DFO: A Framework for Data-driven Decision-making with Endogenous Outliers

Nan Jiang · Weijun Xie

March 19, 2022

Abstract A typical data-driven stochastic program aims to seek 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 outlier scenarios such that their corresponding recourse function values are extremely large or even infinite, the commonly-used DRO framework alone tends to over-emphasize these 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, echoing that the DFO framework might be appropriate for the stochastic recourse function in the presence of endogenous outliers. While being NP-hard, in general, many DFO models are shown to be mixed-integer convex programming representable (MICP-R). 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 performances, and DFO properly measures 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 confirms the promising of the framework.

Keywords. Distributionally Favorable Optimization; Distributionally Robust Optimization; Robust Statistics; Tractability; Mixed-Integer Convex Programming Representability

First Author: Nan Jiang
Affiliation: Virginia Tech, Blacksburg, VA
E-mail: jnan97@vt.edu

Corresponding Author: Weijun Xie
Affiliation: Virginia Tech, Blacksburg, VA
E-mail: wxie@vt.edu

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 a 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

$$\min_{\boldsymbol{x}\in\mathcal{X}} \left\{ \boldsymbol{c}^\top \boldsymbol{x} + \sup_{\mathbb{P}\in\mathcal{P}} \mathbb{E}_{\mathbb{P}} \left[Q(\boldsymbol{x}, \tilde{\boldsymbol{\xi}}) \right] \right\}, \tag{1}$$

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 the support $\mathcal{U} \subseteq \mathbb{R}^m$ (also known as "uncertainty set" throughout this paper). The DRO model (1) has successfully addressed many decision-making problems under uncertainty to achieve decision robustness, and better out-of-sample performances [20]. The inherent assumption in DRO is to assume that the expectation of the recourse function is finite for any distribution from the ambiguity set $\mathcal{P}$. This assumption may not hold when the data used to construct the ambiguity set are contaminated, such as abnormal data measurement or intentional data distortion (i.e., in the presence of "exogenous outliers"), or the recourse function value may be extremely large or even unbounded under some extreme scenarios (i.e., in the presence of "endogenous outliers"). Under such circumstances, the DRO model (1) tends to over-emphasize the outliers and causes undesirable or infeasible decisions[1]. In light of this issue, this paper studies the following Distributionally Favorable Optimization (DFO) by providing a proper measure to mitigate the effect of endogenous outliers

$$v^* = \min_{\boldsymbol{x}\in\mathcal{X}} \left\{ \boldsymbol{c}^\top \boldsymbol{x} + \inf_{\mathbb{P}\in\mathcal{P}} \mathbb{E}_{\mathbb{P}} \left[Q(\boldsymbol{x}, \tilde{\boldsymbol{\xi}}) \right] \right\}, \tag{2}$$

which instead seeks the best decision under the most favorable distribution. We will formally define a notion of decision outlier robustness for selecting a DFO measure in Section 4. It is worthy of mentioning that since DRO can achieve better out-of-sample performances, our Section 5 studies 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 (2) degenerates to a regular one (rDFO), i.e.,

$$v^* = \min_{\boldsymbol{x}\in\mathcal{X}} \left\{ \boldsymbol{c}^\top \boldsymbol{x} + \inf_{\boldsymbol{\xi}\in\mathcal{U}} Q(\boldsymbol{x}, \boldsymbol{\xi}) \right\}. \tag{3}$$

The special cases of the rDFO (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 the 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 naturally recovers many robust statistics, evidencing that DFO might be desirable for stochastic programming under 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.

Throughout the paper, we make the following assumptions for DFO (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 $\boldsymbol{\xi} \in \mathcal{U}$.

[1] Since exogenous outliers can be easily detected by preprocessing via a properly-selected robust statistic, this work mainly focuses on "endogenous outliers."

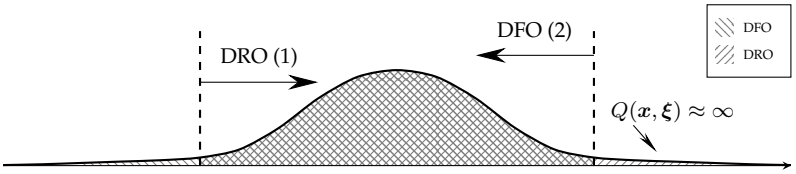


Fig. 1: Illustration of DFO vs. DRO in the Presence of Endogenous Outliers.

Both parts in Assumption 1 are standard in literature (see, e.g., section 5 in [7] and chapter 12 in [55]). Part (i) in Assumption 1 is useful to prove the mixed-integer convex programming representability and 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 4.

1.1 Literature Review

In literature, in contrast to DRO (see more details in [56]), researchers tend to use optimistic optimization (i.e., special cases of DFO) to tackle learning problems in various areas such as reinforcement learning [1, 65], Bayesian optimization [51–53], classification [9], image reconstruction [26], machine learning [54], etc. For instance, the authors in [65] applied the optimistic DRO approach to solve 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 [52], 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 similar to 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 [53] introduced a novel moment-based divergence ambiguity set and proposed a Bayesian contextual classification model using an optimistic score ratio. The researchers in [51] 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, 47].

There are also a few works focusing on special classes of the rDFO problems (see, e.g., [9, 54]). The work [9] proposed a novel formulation of support vector classification and derived a geometric interpretation of the proposed formulation to handle the uncertainty in classification. In [54], the authors argued that the optimistic assumption could be easier to realize regarding the real-world economic resources compared with the pessimistic or worst-case one. However, there either lacks a framework for DFO or optimistic optimization or lacks a study of complexity analysis, and the connection to robust statistics is also missing. This paper will fill the gap.

While this paper was prepared to submit, we became aware of the independent works from [12, 23], which discussed the class of distributionally optimistic optimization problems and their applications to contextual bandit problems. The fundamental difference of this work from theirs is that we focus on data-driven optimization with endogenous outliers, connecting to and motivating from robust statistics.

1.2 Summary of Contributions

In this paper, we study DFO (2) via various perspectives from statistics, machine learning, and optimization. Each perspective justifies and extends DFO. Particularly, we show the following three fundamental aspects of DFO: framework, optimization, 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 outlier robust.
- For the optimization aspect, we show that different from DRO (1) that the inner supremum preserves the convexity of the recourse function, the inner infimum in the DFO (2) often destroys the convexity of the recourse function. Hence, we study and extend the notion of mixed-integer convex programming representability (MICP-R) and apply it to the DFO. That is, although the tractability of DFO is rare, there is a rich family of DFO that can be recast as MICPs, which can be handled by the off-the-shelf solvers.
- For the unification aspect, we integrate DRO with DFO, termed "worst-case DFO," since DRO improves out-of-sample performances given that the sample size is finite. We show a proper way to integrate both. In particular, we focus on data-driven ambiguity set for DRO and decision outlier robust ambiguity set for DFO. The convergence analysis shows that the error of worst-case DFO decreases proportionally to the square root of 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).

Organization. The remainder of the paper is organized as follows. Section 2 shows the equivalence between DFO and many robust statistics. Section 3 conducts complexity analysis and shows the mixed-integer convex programming representability of DFO. Section 4 introduces the DFO framework for data-driven optimization with endogenous outliers and introduces the notion of decision outlier robustness. Section 5 integrates distributional robustness with DFO to leverage the out-of-sample performances. Sections 6 and 7 extend to two-stage cases and numerically illustrate the proposed methods. Section 8 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 $(t)_+ := \max\{t, 0\}$. 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 $\mathrm{supp}(\boldsymbol{x})$ be its support, i.e., $\mathrm{supp}(\boldsymbol{x}) := \{i \in [n] : x_i \neq 0\}$. Given a probability distribution $\mathbb{P}$ defined on Ξ 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 $\mathrm{ess.sup}_{\mathbb{P}}(g(\tilde{\boldsymbol{\xi}}))$ denote the essential supremum of the random 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 $\boldsymbol{\mu}_2 \succeq \boldsymbol{\mu}_1$ 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 notation will be introduced as needed.

2 DFO: A Framework to Recover Robust Statistics and Beyond

Different from DRO, in this section, we show that DFO can recover many robust statistics, demonstrating that the DFO framework can be more desirable for decision-making under uncertainty in the presence of outliers.

2.1 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 (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:

$$\text{mean}(\{s_i\}_{i\in[m]}) \in \arg\min_x \sum_{i\in[m]} \xi^i |x - s_i|^2, \tag{4a}$$

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 (4a) can recover the median of data points $\{s_i\}_{i\in[m]}$.

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

$$\textit{median}(\{s_i\}_{i\in[m]}) \in \arg\min_x \min_{\boldsymbol{\xi}\in\mathcal{U}} \sum_{i\in[m]} \xi^i |x - s_i|^2, \tag{4b}$$

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 (4b) as

$$\min_x \min_{\boldsymbol{\xi}\in\mathcal{U}} \frac{1}{m^2} \sum_{i\in[m]} \frac{1}{\xi^i} \sum_{i\in[m]} \xi^i |x - s_i|^2. \tag{5a}$$

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 (5a) can be written as

$$v^* = \min_x \frac{1}{m^2} \left(\sum_{i\in[m]} |x - s_i| \right)^2 = \left(\min_x \frac{1}{m} \sum_{i\in[m]} |x - s_i| \right)^2, \tag{5b}$$

and the solution of the right-hand problem in (5b) 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. This result also explains why in machine learning, the norm optimization tends to be called a robust model compared to its squared counterpart (see, e.g., [11, 19, 40, 41, 77]). It is worthy of mentioning that since norm squares enjoy better smooth conditions than norms, the DFO interpretation may be useful to design better algorithms for these norm optimization problems (e.g., robust tensor PCA). We remark that using the same uncertainty set $\mathcal{U}$ and following the similar derivation as Proposition 1, we can recover more similar robust statistics, such as median absolute deviation, least absolute deviation, and least median of squares. Interested readers are referred to Section A.1 of Appendix A for the detailed discussions.

2.2 DFO Recovers Least Trimmed Squares

The least trimmed squares (LTS) is a robust regression that learns from a subset of data not being affected by endogenous outliers (see, e.g., [58]). Given N data points $\{\bar{\boldsymbol{x}}_i, y_i\}_{i\in[N]} \subseteq \mathbb{R}^d \times \mathbb{R}$, 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 $\boldsymbol{r}^2(\boldsymbol{\beta})$ are sorted in ascending order $r^2_{(1)}(\boldsymbol{\beta}) := (y_{(1)} -$

$\bar{\boldsymbol{x}}_{(1)}^\top \boldsymbol{\beta})^2 \leq r_{(2)}^2(\boldsymbol{\beta}) \cdots \leq r_{(N)}^2(\boldsymbol{\beta}) := (y_{(N)} - \bar{\boldsymbol{x}}_{(N)}^\top \boldsymbol{\beta})^2$, 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,

$$v^* = \min_{\boldsymbol{\beta}} \min_{\boldsymbol{p}\in\mathcal{P}_I} \sum_{i\in[N]} p_i r_i^2(\boldsymbol{\beta}), \tag{6}$$

where the "interval ambiguity set" $\mathcal{P}_I$ is defined 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 from this example, we will show in Section 4 that DFO with interval ambiguity set is equivalent to favorable conditional value-at-risk (FCVaR).

2.3 DFO Recovers Winsorized Regression

Winsorized regression (see, e.g., [74]), 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 N data points $\{\bar{\boldsymbol{x}}_i, y_i\}_{i\in[N]} \subseteq \mathbb{R}^d \times \mathbb{R}$, let the squared residuals $r_i^2(\boldsymbol{\beta}) := (y_i - \bar{\boldsymbol{x}}_i^\top \boldsymbol{\beta})^2$ for each $i \in [N]$ and let $r_U(\boldsymbol{\beta}) \geq 0$ be the kth smallest residual with an integer number $k \in [N]$. The Winsorized regression can be formulated as

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

The following DFO recovers the Winsorized regression:

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

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

$$\mathcal{P} = \left\{\mathbb{P} = \frac{k}{N}\mathbb{P}^1 + \frac{N-k}{N}\mathbb{P}^2 \colon \mathbb{P}^1 \in \mathcal{P}_I, \mathbb{P}^2\left\{\tilde{\boldsymbol{\xi}} \in \mathcal{U} : \mathcal{L}(\boldsymbol{\beta}, \tilde{\boldsymbol{\xi}}) \geq \operatorname{ess.sup}_{\mathbb{P}^1} L(\boldsymbol{\beta}, \tilde{\boldsymbol{\xi}})\right\} = 1\right\},$$

and the interval ambiguity set $\mathcal{P}_I$ is $\mathcal{P}_I = \{\boldsymbol{p} \in \mathbb{R}_+^N : \sum_{i\in[N]} p_i = 1, 0 \leq p_i \leq 1/k\}$ and the support $\mathcal{U}$ is $\mathcal{U} = \{\boldsymbol{\xi}^i\}_{i\in[N]} = \{\bar{\boldsymbol{x}}_i, y_i\}_{i\in[N]}$.

The result can also be extended to recover the Ramp loss support vector machine, where the latter was studied in work [33].

2.4 DFO Recovers M-Estimators

This subsection uses DFO to recover some useful M-estimators, a broad class of robust estimators. For the sake of page limit, we discuss how to use DFO to recover the Huber-skip estimator [34]. Interested readers are referred to Section A.2 of Appendix A for the discussions of other Huber estimators [34] and Tukey's bisquare estimator [68].

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

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

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

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

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

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

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

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

2.5 DFO Can Recover Many Machine Learning Examples

Phase Retrieval [36, 50]. 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

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

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, 61, 66]. 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,

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

where $\mathcal{U}_I(a) = \{\xi : -\sqrt{(2\log t)/(n_t a)} \leq \xi \leq \sqrt{(2\log t)/(n_t a)}\}$ 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. Due to page limit, we refer the interested readers to Appendix A for more examples, i.e., quantile regression in Section A.3 of Appendix A.

3 Tractability, Complexity Analysis, and Mixed-Integer Convex Programming Representability of DFO

Section 2 shows the importance and practical meaningfulness of DFO by unveiling its close relationship with robust statistics. In this section, we discuss the tractability and complexity of DFO and provide sufficient and necessary conditions under which DFO can be tractable or intractable. We also discuss general conditions under which the DFO problem can be mixed-integer convex programming representable.

3.1 Definitions and Preliminary Results

First, we observe that evaluating the most favorable objective function value of DFO (2) for a given decision, in general, can be NP-hard.

Proposition 2 *Computing the inner infimum of DFO* (2)*, in general, is NP-hard even when the ambiguity set* $\mathcal{P} = \{\mathbb{P}\colon \mathbb{P}\{\tilde{\boldsymbol{\xi}} \in \mathcal{U}\} = 1\}$ *and the recourse function* $Q(\boldsymbol{x}, \boldsymbol{\xi})$ *can be represented as a simple linear program with objective uncertainty.*

Proof: See Appendix B.1. □

Therefore, throughout this section, we consider the case when computing the inner infimum of DFO (2) is tractable and focus on the complexity of DFO (2) and its special case (3) (i.e., the rDFO). When we mention tractability, we adopt the spirit from work [6] and we refer the readers to Definition 5 in Appendix C whenever mentioning the tractability.

We then discuss the conditions under which the DFO (2) can be mixed-integer convex programming (MICP) representable (MICP-R). We follow and generalize the MICP-R notion, introduced by the work [44].

Definition 1 (i) ([44]) Given $n, p, d \in \mathbb{Z}_+$, suppose that sets $\mathcal{S} \subseteq \mathbb{R}^n$ and $\mathcal{M} \subseteq \mathbb{R}^{n+p+d}$ are closed and convex. Then the tuple $(\mathcal{M}, p, d)$ induces an MICP formulation of set $\mathcal{S}$ if

$$\boldsymbol{x} \in \mathcal{S} \Leftrightarrow \exists \boldsymbol{y} \in \mathbb{R}^p, \boldsymbol{z} \in \mathbb{Z}^d, \text{s.t. } (\boldsymbol{x}, \boldsymbol{y}, \boldsymbol{z}) \in \mathcal{M};$$

(ii) (An MICP-R Set, [44]) A set $\mathcal{S} \in \mathbb{R}^n$ is MICP representable (MICP-R) if there exists a closed convex set $\mathcal{M}$ and two positive integers p and d that induce an MICP formulation of set $\mathcal{S}$;
(iii) (An MICP-R Function) A function $f : \mathcal{S} \to \mathbb{R}$ is MICP-R if both its domain $\mathcal{S}$ and its epigraph are MICP-R; and
(iv) (An MICP) A mathematical program is MICP-R if both its feasible region and objective function are MICP-R.

The definition of being not MICP-R is simply the opposite of being MICP-R, which is, unfortunately, difficult to verify in practice. Fortunately, the authors in [44] provided a simple sufficient condition to prove that a set is not MICP-R.

Lemma 1 *(lemma 2, [44]) A set $\mathcal{S} \in \mathbb{R}^n$ is not MICP-R if there exists an infinite sequence $\{\widehat{\boldsymbol{x}}^j\}_j$ such that $\widehat{\boldsymbol{x}}^{j_1} \neq \widehat{\boldsymbol{x}}^{j_2} \in \mathcal{S}$ for all $j_1 \neq j_2$ and $1/2(\widehat{\boldsymbol{x}}^{j_1} + \widehat{\boldsymbol{x}}^{j_2}) \notin \mathcal{S}$.*

For brevity of notation, we also introduce the McCormick representation [49] of a simple bilinear set having a discrete variable.

Definition 2 (McCormick Representation of a Simple Bilinear Set, [49]) The bilinear set $\{(s, \lambda, \gamma) \in \mathbb{R} \times \{\lambda_l, \lambda_u\} \times [\gamma_l, \gamma_u] : s = \lambda\gamma\}$ admits the following mixed-integer linear programming (MILP) McCormick representation:

$$\mathcal{MI}(\lambda_l, \lambda_u, \gamma_l, \gamma_u) = \left\{ \begin{array}{c} s \in \mathbb{R}, \lambda \in \{\lambda_l, \lambda_u\}, \gamma_l \leq \gamma \leq \gamma_u, \\ (s, \lambda, \gamma) : s \geq \lambda_l\gamma + \lambda\gamma_l - \lambda_l\gamma_l, s \geq \lambda_u\gamma + \lambda\gamma_u - \lambda_u\gamma_u, \\ s \leq \lambda_u\gamma + \lambda\gamma_l - \lambda_u\gamma_l, s \leq \lambda\gamma_u + \lambda_l\gamma - \lambda_l\gamma_u \end{array} \right\},$$

where λ_l, λ_u and γ_l, γ_u are the known lower and upper bounds for λ and γ, respectively.

According to Definition 1 and Definition 2, the following result shows the MICP-R of the reverse norm function $f(\boldsymbol{x}) = -\|\boldsymbol{x}\|_p + \chi_{\mathcal{X}}(\boldsymbol{x})$ (recall that set $\mathcal{X}$ is compact and has a non-empty interior).

Lemma 2 *The reverse norm function $f(\boldsymbol{x}) = -\|\boldsymbol{x}\|_p + \chi_{\mathcal{X}}(\boldsymbol{x})$ is MICP-R if $p \in \{1, \infty\}$ and is not MICP-R if $p \in (1, \infty)$.*

Proof: See Appendix B.2. □

Note that when $p = \infty$, although being MICP-R, the optimization over function $f(\boldsymbol{x})$ can be done efficiently by solving $2n$ tractable convex programs. The result in Lemma 2 is useful to prove the MICP-R of the DFO problem.

3.2 DFO (2) with the Ambiguity Set of Finite Support

Many robust statistics recovered by DFO (2) in Section 2 can be considered as the DFO (2) with the finite-support ambiguity set, i.e., when the support $\mathcal{U} := \{\boldsymbol{\xi}^i\}_{i\in[N]}$ is finite. Under this setting, we cannot obtain any nontrivial tractable results for the DFO (2). On the other hand, since evaluating the best case in the DFO (2) with a given decision is, in general, NP-hard (see Proposition 2), we will mainly focus on the MICP-R of DFO (2) with finite support in this subsection. Notably, when the ambiguity set $\mathcal{P}$ is a polytope, we show that DFO (2) is MICP-R by observing that the number of the extreme points of the polyhedral ambiguity set $\mathcal{P}$ is finite.

Theorem 1 *Suppose that the ambiguity set $\mathcal{P} = \left\{\boldsymbol{p} \in \mathbb{R}^N_+ : \boldsymbol{D}\boldsymbol{p} \leq \boldsymbol{d}\right\} \subseteq \left\{\boldsymbol{p} \in \mathbb{R}^N_+ : \mathbf{e}^\top \boldsymbol{p} = 1\right\}$ is a polytope and both set $\mathcal{X}$ and the recourse function $Q(\boldsymbol{x}, \boldsymbol{\xi})$ are MICP-R. Then, the corresponding DFO (2) is MICP-R.*

Proof: Since the ambiguity set $\mathcal{P}$ is a polytope, we can enumerate all its extreme points, i.e., $\boldsymbol{\gamma}^1, \ldots, \boldsymbol{\gamma}^s \in \mathbb{R}^N_+$ are the total s vertices of $\mathcal{P}$. Then, DFO (2) is equivalent to

$$v^* = \min_{j\in[s]} \min_{\boldsymbol{x}\in\mathcal{X}} \left\{ \boldsymbol{c}^\top \boldsymbol{x} + \sum_{i\in[N]} \gamma_i^j Q(\boldsymbol{x}, \boldsymbol{\xi}^i) \right\},$$

which is MICP-R, since both set $\mathcal{X}$ and the recourse function $Q(\boldsymbol{x},\boldsymbol{\xi})$ are MICP-R. □

The proof of Theorem 1 relies on the enumeration of extreme points of the ambiguity set $\mathcal{P}$, which can be computationally inefficient. In the following special case of Theorem 1, we provide the compact MICP-R formulation for the interval ambiguity set; that is, we consider the polyhedral interval ambiguity set as $\mathcal{P}_{PI}=\{\boldsymbol{p}=\boldsymbol{p}^0+\boldsymbol{\psi}\in\mathbb{R}_+^N:\boldsymbol{l}\le\boldsymbol{\psi}\le\boldsymbol{u},\boldsymbol{e}^\top\boldsymbol{\psi}=0\}$, where $\boldsymbol{p}^0$ denotes the nominal probability vector with $\boldsymbol{p}^0\ge\mathbf{0}$ and $\sum_{i\in[N]}p_i^0=1$, the lower bound vector $\boldsymbol{l}\ge-\boldsymbol{p}^0$ and the bounds $l_i=\bar{l}_i/q, u_i=\bar{u}_i/q$ with q being a positive integer and $\bar{l}_i,\bar{u}_i$ being integers for each $i\in[N]$.

