw, and A. Leroy, *Least median of squares: a robust method for outlier and model error detection in regression and calibration*, Analytica Chimica Acta, 187 (1986), pp. 171–179.
[47] P. Mohajerin Esfahani, S. Shafieezadeh-Abadeh, G. A. Hanasusanto, and D. Kuhn, *Data-driven inverse optimization with imperfect information*, Mathematical Programming, 167 (2018), pp. 191–234.
[48] N. Naimipour, S. Khobahi, and M. Soltanalian, *Upr: A model-driven architecture for deep phase retrieval*. arXiv preprint arXiv:2003.04396, 2020.
[49] V. A. Nguyen, S. S. Abadeh, M.-C. Yue, D. Kuhn, and W. Wiesemann, *Calculating optimistic likelihoods using (geodesically) convex optimization*, in Advances in Neural Information Processing Systems, 2019, pp. 13942–13953.
[50] V. A. Nguyen, S. S. Abadeh, M.-C. Yue, D. Kuhn, and W. Wiesemann, *Optimistic distributionally robust optimization for nonparametric likelihood approximation*, in Advances in Neural Information Processing Systems, 2019, pp. 15872–15882.
[51] V. A. Nguyen, N. Si, and J. Blanchet, *Robust bayesian classification using an optimistic score ratio*, in International Conference on Machine Learning, PMLR, 2020, pp. 7327–7337.
[52] M. Norton, A. Takeda, and A. Mafusalov, *Optimistic robust optimization with applications to machine learning*. arXiv preprint arXiv:1711.07511, 2017.
[53] A. Prékopa, *Stochastic programming*, Springer Science & Business Media, 1995.
[54] H. Rahimian and S. Mehrotra, *Frameworks and results in distributionally robust optimization*, Open Journal of Mathematical Optimization, 3 (2022), pp. 1–85.
[55] R. T. Rockafellar, S. Uryasev, et al., *Optimization of conditional value-at-risk*, Journal of risk, 2 (2000), pp. 21–42.
[56] R. T. Rockafellar and R. J. Wets, *Stochastic convex programming: relatively complete recourse and induced feasibility*, SIAM Journal on Control and Optimization, 14 (1976), pp. 574–589.
[57] R. T. Rockafellar and R. J. Wets, *On the interchange of subdifferentiation and conditional expectation for convex functionals*, Stochastics: An International Journal of Probability and Stochastic Processes, 7 (1982), pp. 173–182.
[58] P. J. Rousseeuw and A. M. Leroy, *Robust Regression and Outlier Detection*, John Wiley & Sons, Inc., 1987.
[59] H. L. Royden and P. Fitzpatrick, *Real analysis*, Macmillan New York, 1988.
[60] W. Rudin, *Principles of mathematical analysis*, McGraw-hill New York, 1964.
[61] S. Sarykalin, G. Serraino, and S. Uryasev, *Value-at-risk vs. conditional value-at-risk in risk management and optimization*, in State-of-the-art decision-making tools in the information-intensive age, Informs, 2008, pp. 270–294.
[62] S. Shafieezadeh Abadeh, P. M. Mohajerin Esfahani, and D. Kuhn, *Distributionally robust logistic regression*, Advances in Neural Information Processing Systems, 28 (2015).
[63] S. Shalev-Shwartz et al., *Online learning and online convex optimization*, Foundations and trends in Machine Learning, 4 (2011), pp. 107–194.
[64] A. Shapiro and S. Ahmed, *On a class of minimax stochastic programs*, SIAM Journal on Optimiza-

tion, 14 (2004), pp. 1237–1249.
[65] A. Shapiro, D. Dentcheva, and A. Ruszczyński, *Lectures on stochastic programming: modeling and theory*, SIAM, 2014.
[66] H. Shen and R. Jiang, *Chance-constrained set covering with wasserstein ambiguity*, Mathematical Programming, 198 (2023), pp. 621–674.
[67] J. Song and C. Zhao, *Optimistic distributionally robust policy optimization*. arXiv preprint arXiv:2006.07815, 2020.
[68] L. Sun, W. Xie, and T. Witten, *Distributionally robust fair transit resource allocation during a pandemic*, Transportation science, 57 (2023), pp. 954–978.
[69] R. S. Sutton and A. G. Barto, *Reinforcement learning: An introduction*, MIT press, 2018.
[70] H. C. Tijms, *A first course in stochastic models*, John Wiley and sons, 2003.
[71] J. W. Tukey, *Exploratory data analysis*, Pearson, 1977.
[72] M. J. Wainwright, *High-dimensional statistics: A non-asymptotic viewpoint*, vol. 48, Cambridge University Press, 2019.
[73] A. A. Weiss, *Estimating nonlinear dynamic models using least absolute error estimation*, Econometric Theory, (1991), pp. 46–68.
[74] R. E. Welsch and X. Zhou, *Application of robust statistics to asset allocation models*, REVSTAT-Statistical Journal, 5 (2007), pp. 97–114.
[75] W. Xie, *Tractable reformulations of two-stage distributionally robust linear programs over the type-∞ Wasserstein ball*, Operations Research Letters, 48 (2020), pp. 513–523.
[76] W. Xie, *On distributionally robust chance constrained programs with wasserstein distance*, Mathematical Programming, 186 (2021), pp. 115–155.
[77] W. Xie, J. Zhang, and S. Ahmed, *Distributionally robust bottleneck combinatorial problems: uncertainty quantification and robust decision making*, Mathematical Programming, (2021), pp. 1–44.
[78] C. Yale and A. B. Forsythe, *Winsorized regression*, Technometrics, 18 (1976), pp. 291–300.
[79] K. Yu, Z. Lu, and J. Stander, *Quantile regression: applications and current research areas*, Journal of the Royal Statistical Society: Series D (The Statistician), 52 (2003), pp. 331–350.