Corollary 1 *Suppose both the set $\mathcal{X}$ and the recourse function $Q(\boldsymbol{x},\boldsymbol{\xi})$ are MICP-R. Then under polyhedral interval ambiguity set $\mathcal{P}_{PI}$, the optimal value of the corresponding DFO (2) is $v^*=\min_{j\in[N],\tau\in[\bar{l}_j,\bar{u}_j]}v_{j\tau}^*$ and for each $j\in[N]$ and $\tau\in\{\bar{l}_j,\bar{l}_j+1,\cdots,\bar{u}_j\}$, the value $v_{j\tau}^*$ can be computed via solving the following MICP-R formulation:*

$$\begin{aligned}v_{j\tau}^*=\min_{\substack{\boldsymbol{x}\in\mathcal{X},\boldsymbol{\eta},\boldsymbol{\nu},\\ \boldsymbol{z}^j\in\{0,1\}^N}}\ &\boldsymbol{c}^\top\boldsymbol{x}+\sum_{i\in[N]\setminus\{j\}}\left[p_i^0+\bar{l}_i/q\right]\nu_i+\sum_{i\in[N]\setminus\{j\}}(\bar{u}_i/q-\bar{l}_i/q)\eta_i^j+\left(p_j^0+\tau/q\right)\nu_j,\\ \text{s.t.}\quad&\nu_i\ge Q(\boldsymbol{x},\boldsymbol{\xi}^i),\left(\eta_i^j,z_i^j,\nu_i\right)\in\mathcal{MI}\left(0,1,L_i,U_i\right),\forall i\in[N],j\in[N],\\ &-\sum_{i\in[N]\setminus\{j\}}(\bar{l}_i+\left(\bar{u}_i-\bar{l}_i\right)z_i^j=\tau,\end{aligned}$$

where for each $i\in[N]$, L_i and U_i are the valid lower and upper bounds of the function $Q(\boldsymbol{x},\boldsymbol{\xi}^i)$, respectively.

Proof: The DFO (2) is equivalent to

$$v^*=\min_{\boldsymbol{x}\in\mathcal{X}}\min_{\boldsymbol{\psi}}\left\{\boldsymbol{c}^\top\boldsymbol{x}+\sum_{i\in[N]}\left[(p_i^0+\psi_i)Q(\boldsymbol{x},\boldsymbol{\xi}^i):\sum_{i\in[N]}\psi_i=0,\bar{l}_i/q\le\psi_i\le\bar{u}_i/q,\forall i\in[N]\right]\right\}.\tag{7}$$

According to the extreme point characterization of the ambiguity set $\mathcal{P}_{PI}$, for any extreme point $\widehat{\boldsymbol{\psi}}$, it has at least $N-1$ components taking values from $\boldsymbol{l}$ or $\boldsymbol{u}$ and one component corresponding to equality constraint. Let us assume that component $j\in[N]$ corresponds to the equality constraint. Accordingly, we can define the binary variable $z_i^j\in\{0,1\}$ for each $i\in[N]$ and $\widehat{\psi}_i=l_i+(u_i-l_i)z_i^j$ for each $i\in[N]\setminus\{j\}$. Since we have $\sum_{i\in[N]}\widehat{\psi}_i=0$, thus $\widehat{\psi}_j=-\sum_{i\in[N]\setminus\{j\}}(l_i+(u_i-l_i)z_i^j)$. Plugging the extreme point representation into set $\mathcal{P}_{PI}$, we have

$$\mathcal{P}_{PI}=\operatorname{conv}\left[\bigcup_{j\in[N]}\left\{\boldsymbol{p}=\boldsymbol{p}^0+\widehat{\boldsymbol{\psi}}\in\mathbb{R}_+^N:\ \begin{array}{c}\widehat{\psi}_i=l_i+(u_i-l_i)z_i^j,\forall i\in[N]\setminus\{j\},\\ \widehat{\psi}_j=-\sum_{i\in[N],i\setminus\{j\}}(l_i+(u_i-l_i)z_i^j)\in[l_j,u_j]\end{array}\right\}\right].$$

Plugging the representation of set $\mathcal{P}_I$, the DFO (7) is equivalent to $v^*=\min_{j\in[N]}v_j^*$ and for each $j\in[N]$,

$$\begin{aligned}v_j^*=\min_{\substack{\boldsymbol{x}\in\mathcal{X},\\ \boldsymbol{z}^j\in\{0,1\}^N}}\ &\boldsymbol{c}^\top\boldsymbol{x}+\sum_{i\in[N]\setminus\{j\}}\left[p_i^0+l_i+(u_i-l_i)z_i^j\right]Q(\boldsymbol{x},\boldsymbol{\xi}^i)+\left(p_j^0-\sum_{i\in[N],i\setminus\{j\}}\left(l_i+(u_i-l_i)z_i^j\right)\right)Q(\boldsymbol{x},\boldsymbol{\xi}^j),\\ \text{s.t.}\quad&l_j\le-\sum_{i\in[N]\setminus\{j\}}\left(l_i+(u_i-l_i)z_i^j\right)\le u_j.\end{aligned}$$

Since $l_i=\bar{l}_i/q, u_i=\bar{u}_i/q$ for each $i\in[N]$, then for each $j\in[N]$, the expression $-\sum_{i\in[N]\setminus\{j\}}(l_i+(u_i-l_i)z_i^j)$ can take values from $\{\tau/q\}_{\tau\in[\bar{l}_j,\bar{u}_j]}$ and τ is an integer. This fact allows us to simplify $v_j^*=\min_{\tau\in[\bar{l}_j,\bar{u}_j]}v_{j\tau}^*$, where

$$v_{j\tau}^*=\min_{\substack{\boldsymbol{x}\in\mathcal{X},\boldsymbol{\nu},\\ \boldsymbol{z}^j\in\{0,1\}^N}}\boldsymbol{c}^\top\boldsymbol{x}+\sum_{i\in[N]\setminus\{j\}}\left[p_i^0+\frac{\bar{l}_i}{q}+\frac{1}{q}(\bar{u}_i-\bar{l}_i)z_i^j\right]\nu_i+\left(p_j^0+\frac{\tau}{q}\right)\nu_j,$$

$$\text{s.t.} \quad -\sum_{i\in[N]\setminus\{j\}}\left(\bar{l}_i+(\bar{u}_i-\bar{l}_i)z_i^j\right)=\tau,\nu_i\geq Q(\boldsymbol{x},\boldsymbol{\xi}^i),\forall i\in[N],$$

for each $j\in[N]$ and $\tau\in[\bar{l}_j,\bar{u}_j]$. Since the set $\mathcal{X}$ is compact, we can apply the McCormick inequalities [49] to linearize the bilinear terms $\{z_i^j\nu_i\}_{i\in[N],j\in[N]}$, this completes the proof. □

We remark that as a direct application of Corollary 1, when $l_i=l,u_i=u$ for each $i\in[N]$, the MICP-R formulation of Corollary 1 can be further simplified.

Corollary 2 *Suppose that the premises of Corollary 1 hold and $p_i^0=1/N,l_i=-1/N,u_i=u$ for each $i\in[N]$. Then the optimal value of the corresponding DFO (2) is $v^*=\min_{j\in[N]}v_j^*$ and for each $j\in[N]$, the value v_j^* can be computed via the following MICP-R formulation:*

$$\begin{aligned}v_j^*=\min_{\substack{\boldsymbol{x}\in\mathcal{X},\boldsymbol{\eta},\boldsymbol{\nu},\\ \boldsymbol{z}^j\in\{0,1\}^N}}\ &\boldsymbol{c}^\top\boldsymbol{x}+\sum_{i\in[N]\setminus\{j\}}(u+1/N)\nu_i+(1-\lfloor\kappa\rfloor/\kappa)\nu_j,\\ \text{s.t.}\ &\nu_i\geq Q(\boldsymbol{x},\boldsymbol{\xi}^i),\left(\eta_i^j,z_i^j,\nu_i\right)\in\mathcal{MI}\left(0,1,L_i,U_i\right),\forall i\in[N],j\in[N],\\ &\sum_{i\in[N]\setminus\{j\}}z_i^j=\lfloor\kappa\rfloor,\end{aligned}$$

where $\kappa=N/(uN+1)$ and for each $i\in[N]$, L_i and U_i are the lower and upper bounds of the function $Q(\boldsymbol{x},\boldsymbol{\xi}^i)$, respectively.

This result will be demonstrated in our numerical study section. However, we notice that the MICP-R result in Theorem 1 does not hold when the ambiguity set with finite support is not polyhedral. Below is an illustration.

Proposition 3 *Suppose that the ambiguity set is $\mathcal{P}=\{\boldsymbol{p}\colon\|\boldsymbol{p}-\boldsymbol{p}^0\|_2\leq\theta,\sum_{i\in[N]}p_i=1,\boldsymbol{p}\geq\boldsymbol{0}\}$, where $\boldsymbol{p}^0=\boldsymbol{e}/N$ denotes the nominal probability. When $0<\theta\leq\sqrt{1/(N(N-1))}$, DFO (2) may not be MICP-R.*

Proof: See Appendix B.3. □

This result shows that even for DFO (2) with a finite-support ambiguity set, it may not be MICP-R. Hence, while interesting, it is worthwhile to explore the effective ambiguity set for DFO when performing data-driven decision-making. Due to page limit, we refer the interested readers to Appendix D for more complexity results on DFO, where we follow the discussions in [6, 20, 71] and focus on the recourse function $Q(\boldsymbol{x},\boldsymbol{\xi})$ being convex or concave piecewise affine in $\boldsymbol{x}$, respectively. Besides the tractability and MICP-R, we will show another criterion to select a proper ambiguity set in the next section.

4 Decision Outlier Robustness and Data-driven Stochastic Programs with Endogenous Outliers

The equivalence between DFO (2) and many robust statistics in Section 2 demonstrates that DFO can be useful to mitigate the effect of endogenous outliers. More importantly, our proposed DFO can be optimization-driven and thus is applicable to a data-driven stochastic program with endogenous outliers. This is a different approach from existing ones such as the recent exciting progress made in [21], which focused on using the contextual information to process the contaminated data. It turns out that their framework is equivalent to applying the worst-case CVaR under a general Wasserstein ambiguity set (see Appendix F for a detailed comparison) and thus is quite different from ours. We believe that our results can be further strengthened when contextual information is available, which leaves as a future study due to page limit. In a two-stage stochastic program without relative complete recourse, endogenous outliers can cause the underlying problem to be infeasible. We show that DFO serves as a proper measure to address the infeasibility, reduce the effects of endogenous outliers, and deliver desirable decisions. It is worthy of mentioning that our DFO framework does not remove the endogenous outliers but changes their probability measures to ensure that the corresponding objective function is finite. We provide an example as below.

Example 1 Consider the following two-stage stochastic program:

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

where we assume that there are 5 equiprobable scenarios (i.e., $\mathbb{P}\{\tilde{\xi}=\xi^i\}=1/5$ for all $i\in[5]$), and $\xi^1=-5,\xi^2=-1,\xi^3=0,\xi^4=1,\xi^5=5$, and set $\mathcal{Y}=\{y:0\leq y\leq 1\}$. Under this setting, the two-stage stochastic program is equivalent to

$$\min_{\substack{x\geq 1,\\ \boldsymbol{y}\in[0,1]^5}}\left\{x+\frac{1}{5}\sum_{i\in[5]}y_i : 5y_1\geq x, y_2\geq x, 0y_3\geq x, y_4\geq x, 5y_5\geq x\right\}.$$

In this example, for any first-stage decision $x\geq 1$, there exists a scenario that the second-stage problem is infeasible, i.e., the recourse function is infinity. That is, the entire two-stage stochastic program is infeasible, and the decisions are not comparable.

If the machine learning techniques were employed to preprocess the data $\{\xi^i\}_{i\in[5]}$ to resolve the infeasibility, they would probably choose to remove scenario ξ^1 or ξ^5, since these two scenarios are relatively far away from others based on the Euclidean distance. However, the problem remains infeasible, and the actual endogenous outlier ξ^3 may not be removed unless exploring the optimization problem structure. Thus, the outlier scenario ξ^3 is rather endogenous. On the other hand, applying DFO can properly mitigate the effect of the endogenous outliers and address the infeasibility issue using the interval ambiguity set (i.e., $\mathcal{P}_I=\{\boldsymbol{\mu}:\boldsymbol{\mu}(\mathcal{U})=1,0\preceq\boldsymbol{\mu}\preceq 5/4\mathbb{P}\}$). Thus, let us consider the following DFO

$$v^*=\min_{x\geq 1}\left\{x+\inf_{\mathbb{P}\in\mathcal{P}_I}\mathbb{E}_{\mathbb{P}}\left[Q(x,\tilde{\xi}):=\min_{y\in\mathcal{Y}}\left\{y\colon |\tilde{\xi}|y\geq x\right\}\right]\right\},$$

which is equivalent to

$$v^*=\min_{\substack{x\geq 1,\boldsymbol{y}\in[0,1]^5,\\ \boldsymbol{z}\in\{0,1\}^5}}\left\{x+\frac{\sum_{i\in[5]}y_iz_i}{\sum_{i\in[5]}z_i}\colon 5y_1\geq xz_1, y_2\geq xz_2, 0y_3\geq xz_3, y_4\geq xz_4, 5y_5\geq xz_5, \sum_{i\in[5]}z_i\geq 4\right\}=\frac{8}{5}.$$

This demonstrates that DFO can effectively mitigate the effect of endogenous outliers. ◇

As demonstrated in Example 1, we focus on endogenous outliers, especially when the machine learning-based preprocessing methods (e.g., outlier detection methods) may fail.

4.1 Decision Outlier Robustness

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 For an unamenable decision $\boldsymbol{x}\in\mathcal{X}$, there exists an outlier scenario $\boldsymbol{\xi}^o\in\mathcal{U}$ such that the recourse function $Q(\boldsymbol{x},\boldsymbol{\xi}^o)=+\infty$. The collection of such unamenable decisions is denoted by set $\widehat{\mathcal{X}}$.

Now we are ready to introduce the notion of "*decision outlier robust*," which mainly focuses on unamenable decisions. In this section, we mainly focus on stochastic programs with unamenable decisions.

Definition 4 For any $\lambda\in[0,1]$, the DFO (2) is "decision outlier robust" if the following condition is satisfied:

$$\inf_{\mathbb{P}\in\mathcal{P}}(1-\lambda)\mathbb{E}_{\mathbb{P}}\left[Q(\boldsymbol{x},\tilde{\boldsymbol{\xi}})\right]+\lambda Q(\boldsymbol{x},\boldsymbol{\xi}^o)\mathbb{I}\left(\boldsymbol{\xi}^o\in\operatorname{supp}\{\mathbb{P}\}\right)<\infty, \tag{8a}$$

for each unamenable decision $\boldsymbol{x}\in\widehat{\mathcal{X}}$ and each scenario $\boldsymbol{\xi}^o\in\mathcal{U}$. Here, we let $\infty\times 0=0$.

Note that condition (8a) can also be equivalently written as

$$\inf_{\mathbb{P}\in\mathcal{P}} (1-\lambda)\mathbb{E}_{\mathbb{P}}\left[Q(\boldsymbol{x},\tilde{\boldsymbol{\xi}})\right] + \lambda \text{ess.sup}_{\mathbb{P}}\left\{Q(\boldsymbol{x},\tilde{\boldsymbol{\xi}})\right\} < \infty, \tag{8b}$$

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 (8b) is bounded. We make the following remarks about Definition 4.

(i) In Definition 4, for the DFO (2) to be decision outlier robust, there exists a probability measure $\mathbb{P} \in \mathcal{P}$ such that an unamenable decision with respect to any mixture distribution of $\mathbb{P}$ and a Dirac measure on an outlier scenario $\boldsymbol{\xi}^o \in \mathcal{U}$ yields a bounded objective function value. This should hold for any unamenable decision $\boldsymbol{x} \in \widehat{\mathcal{X}}$.

(ii) This definition is a generalization of outlier robustness in robust statistics, which focuses on the influence curve of an outlier (See Appendix E for a formal definition). The notion of the influence curve has the following 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 7 of Appendix E); (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); and (iii) in many decision-making problems, the objective function may not be necessarily smooth (e.g., two-stage stochastic integer programming in [2]). Thus, the influence curve is not appropriate to analyze the decision outlier robustness of DFO.

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

(iv) Another advantage of using decision outlier robustness is to verify the appropriateness of a DFO ambiguity set (see Section 4.2 for more details).

(V) Although it may require the unnameable decision set beforehand, in practice, one can simply check all the decisions. Besides, the results in Proposition 4 can further help simplify the verification process.

Proposition 4 *The following statements must hold:*

(i) The DFO (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 $\mathbb{P}\{\tilde{\boldsymbol{\xi}} : Q(\boldsymbol{x},\tilde{\boldsymbol{\xi}}) = \infty\} = 0$; and*

(ii) The DFO (2) *is not decision outlier robust if there exists an unamenable decision $\boldsymbol{x} \in \widehat{\mathcal{X}}$ with its outlier scenario $\boldsymbol{\xi}^o$ such that $Q(\boldsymbol{x},\boldsymbol{\xi}^o) = \infty$ and for any probability measure $\mathbb{P} \in \mathcal{P}$, we have $\boldsymbol{\xi}^o \in \text{supp}(\mathbb{P})$.*

Proof: We split the proof into two parts by checking each condition, separately.

(i) According to the definition of the unamenable decision $\boldsymbol{x} \in \widehat{\mathcal{X}}$, there exists an outlier scenario $\boldsymbol{\xi}^o \in \mathcal{U}$ such that the recourse function $Q(\boldsymbol{x},\boldsymbol{\xi}^o) = +\infty$. Based on the assumptions in Proposition 4, there exists a probability measure $\mathbb{P} \in \mathcal{P}$ such that $\mathbb{E}_{\mathbb{P}}[Q(\boldsymbol{x},\tilde{\boldsymbol{\xi}})] < \infty$ and $\mathbb{P}\{\tilde{\boldsymbol{\xi}} : Q(\boldsymbol{x},\boldsymbol{\xi}^o) = \infty\} = 0$. Then, for any $\lambda \in [0,1]$, we have

$$(1-\lambda)\mathbb{E}_{\mathbb{P}}\left[Q(\boldsymbol{x},\tilde{\boldsymbol{\xi}})\right] + \lambda Q(\boldsymbol{x},\boldsymbol{\xi}^o)\mathbb{I}\left(\boldsymbol{\xi}^o \in \text{supp}\{\mathbb{P}\}\right) < \infty,$$

which implies that condition (8a) is satisfied and DFO (2) is decision outlier robust.

(ii) For any unamenable decision $\boldsymbol{x} \in \widehat{\mathcal{X}}$, there exists one outlier scenario $\boldsymbol{\xi}^o$ such that $Q(\boldsymbol{x},\boldsymbol{\xi}^o) = \infty$ with $\boldsymbol{\xi}^o \in \text{supp}(\mathbb{P})$ for any probability measure $\mathbb{P} \in \mathcal{P}$. Then, for any $\lambda \in [0,1]$ and $\mathbb{P} \in \mathcal{P}$, we have

$$(1-\lambda)\mathbb{E}_{\mathbb{P}}\left[Q(\boldsymbol{x},\tilde{\boldsymbol{\xi}})\right] + \lambda Q(\boldsymbol{x},\boldsymbol{\xi}^o)\mathbb{I}\left(\boldsymbol{\xi}^o \in \text{supp}\{\mathbb{P}\}\right) = \infty,$$

which implies that condition (8a) is violated and DFO (2) is not decision outlier robust.

□

Using Proposition 4, we can immediately demonstrate that the expectation operator with a singleton ambiguity set $\mathcal{P}$ is not decision outlier robust.

Corollary 3 *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 relative complete recourse, is not decision outlier robust.*

Proof: Suppose that $\mathcal{P} = \{\mathbb{P}\}$. Since there exists an unamenable decision $\boldsymbol{x} \in \mathcal{X}$, according to Definition 3, there exists an outlier scenario $\boldsymbol{\xi}^o$ such that $Q(\boldsymbol{x}, \boldsymbol{\xi}^o) = \infty$ and $\boldsymbol{\xi}^o \in \operatorname{supp}(\mathbb{P})$. Using Proposition 4, we know that the corresponding DFO is not decision outlier robust. □

Therefore, without relative 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 1). A richer and nontrivial ambiguity set is more desirable and is demonstrated in the following subsections.

4.2 DFO Mitigates the Effect of Outliers for the Optimization with Constraint Uncertainty

Some outlier scenarios can cause the problem infeasible in the robust optimization (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] suggested using the chance constrained programming as a better alternative. The rationality can be well justified from the DFO perspective. In the DFO (2), if the objective of the recourse function is 0 with the uncertain inequalities $G(\boldsymbol{x}, \boldsymbol{\xi}) \leq 0$, where $G(\cdot, \cdot) : \mathbb{R}^n \times \mathbb{R}^m \to \mathbb{R}$ is a continuous function, i.e., $Q(\boldsymbol{x}, \tilde{\boldsymbol{\xi}}) = \min\{0 \colon G(\boldsymbol{x}, \tilde{\boldsymbol{\xi}}) \leq 0\}$, then the corresponding DFO (2) resorts to

$$v^* = \min_{\boldsymbol{x} \in \mathcal{X}} \left\{ \boldsymbol{c}^\top \boldsymbol{x} \colon G(\boldsymbol{x}, \boldsymbol{\xi}) \leq 0, \forall \boldsymbol{\xi} \in \mathcal{U} \right\} = \min_{\boldsymbol{x} \in \mathcal{X}} \left\{ \boldsymbol{c}^\top \boldsymbol{x} \colon \mathbb{E}_{\mathbb{P}} \left[\mathbb{1} \left(G(\boldsymbol{x}, \tilde{\boldsymbol{\xi}}) > 0 \right) \right] \leq 0 \right\}. \tag{9a}$$

where support $\mathcal{U} := \operatorname{supp}(\mathbb{P})$. This is indeed a conventional robust optimization problem. Following the similar interval ambiguity set of the LTS example in Section 2.2 and Example 1, i.e., $\mathcal{P}_I = \{\boldsymbol{\mu} : \boldsymbol{\mu}(\mathcal{U}) = 1, 0 \preceq \boldsymbol{\mu} \preceq \mathbb{P}/(1-\varepsilon)\}$, the DFO counterpart of the robust optimization (9a) can be written as

$$v^* = \min_{\boldsymbol{x} \in \mathcal{X}} \left\{ \boldsymbol{c}^\top \boldsymbol{x} \colon \inf_{\mathbb{P} \in \mathcal{P}_I} \mathbb{E}_{\mathbb{P}} \left[\mathbb{1} \left(G(\boldsymbol{x}, \tilde{\boldsymbol{\xi}}) > 0 \right) \right] \leq 0 \right\}, \tag{9b}$$

and can be further reduced to a regular chance constrained program.

Proposition 5 *Suppose the interval ambiguity set is $\mathcal{P}_I = \{\boldsymbol{\mu} : \boldsymbol{\mu}(\mathcal{U}) = 1, 0 \preceq \boldsymbol{\mu} \preceq \mathbb{P}/(1-\varepsilon)\}$, then the DFO counterpart of a robust optimization* (9a) *is equivalent to a chance constrained program*

$$v^* = \min_{\boldsymbol{x} \in \mathcal{X}} \left\{ \boldsymbol{c}^\top \boldsymbol{x} \colon \mathbb{E}_{\mathbb{P}} \left[\mathbb{1} \left(G(\boldsymbol{x}, \tilde{\boldsymbol{\xi}}) > 0 \right) \right] \leq \varepsilon \right\}. \tag{10}$$

Proof: According to the duality result in [62], we have

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

Since

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

we further have

$$\max_{\lambda_0} F(\boldsymbol{x}, \lambda_0) = \max \left\{ \max_{\lambda_0 \leq 0} F(\boldsymbol{x}, \lambda_0), \max_{0 < \lambda_0 < 1} F(\boldsymbol{x}, \lambda_0), \max_{\lambda_0 \geq 1} F(\boldsymbol{x}, \lambda_0) \right\}$$

$$= \max\left\{0, -\varepsilon + \mathbb{E}_{\mathbb{P}}\left[\mathbb{1}\left(G(\boldsymbol{x}, \tilde{\boldsymbol{\xi}}) > 0\right)\right]\right\}.$$

Therefore, the conclusion follows by substituting the last equality into the left-hand side of the constraint in DFO (9a). □

Proposition 5 shows that by selecting the favorable scenarios, applying the DFO framework reduces the over-conservatism of robust optimization and explains why a chance constrained program can be less conservative. Moreover, we show that the DFO framework (9b) (i.e., the corresponding chance constrained program) is decision outlier robust. In contrast, the robust optimization framework (9a) may not be when there are unamenable decisions.

Theorem 2 *Suppose that the unamenable decision set $\widehat{\mathcal{X}}$ is non-empty and for any $\boldsymbol{x} \in \widehat{\mathcal{X}}$, we have $\mathbb{P}\{\tilde{\boldsymbol{\xi}} : G(\boldsymbol{x}, \tilde{\boldsymbol{\xi}}) > 0\} \leq \varepsilon$, where $\mathbb{P}$ denotes the reference distribution. Then, the DFO* (9b) *is decision outlier robust, while the robust optimization* (9a) *is not.*

Proof: We split the proof into two parts by checking the DFO (9b) and the robust optimization framework (9a) separately.

Part I. According to Proposition 4, for the DFO framework (9b), 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 $\mathbb{E}_{\mathbb{P}^*}[\mathbb{1}(G(\boldsymbol{x}, \tilde{\boldsymbol{\xi}}) > 0)] \leq 0$ 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 $\mathbb{P}^*(d\boldsymbol{\xi}) = \mathbb{P}(d\boldsymbol{\xi})/\mathbb{P}\{\mathcal{U}_1\}$ for each $\boldsymbol{\xi} \in \mathcal{U}_1$, 0, otherwise. Note that by our construction, we have $\mathbb{P}^*(\mathcal{U}_1) = 1, 0 \preceq \mathbb{P}^* \preceq \mathbb{P}/(1-\varepsilon)$. Hence, $\mathbb{P}^* \in \mathcal{P}_I$ and $\mathbb{P}^*\{\tilde{\boldsymbol{\xi}} : \tilde{\boldsymbol{\xi}} = \boldsymbol{\xi}^o\} = 0$. On the other hand, we have

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

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

Part II. For the robust optimization (9a), we have $\mathcal{P} = \{\mathbb{P}\}$. According to Proposition 4, it is sufficient to show that $G(\boldsymbol{x}, \boldsymbol{\xi}^o) > 0$ for some $\boldsymbol{x} \in \widehat{\mathcal{X}}$ and $\boldsymbol{\xi}^o \in \mathrm{supp}(\mathbb{P})$, which holds due to our preassumption. This proves that the robust optimization framework (9a) may not be decision outlier robust. □

We make the following remarks on Theorem 2:

(i) The result of Theorem 2 implies that the value-of-risk (VaR) can also be decision outlier robust, where for a given risk level ε, $(1-\varepsilon)$-VaR of a random variable $\tilde{\boldsymbol{X}}$ is defined as $\mathrm{VaR}_{1-\varepsilon}(\tilde{\boldsymbol{X}}) := \min_s\{s : F_{\tilde{\boldsymbol{X}}}(s) \geq 1-\varepsilon\}$, where $\mathbb{P}$ and $F_{\tilde{\boldsymbol{X}}}(\cdot)$ be its probability distribution and cumulative distribution function. Moreover, letting $\varepsilon = 1/2$ in (10) shows that the median is also decision outlier robust;

(ii) For general quantiles, the notion of "outlier robust" based on influence curve from statistical robustness (See the definition in Appendix E) may not work. We refer the interested readers to Example 7 of Appendix E for an explanation.

4.3 DFO Mitigates the Effect of Outliers for Stochastic Programming

For a stochastic program with outlier scenarios (e.g., there exist endogenous outliers), this subsection focuses on a special family of DFO– 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 $\mathbb{P}$, cumulative distribution function $F_{\tilde{\boldsymbol{X}}}(\cdot)$, and risk level $\varepsilon \in (0,1)$, the FCVaR of $\tilde{\boldsymbol{X}}$ is

$$\mathrm{FCVaR}_{1-\varepsilon}(\tilde{\boldsymbol{X}}) := \max_{\beta}\left\{\beta + \frac{1}{1-\varepsilon}\mathbb{E}_{\mathbb{P}}\left[\tilde{\boldsymbol{X}} - \beta\right]_-\right\}. \tag{12}$$

Roughly speaking, FCVaR (12) can be interpreted as the average of the values no larger than the $(1-\varepsilon)$-VaR.

Proposition 6 *(i) Given an interval ambiguity set* $\mathcal{P}_I = \{\boldsymbol{\mu} : \boldsymbol{\mu}(\mathcal{U}) = 1, 0 \preceq \boldsymbol{\mu} \preceq \mathbb{P}/(1-\varepsilon)\}$ *with support* $\mathcal{U} = \mathrm{supp}(\mathbb{P})$*, we have*

$$\inf_{\boldsymbol{\mu}\in\mathcal{P}_I} \mathbb{E}_{\boldsymbol{\mu}}\left[\tilde{\boldsymbol{X}}\right] = \max_{\beta}\left\{\beta + \frac{1}{1-\varepsilon}\mathbb{E}_{\mathbb{P}}[\tilde{\boldsymbol{X}}-\beta]_-\right\} = \mathrm{FCVaR}_{1-\varepsilon}\left(\tilde{\boldsymbol{X}}\right); \tag{13a}$$

(ii) An optimal solution of the right-hand side optimization problem (12) *is* $\beta^* = \mathrm{VaR}_{1-\varepsilon}(\tilde{\boldsymbol{X}})$*; and*
(iii) The $\mathrm{FCVaR}_{1-\varepsilon}(\tilde{\boldsymbol{X}})$ *can be bounded by two conditional expectations:*

$$\mathbb{E}_{\mathbb{P}}\left[\tilde{\boldsymbol{X}}\middle|\tilde{\boldsymbol{X}} < \mathrm{VaR}_{1-\varepsilon}\left(\tilde{\boldsymbol{X}}\right)\right] \le \mathrm{FCVaR}_{1-\varepsilon}\left(\tilde{\boldsymbol{X}}\right) \le \mathbb{E}_{\mathbb{P}}\left[\tilde{\boldsymbol{X}}\middle|\tilde{\boldsymbol{X}} \le \mathrm{VaR}_{1-\varepsilon}\left(\tilde{\boldsymbol{X}}\right)\right]. \tag{13b}$$

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 5 and thus is omitted.

(ii) Since the right-hand side optimization problem (12) is an unconstrained concave minimization, let us consider the first-order condition of FCVaR (12) for an optimal solution β^*, that is,

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

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,

$$0 = 1 + \frac{1}{1-\varepsilon}\mathbb{E}_{\mathbb{P}}\left[\partial_\beta\left[\tilde{\boldsymbol{X}}-\beta\right]_-\Big|_{\beta=\beta^*}\right] = 1 - \frac{1}{1-\varepsilon}\mathbb{P}\left\{\tilde{\boldsymbol{X}} < \beta^*\right\} - \frac{\omega}{1-\varepsilon}\mathbb{P}\left\{\tilde{\boldsymbol{X}} = \beta^*\right\}, \tag{14}$$

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

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

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

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

Thus, all inequalities become equalities. Letting $\omega = 1$ in the optimality condition (14), we have

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

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

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

$$\mathbb{E}_{\mathbb{P}}\left[\tilde{\boldsymbol{X}}\middle|\tilde{\boldsymbol{X}} < \mathrm{VaR}_{1-\varepsilon}\left(\tilde{\boldsymbol{X}}\right)\right] = \mathrm{VaR}_{1-\varepsilon}(\tilde{\boldsymbol{X}}) + \frac{\mathbb{E}_{\mathbb{P}}\left[\left(\tilde{\boldsymbol{X}} - \mathrm{VaR}_{1-\varepsilon}(\tilde{\boldsymbol{X}})\right)\mathbb{1}\{\tilde{\boldsymbol{X}} < \mathrm{VaR}_{1-\varepsilon}(\tilde{\boldsymbol{X}})\}\right]}{\mathbb{P}\left\{\tilde{\boldsymbol{X}} < \mathrm{VaR}_{1-\varepsilon}(\tilde{\boldsymbol{X}})\right\}}.$$

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

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

Thus, the lower bound is valid.

Similarly, we can write the upper bound as

$$\mathbb{E}_{\mathbb{P}}\left[\tilde{\boldsymbol{X}}\middle|\tilde{\boldsymbol{X}}\leq \mathrm{VaR}_{1-\varepsilon}\left(\tilde{\boldsymbol{X}}\right)\right]=\mathrm{VaR}_{1-\varepsilon}(\tilde{\boldsymbol{X}})+\frac{\mathbb{E}_{\mathbb{P}}\left[\left(\tilde{\boldsymbol{X}}-\mathrm{VaR}_{1-\varepsilon}(\tilde{\boldsymbol{X}})\right)\mathbb{1}\{\tilde{\boldsymbol{X}}\leq \mathrm{VaR}_{1-\varepsilon}(\tilde{\boldsymbol{X}})\}\right]}{\mathbb{P}\left\{\tilde{\boldsymbol{X}}\leq \mathrm{VaR}_{1-\varepsilon}(\tilde{\boldsymbol{X}})\right\}}.$$

Since $\mathbb{P}\{\tilde{\boldsymbol{X}}\leq \mathrm{VaR}_{1-\varepsilon}(\tilde{\boldsymbol{X}})\}\geq 1-\varepsilon$ and $\mathbb{E}_{\mathbb{P}}[(\tilde{\boldsymbol{X}}-\mathrm{VaR}_{1-\varepsilon}(\tilde{\boldsymbol{X}}))\mathbb{1}\{\tilde{\boldsymbol{X}}\leq \mathrm{VaR}_{1-\varepsilon}(\tilde{\boldsymbol{X}})\}]=\mathbb{E}_{\mathbb{P}}[\min\{\tilde{\boldsymbol{X}}-\mathrm{VaR}_{1-\varepsilon}(\tilde{\boldsymbol{X}}),0\}]$, we have

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

This completes the proof. □

The equivalence (13a) shows that FCVaR (12) can be a special case of DFO (2). That is, letting $\tilde{\boldsymbol{X}}:=Q(\boldsymbol{x},\tilde{\boldsymbol{\xi}})$ and choosing the same interval ambiguity set as Proposition 6, DFO (2) reduces to the following FCVaR optimization

$$v^*=\min_{\boldsymbol{x}\in\mathcal{X}}\inf_{\mathbb{P}\in\mathcal{P}_I}\mathbb{E}_{\mathbb{P}}\left[Q(\boldsymbol{x},\tilde{\boldsymbol{\xi}})\right]=\min_{\boldsymbol{x}\in\mathcal{X}}\mathrm{FCVaR}_{1-\varepsilon}\left[Q(\boldsymbol{x},\tilde{\boldsymbol{\xi}})\right]. \tag{15}$$

We remark that the LTS introduced in Section 2.2 can be viewed as a special case of FCVaR (15). That is, suppose that the random vector $\tilde{\boldsymbol{\xi}}$ has an equiprobable distribution over a finite support $\mathcal{U}=\{\boldsymbol{\xi}^i\}_{i\in[N]}=\{\bar{\boldsymbol{x}}_i,y_i\}_{i\in[N]}\subseteq\mathbb{R}^d\times\mathbb{R}$. Let $\varepsilon=(N-k)/N$ with an integer $k\in[N]$ and the recourse function be $Q(\boldsymbol{x},\boldsymbol{\xi}^i)=(y_i-\bar{\boldsymbol{x}}_i^\top\boldsymbol{x})^2$ for each $i\in[N]$. Then the interval ambiguity set in Proposition 6 reduces to $\mathcal{P}_I=\{\boldsymbol{p}\in\mathbb{R}_+^N:\sum_{i\in[N]}p_i=1,0\leq p_i\leq 1/k\}$ and DFO (15) reduces to LTS (6).

Interestingly, if one replaces the inner infimum operator with the supremum operator on the left-hand side (15), then the left-hand side reduces to the CVaR minimization problem, a well-known DRO model, i.e.,

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

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. For example, instead of using LTS, we can use winsorized approach, e.g., replacing the recourse function values of unfavorable scenarios with the $(1-\epsilon)$-quantile $\mathrm{VaR}_{1-\varepsilon}(\cdot)$. 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:

$$\mathrm{WCVaR}_{1-\varepsilon}(\tilde{\boldsymbol{X}}):=(1-\varepsilon)\mathrm{FCVaR}_{1-\varepsilon}\left(\tilde{\boldsymbol{X}}\right)+\varepsilon\mathrm{VaR}_{1-\varepsilon}\left(\tilde{\boldsymbol{X}}\right), \tag{16}$$

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 $\mathrm{VaR}_{1-\varepsilon}(\tilde{\boldsymbol{X}})$, 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.,

$$\mathrm{HCVaR}(\tilde{\boldsymbol{X}}, H) \colon = \mathbb{E}_{\mathbb{P}}\left[\min\left\{\tilde{\boldsymbol{X}}, H\right\}\right]. \tag{17}$$

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 $\mathrm{VaR}_{1-\varepsilon}(\cdot)$, then HCVaR (17) and WCVaR (16) coincide. We use the following example to illustrate HCVaR.

Example 2 In Example 1, we can also resolve the infeasibility issue by applying the Huber-skip CVaR, which admits the following equivalent reformulation:

$$\min_{\substack{x\geq 1, \boldsymbol{y}\in[0,1]^5,\\ \boldsymbol{z}\in\{0,1\}^5}} \left\{ x + \frac{1}{5}\sum_{i\in[5]}[z_i y_i + (1-z_i)H] : 5y_1 \geq xz_1, y_2 \geq xz_2, 0y_3 \geq xz_3, y_4 \geq xz_4, 5y_5 \geq xz_5 \right\},$$

For instance, letting $H = 10$, this problem can be solved as $z_1 = z_2 = z_4 = z_5 = 1$ and $z_3 = 0$ with the objective value $87/25$. Here the value of H can be interpreted as the emergency cost in the two-stage stochastic programming literature (see, e.g., [18, 43]). ◇

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 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 with 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 $\mathrm{FCVaR}_{1-\varepsilon}$, $\mathrm{WCVaR}_{1-\varepsilon}$, $\mathrm{VaR}_{1-\varepsilon}$, and $\mathrm{HCVaR}(\cdot, H)$ with $H = 3$, while the expectation and $\mathrm{CVaR}_{1-\varepsilon}$ do not exist. Please see Figure 2 for an illustration. ◇

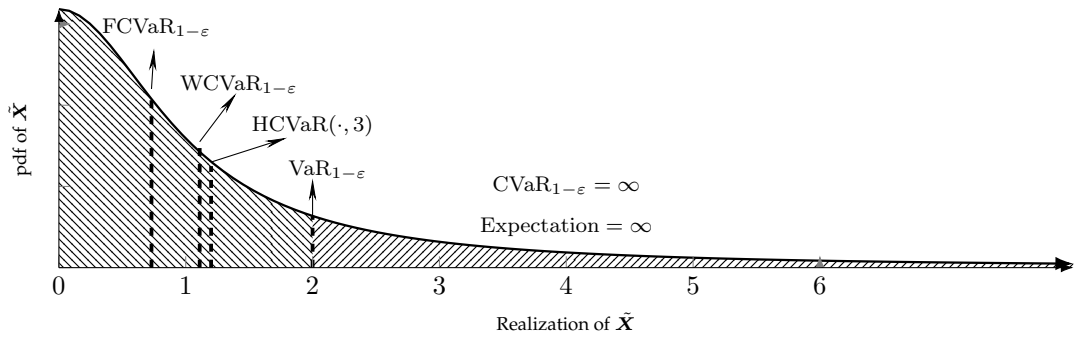


Fig. 2: Illustration of Expectation, FCVaR, WCVaR, HCVaR, VaR, and CVaR with Truncated Cauchy Distribution.

Example 4 Let us assume $\tilde{\boldsymbol{X}}$ to be truncated Gaussian distribution with 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, $\mathrm{CVaR}_{1-\varepsilon}$, $\mathrm{FCVaR}_{1-\varepsilon}$, $\mathrm{WCVaR}_{1-\varepsilon}$, $\mathrm{VaR}_{1-\varepsilon}$, and $\mathrm{HCVaR}(\cdot, H)$ with $H = 2$, which are illustrated in Figure 3. ◇

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

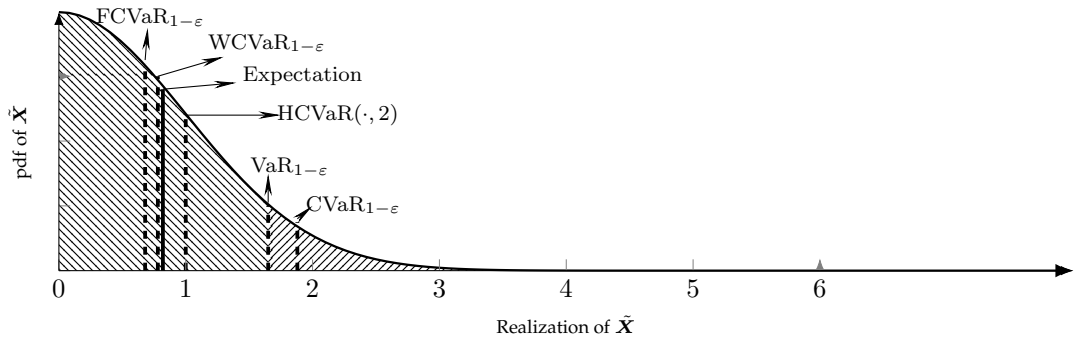


Fig. 3: Illustration of Expectation (solid line), FCVaR, WCVaR, HCVaR, VaR, and CVaR with Truncated Gaussian Distribution.

Theorem 3 *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 $\mathbb{P}\{\tilde{\boldsymbol{\xi}} : Q(\boldsymbol{x}, \tilde{\boldsymbol{\xi}}) > M\} \leq \varepsilon$, where $\mathbb{P}$ denotes the reference distribution. Then $\mathrm{FCVaR}_{1-\varepsilon}\{Q(\boldsymbol{x}, \tilde{\boldsymbol{\xi}})\}$ is decision outlier robust.*

Proof: Based on Proposition 4, for $\mathrm{FCVaR}_{1-\varepsilon}\{Q(\boldsymbol{x}, \tilde{\boldsymbol{\xi}})\}$ defined in (13a), 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 $\mathcal{U}_1 = \{\tilde{\boldsymbol{\xi}} : Q(\boldsymbol{x}, \tilde{\boldsymbol{\xi}}) \leq M\}$, which is $\mathbb{P}$-measurable (see, e.g., proposition 1 in section 3.1 of [59]). Given the presumption, we have $\mathbb{P}\{\mathcal{U}_1\} \geq 1 - \varepsilon$. Let us construct $\mathbb{P}^*(d\boldsymbol{\xi}) = \mathbb{P}(d\boldsymbol{\xi})/\mathbb{P}\{\mathcal{U}_1\}$ for each $\boldsymbol{\xi} \in \mathcal{U}_1$, 0, otherwise. Note that by our construction, we have $\mathbb{P}^*(\mathcal{U}_1) = 1, 0 \preceq \mathbb{P}^* \preceq \mathbb{P}/(1-\varepsilon)$ and hence $\mathbb{P}^* \in \mathcal{P}_I$. On the other hand, we also have

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

This proves that $\mathbb{P}^*$ is a desirable probability measure. Hence, $\mathrm{FCVaR}_{1-\varepsilon}\{Q(\boldsymbol{x}, \tilde{\boldsymbol{\xi}})\}$ is decision outlier robust. □

We make the following remarks about Theorem 3:

(i) The assumption that $\mathbb{P}\{\tilde{\boldsymbol{\xi}} : Q(\boldsymbol{x}, \tilde{\boldsymbol{\xi}}) > M\} \leq \varepsilon$ 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 (10), when the reference distribution is discrete, outlier robustness using the influence curve may not work. We refer the readers to Example 8 of Appendix E for an explanation.