Appendix A. Formal Proof of the Connections Between Chance Constrained Programming and Robust Optimization Using DFO (1.2).

Proposition A.1. *Suppose the interval ambiguity set is $\mathcal{P}_I = \{\boldsymbol{\mu} : \boldsymbol{\mu}(\mathcal{U}) = 1, 0 \preceq \boldsymbol{\mu} \preceq \mathbb{P}_0/(1-\varepsilon)\}$, then the DFO counterpart of a robust optimization (1.4a) is equivalent to a chance constrained program*

$$v^* = \min_{\boldsymbol{x}\in\mathcal{X}} \left\{ \boldsymbol{c}^\top \boldsymbol{x} \colon \mathbb{E}_{\mathbb{P}_0}\left[\mathbb{I}\left(G(\boldsymbol{x},\tilde{\boldsymbol{\xi}}) > 0\right)\right] \le \varepsilon \right\}. \tag{A.1}$$

Proof. According to the duality result in [64], we have

$$\inf_{\boldsymbol{\mu}\in\mathcal{P}_I} \mathbb{E}_{\boldsymbol{\mu}}\left[\mathbb{I}\left(G(\boldsymbol{x},\tilde{\boldsymbol{\xi}})>0\right)\right] = \max_{\lambda_0}\left\{F(\boldsymbol{x},\lambda_0) := \lambda_0 + \frac{1}{1-\varepsilon}\mathbb{E}_{\mathbb{P}_0}\left[\left(\mathbb{I}\left(G(\boldsymbol{x},\tilde{\boldsymbol{\xi}}) > 0\right) - \lambda_0\right)_-\right]\right\}.$$

Since

$$F(\boldsymbol{x},\lambda_0) = \begin{cases} \lambda_0, & \text{if } \lambda_0 \le 0,\\ \lambda_0 + \frac{1-\lambda_0}{1-\varepsilon}\mathbb{E}_{\mathbb{P}_0}\left[\mathbb{I}\left(G(\boldsymbol{x},\tilde{\boldsymbol{\xi}}) > 0\right)\right], & \text{if } 0 < \lambda_0 < 1,\\ -\frac{\varepsilon\lambda_0}{1-\varepsilon} + \frac{1}{1-\varepsilon}\mathbb{E}_{\mathbb{P}_0}\left[\mathbb{I}\left(G(\boldsymbol{x},\tilde{\boldsymbol{\xi}}) > 0\right)\right], & \text{if } \lambda_0 \ge 1, \end{cases}$$

by optimizing over λ_0, we further have

$$\begin{aligned} \max_{\lambda_0} F(\boldsymbol{x},\lambda_0) &= \max\left\{\max_{\lambda_0\le 0} F(\boldsymbol{x},\lambda_0), \max_{0<\lambda_0<1} F(\boldsymbol{x},\lambda_0), \max_{\lambda_0\ge 1} F(\boldsymbol{x},\lambda_0)\right\}\\ &= \max\left\{0, -\varepsilon + \mathbb{E}_{\mathbb{P}_0}\left[\mathbb{I}\left(G(\boldsymbol{x},\tilde{\boldsymbol{\xi}}) > 0\right)\right]\right\}. \end{aligned}$$

Therefore, the conclusion follows by substituting the last equality into the left-hand side of the constraint in DFO (1.4b). □

A.1 Proof of Proposition 2.1

Proposition A.2. *(i) Given an interval ambiguity set $\mathcal{P}_I = \{\mathbb{P} : \mathbb{P}(\mathcal{U}) = 1, 0 \preceq \mathbb{P} \preceq \mathbb{P}_0/(1-\varepsilon)\}$*

with support $\mathcal{U} = \operatorname{supp}(\mathbb{P}_0)$*, we have*

$$\text{(2.3a)} \qquad \inf_{\mathbb{P}\in\mathcal{P}_I} \mathbb{E}_{\mathbb{P}}\left[\tilde{\boldsymbol{X}}\right] = \max_{\beta}\left\{\beta + \frac{1}{1-\varepsilon}\mathbb{E}_{\mathbb{P}_0}\left[\left(\tilde{\boldsymbol{X}}-\beta\right)_-\right]\right\} = \mathbb{P}_0\text{-FCVaR}_{1-\varepsilon}\left(\tilde{\boldsymbol{X}}\right);$$

(ii) An optimal solution of the right-hand side optimization problem (2.2) *is* $\beta^* = \mathbb{P}_0\text{-VaR}_{1-\varepsilon}(\tilde{\boldsymbol{X}})$*; and*

(iii) The $\mathbb{P}_0\text{-FCVaR}_{1-\varepsilon}(\tilde{\boldsymbol{X}})$ *can be bounded by two conditional expectations:*

$$\text{(2.3b)} \qquad \mathbb{E}_{\mathbb{P}}\left[\tilde{\boldsymbol{X}}\middle|\tilde{\boldsymbol{X}} < \mathbb{P}_0\text{-VaR}_{1-\varepsilon}\left(\tilde{\boldsymbol{X}}\right)\right] \le \mathbb{P}_0\text{-FCVaR}_{1-\varepsilon}\left(\tilde{\boldsymbol{X}}\right) \le \mathbb{E}_{\mathbb{P}}\left[\tilde{\boldsymbol{X}}\middle|\tilde{\boldsymbol{X}} \le \mathbb{P}_0\text{-VaR}_{1-\varepsilon}\left(\tilde{\boldsymbol{X}}\right)\right].$$

Proof. We split the proof into three parts by checking these three statements separately.

(i) The proof of the first statement is similar to that of Proposition A.1 and thus is omitted.

(ii) Since the right-hand side optimization problem (2.2) is an unconstrained concave minimization, let us consider the first-order condition of FCVaR (2.2) for an optimal solution β^*, that is,

$$0 \in \frac{\partial \mathbb{P}_0\text{-FCVaR}_{1-\varepsilon}(\tilde{\boldsymbol{X}})}{\partial\beta}\Big|_{\beta=\beta^*} = 1 + \frac{1}{1-\varepsilon}\partial_\beta\left[\mathbb{E}_{\mathbb{P}_0}\left[\left(\tilde{\boldsymbol{X}}-\beta\right)_-\right]\right]\Bigg|_{\beta=\beta^*}.$$

According to the continuity of function $f(t)=\min(t,0)$ and theorem 1 in [57], we can interchange the subdifferential operator and expectation, that is,

$$\text{(A.2)}\quad 0 = 1 + \frac{1}{1-\varepsilon}\mathbb{E}_{\mathbb{P}_0}\left[\partial_\beta\left[\left(\tilde{\boldsymbol{X}}-\beta\right)_-\right]\Big|_{\beta=\beta^*}\right] = 1 - \frac{1}{1-\varepsilon}\mathbb{P}_0\left\{\tilde{\boldsymbol{X}} < \beta^*\right\} - \frac{\omega}{1-\varepsilon}\mathbb{P}_0\left\{\tilde{\boldsymbol{X}} = \beta^*\right\},$$

for some $\omega\in[0,1]$. Letting $\omega=0$ and 1, we have the following inequalities

$$1-\varepsilon \ge \mathbb{P}_0\left\{\tilde{\boldsymbol{X}} < \beta^*\right\}, \quad 1-\varepsilon \le \mathbb{P}_0\left\{\tilde{\boldsymbol{X}} \le \beta^*\right\}.$$

Above, the second inequality implies that $\beta^* \ge \mathbb{P}_0\text{-VaR}_{1-\varepsilon}(\tilde{\boldsymbol{X}})$. Suppose that $\beta^* > \mathbb{P}_0\text{-VaR}_{1-\varepsilon}(\tilde{\boldsymbol{X}})$. Then the first inequality together and the definition of $\mathbb{P}_0\text{-VaR}_{1-\varepsilon}(\tilde{\boldsymbol{X}})$ implies that

$$1-\varepsilon \ge \mathbb{P}_0\left\{\tilde{\boldsymbol{X}} < \beta^*\right\} \ge \mathbb{P}_0\left\{\tilde{\boldsymbol{X}} \le \mathbb{P}_0\text{-VaR}_{1-\varepsilon}(\tilde{\boldsymbol{X}})\right\} \ge 1-\varepsilon.$$