We conclude this section by remarking that the result in Theorem 3 can be extended to Winsorized CVaR and Huber-skip CVaR. The proofs are similar and thus are omitted.

Corollary 4 *Suppose that the unamenable decision set $\widehat{\mathcal{X}}$ is non-empty. Then*

(i) the $\mathrm{WCVaR}_{1-\varepsilon}\{Q(\boldsymbol{x}, \tilde{\boldsymbol{\xi}})\}$ is decision outlier robust if for any $\boldsymbol{x} \in \widehat{\mathcal{X}}$, there exists an $M \in \mathbb{R}$ such that $\mathbb{P}\{\tilde{\boldsymbol{\xi}} : Q(\boldsymbol{x}, \tilde{\boldsymbol{\xi}}) > M\} \leq \varepsilon$; and
(ii) the Huber-skip $\mathrm{HCVaR}\{Q(\boldsymbol{x}, \tilde{\boldsymbol{\xi}}), H\}$ is outlier robust.

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

5 Achieving Out-of-Sample Performances: Worst-case DFO

In this section, we study the DFO models in Section 4.2 when the reference distribution is not finitely supported. To effectively use i.i.d. samples to approximate the DFO models and achieve better out-of-sample performances, 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 4.2 (i.e., a chance constrained program), its worst-case counterpart, known as distributionally robust chance constrained programs (DRCCP), has been studied in the literature (see more discussions in [15, 28, 37, 63, 72]). 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. Particularly, we study the worst-case FCVaR of the form

$$v_W^* = \min_{\boldsymbol{x}\in\mathcal{X}} \sup_{\mathbb{P}\in\mathcal{P}_\infty^W} \mathrm{FCVaR}_{1-\varepsilon}\left[Q(\boldsymbol{x},\tilde{\boldsymbol{\xi}})\right]. \tag{18}$$

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

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

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

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

5.1 Worst-case FCVaR is Equivalent to DRO with Favorable Sample-selection

We first show that the worst-case FCVaR (18) admits a neat representation.

Theorem 4 *The worst-case* FCVaR (18) *is equivalent to*

$$v_W^* = \min_{\boldsymbol{x}\in\mathcal{X}} \mathrm{FCVaR}_{1-\varepsilon}\left[\bar{Q}(\boldsymbol{x},\tilde{\boldsymbol{\zeta}})\right], \tag{19}$$

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

Proof: According to the definition of $\mathrm{FCVaR}_{1-\varepsilon}$ (12), the worst-case FCVaR (18) is equivalent to

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

Interchanging the supremum operator and the maximum operator, we have

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

Recall the following equivalent representation in type$-\infty$ Wasserstein ambiguity set (see, e.g., [8, 71]):

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

which implies that

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

Plugging back the definition of $\mathrm{FCVaR}_{1-\varepsilon}$ (12), 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 worst-case FCVaR (18) in fact can be interpreted as 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 is to optimally select the most favorable $k := N - N\varepsilon$ out of N empirical samples. Given a collection S of k samples, we denote its corresponding type$-\infty$ Wasserstein ambiguity set as $\mathcal{P}_\infty^W(S)$. Intuitively, the DRO with sample-selection Wasserstein is defined as

$$v_W^* = \min_{\substack{\boldsymbol{x}\in\mathcal{X},\\ S\in\mathcal{S}}} \sup_{\mathbb{P}\in\mathcal{P}_\infty^W(S)} \mathbb{E}_{\mathbb{P}}\left[Q(\boldsymbol{x},\tilde{\boldsymbol{\xi}})\right], \tag{20}$$

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

Letting the binary variable z_i indicate whether the ith sample is selected or not, according to the result in [8, 71], under $\infty-$Wasserstein ambiguity set, the problem (20) can be represented as

$$v_W^* = \min_{\substack{\boldsymbol{x}\in\mathcal{X},\\ S\in\mathcal{S}}} \sup_{\mathbb{P}\in\mathcal{P}_\infty^W(S)} \mathbb{E}_{\mathbb{P}}\left[Q(\boldsymbol{x},\tilde{\boldsymbol{\xi}})\right] = \min_{\substack{\boldsymbol{x}\in\mathcal{X},\\ \boldsymbol{z}\in\{0,1\}^N}} \left\{\frac{1}{k}\sum_{i\in[N]} z_i\bar{Q}(\boldsymbol{x},\boldsymbol{\zeta}^i)\colon \sum_{i\in[N]} z_i = k\right\}, \tag{21}$$

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

Proposition 7 *Given that the type$-\infty$ Wasserstein ambiguity set is considered and $N\varepsilon$ is an integer, the worst-case* FCVaR (18) *is equivalent to the DRO with a favorable sample-selection Wasserstein ambiguity set.*

This result shows that applying distributional robustness is essentially to select favorable samples in an optimal manner, consistent with the findings in the previous sections that are beyond the simple preprocessing and are important to get rid of endogenous outliers.

5.2 Confidence Bounds and Decision Outlier Robustness of the Worst-case FCVaR

Before deriving the confidence bounds, we define the following important quantities. We let v^T denote the FCVaR under the true distribution $\mathbb{P}^T$, that is,

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

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 6, we have $\beta^*(\boldsymbol{x}) = \mathrm{VaR}_{1-\varepsilon}\{Q(\boldsymbol{x},\tilde{\boldsymbol{\xi}})\}$.

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 Smooothness of True Cumulative Distribution Function (CDF) around Quantile** $\beta^*(\mathbf{x})$) 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 $\widehat{\Delta}\in[-\Delta_2,\Delta_2]$.*

Note that in Assumption 2, Part (i) is standard in the concentration inequality literature (see, e.g., chapter 2 of [69]). 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, 46]), 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 worst-case FCVaR under type$-\infty$ Wasserstein ambiguity set.

Theorem 5 *(Confidence Bounds) Suppose that Assumption 2 holds. Then for any given $\gamma \in (0,1)$, we have: (i) $\mathbb{P}^T\{v_W^* \leq v^T + 2L\theta\} \geq 1-\gamma$; and (ii) $\mathbb{P}^T\{v_W^* \geq v^T - L\theta\} \geq 1-\gamma$, where $\theta = \mathcal{O}(1)N^{-1/2}\sqrt{n\log(\gamma^{-1})}$ for a discrete compact set $\mathcal{X}$, and $\theta = \mathcal{O}(1)N^{-1/2}\sqrt{n\log(nN)\log(\gamma^{-1})}$ for a general compact set $\mathcal{X}$.*

Proof: See Appendix B.4. □

We make the following remarks on Theorem 5:

(i) Parts (i) and (ii) together show that with high probability, the value 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., $\mathcal{O}(N^{-1/2})$).

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 properties in Part (ii) of Assumption 2.

Theorem 6 *(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 $N \geq \log(\gamma^{-1})/(2\alpha_1^2)$, then with probability $1-\gamma$, 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 $1-\gamma$, the number of N i.i.d. empirical samples $\{\boldsymbol{\zeta}^i\}_{i\in[N]}$ is large enough, such that the number of the samples which falls 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

$$\mathbb{P}^T\left\{\sum_{i\in[N]} \mathbb{1}\left(\boldsymbol{\zeta}^i \notin \widehat{\mathcal{U}}(\Delta_1)\right) \leq \lfloor N\varepsilon \rfloor\right\} \leq \exp\left\{-2N\left(\alpha_1 + \frac{\lfloor N\varepsilon \rfloor}{N} - \varepsilon\right)^2\right\} \approx \exp\left\{-2N\alpha_1^2\right\}.$$

Letting $\exp\left\{-2N\alpha_1^2\right\} \leq \gamma$, the sample size is at least $N \geq \log(\gamma^{-1})/(2\alpha_1^2)$.

Step II. Note that $\Delta_1^L < \Delta_1 + L\theta < \Delta_1^U$ and the function $\bar{Q}(\boldsymbol{x},\boldsymbol{\zeta})$ is defined as

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

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

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

Step III. For the worst distribution $\mathbb{P}\in\mathcal{P}_\infty^W$, according to [8], it can be represented as $\mathbb{P}=1/N\sum_{i\in[N]}\delta_{(\tilde{\boldsymbol{\xi}}=\widehat{\boldsymbol{\xi}}^i)}$ with $\widehat{\boldsymbol{\xi}}^i\in\operatorname{argmax}_{\boldsymbol{\xi}}\{Q(\boldsymbol{x},\boldsymbol{\xi})\colon\|\boldsymbol{\xi}-\boldsymbol{\zeta}^i\|_p\leq\theta\}$ for each $i\in[N]$.

Next, we construct the favorable distribution $\mathbb{P}^*$ such that $\mathbb{P}^*\{\tilde{\boldsymbol{\xi}}=\widehat{\boldsymbol{\xi}}^i\}=\mathbb{1}\{\bar{Q}(\boldsymbol{x},\boldsymbol{\zeta}^i)<\infty\}/(N\eta)$ for each $i\in[N]$. By our construction, we have $\mathbb{P}^*\{\mathcal{U}\}=1,0\preceq\mathbb{P}^*\preceq\mathbb{P}/(1-\varepsilon)$. 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 4 is satisfied. Hence, we conclude that with probability $1-\gamma$, the worst-case FCVaR is decision outlier robust. □

According to Theorem 6, 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 and decision outlier robustness, since $\theta\propto 1/\sqrt{N}$ according to Theorem 5, it is expected that the sample size should not be too small.

We conclude this section by remarking that the results in Theorem 5 and Theorem 6 can be extended to Winsorized CVaR and Huber-skip CVaR. The proofs are similar and thus are omitted.

6 Application to the Two-stage Stochastic Program Without Relatively Complete Recourse

We extend the proposed DFO framework to the two-stage stochastic programs without relatively complete recourse and apply the proposed DFO framework, i.e., we provide Big-M free formulations and discuss how to achieve the out-of-sample performances.

6.1 Two-stage Stochastic Programs without Relatively Complete Recourse

Motivated from the discussions in Section 4, this section focuses on a two-stage stochastic program, which, in general, is defined as $\min_{\boldsymbol{x}\in\mathcal{X}}\boldsymbol{c}^\top\boldsymbol{x}+\mathbb{E}_{\mathbb{P}}[Q(\boldsymbol{x},\tilde{\boldsymbol{\xi}})]$, where for a realization $\boldsymbol{\xi}$ of $\tilde{\boldsymbol{\xi}}$, the recourse function $Q(\boldsymbol{x},\boldsymbol{\xi})$ is defined as

$$Q(\boldsymbol{x},\boldsymbol{\xi})=\inf_{\boldsymbol{y}\in\mathcal{Y}}\left[(\boldsymbol{Q}\boldsymbol{\xi}_q+\boldsymbol{q})^\top\boldsymbol{y}\colon\boldsymbol{T}(\boldsymbol{x})\boldsymbol{\xi}_T+\boldsymbol{\xi}_W\boldsymbol{y}\geq\boldsymbol{h}(\boldsymbol{x})\right],\tag{22}$$

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 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 $\inf_{\boldsymbol{y}\in\mathcal{Y}}[(\boldsymbol{Q}\tilde{\boldsymbol{\xi}}_q+\boldsymbol{q})^\top\boldsymbol{y}]>-\infty$ almost surely.

Complete recourse is a sufficient condition to ensure the feasibility of the second-stage problem. However, many problems in practice genuinely fail to have complete recourse, i.e., warehouses may not fulfill the demand due to the disruptions of extreme scenarios. When the second-stage problem is possible to 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 $\mathbb{E}_{\mathbb{P}}[Q(\boldsymbol{x},\tilde{\boldsymbol{\xi}})]=\infty$. Following the discussions in Section 4, we apply DFO to select favorable scenarios, where the distributionally favorable counterpart of the two-stage programs is defined in (2) and $Q(\boldsymbol{x},\tilde{\boldsymbol{\xi}})$ is defined in (22).

Suppose that the empirical distribution $\mathbb{P}_{\tilde{\zeta}}$ of the second-stage problem with N i.i.d. samples $\{\boldsymbol{\zeta}^i\}_{i\in[N]}$ with the ambiguity set introduced in Section 4 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

$$v^* = \min_{\boldsymbol{x}\in\mathcal{X},\boldsymbol{z}} \left\{ \boldsymbol{c}^\top \boldsymbol{x} + \frac{1}{N-N\varepsilon} \sum_{i\in[N]} z_i Q(\boldsymbol{x},\boldsymbol{\zeta}^i) \colon \sum_{i\in[N]} z_i = N - N\varepsilon, \boldsymbol{z}\in\{0,1\}^N \right\}, \tag{23}$$

where we assume that $\infty\times 0 = 0$. In problem (23), for each $i\in[N]$, the product $z_i Q(\boldsymbol{x},\boldsymbol{\zeta}^i)$ can be represented as the following MILP

$$z_i Q(\boldsymbol{x},\boldsymbol{\zeta}^i) = \min_{\boldsymbol{y}^i\in\mathcal{Y}} \left[(\boldsymbol{Q}\boldsymbol{\zeta}_q^i + \boldsymbol{q})^\top \boldsymbol{y}^i - L_i(1-z_i) \colon \boldsymbol{T}(\boldsymbol{x})\boldsymbol{\zeta}_T^i + \boldsymbol{\zeta}_W^i \boldsymbol{y}^i \geq \boldsymbol{h}(\boldsymbol{x}) - \boldsymbol{M}^i(1-z_i) \right]. \tag{24}$$

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

$$M_j^i \geq \max_{\boldsymbol{x}\in\mathcal{X},\boldsymbol{y}^i\in\mathcal{Y}} h_j(\boldsymbol{x}) - (\boldsymbol{T}(\boldsymbol{x})\boldsymbol{\zeta}_T^i + \boldsymbol{\zeta}_W^i \boldsymbol{y}^i)_j$$

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

The purpose of using $\boldsymbol{z}$ variables in the constraints of the second-stage problem (23) 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, in the next subsection, we further explore the structure of the problem and discuss sufficient conditions under which we can obtain the big-M free formulations.

6.2 Big-M Free Formulations

In this subsection, we show that under some conditions, we are able to represent the bilinear terms $\{z_i Q(\boldsymbol{x},\boldsymbol{\zeta}^i)\}_{i\in[N]}$ in problem (23) using the big-M free formulations.

Theorem 7 *Suppose that the set $\mathcal{Y} := \{\boldsymbol{y} : \boldsymbol{D}\boldsymbol{y}\geq\boldsymbol{d}\}$ and $\boldsymbol{T}(\boldsymbol{x}) = \widehat{\boldsymbol{T}}_1\boldsymbol{x}\otimes\mathbf{e} + \widehat{\boldsymbol{T}}_2$, $\boldsymbol{h}(\boldsymbol{x}) = \widehat{\boldsymbol{H}}\boldsymbol{x} + \widehat{\boldsymbol{h}}$, $\widehat{\boldsymbol{T}}_1\in\mathbb{R}^{\ell\times n}$, $\widehat{\boldsymbol{T}}_2\in\mathbb{R}^{\ell\times m_2}$, $\widehat{\boldsymbol{H}}\in\mathbb{R}^{\ell\times n}$, $\widehat{\boldsymbol{h}}\in\mathbb{R}^{\ell}$, and vector $\mathbf{0}$ is contained in the polyhedron $\{\boldsymbol{y}^i : \widehat{\boldsymbol{T}}_1\boldsymbol{x}\otimes\mathbf{e}\boldsymbol{\zeta}_T^i + \boldsymbol{\zeta}_W^i\boldsymbol{y}^i - \widehat{\boldsymbol{H}}\boldsymbol{x}\geq\mathbf{0}\}$ for each $\boldsymbol{x}\in\mathcal{X}$ and $i\in[N]$. Then, the favorable two-stage stochastic problem* (23) *is equivalent to*

$$v^* = \min_{\boldsymbol{x}\in\mathcal{X},\boldsymbol{z}} \left\{ \boldsymbol{c}^\top \boldsymbol{x} + \frac{1}{N-N\varepsilon} \sum_{i\in[N]} \widehat{Q}(\boldsymbol{x},z_i,\boldsymbol{\zeta}^i) \colon \sum_{i\in[N]} z_i \geq N - N\varepsilon, \boldsymbol{z}\in\{0,1\}^N \right\}, \tag{25}$$

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

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

Proof: In problem (25), we first consider $z_i = 0$. Since the vector $\mathbf{0}$ is contained in the polyhedron $\{\boldsymbol{y}^i : \widehat{\boldsymbol{T}}_1\boldsymbol{x}\otimes\mathbf{e}\boldsymbol{\zeta}_T^i + \boldsymbol{\zeta}_W^i\boldsymbol{y}^i - \widehat{\boldsymbol{H}}\boldsymbol{x}\geq\mathbf{0}\}$ for each $\boldsymbol{x}\in\mathcal{X}$ and $i\in[N]$, then the optimal value of the second-stage problem $\widehat{Q}(\boldsymbol{x},z_i,\boldsymbol{\zeta}^i)$ is 0, which is as the same as the value of $z_i Q(\boldsymbol{x},\boldsymbol{\zeta}^i)$. If $z_i = 1$, then $\widehat{Q}(\boldsymbol{x},z_i,\boldsymbol{\zeta}^i)$ is identical to $Q(\boldsymbol{x},\boldsymbol{\zeta}^i)$. □

Notice that there is no big-M coefficient in the formulation (25). The condition that the vector $\mathbf{0}$ is contained in the polyhedron $\{\boldsymbol{y}^i : \widehat{\boldsymbol{T}}_1\boldsymbol{x}\otimes\mathbf{e}\boldsymbol{\zeta}_T^i + \boldsymbol{\zeta}_W^i\boldsymbol{y}^i - \widehat{\boldsymbol{H}}\boldsymbol{x}\geq\mathbf{0}\}$ for each $\boldsymbol{x}\in\mathcal{X}$ and $i\in[N]$ can be further simplified as $\widehat{\boldsymbol{T}}_1\boldsymbol{x}\otimes\mathbf{e}\boldsymbol{\zeta}_T^i \geq \widehat{\boldsymbol{H}}\boldsymbol{x}$ for all $\boldsymbol{x}\in\mathcal{X}$ and $i\in[N]$.

We use the following example to illustrate Theorem 7.

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, 45]). Following the notation, the TRP problem can be formulated as

$$\min_{\boldsymbol{x}\geq\mathbf{0},\boldsymbol{z}} \left\{ \boldsymbol{c}^\top \boldsymbol{x} + \frac{1}{N-N\varepsilon} \sum_{i\in[N]} z_i Q(\boldsymbol{x},\boldsymbol{\zeta}^i) \colon \sum_{i\in[N]} z_i \geq N - N\varepsilon, \boldsymbol{z}\in\{0,1\}^N \right\}, \tag{26a}$$

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

$$Q(\boldsymbol{x},\boldsymbol{\zeta}^i) = \min_{\boldsymbol{y}^i\geq\mathbf{0}} \left\{ \sum_{s\in[n]}\sum_{j\in[n_1]} q^i_{sj} y^i_{sj} \colon \sum_{j\in[n_1]} y^i_{sj} \leq p^i_s x_s, \forall s\in[n], \sum_{s\in[n]} u^i_{sj} y^i_{sj} \geq \lambda^i_j, \forall j\in[n_1] \right\}. \tag{26b}$$

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 $\tilde{\boldsymbol{q}}, \tilde{\boldsymbol{p}}, \tilde{\boldsymbol{\mu}}, \tilde{\boldsymbol{\lambda}}$ 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 $s \in [n]$, $\tilde{\mu}_{sj}$ 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 is a two-stage stochastic program without relatively complete recourse, and for any $\boldsymbol{x} \geq \mathbf{0}$, $\boldsymbol{y}^i = \mathbf{0}$ is always feasible to (26b) for each $i \in [N]$. Hence, we can apply the result in Theorem 7. Using the binary variables $\boldsymbol{z}$, we can rewrite the bilinear term as

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

Thus, we arrive at a big-M free formulation for (26a). ◇

As a direct corollary of Theorem 7, we can provide big-M free formulations for the Winsorized CVaR and the Huber-skip CVaR type of the two-stage problem.