Thus, all inequalities become equalities. Letting $\omega=1$ in the optimality condition (A.2), we have

$$0 = 1 - \frac{1}{1-\varepsilon}\mathbb{P}_0\left\{\tilde{\boldsymbol{X}} < \mathbb{P}_0\text{-VaR}_{1-\varepsilon}(\tilde{\boldsymbol{X}})\right\} - \frac{1}{1-\varepsilon}\mathbb{P}_0\left\{\tilde{\boldsymbol{X}} = \mathbb{P}_0\text{-VaR}_{1-\varepsilon}(\tilde{\boldsymbol{X}})\right\},$$

which implies that $\beta^* = \mathbb{P}_0\text{-VaR}_{1-\varepsilon}(\tilde{\boldsymbol{X}})$ is another optimal solution.

(iii) Let us first prove the lower bound. According to the definition of conditional expectation, we have

$$\begin{aligned}&\mathbb{E}_{\mathbb{P}_0}\left[\tilde{\boldsymbol{X}}\middle|\tilde{\boldsymbol{X}} < \mathbb{P}_0\text{-VaR}_{1-\varepsilon}(\tilde{\boldsymbol{X}})\right]\\ =&\mathbb{P}_0\text{-VaR}_{1-\varepsilon}(\tilde{\boldsymbol{X}}) + \frac{\mathbb{E}_{\mathbb{P}_0}\left[\left(\tilde{\boldsymbol{X}} - \mathbb{P}_0\text{-VaR}_{1-\varepsilon}(\tilde{\boldsymbol{X}})\right)\mathbb{I}\{\tilde{\boldsymbol{X}} < \mathbb{P}_0\text{-VaR}_{1-\varepsilon}(\tilde{\boldsymbol{X}})\}\right]}{\mathbb{P}_0\left\{\tilde{\boldsymbol{X}} < \mathbb{P}_0\text{-VaR}_{1-\varepsilon}(\tilde{\boldsymbol{X}})\right\}}.\end{aligned}$$

Since $\mathbb{P}_0\{\tilde{\boldsymbol{X}} < \mathbb{P}_0\text{-VaR}_{1-\varepsilon}(\tilde{\boldsymbol{X}})\} \le 1-\varepsilon$ and $\mathbb{E}_{\mathbb{P}_0}[(\tilde{\boldsymbol{X}} - \mathbb{P}_0\text{-VaR}_{1-\varepsilon}(\tilde{\boldsymbol{X}}))\mathbb{I}\{\tilde{\boldsymbol{X}} < \mathbb{P}_0\text{-VaR}_{1-\varepsilon}(\tilde{\boldsymbol{X}})\}] = \mathbb{E}_{\mathbb{P}_0}[\min\{\tilde{\boldsymbol{X}} - \mathbb{P}_0\text{-VaR}_{1-\varepsilon}(\tilde{\boldsymbol{X}}), 0\}] \le 0$, we have

$$\begin{aligned}&\mathbb{E}_{\mathbb{P}_0}\left[\tilde{\boldsymbol{X}}\middle|\tilde{\boldsymbol{X}} < \mathbb{P}_0\text{-VaR}_{1-\varepsilon}(\tilde{\boldsymbol{X}})\right]\\ \le&\frac{\mathbb{E}_{\mathbb{P}_0}\left[\min\left\{\tilde{\boldsymbol{X}} - \mathbb{P}_0\text{-VaR}_{1-\varepsilon}(\tilde{\boldsymbol{X}}), 0\right\}\right]}{1-\varepsilon} + \mathbb{P}_0\text{-VaR}_{1-\varepsilon}\left(\tilde{\boldsymbol{X}}\right) = \mathbb{P}_0\text{-FCVaR}_{1-\varepsilon}\left(\tilde{\boldsymbol{X}}\right).\end{aligned}$$

Thus, the lower bound is valid.

Similarly, we can write the upper bound as

$$\begin{aligned}&\mathbb{E}_{\mathbb{P}_0}\left[\tilde{\boldsymbol{X}}\middle|\tilde{\boldsymbol{X}} \le \mathbb{P}_0\text{-VaR}_{1-\varepsilon}(\tilde{\boldsymbol{X}})\right]\\ =&\mathbb{P}_0\text{-VaR}_{1-\varepsilon}(\tilde{\boldsymbol{X}}) + \frac{\mathbb{E}_{\mathbb{P}_0}\left[\left(\tilde{\boldsymbol{X}} - \mathbb{P}_0\text{-VaR}_{1-\varepsilon}(\tilde{\boldsymbol{X}})\right)\mathbb{I}\{\tilde{\boldsymbol{X}} \le \mathbb{P}_0\text{-VaR}_{1-\varepsilon}(\tilde{\boldsymbol{X}})\}\right]}{\mathbb{P}_0\left\{\tilde{\boldsymbol{X}} \le \mathbb{P}_0\text{-VaR}_{1-\varepsilon}(\tilde{\boldsymbol{X}})\right\}}.\end{aligned}$$

Since $\mathbb{P}_0\{\tilde{\boldsymbol{X}} \le \mathbb{P}_0\text{-VaR}_{1-\varepsilon}(\tilde{\boldsymbol{X}})\} \ge 1-\varepsilon$ and $\mathbb{E}_{\mathbb{P}_0}[(\tilde{\boldsymbol{X}} - \mathbb{P}_0\text{-VaR}_{1-\varepsilon}(\tilde{\boldsymbol{X}}))\mathbb{I}\{\tilde{\boldsymbol{X}} \le \mathbb{P}_0\text{-VaR}_{1-\varepsilon}(\tilde{\boldsymbol{X}})\}] =$

$\mathbb{E}_{\mathbb{P}_0}[\min\{\tilde{\boldsymbol{X}} - \mathbb{P}_0\text{-VaR}_{1-\varepsilon}(\tilde{\boldsymbol{X}}), 0\}]$, we have

$$\mathbb{E}_{\mathbb{P}_0}\left[\tilde{\boldsymbol{X}} \middle| \tilde{\boldsymbol{X}} \leq \mathbb{P}_0\text{-VaR}_{1-\varepsilon}(\tilde{\boldsymbol{X}})\right] \geq \mathbb{P}_0\text{-VaR}_{1-\varepsilon}(\tilde{\boldsymbol{X}}) + \frac{1}{1-\varepsilon}\mathbb{E}_{\mathbb{P}_0}\left[\min\left\{\tilde{\boldsymbol{X}} - \mathbb{P}_0\text{-VaR}_{1-\varepsilon}(\tilde{\boldsymbol{X}}), 0\right\}\right]$$

$$= \mathbb{P}_0\text{-FCVaR}_{1-\varepsilon}\left(\tilde{\boldsymbol{X}}\right).$$

This completes the proof. □

We remark that existing works (see, e.g., [55] and [61]) focus on CVaR, while our result in the proof above holds for a distinct notion FCVaR. Our proof is also different from the CVaR literature.

Appendix B. More Robust Statistics that DFO Can Recover and Beyond

B.1 DFO Recovers Median It is well-known that the median of a dataset is much less sensitive to outliers than the mean (see more discussions in [35]). For example, one or two outlier data points with large values may change the mean dramatically, while the median may not even change. By choosing a proper uncertainty set, we observe that the rDFO (1.3) can recover the median of a dataset. That is, given m data points $\{s_i\}_{i\in[m]} \in \mathbb{R}$, it is well known that the mean of $\{s_i\}_{i\in[m]}$ is achieved by solving the following least-square optimization:

$$\text{(B.1a)} \qquad \text{mean}(\{s_i\}_{i\in[m]}) \in \arg\min_x \sum_{i\in[m]} \xi^i |x - s_i|^2,$$

which places equal weight $\xi^i = 1/m$ on each data point for all $i \in [m]$. If we consider the weight uncertainty set $\mathcal{U} = \{\boldsymbol{\xi} \in \mathbb{R}_+^m : \sum_{i\in[m]} 1/\xi^i = m^2\}$, applying rDFO to the problem (B.1a) can recover the median of data points $\{s_i\}_{i\in[m]}$.

Proposition B.1. *The median of data points $\{s_i\}_{i\in[m]} \in \mathbb{R}$ can be found by*

$$\text{(B.1b)} \qquad \textit{median}(\{s_i\}_{i\in[m]}) \in \arg\min_x \min_{\boldsymbol{\xi}\in\mathcal{U}} \sum_{i\in[m]} \xi^i |x - s_i|^2,$$

where $\mathcal{U} = \{\boldsymbol{\xi} \in \mathbb{R}_+^m : \sum_{i\in[m]} 1/\xi^i = m^2\}$.

Proof. From the definition of the weight uncertainty set $\mathcal{U}$, we can rewrite problem (B.1b) as

$$\text{(B.2a)} \qquad \min_x \min_{\boldsymbol{\xi}\in\mathcal{U}} \frac{1}{m^2} \sum_{i\in[m]} \frac{1}{\xi^i} \sum_{i\in[m]} \xi^i |x - s_i|^2.$$

According to Cauchy-Schwarz inequality (see, e.g., thereon 1.37 in [60]), we have

$$\sum_{i\in[m]} \frac{1}{\xi^i} \sum_{i\in[m]} \xi^i |x - s_i|^2 \geq \left(\sum_{i\in[m]} |x - s_i|\right)^2,$$

and the equality can be achieved when $\xi^{i^*} = c/|x - s_i|$ for each $i \in [m]$ and $c = \sum_{j\in[m]} |x - s_j|/m^2$.