Corollary 5 *Under the same assumptions in Theorem 7:*

(i) WCVaR *of the objective function* (23) *admits the following formulation*

$$\min_{\substack{\boldsymbol{x}\in\mathcal{X},\\ \boldsymbol{z}\in\{0,1\}^N}} \left\{ \boldsymbol{c}^\top \boldsymbol{x} + \frac{1}{N}\sum_{i\in[N]} z_i Q(\boldsymbol{x},\boldsymbol{\zeta}^i) + \eta\varepsilon \colon \eta \geq z_i Q(\boldsymbol{x},\boldsymbol{\zeta}^i) + (1-z_i)L_i, \forall i\in[N], \sum_{i\in[N]} z_i \geq N - N\varepsilon \right\}; \tag{27a}$$

(ii) HCVaR *of the objective function* (23) *admits the following formulation*

$$\min_{\boldsymbol{x}\geq\mathbf{0},\boldsymbol{z}\in\{0,1\}^N} \left\{ \boldsymbol{c}^\top \boldsymbol{x} + \frac{1}{N}\sum_{i\in[N]} \big(z_i Q(\boldsymbol{x},\boldsymbol{\zeta}^i) + (1-z_i)H\big) \right\}, \tag{27b}$$

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

Notice that the bilinear terms $\{z_i Q(\boldsymbol{x},\boldsymbol{\zeta}^i)\}_{i\in[N]}$ in (27a) *and* (27b) *can be linearized by applying the result in* (24) *or using Theorem 7.*

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

6.3 Achieving Out-of-Sample Performance in Favorable Two-stage Stochastic Programs

In this subsection, to overcome 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 FCVaR to a two-stage stochastic program, we have

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

which can be written as

$$v_W^* = \min_{\substack{\boldsymbol{x}\in\mathcal{X},\\ \boldsymbol{z}\in\{0,1\}^N}} \boldsymbol{c}^\top \boldsymbol{x} + \left\{ \frac{1}{N-N\varepsilon} \sum_{i\in[N]} z_i \max_{\boldsymbol{\xi}} \left\{ Q(\boldsymbol{x},\boldsymbol{\xi}) : \|\boldsymbol{\xi} - \boldsymbol{\zeta}^i\| \le \theta \right\} : \sum_{i\in[N]} z_i = N - N\varepsilon \right\}, \tag{28}$$

Notice that in general, for a given $\boldsymbol{z}$, the optimization problem above is NP-hard (see the details in [71]). Therefore, instead of focusing on (28), 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 (28) is equivalent to

$$v_W^* = \min_{\substack{\boldsymbol{x}\in\mathcal{X},\\ \boldsymbol{z}\in\{0,1\}^N}} \boldsymbol{c}^\top \boldsymbol{x} + \left\{ \frac{1}{N-N\varepsilon} \sum_{i\in[N]} z_i Q(\boldsymbol{x}, \boldsymbol{\zeta}^i \pm \theta\mathbf{e}) : \sum_{i\in[N]} z_i = N - N\varepsilon \right\}, \tag{29}$$

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 7 or the MILP (24) to derive a proper formulation. Notice that this monotonicity structure has been studied in several recent works (see, e.g., [16, 71, 73]). We use Example 5 to illustrate the formulation (29).

Example 6 Consider Example 5 under the 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 $\{\boldsymbol{\zeta}^i\}_{i\in[N]}$ on the nonnegative support $\mathcal{U}$. Then, the worst-case DFO (29) is equivalent to

$$v_W^* = \min_{\boldsymbol{x}\ge\mathbf{0},\boldsymbol{z}} \left\{ \boldsymbol{c}^\top \boldsymbol{x} + \frac{1}{N-N\varepsilon} \sum_{i\in[N]} z_i \bar{Q}(\boldsymbol{x}, \boldsymbol{\zeta}^i) : \sum_{i\in[N]} z_i \ge N - N\varepsilon, \boldsymbol{z}\in\{0,1\}^N \right\},$$

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

$$z_i \bar{Q}(\boldsymbol{x}, \boldsymbol{\zeta}^i) = \min_{\boldsymbol{y}^i\ge\mathbf{0}} \left\{ \sum_{s\in[n]} \sum_{j\in[n_1]} (q_{sj}^i + \theta/w_q) y_{sj}^i : \begin{array}{c} \sum\limits_{j\in[n_1]} y_{sj}^i \le (p_s^i - \theta/w_p)_+ x_s, \forall s\in[n], \\ \sum\limits_{s\in[n]} (u_{sj}^i - \theta/w_u)_+ y_{sj}^i \ge (\lambda_j^i + \theta/w_\lambda) z_i, \forall j\in[n_1] \end{array} \right\}.$$

$\diamond$

We remark that interested readers are referred to [71] 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.

7 Numerical Study

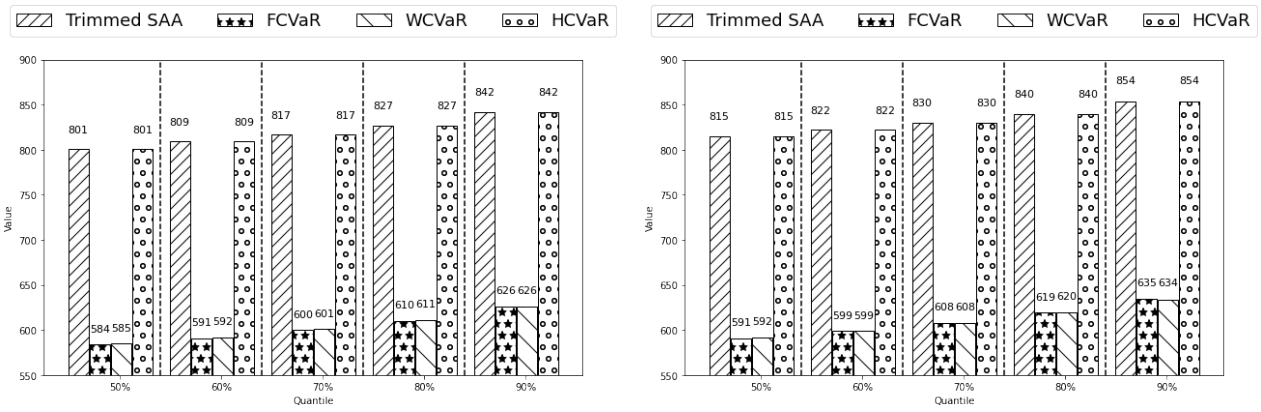
For the demonstration purpose, this section presents the numerical results to compare the strengths of FCVaR and its alternatives based on Example 5 in Section 6, where the relatively complete recourse assumption may not be satisfied.

We generate random instances with varying sample size 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 $\tilde{\boldsymbol{\mu}}$) 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 mean 1 and variance 0.2, the components of random service rate $\tilde{\boldsymbol{\mu}}$ are i.i.d. truncated Gaussian ones with mean 1 and variance 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 mean 10 and variance 0.2. We also assume that there exist some outliers in the customer demand information, denoted by $\tilde{\boldsymbol{\lambda}}^o$, where its components are i.i.d. truncated Gaussian distributed with mean 40 and variance 10, and may cause the underlying two-stage problem infeasible. The observed demand follows the following distribution $0.9\tilde{\boldsymbol{\lambda}} + 0.1\tilde{\boldsymbol{\lambda}}^o$. 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 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 [17]). 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 6 and focus on type$-\infty$ Wasserstein ambiguity set equipped with weighted infinity norm. All the instances are coded in Python 3.7.0 with calls to Gurobi 9.0.1 on a personal computer with a 2.8GHz Intel Core i5 processor and 8G 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, and HCVaR), when evaluating the first-stage decision using 50 test instances, we record all the $50\%, 60\%, 70\%, 80\%, 90\%$ quantiles of the second-stage values, respectively. We then report the average of each quantile among these 50 testing instances. We set $\varepsilon = 0.15$ and consider the sample size with $N = \{100, 200\}$. For each testing instance, we assume the components of customer demand $\tilde{\boldsymbol{\lambda}}$ are i.i.d. truncated Gaussian ones with mean 10 and variance 0.2 and the remaining parameters follow the same assumptions described in the training procedure. The result is shown in Figure 4. It is seen that both FCVaR and WCVaR can consistently provide a favorable solution with a lower cost than the trimmed SAA. However, HCVaR performs worse than the other two. We continue to discuss the performance of HCVaR in the next experiment.

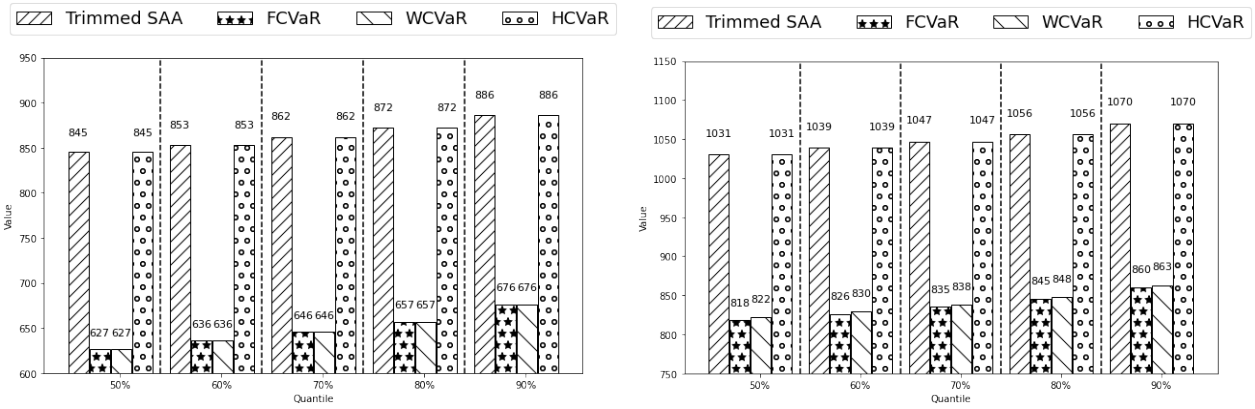
Experiment 2. Model Comparisons When the Testing Distribution is Different From the Training one. We follow the same procedure described in Experiment 1, i.e., we record all the $50\%, 60\%, 70\%, 80\%, 90\%$ quantiles in the second-stage scenarios for each method (e.g., Trimmed SAA, FCVaR, WCVaR, and HCVaR) in each testing instance, respectively, and report the average of each quantile among these 50 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 mean 0.8 and variance 0.4. The result is shown in Figure 5. 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 the other three. 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 1500 to avoid any trivial solution; i.e., if H is small, e.g., H is less than the first-stage cost, it provides a trivial solution is $\boldsymbol{x} = \boldsymbol{0}, \boldsymbol{z} = \boldsymbol{0}$. Meanwhile, to choose a reasonable value of H, one can solve the trimmed SAA model first and then select its $(1-\varepsilon)-$quantile. In this way, the value of H is larger than the first-stage cost.



(a) $N = 100$ and $\varepsilon = 0.15$ (b) $N = 200$ and $\varepsilon = 0.15$

Fig. 4: Quantile Comparisons among Trimmed SAA, FCVaR, WCVaR, and HCVaR in Experiment 1.

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.



(a) $N = 100$ and $\varepsilon = 0.15$ (b) $N = 200$ and $\varepsilon = 0.15$

Fig. 5: Quantile Comparisons among Trimmed SAA, FCVaR, WCVaR, and HCVaR in Experiment 2.

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 will focus on FCVaR and WCVaR for the remaining experiments. We follow the same setting and derivation of Example 6 in Section 6 for both worst-case FCVaR and 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 $\varepsilon = 0.15$, sample size $N = 100$, and report the average of each quantile of the testing instances. To choose a proper Wasserstein radius θ, we suggest selecting the smallest θ such that its corresponding training costs of FCVaR and WCVaR are beyond the 95% testing confidence interval. 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. Then, following the same procedure as described in Experiment 1, the result is shown in Table 1. The optimal Wasserstein radius is $\theta = 0.02$ for FCVaR and $\theta = 0.01$ for WCVaR, and we observe that all FCVaR and WCVaR can still

consistently provide a favorable solution with a lower cost, and the running time of FCVaR is slightly less than that of WCVaR.

Table 1: Comparisons in the Worst Case and θ Selection in Experiment 3

N	ε	θ	FCVaR (26a)		WCVaR (27a)		Testing	
			Opt. Val.	Time (s)	Opt. Val.	Time (s)	FCVaR (26a) C.I.	WCVaR (27a) C.I.
100	0.15	0.00	593.79	20.34	601.43	22.93	[592.92, 597.02]	[599.28, 605.38]
		0.01	596.12	22.19	607.34	26.03	[594.28, 598.06]	[601.82, 606.99]
		0.02	602.89	22.65	609.38	27.85	[596.92, 601.73]	[603.30, 608.55]
		0.03	609.45	22.47	623.25	29.28	[599.81, 604.75]	[607.57, 612.84]
		0.04	618.92	22.92	637.91	27.41	[604.12, 609.93]	[612.78, 618.93]
		0.05	635.01	24.56	652.72	28.75	[612.77, 617.92]	[624.22, 631.92]
		0.06	652.19	23.81	668.91	29.21	[620.89, 626.83]	[635.97, 643.03]
		0.07	663.02	25.72	672.91	29.07	[624.25, 630.49]	[642.74, 650.62]
		0.08	671.44	25.21	681.88	31.99	[628.59, 634.26]	[651.07, 659.27]
		0.09	682.89	27.31	693.90	29.34	[633.82, 640.92]	[660.38, 668.61]
		0.10	705.00	29.75	715.82	30.81	[640.28, 647.87]	[668.90, 675.92]

Experiment 4. Value of MICP-R: Computing FCVaR when $N\varepsilon$ is Not an Integer. In this experiment, we follow the same setting as Experiment 1 and demonstrate the value of using an MICP-R formulation. For the demonstration purpose, we focus on the FCVaR formulation. When $N\varepsilon$ is a rational number but not an integer, according to Corollary 2, the proposed FCVaR formulation (26a) is still MICP-R. On the other hand, the two-stage program (23) admits a naive bilinear formulation, which can be solved directly by Gurobi. Since we cannot solve the bilinear model to optimality within the time limit, we use GAP to denote its optimality gap as $\text{GAP}(\%) = (|\text{UB} - \text{LB}|)/|\text{LB}| \times 100$, where "UB" and "LB" to denote the best upper bound and the best lower bound found by Gurobi. The result is displayed in Table 2. We find that the MICP-R formulation can improve the running time even for a small-scale instance, which shows the effectiveness of exploring the MICP-R formulation.

Table 2: Comparisons Between FCVaR (26a) and Its Bilinear Counterpart in Experiment 4

N	$\varepsilon = 0.15$			$\varepsilon = 0.18$		
	Bilinear		FCVaR (26a)	Bilinear		FCVaR (26a)
	Time (s)	GAP	Time (s)	Time (s)	GAP	Time (s)
25	3600	5.02	62.82	3600	5.39	71.35
30	3600	5.29	71.57	3600	5.88	84.48
35	3600	5.87	129.14	3600	6.29	159.83
55	3600	7.92	160.02	3600	8.55	183.71

Experiment 5. Value of Confidence Bound. In this experiment, we test the order magnitude of the proposed confidence bound presented in Section 5.2. In this experiment, we follow the same setting as that in Experiment 3, except that we let the risk parameter be 0.10 and assume that the observed demand follows the distribution $0.95\tilde{\boldsymbol{\lambda}} + 0.05\tilde{\boldsymbol{\lambda}}^o$. Then, we follow the same procedure described in Experiment 3 to choose a proper θ for each sample size. We repeat this process 5 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 5.2.

Experiment 6. 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 6. The proposed big-M free formulations can effectively identify better feasible solutions than the exact Big-M model with a much shorter solution time. To illustrate this, we use "UB" and "LB" to 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

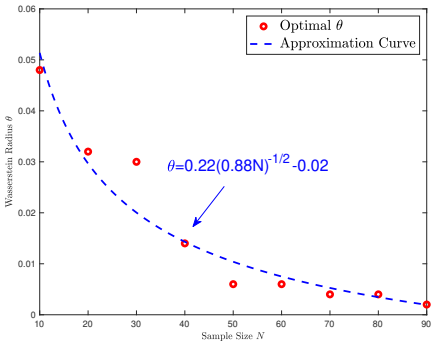


Fig. 6: Optimal θ vs. Sample Size N in Experiment 5.

then let the value of the big-M coefficient be the feasible scenario with the largest recourse value. We repeat this process for 5 times, and the average performance can be found in Table 3. 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.

Table 3: Comparisons Between Big-M and Big-M Free Formulations of FCVaR and WCVaR in Experiment 6

ε	θ	$N=300$						$N=400$					
		FCVaR			WCVaR			FCVaR			WCVaR		
		Big-M		Big-M Free (26a)	Big-M		Big-M Free (27a)	Big-M		Big-M Free (26a)	Big-M		Big-M Free (27a)
		Time (s)	GAP	Time (s)	Time (s)	GAP	Time (s)	Time (s)	GAP	Time (s)	Time (s)	GAP	Time (s)
0.15	0.00	3234.38	0.00	1883.34	3545.11	0.00	2043.45	3600	2.06	3023.15	3600	2.22	3357.91
	0.01	3291.73	0.00	1921.49	3491.02	0.00	2065.98	3600	1.92	3109.76	3600	2.03	3490.14
	0.02	3165.29	0.00	1839.34	3587.15	0.00	2239.11	3600	1.93	3091.34	3600	1.68	3418.67
	0.03	3314.42	0.00	1973.68	3602.89	0.00	2284.37	3600	1.96	3128.56	3600	2.06	3301.73
	0.04	3253.43	0.00	1890.01	3432.12	0.00	2144.83	3600	1.83	3185.89	3600	1.94	3589.18
	0.05	3136.54	0.00	1965.33	3519.36	0.00	2198.63	3600	1.78	3098.68	3600	1.95	3472.47
	0.06	3254.17	0.00	1973.59	3492.10	0.00	2209.68	3600	1.81	2964.01	3600	2.04	3371.29
	0.07	3109.76	0.00	1872.22	3449.72	0.00	2158.93	3600	1.82	3072.66	3600	2.08	3491.02
	0.08	3287.51	0.00	1820.74	3573.29	0.00	2239.07	3600	1.69	3068.48	3600	1.98	3311.29
	0.09	3189.73	0.00	1845.27	3453.55	0.00	2061.84	3600	1.72	3279.91	3600	1.82	3518.90
	0.10	3296.47	0.00	1792.46	3511.34	0.00	2013.72	3600	1.61	2815.92	3600	1.74	3320.91

8 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, proved mixed-integer convex programming representability, established decision outlier robustness concept, and integrated distributional robustness to achieve out-of-sample performances. Exploring the contextual information in DFO can be a promising future research direction.

Acknowledgment

This research has been supported in part by the National Science Foundation grants 2046426 and 2153607.

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(3):169–175, 2002.
4. P. Auer, N. Cesa-Bianchi, and P. Fischer. Finite-time analysis of the multiarmed bandit problem. *Machine learning*, 47(2):235–256, 2002.
5. E. Balas. Disjunctive programming. *Annals of discrete mathematics*, 5:3–51, 1979.
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. D. Bertsimas, S. Shtern, and B. Sturt. A data-driven approach to multi-stage stochastic linear optimization. Preprint, 2018.
9. J. Bi and T. Zhang. Support vector classification with input data uncertainty. In *Advances in neural information processing systems*, pages 161–168, 2005.
10. G. Boole. *The mathematical analysis of logic*. Philosophical Library, 1847.
11. E. J. Candès, X. Li, Y. Ma, and J. Wright. Robust principal component analysis? *Journal of the ACM (JACM)*, 58(3):1–37, 2011.
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*, pages 1–36, 2022.
15. Z. Chen, D. Kuhn, and W. Wiesemann. Data-driven chance constrained programs over Wasserstein balls. arXiv preprint arXiv:1809.00210, 2018.
16. Z. Chen and W. Xie. Regret in the newsvendor model with demand and yield randomness. *Production and Operations Management*, 2021.
17. J. W. Chinneck. *Feasibility and Infeasibility in Optimization:: Algorithms and Computational Methods*, volume 118. Springer Science & Business Media, 2007.
18. T. Cui, Y. Ouyang, and Z.-J. M. Shen. Reliable facility location design under the risk of disruptions. *Operations research*, 58(4-part-1):998–1011, 2010.
19. C. Ding, D. Zhou, X. He, and H. Zha. R1 - pca: rotational invariant l1-norm principal component analysis for robust subspace factorization. In *Proceedings of the 23rd international conference on Machine learning*, pages 281–288, 2006.
20. P. M. Esfahani and D. Kuhn. Data-driven distributionally robust optimization using the Wasserstein metric: Performance guarantees and tractable reformulations. *Mathematical Programming*, 171(1):115–166, 2018.
21. A. Esteban-Pérez and J. M. Morales. Distributionally robust stochastic programs with side information based on trimmings. arXiv preprint arXiv:2009.10592, 2020.
22. D. Ge, X. Jiang, and Y. Ye. A note on the complexity of l p minimization. *Mathematical programming*, 129(2):285–299, 2011.
23. J.-y. Gotoh, M. J. Kim, and A. Lim. A data-driven approach to beating saa out-of-sample. Available at SSRN 3853493, 2021.
24. V. Guigues, A. Juditsky, and A. Nemirovski. Non-asymptotic confidence bounds for the optimal value of a stochastic program. *Optimization Methods and Software*, 32(5):1033–1058, 2017.
25. F. R. Hampel. The influence curve and its role in robust estimation. *Journal of the american statistical association*, 69(346):383–393, 1974.
26. G. A. Hanasusanto, V. Roitch, D. Kuhn, and W. Wiesemann. Ambiguous joint chance constraints under mean and dispersion information. *Operations Research*, 65(3):751–767, 2017.
27. J. Heinonen. *Lectures on Lipschitz analysis*. University of Jyväskylä, 2005.