Thus, problem (B.2a) can be written as

$$\text{(B.2b)} \qquad v^* = \min_x \frac{1}{m^2}\left(\sum_{i\in[m]} |x - s_i|\right)^2 = \left(\min_x \frac{1}{m} \sum_{i\in[m]} |x - s_i|\right)^2,$$

and the solution of the right-hand problem in (B.2b) can be interpreted as the median of $\{s_i\}_{i\in[m]}$. This completes the proof. □

This result shows that in the presence of endogenous outliers, the DFO framework, weighing more on the favorable data points, can be more desirable than its risk-neutral counterpart.

B.2 DFO Recovers More Robust Statistics Based on Proposition B.1 Using the same weight uncertainty set $\mathcal{U}$ and following the similar derivation as Proposition B.1, we are able to recover more similar robust statistics, such as median absolute deviation (MAD), least absolute deviation (LAD), and least median of squares (LMS).

(i) Median absolute deviation (MAD), a robust measure of the variability of the data (see, e.g., [32]), can be represented as the median of the absolute deviations from the median of the data. That is, given data points $\{s_i\}_{i\in[m]} \in \mathbb{R}$ and their median $\widehat{s}$, the MAD can be interpreted as

$$\min_{x} \min_{\boldsymbol{\xi}\in\mathcal{U}} \sum_{i\in[m]} \xi^i \left(x - |s_i - \widehat{s}|\right)^2 = \left(\min_{x} \frac{1}{m} \sum_{i\in[m]} |x - |s_i - \widehat{s}|| \right)^2 .$$

Here, applying DFO converts the less reliable average absolute deviation (i.e., $\xi^i = 1/m$ in the above left-hand problem) to the desirable MAD;

(ii) Least absolute deviation (LAD), a special case of robust regression (see, e.g., [40]), minimizes the L_1 norm of the residuals. That is, given m data points $\{\bar{\boldsymbol{x}}_i, y_i\}_{i\in[m]} \subseteq \mathbb{R}^d \times \mathbb{R}$, suppose that the residual function is defined as $r_i(\boldsymbol{\beta}) = (y_i - \bar{\boldsymbol{x}}_i^\top \boldsymbol{\beta})$, for each $i \in [m]$. Then, applying the DFO converts the least-square regression problem to the LAD regression problem

$$v^* = \min_{\boldsymbol{\beta}} \min_{\boldsymbol{\xi}\in\mathcal{U}} \sum_{i\in[m]} \xi^i (r_i(\boldsymbol{\beta}))^2 = \left(\min_{\boldsymbol{\beta}} \frac{1}{m} \sum_{i\in[m]} |r_i(\boldsymbol{\beta})| \right)^2 ;$$

(iii) Least median of squares (LMS) is another known robust regression (see, e.g., [46]), which minimizes the median of the squared residuals. Given m data points $\{\bar{\boldsymbol{x}}_i, y_i\}_{i\in[m]} \subseteq \mathbb{R}^d \times \mathbb{R}$, suppose the residual $r_i(\boldsymbol{\beta}) = (y_i - \bar{\boldsymbol{x}}_i^\top \boldsymbol{\beta})$ for each $i \in [m]$. Then LMS can be interpreted as applying DFO to the average squared residuals:

$$\min_{x,\boldsymbol{\beta}} \min_{\boldsymbol{\xi}\in\mathcal{U}} \sum_{i\in[m]} \xi^i |x - r_i^2(\boldsymbol{\beta})|^2 = \left(\min_{x,\boldsymbol{\beta}} \frac{1}{m} \sum_{i\in[m]} |x - r_i^2(\boldsymbol{\beta})| \right)^2 ;$$

(iv) Least Absolute Error Estimation (LAEE) is an alternative to LAD when the size of the relative error is a severe concern (see, e.g., [73]). Given m data points $\{\bar{\boldsymbol{x}}_i, y_i\}_{i\in[m]} \subseteq \mathbb{R}^d \times \mathbb{R}$, suppose that the residual $r_i(\boldsymbol{\beta}) = (y_i - \bar{\boldsymbol{x}}_i^\top \boldsymbol{\beta})$ for each $i \in [m]$. Then LAEE can be interpreted as applying DFO to the average squared relative residuals:

$$v^* = \min_{\boldsymbol{\beta}} \min_{\boldsymbol{\xi}\in\mathcal{U}} \sum_{i\in[m]} \xi^i \left(\frac{r_i(\boldsymbol{\beta})}{y_i} \right)^2 = \left(\min_{\boldsymbol{\beta}} \frac{1}{m} \sum_{i\in[m]} \left| \frac{r_i(\boldsymbol{\beta})}{y_i} \right| \right)^2 .$$

B.3 DFO Recovers More M-Estimators We use DFO to recover the Huber estimator [34] and Tukey's bisquare estimator [71].