28. N. Ho-Nguyen, F. Kılınç-Karzan, S. Küçükyavuz, and D. Lee. Distributionally robust chance-constrained programs with right-hand side uncertainty under Wasserstein ambiguity. *Mathematical Programming*, pages 1–32, 2021.
29. V. Hodge and J. Austin. A survey of outlier detection methodologies. *Artificial intelligence review*, 22(2):85–126, 2004.
30. W. Hoeffding. Probability inequalities for sums of bounded random variables. In *The Collected Works of Wassily Hoeffding*, pages 409–426. Springer, 1994.
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(4):1–37, 2014.
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(1):2185–2211, 2014.
34. P. J. Huber. Robust estimation of a location parameter. In *Breakthroughs in statistics*, pages 492–518. Springer, 1992.
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(4):779–811, 2021.
38. R. Koenker and K. F. Hallock. Quantile regression. *Journal of economic perspectives*, 15(4):143–156, 2001.
39. G. Li. Robust regression. *Exploring data tables, trends, and shapes*, 281:U340, 1985.
40. X. Li, Z. Zhu, A. Man-Cho So, and R. Vidal. Nonconvex robust low-rank matrix recovery. *SIAM Journal on Optimization*, 30(1):660–686, 2020.
41. Y. Li, Y. Sun, and Y. Chi. Low-rank positive semidefinite matrix recovery from corrupted rank-one measurements. *IEEE Transactions on Signal Processing*, 65(2):397–408, 2016.
42. X. Liu, S. Küçükyavuz, and J. Luedtke. Decomposition algorithms for two-stage chance-constrained programs. *Mathematical Programming*, 157(1):219–243, 2016.
43. M. Lu, L. Ran, and Z.-J. M. Shen. Reliable facility location design under uncertain correlated disruptions. *Manufacturing & Service Operations Management*, 17(4):445–455, 2015.
44. M. Lubin, I. Zadik, and J. P. Vielma. Mixed-integer convex representability. In *International Conference on Integer Programming and Combinatorial Optimization*, pages 392–404. Springer, 2017.
45. J. Luedtke. A branch-and-cut decomposition algorithm for solving chance-constrained mathematical programs with finite support. *Mathematical Programming*, 146(1):219–244, 2014.
46. J. Luedtke and S. Ahmed. A sample approximation approach for optimization with probabilistic constraints. *SIAM Journal on Optimization*, 19(2):674–699, 2008.
47. R. A. Maronna, R. D. Martin, V. J. Yohai, and M. Salibián-Barrera. *Robust statistics: theory and methods (with R)*. John Wiley & Sons, 2019.
48. 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*, 187:171–179, 1986.
49. G. P. McCormick. Computability of global solutions to factorable nonconvex programs: Part i—convex underestimating problems. *Mathematical programming*, 10(1):147–175, 1976.
50. N. Naimipour, S. Khobahi, and M. Soltanalian. Upr: A model-driven architecture for deep phase retrieval. arXiv preprint arXiv:2003.04396, 2020.
51. 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*, pages 13942–13953, 2019.
52. 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*, pages 15872–15882, 2019.

53. V. A. Nguyen, N. Si, and J. Blanchet. Robust bayesian classification using an optimistic score ratio. In *International Conference on Machine Learning*, pages 7327–7337. PMLR, 2020.
54. M. Norton, A. Takeda, and A. Mafusalov. Optimistic robust optimization with applications to machine learning. arXiv preprint arXiv:1711.07511, 2017.
55. A. Prékopa. *Stochastic programming*. Springer Science & Business Media, 1995.
56. H. Rahimian and S. Mehrotra. Distributionally robust optimization: A review. arXiv preprint arXiv:1908.05659, 2019.
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(3):173–182, 1982.
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. Shalev-Shwartz et al. Online learning and online convex optimization. *Foundations and trends in Machine Learning*, 4(2):107–194, 2011.
62. A. Shapiro and S. Ahmed. On a class of minimax stochastic programs. *SIAM Journal on Optimization*, 14(4):1237–1249, 2004.
63. H. Shen and R. Jiang. Chance-constrained set covering with Wasserstein ambiguity. arXiv preprint arXiv:2010.05671, 2020.
64. M. Sion. On general minimax theorems. *Pacific Journal of mathematics*, 8(1):171–176, 1958.
65. J. Song and C. Zhao. Optimistic distributionally robust policy optimization. arXiv preprint arXiv:2006.07815, 2020.
66. R. S. Sutton and A. G. Barto. *Reinforcement learning: An introduction*. MIT press, 2018.
67. H. C. Tijms. *A first course in stochastic models*. John Wiley and sons, 2003.
68. J. W. Tukey. *Exploratory data analysis*. Pearson, 1977.
69. M. J. Wainwright. *High-dimensional statistics: A non-asymptotic viewpoint*, volume 48. Cambridge University Press, 2019.
70. A. A. Weiss. Estimating nonlinear dynamic models using least absolute error estimation. *Econometric Theory*, pages 46–68, 1991.
71. W. Xie. Tractable reformulations of two-stage distributionally robust linear programs over the type-∞ Wasserstein ball. *Operations Research Letters*, 48(4):513–523, 2020.
72. W. Xie. On distributionally robust chance constrained programs with wasserstein distance. *Mathematical Programming*, 186(1):115–155, 2021.
73. W. Xie, J. Zhang, and S. Ahmed. Distributionally robust bottleneck combinatorial problems: uncertainty quantification and robust decision making. *Mathematical Programming*, pages 1–44, 2021.
74. C. Yale and A. B. Forsythe. Winsorized regression. *Technometrics*, 18(3):291–300, 1976.
75. 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(3):331–350, 2003.
76. M.-C. Yue, D. Kuhn, and W. Wiesemann. On linear optimization over wasserstein balls. *Mathematical Programming*, pages 1–16, 2021.
77. J. Zhu, S. Rosset, R. Tibshirani, and T. J. Hastie. 1-norm support vector machines. In *Advances in neural information processing systems*, pages 49–56, 2004.

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

A.1 DFO Recovers More Robust Statistics Based on Proposition 1

Using the same weight uncertainty set $\mathcal{U}$ and following the similar derivation as Proposition 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., [39]), 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., [48]), 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., [70]). 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:

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

A.2 DFO Recovers More M-Estimators

Based on the discussions in Section 2.4, we use DFO with a similar uncertainty set to recover the Huber estimator [34] and Tukey's bisquare estimator [68].

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

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

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

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

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

A.3 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., [38, 75]). 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

$$\min_{\boldsymbol{\beta}} \left\{ \tau \sum_{i\in[m]} (y_i - \bar{\boldsymbol{x}}_i^\top \boldsymbol{\beta})_+ + (1-\tau) \sum_{i\in[m]} (\bar{\boldsymbol{x}}_i^\top \boldsymbol{\beta} - y_i)_+ \right\}, \tag{30a}$$

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

$$v^* = \min_{\boldsymbol{\beta}} \min_{\boldsymbol{\xi}\in\mathcal{U}_I} \sum_{i\in[m]} \xi^i (y_i - \bar{\boldsymbol{x}}_i^\top \boldsymbol{\beta}) + \sum_{i\in[m]} |y_i - \bar{\boldsymbol{x}}_i^\top \boldsymbol{\beta}|, \tag{30b}$$

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 (30b), letting $\xi^i = 0$ for all $i \in [m]$, it reduces to LAD.

Appendix B. Proofs

Proofs in Section 3

B.1 Proof of Proposition 2

Proposition 2 *Computing the inner infimum of DFO* (2)*, in general, is NP-hard even when the ambiguity set* $\mathcal{P} = \{\mathbb{P}\colon \mathbb{P}\{\tilde{\boldsymbol{\xi}} \in \mathcal{U}\} = 1\}$ *and the recourse function* $Q(\boldsymbol{x}, \boldsymbol{\xi})$ *can be represented as a simple linear program with objective uncertainty.*

Proof: Let us consider the NP-complete problem — set partition problem, which asks

> ***Set partition problem.*** *Given* N *nonnegative integers* $w_1, w_2, \cdots, w_N$*, does there exist one set partition* S*, such that* $\sum_{i\in S} w_i = \sum_{i\in[N]\setminus S} w_i$*?*

In DFO (2), let the ambiguity set $\mathcal{P} = \{\mathbb{P}\colon \mathbb{P}\{\tilde{\boldsymbol{\xi}} \in \mathcal{U}_I\} = 1\}$ with interval uncertainty set $\mathcal{U}_I = [-1, 1]^N$, and let the recourse function be

$$Q(\boldsymbol{x}, \boldsymbol{\xi}) = \min_{\boldsymbol{y}\in\mathcal{Y}} \sum_{i\in[N]} \xi^i(\boldsymbol{a}_i^\top \boldsymbol{y} - b_i),$$

where $\boldsymbol{a}_i = \boldsymbol{e}_i$ and $b_i = 0$ for each $i \in [N]$, and set $\mathcal{Y} = \{\boldsymbol{y} \in \mathbb{R}^N : -1 \le y_i \le 1, \forall j \in [N], \sum_{i\in[N]} w_j y_j = 0\}$. Under this setting, the inner infimum of DFO (2) reduces to

$$v^* = \min_{\boldsymbol{\xi},\boldsymbol{y}} \left\{ \sum_{i\in[N]} \xi^i y_i\colon\ -1 \le \xi^i \le 1, \forall i \in [N], -1 \le y_i \le 1, \forall i \in [N],\ \sum_{i\in[N]} w_i y_i = 0 \right\}. \tag{31a}$$

Above, optimizing over $\boldsymbol{\xi}$ first, problem (31a) reduces to

$$v^* = \min_{\boldsymbol{y}} \left\{ -\sum_{i\in[N]} \max(y_i, 0) + \sum_{i\in[N]} \min(y_i, 0) :\ -1 \le y_i \le 1, \forall i \in [N],\ \sum_{i\in[N]} w_i y_i = 0 \right\}. \tag{31b}$$

Then, we observe that the optimal value $v^* = -N$ in (31b) if and only if there exists an optimal solution $\boldsymbol{y}^* \in \{-1, 1\}^N$, i.e., the optimal value $v^* = -N$ in (31b) if and only if there exists a set partition such that $\sum_{i\in S} w_i = \sum_{i\in[N]\setminus S} w_i$. Since the set partition problem is NP-hard, solving problem (31b) is NP-hard, so is computing the inner infimum of DFO (2). □

B.2 Proof of Lemma 2

Lemma 2 *The reverse norm function* $f(\boldsymbol{x}) = -\|\boldsymbol{x}\|_p + \chi_{\mathcal{X}}(\boldsymbol{x})$ *is MICP-R if* $p \in \{1, \infty\}$ *and is not MICP-R if* $p \in (1, \infty)$*.*

Proof: We focus on the MICP-R of the epigraph of function $f(\cdot)$, which reads as

$$\mathrm{epi}(f) = \{(\boldsymbol{x}, t)\colon\ -\|\boldsymbol{x}\|_p \le t, \boldsymbol{x} \in \mathcal{X}\}. \tag{32a}$$

Next, we split the proof into three cases based on the choice of p.
Case 1: When $p = 1$, i.e., the norm is L_1, we have

$$\mathrm{epi}(f) = \left\{ (\boldsymbol{x}, t)\colon \sum_{i\in[n]} |x_i| \ge -t, \boldsymbol{x} \in \mathcal{X} \right\},$$

which is equivalent to

$$\mathrm{epi}(f)=\left\{(\boldsymbol{x},t)\colon \max_{\boldsymbol{z}\in\{-1,1\}^n}\sum_{i\in[n]}x_iz_i\geq -t,\boldsymbol{x}\in\mathcal{X}\right\},$$

or

$$\mathrm{epi}(f)=\left\{(\boldsymbol{x},t)\colon \sum_{i\in[n]}x_iz_i\geq -t,\boldsymbol{x}\in\mathcal{X},\boldsymbol{z}\in\{-1,1\}^n\right\}.$$

Since set $\mathcal{X}$ is compact, thus, we can assume that $\mathcal{X}\subseteq[\boldsymbol{l},\boldsymbol{u}]$, i.e., given $\boldsymbol{x}\in\mathcal{X}$, we have $x_i\in[l_i,u_i]$ for each $i\in[n]$. We can apply the following McCormick inequalities (see more details in Definition 2) to linearize the bilinear term $\{s_i:=x_iz_i\}_{i\in[n]}$, i.e., $(s_i,z_i,x_i)\in\mathcal{MI}(-1,1,l_i,u_i)$ for each $i\in[n]$. Thus, $\mathrm{epi}(f)$ is MICP-R, i.e.,

$$\mathrm{epi}(f)=\left\{(\boldsymbol{x},t)\colon \exists\boldsymbol{s}\in\mathbb{R}^n,\sum_{i\in[n]}s_i\geq -t,\boldsymbol{x}\in\mathcal{X},\boldsymbol{z}\in\{-1,1\}^n,(s_i,z_i,x_i)\in\mathcal{MI}\left(-1,1,l_i,u_i\right),\forall i\in[n]\right\}.$$

Case 2: When $p\in(1,\infty)$, i.e., the norm is neither L_1 nor L_∞, since the set $\mathcal{X}$ is convex, compact, and has a non-empty interior, there exists an open ball $B(\bar{\boldsymbol{x}},r)$ centered at $\bar{\boldsymbol{x}}$ and a positive radius $r>0$ such that $B(\bar{\boldsymbol{x}},r)\subseteq\mathcal{X}$. Therefore, set

$$\mathcal{S}:=\{\boldsymbol{x}\in X:\|\boldsymbol{x}\|_p=\|\bar{\boldsymbol{x}}\|_p\}$$

has a non-empty relative interior. Thus, we can pick a sequence of distinct elements from set $\mathcal{S}$ (e.g., all the possible rational elements) $\{\widehat{\boldsymbol{x}}^j\}_j$. Since $(\widehat{\boldsymbol{x}}^j,\|\bar{\boldsymbol{x}}\|_p)\in\mathrm{epi}(f)$ for each j and function $\|\boldsymbol{x}\|_p$ is strictly convex for any $p\in(1,\infty)$, for any pair (j_1,j_2) with $j_1\neq j_2$, we must have

$$\frac{1}{2}(\widehat{\boldsymbol{x}}^{j_1},\|\bar{\boldsymbol{x}}\|_p)+\frac{1}{2}(\widehat{\boldsymbol{x}}^{j_2},\|\bar{\boldsymbol{x}}\|_p)\notin\mathrm{epi}(f).$$

Since $\{\widehat{\boldsymbol{x}}^j\}_j$ is an infinite sequence, according to Lemma 1, set $\mathrm{epi}(f)$ is not MICP-R.
Case 3: When $p=\infty$, i.e., the norm is L_∞, set $\mathrm{epi}(f)$ reduces to

$$\mathrm{epi}(f)=\left\{(\boldsymbol{x},t)\colon \max_{i\in[n]}|x_i|\geq -t,\forall i\in[n],\boldsymbol{x}\in\mathcal{X}\right\},$$

which can be reformulated in the form of the following disjunction [5]

$$\mathrm{epi}(f)=\bigvee_{i\in[n]}\{(\boldsymbol{x},t)\colon x_i\geq -t,\boldsymbol{x}\in\mathcal{X}\}\bigvee_{i\in[n]}\{(\boldsymbol{x},t)\colon -x_i\geq -t,\boldsymbol{x}\in\mathcal{X}\}.$$

Since set $\mathcal{X}$ is compact, MICP-R of set $\mathrm{epi}(f)$ follows from the well-known results from disjunctive programming [5]. □

B.3 Proof of Proposition 3

Proposition 3 *Suppose that the ambiguity set is $\mathcal{P}=\{\boldsymbol{p}\colon\|\boldsymbol{p}-\boldsymbol{p}^0\|_2\leq\theta,\sum_{i\in[N]}p_i=1,\boldsymbol{p}\geq\boldsymbol{0}\}$, where $\boldsymbol{p}^0=\boldsymbol{e}/N$ denotes the nominal probability. When $0<\theta\leq\sqrt{1/(N(N-1))}$, DFO (2) may not be MICP-R.*

Proof: Let us consider a simple recourse function $Q(\boldsymbol{x}, \boldsymbol{\xi}^i) = x_i$ for each $i \in [N]$ and $\boldsymbol{c} = \boldsymbol{0}$. Then, the DFO (2) is equivalent to

$$v^* = \min_{\boldsymbol{x}\in\mathcal{X}} \min_{\boldsymbol{p}\geq\boldsymbol{0}} \left[\sum_{i\in[N]} p_i x_i \colon \sum_{i\in[N]} p_i = 1, \left\| \boldsymbol{p} - \frac{1}{N}\boldsymbol{e} \right\|_2 \leq \theta \right]. \tag{33a}$$

Let us focus on simplifying the inner minimization of DFO (33a) and define $\boldsymbol{y} = \boldsymbol{x} - (\boldsymbol{x}^\top \boldsymbol{e})\boldsymbol{e}/N$. Then we must have $\sum_{i\in[N]} y_i = 0$ and

$$\sum_{i\in[N]} p_i y_i = \sum_{i\in[N]} p_i x_i - \frac{1}{N} \sum_{i\in[N]} (x_i - 1).$$

The DFO (33a) is equivalent to

$$v^* = \min_{\substack{\boldsymbol{x}\in\mathcal{X},\\ \boldsymbol{y}=\boldsymbol{x}-(\boldsymbol{x}^\top\boldsymbol{e})\boldsymbol{e}/N}} \frac{1}{N} \sum_{i\in[N]} (x_i - 1) + \min_{\boldsymbol{p}\geq\boldsymbol{0}} \left\{ \sum_{i\in[N]} p_i y_i \colon \sum_{i\in[N]} p_i = 1, \left\| \boldsymbol{p} - \frac{1}{N}\boldsymbol{e} \right\|_2 \leq \theta \right\}. \tag{33b}$$

Letting $\widehat{\boldsymbol{p}} = \boldsymbol{p} - \boldsymbol{e}/N$, (33b) is simplified as

$$\min_{\boldsymbol{y}} \min_{\widehat{\boldsymbol{p}}\geq -\boldsymbol{e}/N} \left\{ \sum_{i\in[N]} \widehat{p}_i y_i \colon \sum_{i\in[N]} \widehat{p}_i = 0, \|\widehat{\boldsymbol{p}}\|_2 \leq \theta, \sum_{i\in[N]} y_i = 0 \right\}. \tag{33c}$$

According to the Hölder's inequality, the inner minimization of (33c) can be lower bounded by

$$\min_{\widehat{\boldsymbol{p}}\geq -\boldsymbol{e}/N} \left\{ \sum_{i\in[N]} \widehat{p}_i y_i \colon \sum_{i\in[N]} \widehat{p}_i = 0, \|\widehat{\boldsymbol{p}}\|_2 \leq \theta \right\} \geq -\theta \|\boldsymbol{y}\|_2.$$

In fact, the above equality can be achieved by the solution $\widehat{p}_i^* = -\theta y_i / \|\boldsymbol{y}\|_2$ for all $i \in [N]$. Since $\sum_{i\in[N]} \widehat{p}_i^* = 0$ and $\|\widehat{\boldsymbol{p}}^*\|_2 = \theta$, it suffices to show that

$$-|\widehat{p}_i^*| = -\theta |y_i| / \|\boldsymbol{y}\|_2 \geq -\frac{1}{N}, \tag{33d}$$

for all $\sum_{i\in[N]} \widehat{p}_i^* = 0$. That is, we need to show that

$$\max_{\sum_{i\in[N]} y_i = 0} \frac{|y_i|}{\|\boldsymbol{y}\|_2} \leq \frac{1}{N\theta}. \tag{33e}$$

Without loss of generality, suppose that $y_\ell \neq 0$. Letting $y_i' = y_i / y_\ell$ for each $i \in [N] \setminus \{\ell\}$, the condition $\sum_{i\in[N]} y_i = 0$ is equivalent to $\sum_{i\in[N]} y_i' = -1$. Thus,

$$\max_{\sum_{i\in[N]} y_i = 0} \frac{|y_i|}{\|\boldsymbol{y}\|_2} = \min_{\boldsymbol{y}'} \left\{ \sqrt{1 + \sum_{i\in[N]\setminus\ell} y_i'^2} \colon \sum_{i\in[N]} y_i' = -1 \right\} = \sqrt{1 + \left(\frac{1}{\sqrt{N-1}} \right)^2} = \sqrt{\frac{N}{N-1}}.$$

Hence, the inequality in (33e) must be satisfied since

$$\max_{\sum_{i\in[N]} y_i' = -1} \frac{1}{\sqrt{1 + \sum_{i\in[N]\setminus\ell} y_i'^2}} = \sqrt{\frac{N-1}{N}} \leq \frac{1}{N\theta}$$

and $0 < \theta \leq \sqrt{1/(N(N-1))}$. Therefore, plugging in $\boldsymbol{y} = \boldsymbol{x} - (\boldsymbol{x}^\top \boldsymbol{e})\boldsymbol{e}/N$, the DFO (33c) is equivalent to

$$v^* = \min_{\boldsymbol{x}\in\mathcal{X}} \frac{1}{N}\sum_{i\in[N]}(x_i - 1) - \frac{\theta}{N}\left\|N\boldsymbol{x} - (\boldsymbol{x}^\top \boldsymbol{e})\boldsymbol{e}\right\|_2.$$

Using the fact that set $\mathcal{X}$ has a non-empty interior and following the similar proof as that of Lemma 2, we conclude that this DFO cannot be MICP-R. □

Proofs in Section 5

B.4 Proof of Theorem 5

Theorem 5 *(Confidence Bounds) Suppose that Assumption 2 holds. Then for any given $\gamma \in (0,1)$, we have: (i) $\mathbb{P}^T\{v_W^* \leq v^T + 2L\theta\} \geq 1-\gamma$; and (ii) $\mathbb{P}^T\{v_W^* \geq v^T - L\theta\} \geq 1-\gamma$, where $\theta = \mathcal{O}(1)N^{-1/2}\sqrt{n\log(\gamma^{-1})}$ for a discrete compact set $\mathcal{X}$, and $\theta = \mathcal{O}(1)N^{-1/2}\sqrt{n\log(nN)\log(\gamma^{-1})}$ 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 $\{\boldsymbol{\zeta}^i\}_{i\in[N]}$, which admits the following form

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

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

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

Recall that an optimal $\beta^*(\boldsymbol{x}) = F^{-1}(1-\varepsilon)$, where we let $F(\cdot)$ denotes 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

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

with an optimal $\beta_N^*(\boldsymbol{x}) = F_N^{-1}(1-\epsilon)$, where $F_N(\cdot)$ denote the CDF of random parameter $Q(\boldsymbol{x},\tilde{\boldsymbol{\zeta}})$ with respect to empirical distribution $\mathbb{P}_{\tilde{\boldsymbol{\zeta}}}$.

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

$$\mathbb{P}^T\left\{F_N\left(\beta^*(\boldsymbol{x}) + \widehat{\Delta}_N\right) - F\left(\beta^*(\boldsymbol{x}) + \widehat{\Delta}_N\right) \geq -\bar{\Delta}\right\} \geq 1 - \exp\{-2N\bar{\Delta}^2\}. \tag{34a}$$

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, the inequality (34a) 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

$$\begin{aligned}
&\mathbb{P}^T\left\{|\beta_N^*(\boldsymbol{x})-\beta^*(\boldsymbol{x})|\leq\widehat{\Delta}_N\right\}=\mathbb{P}^T\left\{\beta_N^*(\boldsymbol{x})\leq\beta^*(\boldsymbol{x})+\widehat{\Delta}_N,\beta_N^*(\boldsymbol{x})\geq\beta^*(\boldsymbol{x})-\widehat{\Delta}_N\right\}\\
=&\mathbb{P}^T\left\{F_N\left(\beta^*(\boldsymbol{x})+\Delta\right)\geq1-\varepsilon,F_N\left(\beta^*(\boldsymbol{x})-\Delta\right)\leq1-\varepsilon\right\}\\
\geq&1-\mathbb{P}^T\left\{F_N\left(\beta^*(\boldsymbol{x})+\widehat{\Delta}_N\right)<1-\varepsilon\right\}-\mathbb{P}^T\left\{F_N\left(\beta^*(\boldsymbol{x})-\widehat{\Delta}_N\right)>1-\varepsilon\right\}\\
\geq&1-2\exp\left\{-2N(\ell\widehat{\Delta}_N)^2\right\}.
\end{aligned}\tag{34b}$$

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

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

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

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

Then, it is sufficient to prove

$$\mathbb{P}^T\left\{v_N^{SAA}\leq v^T+L\theta\right\}\geq1-\gamma.$$

Step III. Given that the quantile is close to the true quantile (i.e., the inequalities from Step I hold), we will 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

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

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

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

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

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

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

$$\begin{aligned}&\beta_N^*(\boldsymbol{x})-\beta^*(\boldsymbol{x})+\frac{1}{N-N\varepsilon}\sum_{i\in[N]}\left[[Q(\boldsymbol{x},\boldsymbol{\zeta}^i)-\beta_N^*(\boldsymbol{x})]_- - \mathbb{E}_{\mathbb{P}^T}[Q(\boldsymbol{x},\tilde{\boldsymbol{\xi}})-\beta^*(\boldsymbol{x})]_-\right]\\ \leq&\frac{1}{N-N\varepsilon}\sum_{i\in[N]}\left[[Q(\boldsymbol{x},\boldsymbol{\zeta}^i)-\beta^*(\boldsymbol{x})]_- - \mathbb{E}_{\mathbb{P}^T}[Q(\boldsymbol{x},\tilde{\boldsymbol{\xi}})-\beta^*(\boldsymbol{x})]_-\right]+\frac{\widehat{\Delta}_N}{1-\varepsilon}.\end{aligned}\tag{34c}$$

Now, we are going to apply lemma A.1 in [24] 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,

$$\frac{\widehat{\Delta}_N}{1-\varepsilon}=\frac{\lambda_1\sigma}{(1-\varepsilon)\sqrt{N}}=\frac{\lambda_2\sigma}{2\sqrt{N}}.\tag{34d}$$

According to equation (34d), we have

$$\mathbb{P}^T\left\{v_N(\boldsymbol{x})-\frac{\lambda_2\sigma}{\sqrt{N}}\leq v^T(\boldsymbol{x})\right\}=\mathbb{P}^T\left\{v_N(\boldsymbol{x})-\frac{\lambda_1\sigma}{(1-\varepsilon)\sqrt{N}}-\frac{\widehat{\Delta}_N}{1-\varepsilon}\leq v^T(\boldsymbol{x})\right\}.\tag{34e}$$

Invoking the definition of $v^T(\boldsymbol{x})$ and $v_N(\boldsymbol{x})$, we can rewrite (34e) as

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

By the law of total probability (see, e.g., appendix A of [67]), we have

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

According to the inequality (34c), we have

$$\mathbb{P}^T\left\{v_N(\boldsymbol{x})-\frac{\lambda_2\sigma}{\sqrt{N}}\leq v^T(\boldsymbol{x})\right\}$$

$$\geq \mathbb{P}^T\left\{\frac{1}{N-N\varepsilon}\sum_{i\in[N]}\left[[Q(\boldsymbol{x},\boldsymbol{\zeta}^i)-\beta^*(\boldsymbol{x})]_- - \mathbb{E}_{\mathbb{P}^T}[Q(\boldsymbol{x},\tilde{\boldsymbol{\xi}})-\beta^*(\boldsymbol{x})]_-\right] \leq \frac{\lambda_1\sigma}{(1-\varepsilon)\sqrt{N}}, |\beta_N^*(\boldsymbol{x})-\beta^*(\boldsymbol{x})| \leq \widehat{\Delta}_N\right\}$$

$$\geq \mathbb{P}^T\left\{\frac{1}{N}\sum_{i\in[N]}\left[[Q(\boldsymbol{x},\boldsymbol{\zeta}^i)-\beta^*(\boldsymbol{x})]_- - \mathbb{E}_{\mathbb{P}^T}[Q(\boldsymbol{x},\tilde{\boldsymbol{\xi}})-\beta^*(\boldsymbol{x})]_-\right] \leq \frac{\lambda_1\sigma}{\sqrt{N}}\right\} + \mathbb{P}^T\left\{|\beta_N^*(\boldsymbol{x})-\beta^*(\boldsymbol{x})| \leq \widehat{\Delta}_N\right\} - 1,$$

where the second equality is due to the union bound (see, e.g., [10]).

Defining $c^i = [Q(\boldsymbol{x},\boldsymbol{\zeta}^i)-\beta^*(\boldsymbol{x})]_-$ and $c^T = \mathbb{E}_{\mathbb{P}^T}[Q(\boldsymbol{x},\tilde{\boldsymbol{\xi}})-\beta^*(\boldsymbol{x})]_-$ and applying lemma A.1 in [24] with $d_i = c^i - c^T$ for each $i \in [N]$, together with the inequalities in (34b), 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. If 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 $\log(2/\gamma)/(2(\ell\Delta_N)^2)$.

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

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

By simple calculation, we have

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

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

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

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 of the 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

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

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

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

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

where the first inequality is due to Part (ii) of Assumption 2 and discretization, the second one is based on $\beta^*(\widehat{\boldsymbol{x}})<\beta^*(\widehat{\boldsymbol{y}})$, the third 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

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

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

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

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

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

and we have

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

Letting $\lambda_2\sigma/\sqrt{N}=L\nu(1+\sqrt[q]{n})(1-\varepsilon)$ and setting $\theta:=4\lambda_1\sigma L^{-1}N^{-1/2}(1-\varepsilon)^{-1}=\mathcal{O}(1)N^{-1/2}\sqrt{n\log(nN)\log(\gamma^{-1})}$, we arrive at the conclusion. □

Appendix C. Tractability of DFO

Definition 5 (Tractability, [6]) Suppose that for any given compact set $\mathcal{X} \subseteq \mathbb{R}^n$, which has a non-empty interior, and is contained in a Euclidean ball with radius R and is containing a Euclidean ball with radius r. Then there exists an efficient algorithm to solve the favorable problem $\min_{\boldsymbol{x}\in\mathcal{X}} \inf_{\mathbb{P}\in\mathcal{P}} \mathbb{E}_{\mathbb{P}}[Q(\boldsymbol{x}, \tilde{\boldsymbol{\xi}})]$ to $\widehat{\varepsilon} > 0$ accuracy, whose running time is polynomial in $n, m, \ln(R/r), \ln(1/\widehat{\varepsilon})$, and the encoding length of $\min_{\boldsymbol{x}\in\mathcal{X}} \inf_{\mathbb{P}\in\mathcal{P}} \mathbb{E}_{\mathbb{P}}[Q(\boldsymbol{x}, \tilde{\boldsymbol{\xi}})]$.

Appendix D. More Results on Complexity Analysis, and Mixed-Integer Convex Programming Representability of DFO

In this section, similar to the discussions in [6, 20, 71], we focus on the recourse function $Q(\boldsymbol{x}, \boldsymbol{\xi})$ being convex or concave piecewise affine in $\boldsymbol{x}$, respectively.

D.1 rDFO (3) with Concave Piecewise Affine Functions

We first consider that the function $\boldsymbol{c}^\top \boldsymbol{x} + Q(\boldsymbol{x}, \boldsymbol{\xi})$ is the minimum of K piecewise affine functions $\boldsymbol{\xi}^\top \boldsymbol{a}_k(\boldsymbol{x}) + b_k(\boldsymbol{x})$ with affine mappings $\boldsymbol{a}_k(\boldsymbol{x}) = \widehat{\boldsymbol{A}}_k \boldsymbol{x} + \widehat{\boldsymbol{a}}_k \in \mathbb{R}^m$ with $\widehat{\boldsymbol{A}}_k \in \mathbb{R}^{m\times n}, \widehat{\boldsymbol{a}}_k \in \mathbb{R}^m$ and $b_k(\boldsymbol{x}) = \widehat{\boldsymbol{B}}_k^\top \boldsymbol{x} + \widehat{b}_k \in \mathbb{R}$ with $\widehat{\boldsymbol{B}}_k \in \mathbb{R}^n, \widehat{b}_k \in \mathbb{R}$ for each $k \in [K]$, i.e., $\boldsymbol{c}^\top \boldsymbol{x} + Q(\boldsymbol{x}, \boldsymbol{\xi}) = \min_{k\in[K]} [\boldsymbol{\xi}^\top \boldsymbol{a}_k(\boldsymbol{x}) + b_k(\boldsymbol{x})]$. Suppose that the uncertainty set is defined as $\mathcal{U} = \{\boldsymbol{\xi} : \|\boldsymbol{\xi} - \boldsymbol{\xi}^0\|_p \leq \theta\}$ with the known parameter $\boldsymbol{\xi}^0$ and $\theta \geq 0$. It is worthy of mentioning that for more general uncertainty sets such as polyhedral uncertainty set and ellipsoidal uncertainty set (see, e.g., [6]), the reformulation and complexity analyses can be simply extended. In this section, for the sake of page limit, we focus on the most common uncertainty set $\mathcal{U}$.

Under this setting, the rDFO (3) can be recast as

$$v^* = \min_{\boldsymbol{x}\in\mathcal{X}} \min_{\boldsymbol{\xi}\in\mathcal{U}} \min_{k\in[K]} \left\{\boldsymbol{\xi}^\top \boldsymbol{a}_k(\boldsymbol{x}) + b_k(\boldsymbol{x})\right\}. \tag{35}$$

Switching the second minimum operator with the third one and invoking the definition of dual norm, problem (35) is further equivalent to

$$v^* = \min_{\boldsymbol{x}\in\mathcal{X}} \min_{k\in[K]} \left\{{\boldsymbol{\xi}^0}^\top \boldsymbol{a}_k(\boldsymbol{x}) + b_k(\boldsymbol{x}) - \theta\|\boldsymbol{a}_k(\boldsymbol{x})\|_{p^*}\right\},$$

which can be solved by selecting the lowest objective value within these K mathematical programs, that is,

$$v^* = \min_{k\in[K]} \left\{ v_k^* := \min_{\boldsymbol{x}\in\mathcal{X}} \left\{{\boldsymbol{\xi}^0}^\top \boldsymbol{a}_k(\boldsymbol{x}) + b_k(\boldsymbol{x}) - \theta\|\boldsymbol{a}_k(\boldsymbol{x})\|_{p^*}\right\}\right\}. \tag{36}$$

Note that the inner minimization of the rDFO (36) is a concave minimization problem and, in general, can be difficult. However, by exploring the properties of the dual norm, there are some conditions under which the rDFO (36) can be tractable.

Theorem 8 *The rDFO* (36) *can be tractable if either condition holds:*

(i) If $\|\boldsymbol{a}_k(\boldsymbol{x})\|_{p^*} := C_k$ is constant for each $k \in [K]$ and $\boldsymbol{x} \in \mathcal{X}$, the rDFO (36) *is equivalent to solving K tractable convex programs, i.e., $v^* = \min_{k\in[K]} v_k^*$, where for each $k \in [K]$, we have*

$$v_k^* = \min_{\boldsymbol{x}\in\mathcal{X}} \left\{{\boldsymbol{\xi}^0}^\top \boldsymbol{a}_k(\boldsymbol{x}) + b_k(\boldsymbol{x}) - \theta C_k\right\};$$

(ii) If $p = 1$, the rDFO (36) *is equivalent to solving $2mK$ tractable convex programs, i.e., $v^* = \min_{k\in[K], i\in[m], \ell\in[2]} v_{ik\ell}^*$, where for each $k \in [K]$ and $i \in [m]$, we have*

$$v_{ik1}^* = \min_{\boldsymbol{x}\in\mathcal{X}} \left\{{\boldsymbol{\xi}^0}^\top \boldsymbol{a}_k(\boldsymbol{x}) + b_k(\boldsymbol{x}) + \theta a_{ki}(\boldsymbol{x})\right\}, \quad v_{ik2}^* = \min_{\boldsymbol{x}\in\mathcal{X}} \left\{{\boldsymbol{\xi}^0}^\top \boldsymbol{a}_k(\boldsymbol{x}) + b_k(\boldsymbol{x}) - \theta a_{ki}(\boldsymbol{x})\right\}.$$

Proof: We split the proof into two parts by checking these two conditions separately.

(i) When $\|\boldsymbol{a}_k(\boldsymbol{x})\|_{p^*}$ is a constant for each $k \in [K]$, i.e., $\|\boldsymbol{a}_k(\boldsymbol{x})\|_{p^*} = C_k$ for $k \in [K]$, then the objective function of the rDFO (36) is linear and optimizing it is equivalent to solving K convex programs;

(ii) When $p = 1$, i.e., when the dual norm is L_∞, then $\theta\|\boldsymbol{a}_k(\boldsymbol{x})\|_\infty = \theta \max_{i\in[m]} \max\{a_{ki}(\boldsymbol{x}), -a_{ki}(\boldsymbol{x})\}$. That is, the rDFO (36) can be simplified as

$$\min_{k\in[K]} \min_{i\in[m]} \min \left\{ \min_{\boldsymbol{x}\in\mathcal{X}} \left\{ {\boldsymbol{\xi}^0}^\top \boldsymbol{a}_k(\boldsymbol{x}) + b_k(\boldsymbol{x}) + \theta a_{ki}(\boldsymbol{x}) \right\}, \min_{\boldsymbol{x}\in\mathcal{X}} \left\{ {\boldsymbol{\xi}^0}^\top \boldsymbol{a}_k(\boldsymbol{x}) + b_k(\boldsymbol{x}) - \theta a_{ki}(\boldsymbol{x}) \right\} \right\},$$

which is equivalent to solving $2mK$ convex programs and selecting the best one with the lowest optimal value.

□

In the following complexity analysis, we focus on the non-trivial cases where $\|\boldsymbol{a}_k(\boldsymbol{x})\|_{p^*}$ is not a constant for some $k \in [K]$. Unfortunately, when $p \in (1, \infty]$, solving the rDFO (36), in general, is NP-hard with reduction to the well-known NP-hard problem – maximizing a norm over a polytope.

Proposition 8 *For any $p \in (1, \infty]$, solving the rDFO* (36)*, in general, is NP-hard even with $K = 1$.*

Proof: Let us consider an NP-hard problem — Norm maximization over a polytope (see theorem 1 in [22]), which asks

> ***Norm maximization over a polytope.*** *Given the polytope $\{\boldsymbol{x} : \boldsymbol{D}\boldsymbol{x} \le \boldsymbol{d}\}$, where $\boldsymbol{D} \in \mathbb{R}^{\tau\times n}$ and $\boldsymbol{d} \in \mathbb{R}^\tau$, what is the optimal value of the problem $\max_{\boldsymbol{x}} \{\|\boldsymbol{x}\|_{p^*} : \boldsymbol{D}\boldsymbol{x} \le \boldsymbol{d}\}$ with $p \in (1, \infty]$?*

Consider a special case of the rDFO (36), where $\boldsymbol{a}_k(\boldsymbol{x}) = \boldsymbol{x}$ and $b_k(\boldsymbol{x}) = 0$ for each $k \in [K]$ and $\boldsymbol{\xi}^0 = \boldsymbol{0}$, $\theta = 1$, and set $\mathcal{X} = \{\boldsymbol{x} : \boldsymbol{D}\boldsymbol{x} \le \boldsymbol{d}\}$. In this case, the rDFO (36) can be written as

$$\max_{\boldsymbol{x}} \{\|\boldsymbol{x}\|_{p^*} : \boldsymbol{D}\boldsymbol{x} \le \boldsymbol{d}\},$$

which is exactly the desirable norm maximization problem over a polytope for any $p \in (1, \infty]$. Thus, solving the rDFO (36), in general, is NP-hard for any $p \in (1, \infty]$. □

The complexity result suggests that the tractable result in Theorem 8 with $p = 1$ is the best that one can expect.

Next, for the intractable case, we study the MICP-R of the objective function of the rDFO (36). As an extension of Lemma 2, we notice that when $p = \infty$, the objective function of the rDFO (36) is MICP-R; otherwise, when $p \in (1, \infty)$, it is not.

Theorem 9 *When $p = \infty$, the objective function of the rDFO* (36) *with domain $\mathcal{X}$ is MICP-R; otherwise, when $p \in (1, \infty)$, the objective function of the rDFO* (36) *with domain $\mathcal{X}$ may not be MICP-R.*

Proof: Let us write the objective function of the rDFO (36) with domain $\mathcal{X}$ as $f_k(\boldsymbol{x}) = {\boldsymbol{\xi}^0}^\top \boldsymbol{a}_k(\boldsymbol{x}) + b_k(\boldsymbol{x}) - \|\boldsymbol{a}_k(\boldsymbol{x})\|_{p^*} + \chi_{\mathcal{X}}(\boldsymbol{x})$ for each $k \in [K]$. We focus on the MICP-R of the epigraph of function $f_k(\cdot)$, which reads as

$$\operatorname{epi}(f_k) = \left\{ (\boldsymbol{x}, t) : {\boldsymbol{\xi}^0}^\top \boldsymbol{a}_k(\boldsymbol{x}) + b_k(\boldsymbol{x}) - \|\boldsymbol{a}_k(\boldsymbol{x})\|_{p^*} \le t, \boldsymbol{x} \in \mathcal{X} \right\}. \tag{37}$$

Next, we split the proof into two cases based on the choice of p.

Case 1: When $p = \infty$, i.e., the dual norm is L_1, set $\operatorname{epi}(f_k)$ can be written as

$$\operatorname{epi}(f_k) = \left\{ (\boldsymbol{x}, t) : \sum_{i\in[m]} |a_{ki}(\boldsymbol{x})| - {\boldsymbol{\xi}^0}^\top \boldsymbol{a}_k(\boldsymbol{x}) - b_k(\boldsymbol{x}) \ge -t, \boldsymbol{x} \in \mathcal{X} \right\}.$$

Then we have

$$\mathrm{epi}(f_k) = \left\{ (\boldsymbol{x}, t) \colon \max_{\boldsymbol{z} \in \{-1,1\}^m} \sum_{i \in [m]} a_{ki}(\boldsymbol{x}) z_i - {\boldsymbol{\xi}^0}^\top \boldsymbol{a}_k(\boldsymbol{x}) - b_k(\boldsymbol{x}) \geq -t, \boldsymbol{x} \in \mathcal{X} \right\},$$

which is equivalent to

$$\mathrm{epi}(f_k) = \left\{ (\boldsymbol{x}, t) \colon \sum_{i \in [m]} a_{ki}(\boldsymbol{x}) z_i - {\boldsymbol{\xi}^0}^\top \boldsymbol{a}_k(\boldsymbol{x}) - b_k(\boldsymbol{x}) \geq -t, \boldsymbol{x} \in \mathcal{X}, \boldsymbol{z} \in \{-1,1\}^m \right\}.$$

Since set $\mathcal{X}$ is compact, we can apply the McCormick inequalities (see Definition 2) to linearize the bilinear terms $\{a_{ki}(\boldsymbol{x}) z_i\}_{i \in [m]}$. Thus, $\mathrm{epi}(f_k)$ is MICP-R for each $k \in [K]$.

Case 2: When $p \in (1, \infty)$, i.e., when the dual norm is neither L_1 nor L_∞, suppose that $K = 1$, $\boldsymbol{\xi}^0 = \boldsymbol{0}$, $\boldsymbol{a}_1(\boldsymbol{x}) = \boldsymbol{x}$, and $b_1(\boldsymbol{x}) = 0$, then set (37) reduces to

$$\mathrm{epi}(f_1) = \{ (\boldsymbol{x}, t) \colon -\|\boldsymbol{x}\|_{p^*} \leq t, \boldsymbol{x} \in \mathcal{X} \},$$

which is identical to (32a). According to the result in Lemma 2, when $p \in (1, \infty)$, the objective function of the rDFO (36) with domain $\mathcal{X}$, in general, may not be MICP-R. □

Theorem 9 suggests that the objective function of the rDFO (36) with domain $\mathcal{X}$ may not be MICP-R with a general norm, but it is MICP-R when the norm is L_∞. As a direct corollary of Theorem 9, when $p = \infty$, the MICP-R formulation of the rDFO (36) can be summarized as follows.