Huber Estimator [34]. The Huber loss function is defined as

$$L_\delta(x) = \begin{cases} \frac{1}{2}x^2, & |x| \le \delta \\ \delta \left(|x| - \frac{1}{2}\delta \right), & \text{otherwise} \end{cases} .$$

The following DFO can recover the Huber estimator:

$$v^* = \min_{\boldsymbol{\beta}} \min_{\mathbb{P}\in\mathcal{P}} \mathbb{E}_{\mathbb{P}} \left[\mathcal{L}(\boldsymbol{\beta}, \tilde{\boldsymbol{\xi}}) \right]$$

where the ambiguity set $\mathcal{P}$ is decision-dependent as below

$$\mathcal{P} = \left\{ \frac{1}{N} \sum_{i\in[N]} \mathbb{P}_i : \mathbb{P}_i \left\{ \tilde{\boldsymbol{\xi}} : \mathcal{L}(\boldsymbol{\beta}, \tilde{\boldsymbol{\xi}}) = \frac{1}{2} r_i^2(\boldsymbol{\beta}) \right\} + \mathbb{P}_i \left\{ \tilde{\boldsymbol{\xi}} : \mathcal{L}(\boldsymbol{\beta}, \tilde{\boldsymbol{\xi}}) = \delta \left(|r_i(\boldsymbol{\beta})| - \frac{1}{2}\delta \right) \right\} = 1 \right\},$$

with support $\mathcal{U} = \{\boldsymbol{\xi}^i\}_{i\in[N]} = \{\bar{\boldsymbol{x}}_i, y_i\}_{i\in[N]}$.

Tukey's Bisquare Estimator [71]. Similarly, we can use the DFO to recover the Tukey's bisquare estimator, where Tukey's bisquare loss function is defined as

$$L_\delta(x) = \begin{cases} \frac{x^2}{2} - \frac{x^4}{2\delta^2} + \frac{x^6}{6\delta^4}, & |x| \le \delta \\ \frac{\delta^2}{6}, & \text{otherwise} \end{cases} .$$

The Tukey's bisquare estimator can be recovered as

$$v^* = \min_{\boldsymbol{\beta}} \min_{\mathbb{P}\in\mathcal{P}} \mathbb{E}_{\mathbb{P}} \left[\mathcal{L}(\boldsymbol{\beta}, \tilde{\boldsymbol{\xi}})\right]$$

where the ambiguity set $\mathcal{P}$ is decision-dependent as below

$$\mathcal{P} = \left\{ \frac{1}{N} \sum_{i\in[N]} \mathbb{P}_i : \mathbb{P}_i \left\{ \tilde{\boldsymbol{\xi}} : \mathcal{L}(\boldsymbol{\beta}, \tilde{\boldsymbol{\xi}}) = \frac{r_i^2(\boldsymbol{\beta})}{2} - \frac{r_i^4(\boldsymbol{\beta})}{2\delta^2} + \frac{r_i^6(\boldsymbol{\beta})}{6\delta^4} \right\} + \mathbb{P}_i \left\{ \tilde{\boldsymbol{\xi}} : \mathcal{L}(\boldsymbol{\beta}, \tilde{\boldsymbol{\xi}}) = \frac{\delta^2}{6} \right\} = 1 \right\},$$

with support $\mathcal{U} = \{\boldsymbol{\xi}^i\}_{i\in[N]} = \{\bar{\boldsymbol{x}}_i, y_i\}_{i\in[N]}$.

B.4 DFO Recovers Quantile Regression Quantile regression can be used to estimate and conduct inference on the conditional quantile functions, which is more robust against outliers in the response measurements (see, e.g., [39, 79]). Given n data points $\{\bar{\boldsymbol{x}}_i, y_i\}_{i\in[m]} \subseteq \mathbb{R}^d \times \mathbb{R}$, the quantile regression problem can be modeled as

$$\text{(B.3a)} \qquad \min_{\boldsymbol{\beta}} \left\{ \tau \sum_{i\in[m]} (y_i - \bar{\boldsymbol{x}}_i^\top \boldsymbol{\beta})_+ + (1-\tau) \sum_{i\in[m]} (\bar{\boldsymbol{x}}_i^\top \boldsymbol{\beta} - y_i)_+ \right\},$$

where $\tau \in (0,1)$ is the given quantile. Similarly, we can recover the quantile regression problem with the following DFO:

$$\text{(B.3b)} \qquad v^* = \min_{\boldsymbol{\beta}} \min_{\boldsymbol{\xi}\in\mathcal{U}_I} \sum_{i\in[m]} \xi^i (y_i - \bar{\boldsymbol{x}}_i^\top \boldsymbol{\beta}) + \sum_{i\in[m]} |y_i - \bar{\boldsymbol{x}}_i^\top \boldsymbol{\beta}|,$$

where the "interval uncertainty set" $\mathcal{U}_I$ is defined as

$$\mathcal{U}_I = \left\{ \boldsymbol{\xi} \in \mathbb{R}^m : \tau - 1 \le \xi^i \le \tau, \forall i \in [m] \right\}.$$

Note that in (B.3b), letting $\xi^i = 0$ for all $i \in [m]$, it reduces to LAD.

B.5 DFO Can Recover Many Machine Learning Examples Phase Retrieval [36, 48]. Considering the least-square criterion, the task of recovering the signal from the measurements vector in phase retrieval admits the following form

$$v^* = \min_{\boldsymbol{x}} \frac{1}{n} \sum_{i\in[n]} \left(y_i - |\boldsymbol{a}_i^\top \boldsymbol{x}|\right)^2,$$

where $\boldsymbol{A} \in \mathbb{R}^{n\times d}$ is the sensing matrix with $\boldsymbol{a}_i$ denoting its ith row, $\boldsymbol{x}$ is the task of recovering the signal of interest, and $\boldsymbol{y} \in \mathbb{R}_+^n$ is the measurement.

Using the uncertainty $\mathcal{U} = \{-1, 1\}^n$, we can rewrite the phase retrieval problem as an equivalent DFO

$$v^* = \min_{\boldsymbol{x}} \min_{\boldsymbol{\xi}\in\mathcal{U}} \frac{1}{n} \sum_{i\in[n]} \left(y_i - \xi^i \boldsymbol{a}_i \boldsymbol{x}\right)^2,$$

which can be formulated as a mixed-integer program.

Clusterwise Linear Regression [3]. For a given dataset with N data points and d features $\{\bar{\boldsymbol{x}}_i, y_i\}_{i\in[N]} \subseteq \mathbb{R}^d \times \mathbb{R}$, for an integer $k \in [N]$, clusterwise linear regression (CLR) aims to find the partition of the data into k disjoint clusters such that each cluster subjects to a linear model and the overall sum of squared errors of linear regression models within each cluster is minimized. That is, CLR is equivalent to

$$\min_{\boldsymbol{\beta}, C_i} \left\{ \sum_{i\in[k]} \sum_{j\in C_i} \left(y_j - \bar{\boldsymbol{x}}_j^\top \boldsymbol{\beta}_i\right)^2 : \cup_{i\in[k]} C_i = [N], C_i \cap C_j = \emptyset, \forall i \ne j \right\}.$$

We can recast CLR problem as a DFO one. That is, suppose we choose the most favorable clusters, each with the least sum of squares. That is, we can rewrite the problem as the following DFO

$$v^* = \min_{\boldsymbol{\beta}} \min_{\boldsymbol{\xi}\in\mathcal{U}} \left\{ \sum_{i\in[k]} \sum_{j\in[N]} \xi^{ij} \left(y_j - \bar{\boldsymbol{x}}_j^\top \boldsymbol{\beta}_i\right)^2 \right\},$$

where $\mathcal{U} = \{\boldsymbol{\xi} : \sum_{i\in[k]} \xi^{ij} = 1, \xi^{ij} \in [0,1], \forall i \in [k], j \in [N]\}$.

The Upper Confidence Bound (UCB) Algorithm [4]. The UCB algorithm has been widely used in online learning [13, 63, 69]. The UCB algorithm aims to explore the most favorable action when facing uncertainty, i.e., choose the most plausibly possible payoffs. The essence of the UCB algorithm is coincident with what we propose in DFO, that is,

$$a_t = \text{argmax}_{a\in\mathcal{A}} \max_{\xi\in\mathcal{U}_I(a)} Q(a) + \xi,$$

where $\mathcal{U}_I(a) = \{\xi : -\sqrt{(2\text{log}t)/(n_t a)} \leq \xi \leq \sqrt{(2\text{log}t)/(n_t a)}\}$ denotes the action-dependent interval uncertainty set with n_t being the number of the action a that has been selected at time epoch t, $Q(a)$ is the expected reward with decision a, and $\mathcal{A}$ is the action set.

We conclude this section by remarking that DFO can recover many other robust statistics.