Corollary 6 *When $p = \infty$, suppose that $\mathcal{X} \subseteq [\boldsymbol{l}, \boldsymbol{u}]$ and let $\widehat{l}_{ki} = \sum_{j \in [n]} \min\{\widehat{A}_{kij} l_j, \widehat{A}_{kij} u_j\} + \widehat{a}_{ki}$ and $\widehat{u}_{ki} = \sum_{j \in [n]} \max\{\widehat{A}_{kij} l_j, \widehat{A}_{kij} u_j\} + \widehat{a}_{ki}$ for each $i \in [m]$ such that $\boldsymbol{a}_k(\boldsymbol{x}) \in [\widehat{\boldsymbol{l}}_k, \widehat{\boldsymbol{u}}_k]$ for each $k \in [K]$. Then, the rDFO (36) is equivalent to solving the following K MICPs, i.e., $v^* = \min_{k \in [K]} v_k^*$, where for each $k \in [K]$, we have*

$$v_k^* = \min_{\boldsymbol{x} \in \mathcal{X}} \left\{ {\boldsymbol{\xi}^0}^\top \boldsymbol{a}_k(\boldsymbol{x}) + b_k(\boldsymbol{x}) - \theta \sum_{i \in [m]} s_{ki} : (s_{ki}, z_{ki}, a_{ki}(\boldsymbol{x})) \in \mathcal{MI}\left(-1, 1, \widehat{l}_{ki}, \widehat{u}_{ki}\right), \forall i \in [m] \right\}.$$

D.2 rDFO (3) with Convex Piecewise Affine Functions

In this subsection, we follow the same notation and uncertainty set as the previous subsection and consider the maximum of piecewise affine function, that is, the function $\boldsymbol{c}^\top \boldsymbol{x} + Q(\boldsymbol{x}, \boldsymbol{\xi})$ is defined as the maximum of K piecewise affine function $\boldsymbol{\xi}^\top \boldsymbol{a}_k(\boldsymbol{x}) + b_k(\boldsymbol{x})$, i.e., $\boldsymbol{c}^\top \boldsymbol{x} + Q(\boldsymbol{x}, \boldsymbol{\xi}) = \max_{k \in [K]} [\boldsymbol{\xi}^\top \boldsymbol{a}_k(\boldsymbol{x}) + b_k(\boldsymbol{x})]$. Under this setting, the rDFO (3) can be recast as

$$v^* = \min_{\boldsymbol{x} \in \mathcal{X}} \min_{\boldsymbol{\xi} \in \mathcal{U}} \max_{k \in [K]} \left\{ \boldsymbol{\xi}^\top \boldsymbol{a}_k(\boldsymbol{x}) + b_k(\boldsymbol{x}) \right\}. \tag{38}$$

Let us first provide an equivalent reformulation of the rDFO (38), which inspires us to study the tractability and MICP-R of the rDFO (38).

Lemma 3 *The rDFO* (38) *is equivalent to*

$$v^* = \min_{\boldsymbol{x} \in \mathcal{X}} \max_{\boldsymbol{\lambda} \geq \boldsymbol{0}} \left\{ \sum_{k \in [K]} \lambda_k \left[{\boldsymbol{\xi}^0}^\top \boldsymbol{a}_k(\boldsymbol{x}) + b_k(\boldsymbol{x}) \right] - \theta \left\| \sum_{k \in [K]} \lambda_k \boldsymbol{a}_k(\boldsymbol{x}) \right\|_{p^*} : \sum_{k \in [K]} \lambda_k = 1 \right\}. \tag{39}$$

Proof: Let us first consider the inner minimax of the rDFO (38) as

$$\min_{\boldsymbol{\xi}} \left\{ \max_{k\in[K]} \boldsymbol{\xi}^\top \boldsymbol{a}_k(\boldsymbol{x}) + b_k(\boldsymbol{x}) \colon \|\boldsymbol{\xi} - \boldsymbol{\xi}^0\|_p \le \theta \right\}.$$

Introducing auxiliary nonnegative variables $\boldsymbol{\lambda}$, the inner minimax of the rDFO (38) is equivalent to

$$\min_{\boldsymbol{\xi}} \left\{ \max_{\boldsymbol{\lambda}\ge\mathbf{0}} \left\{ \sum_{k\in[K]} \lambda_k \left[\boldsymbol{\xi}^\top \boldsymbol{a}_k(\boldsymbol{x}) + b_k(\boldsymbol{x})\right] : \sum_{k\in[K]} \lambda_k = 1 \right\} : \|\boldsymbol{\xi} - \boldsymbol{\xi}^0\|_p \le \theta \right\}.$$

According to Sion's minimax theorem [64], we can interchange the maximum operator with the minimum one as

$$\max_{\boldsymbol{\lambda}\ge\mathbf{0}} \left\{ \min_{\boldsymbol{\xi}} \left\{ \sum_{k\in[K]} \lambda_k \left[\boldsymbol{\xi}^\top \boldsymbol{a}_k(\boldsymbol{x}) + b_k(\boldsymbol{x})\right] : \|\boldsymbol{\xi} - \boldsymbol{\xi}^0\|_p \le \theta \right\} : \sum_{k\in[K]} \lambda_k = 1 \right\}.$$

Invoking the definition of dual norm, we have

$$\max_{\boldsymbol{\lambda}\ge\mathbf{0}} \left\{ \sum_{k\in[K]} \lambda_k \left[{\boldsymbol{\xi}^0}^\top \boldsymbol{a}_k(\boldsymbol{x}) + b_k(\boldsymbol{x})\right] - \theta \left\| \sum_{k\in[K]} \lambda_k \boldsymbol{a}_k(\boldsymbol{x}) \right\|_{p^*} : \sum_{k\in[K]} \lambda_k = 1 \right\}.$$

This completes the proof. □

Due to the bilinear terms in the reformulation (39), the rDFO (38), in general, can be difficult to solve. However, by exploring the objective function and the properties of the dual norm, we are able to prove conditions under which the rDFO (38) can be tractable.

Theorem 10 *The rDFO* (38) *can be tractable if any of the following conditions holds:*

(i) When $\boldsymbol{a}_k(\boldsymbol{x}) := \bar{\boldsymbol{a}}_k$ is constant for all $k \in [K]$ and $\boldsymbol{x} \in \mathcal{X}$, the rDFO (38) *is equivalent to*

$$v^* = \min_{\boldsymbol{x}\in\mathcal{X},\boldsymbol{\xi},\eta} \left\{ \eta \colon \eta \ge {\boldsymbol{\xi}}^\top \bar{\boldsymbol{a}}_k + b_k(\boldsymbol{x}), \forall k \in [K], \|\boldsymbol{\xi} - \boldsymbol{\xi}^0\|_p \le \theta \right\};$$

(ii) When $p = 1$ and $a_k(\boldsymbol{x}) = a_1(\boldsymbol{x})$ for each $k \in [K]$, the rDFO (38) *is equivalent to solving $2m$ tractable convex programs, and selecting the lowest optimal value, i.e., $v^* = \min_{i\in[m],\ell\in[2]} v^*_{i\ell}$, where for each $i \in [m]$, we have*

$$v^*_{i1} = \min_{\boldsymbol{x}\in\mathcal{X},\eta} \left\{ \eta \colon \eta \ge b_k(\boldsymbol{x}) + {\boldsymbol{\xi}^0}^\top \boldsymbol{a}_1(\boldsymbol{x}) - \theta a_{1i}(\boldsymbol{x}), \forall k \in [K] \right\},$$
$$v^*_{i2} = \min_{\boldsymbol{x}\in\mathcal{X},\eta} \left\{ \eta \colon \eta \ge b_k(\boldsymbol{x}) + {\boldsymbol{\xi}^0}^\top \boldsymbol{a}_1(\boldsymbol{x}) + \theta a_{1i}(\boldsymbol{x}), \forall k \in [K] \right\};$$

(iii) Suppose that $p = 1$, and $\boldsymbol{\xi} := [\boldsymbol{\xi}_1, \ldots, \boldsymbol{\xi}_K]$ such that $\boldsymbol{\xi}^i$ and $\boldsymbol{\xi}^j$ do not overlap for each $i \ne j$, and $\boldsymbol{\xi}^\top \boldsymbol{a}_k(\boldsymbol{x}) = {\boldsymbol{\xi}^k}^\top \bar{\boldsymbol{a}}_k(\boldsymbol{x})$ for each $k \in [K]$ such that $\|\bar{\boldsymbol{a}}_k(\boldsymbol{x})\|_\infty = C_k$ is constant for each $k \in [K]$, where $\boldsymbol{\xi}_k \in \mathbb{R}^{m_k}$, $\bar{\boldsymbol{a}}_k(\boldsymbol{x}) = \bar{\boldsymbol{A}}_k \boldsymbol{x} + \bar{\boldsymbol{a}}_k \in \mathbb{R}^{m_k}$ with $\bar{\boldsymbol{A}}_k \in \mathbb{R}^{m_k\times n}, \bar{\boldsymbol{a}}_k \in \mathbb{R}^{m_k}$ such that $\bar{\boldsymbol{A}}_i$ and $\bar{\boldsymbol{A}}_j$, $\bar{\boldsymbol{a}}_i$ and $\bar{\boldsymbol{a}}_j$ do not overlap each $i \ne j$ with $\sum_{k\in[K]} m_k = m$ and each m_k is nonnegative. Then, the rDFO (38) *is equivalent to solving*

$$v^* = \min_{\boldsymbol{x}\in\mathcal{X},\beta,\boldsymbol{\gamma}\ge\mathbf{0}} \left\{ \beta \colon \sum_{k\in[K]} \gamma_k = \theta, \beta \ge {\boldsymbol{\xi}_k^0}^\top \bar{\boldsymbol{a}}_k(\boldsymbol{x}) + b_k(\boldsymbol{x}) - \gamma_k C_k, \forall k \in [K] \right\}.$$

Proof: We split the proof into three parts accordingly.

(i) When $\boldsymbol{a}_k(\boldsymbol{x}) = \bar{\boldsymbol{a}}_k$ is constant for all $k \in [K]$, the rDFO (38) can be written as

$$v^* = \min_{\boldsymbol{x}\in\mathcal{X}} \min_{\boldsymbol{\xi}} \max_{k\in[K]} \left\{ \boldsymbol{\xi}^\top \bar{\boldsymbol{a}}_k + b_k(\boldsymbol{x}) \colon \|\boldsymbol{\xi} - \boldsymbol{\xi}^0\|_p \le \theta \right\}.$$

Introducing an auxiliary variable η to linearize the inner maximum, we arrive at Part (i).

(ii) When $\boldsymbol{a}_k(\boldsymbol{x}) = \boldsymbol{a}_1(\boldsymbol{x})$ for each $k \in [K]$, we rewrite the rDFO (39) as

$$v^* = \min_{\boldsymbol{x}\in\mathcal{X}} \max_{\boldsymbol{\lambda}\ge\boldsymbol{0}} \left\{ \sum_{k\in[K]} \lambda_k b_k(\boldsymbol{x}) + {\boldsymbol{\xi}^0}^\top \boldsymbol{a}_1(\boldsymbol{x}) - \theta \|\boldsymbol{a}_1(\boldsymbol{x})\|_{p^*} \colon \sum_{k\in[K]} \lambda_k = 1 \right\}.$$

Taking the dual of the inner maximization problem and using strong duality from linear programming, we have

$$v^* = \min_{\boldsymbol{x}\in\mathcal{X},\eta} \left\{ \eta \colon \eta \ge b_k(\boldsymbol{x}) + {\boldsymbol{\xi}^0}^\top \boldsymbol{a}_1(\boldsymbol{x}) - \theta \|\boldsymbol{a}_1(\boldsymbol{x})\|_{p^*}, \forall k \in [K] \right\}.$$

When $p = 1$, i.e., when the dual norm is L_∞, then $\theta\|\boldsymbol{a}_1(\boldsymbol{x})\|_\infty = \theta \max_{i\in[m]} \max\{a_{1i}(\boldsymbol{x}), -a_{1i}(\boldsymbol{x})\}$. Therefore, the rDFO (39) is equivalent to solving $2m$ tractable problems, and selecting the best one with the lowest optimal value, i.e., $v^* = \min_{i\in[m],\ell\in[2]} v^*_{i\ell}$, where for each $i \in [m]$, we have

$$v^*_{i1} = \min_{\boldsymbol{x}\in\mathcal{X},\eta} \left\{ \eta \colon \eta \ge b_k(\boldsymbol{x}) + {\boldsymbol{\xi}^0}^\top \boldsymbol{a}_1(\boldsymbol{x}) - \theta a_{1i}(\boldsymbol{x}), \forall k \in [K] \right\},$$
$$v^*_{i2} = \min_{\boldsymbol{x}\in\mathcal{X},\eta} \left\{ \eta \colon \eta \ge b_k(\boldsymbol{x}) + {\boldsymbol{\xi}^0}^\top \boldsymbol{a}_1(\boldsymbol{x}) + \theta a_{1i}(\boldsymbol{x}), \forall k \in [K] \right\}.$$

(iii) Since $\boldsymbol{\xi}^i$ and $\boldsymbol{\xi}^j$ do not overlap for each $i \ne j$ and $\bar{\boldsymbol{a}}_j(\boldsymbol{x})$ and $\bar{\boldsymbol{a}}_j(\boldsymbol{x})$ do not overlap for each $i \ne j$ as well, when $p = 1$ and $\|\bar{\boldsymbol{a}}_k(\boldsymbol{x})\|_\infty = C_k$ is constant for each $k \in [K]$, the dual norm term in (39) can be simplified as

$$\left\| \sum_{k\in[K]} \lambda_k \boldsymbol{a}_k(\boldsymbol{x}) \right\|_\infty = \max_{k\in[K]} \lambda_k \left\| \bar{\boldsymbol{a}}_k(\boldsymbol{x}) \right\|_\infty = \max_{k\in[K]} \lambda_k C_k.$$

Then, the rDFO (39) can be written as

$$\min_{\boldsymbol{x}\in\mathcal{X}} \max_{\boldsymbol{\lambda}\ge\boldsymbol{0}} \left\{ \sum_{k\in[K]} \lambda_k \left[{\boldsymbol{\xi}_k^0}^\top \bar{\boldsymbol{a}}_k(\boldsymbol{x}) \right] + \sum_{k\in[K]} \lambda_k b_k(\boldsymbol{x}) - \theta \max_{k\in[K]} \lambda_k C_k \colon \sum_{k\in[K]} \lambda_k = 1 \right\}.$$

Introducing one variable η to linearize the term $\max_{k\in[K]} \lambda_k C_k$, then we have

$$\max_{\boldsymbol{\lambda}\ge\boldsymbol{0},\eta} \left\{ \sum_{k\in[K]} \lambda_k \left[{\boldsymbol{\xi}_k^0}^\top \bar{\boldsymbol{a}}_k(\boldsymbol{x}) \right] + \sum_{k\in[K]} \lambda_k b_k(\boldsymbol{x}) - \theta\eta \colon \sum_{k\in[K]} \lambda_k = 1, \lambda_k C_k - \eta \le 0, \forall k \in [K] \right\}.$$

Taking the dual of the inner maximization problem with dual variables $\beta, \boldsymbol{\gamma}$ and using strong duality from linear programming, we have

$$\min_{\beta,\boldsymbol{\gamma}\ge\boldsymbol{0}} \left\{ \beta \colon \sum_{k\in[K]} \gamma_k = \theta, \beta \ge {\boldsymbol{\xi}_k^0}^\top \bar{\boldsymbol{a}}_k(\boldsymbol{x}) + b_k(\boldsymbol{x}) - \gamma_k C_k, \forall k \in [K] \right\}.$$

This completes the proof. □

It turns out that the results in Theorem 10 may be the best ones that we could expect. In general, solving the rDFO (38) is NP-hard for any convex L_p norm.

Proposition 9 *For any $p \in [1, \infty]$, solving the rDFO* (38)*, in general, is NP-hard even with $K = 1$.*

Proof: Note that when $K = 1$, the regular FO (38) is equivalent to formulation (36). Thus, the complexity results in Proposition 8 hold, i.e., solving the regular FO (38) is, in general, NP-hard for $p \in (1, \infty]$.

It remains to show that solving the regular FO (38) is also NP-hard when $p = 1$.

Let us consider the NP-complete problem — feasibility problem of a general binary program, which asks

Feasibility of a binary program. *Given an integer matrix $\boldsymbol{D} \in \mathbb{Z}^{\tau \times K}$, and integer vector $\boldsymbol{d} \in \mathbb{Z}^{\tau}$, is there a vector $\boldsymbol{x} \in \{-1, 1\}^K$ such that $\boldsymbol{D}\boldsymbol{x} \leq \boldsymbol{d}$?*

Let us consider the following special case of the regular FO (38). We first suppose $\theta = 1$ and $\boldsymbol{\xi}^0 = \boldsymbol{0}$, then the uncertainty set becomes

$$\mathcal{U} = \{\boldsymbol{\xi} \colon \|\boldsymbol{\xi}\|_1 \leq 1\}.$$

Next, let us consider the following recourse function

$$Q(\boldsymbol{x}, \boldsymbol{\xi}) = \max_{k \in [K]} \max\left\{\xi^k x_k - 1, -\xi^k x_k + 1\right\},$$

and the set $\mathcal{X} = \{\boldsymbol{x} : \boldsymbol{D}\boldsymbol{x} \leq \boldsymbol{d}, -1 \leq x_k \leq 1, \forall k \in [K]\}$. Under this special setting, the regular FO (38) reduces to

$$v^* = \min_{\boldsymbol{x}, \boldsymbol{\xi}} \left\{ \max_{k \in [K]} \left\{|\xi^k x_k - 1|\right\} : \boldsymbol{D}\boldsymbol{x} \leq \boldsymbol{d}, \boldsymbol{x} \in [-1, 1]^K, \boldsymbol{\xi} \in [-1, 1]^K \right\}. \tag{40}$$

We observe that the optimal value $v^* = 0$ in (40) if and only if $\xi^k x_k = 1$ for all $k \in [K]$, i.e., if and only if there exists a binary feasible solution $\boldsymbol{x} \in \{-1, 1\}^K$ such that $\boldsymbol{D}\boldsymbol{x} \leq \boldsymbol{d}$. Thus, solving problem (40) is NP-hard, so is the regular FO (38). □

Next, we discuss the MICP-R of the objective function of the rDFO (38) with domain $\mathcal{X}$. Unfortunately, for most cases, the objective function of the rDFO (38) with domain $\mathcal{X}$ may not be MICP-R.

Proposition 10 *For any $p \in (1, \infty)$, the objective function of the rDFO* (38) *with domain $\mathcal{X}$ may not be MICP-R.*

Proof: Note that when $K = 1$, the rDFO (38) is equivalent to the rDFO (36). Therefore, according to the result in Theorem 9, when $p \in (1, \infty)$, the objective function of the rDFO (38) with domain $\mathcal{X}$ may not be MICP-R. □

When $p \in \{1, \infty\}$, although we strongly believe that the rDFO (38) with domain $\mathcal{X}$ is either MICP-R or tractable, we are not able to prove this result. Instead, in Theorem 10, we show that when $p = 1$, there exist special cases such that the rDFO (38) can be tractable. Next, we show special cases under which the objective function of the rDFO (38) with domain $\mathcal{X}$ can be MICP-R when $p = \infty$.

Theorem 11 *When $p = \infty$, the objective function of the rDFO* (38) *with domain $\mathcal{X}$ is MICP-R if one of the following conditions holds:*

(i) When $\boldsymbol{a}_k(\boldsymbol{x}) = \boldsymbol{a}_1(\boldsymbol{x})$ for each $k \in [K]$; or

(ii) When $\boldsymbol{\xi} := [\boldsymbol{\xi}_1, \ldots, \boldsymbol{\xi}_K]$ such that $\boldsymbol{\xi}_i$ and $\boldsymbol{\xi}_j$ do not overlap for each $i \neq j$, and $\boldsymbol{\xi}^\top \boldsymbol{a}_k(\boldsymbol{x}) = \boldsymbol{\xi}_k{}^\top \bar{\boldsymbol{a}}_k(\boldsymbol{x})$ for each $k \in [K]$, where $\boldsymbol{\xi}_k \in \mathbb{R}^{m_k}$, $\bar{\boldsymbol{a}}_k(\boldsymbol{x}) = \bar{\boldsymbol{A}}_k \boldsymbol{x} + \bar{\boldsymbol{a}}_k \in \mathbb{R}^{m_k}$ with $\bar{\boldsymbol{A}}_k \in \mathbb{R}^{m_k \times n}, \bar{\boldsymbol{a}}_k \in \mathbb{R}^{m_k}$ such that $\bar{\boldsymbol{A}}_i$ and $\bar{\boldsymbol{A}}_j$, $\bar{\boldsymbol{a}}_i$ and $\bar{\boldsymbol{a}}_j$ do not overlap each $i \neq j$ with $\sum_{k \in [K]} m_k = m$.

Proof: We split the proof into two parts accordingly.

(i) When $\boldsymbol{a}_k(\boldsymbol{x}) = \boldsymbol{a}_1(\boldsymbol{x})$ for each $k \in [K]$, we can rewrite the rDFO (39) as

$$v^* = \min_{\boldsymbol{x} \in \mathcal{X}} \max_{\boldsymbol{\lambda} \geq \boldsymbol{0}} \left\{ \sum_{k \in [K]} \lambda_k b_k(\boldsymbol{x}) + \boldsymbol{\xi}^0{}^\top \boldsymbol{a}_1(\boldsymbol{x}) - \theta \|\boldsymbol{a}_1(\boldsymbol{x})\|_1 \colon \sum_{k \in [K]} \lambda_k = 1 \right\}.$$

Taking the dual of the inner maximization problem and using strong duality from linear programming, we have

$$v^* = \min_{\boldsymbol{x}\in\mathcal{X},\eta} \left\{ \eta \colon \eta \geq {\boldsymbol{\xi}^0}^\top \boldsymbol{a}_1(\boldsymbol{x}) + b_k(\boldsymbol{x}) - \theta \|\boldsymbol{a}_1(\boldsymbol{x})\|_1, \forall k \in [K] \right\}. \tag{41}$$

According to Part (i) in Theorem 9, the rDFO (41) is MICP-R.

(ii) When $p = \infty$ and ${\boldsymbol{\xi}}^\top \boldsymbol{a}_k(\boldsymbol{x}) = {\boldsymbol{\xi}_k}^\top \bar{\boldsymbol{a}}_k(\boldsymbol{x})$ for each $k \in [K]$, we can rewrite the rDFO (39) as

$$\min_{\boldsymbol{x}\in\mathcal{X}} \max_{\boldsymbol{\lambda}\geq\boldsymbol{0}} \left\{ \sum_{k\in[K]} \lambda_k \left[{\boldsymbol{\xi}_k^0}^\top \bar{\boldsymbol{a}}_k(\boldsymbol{x}) \right] + \sum_{k\in[K]} \lambda_k b_k(\boldsymbol{x}) - \theta \sum_{k\in[K]} \lambda_k \left\| \bar{\boldsymbol{a}}_k(\boldsymbol{x}) \right\|_1 : \sum_{k\in[K]} \lambda_k = 1 \right\},$$

which can be simplified as

$$\min_{\boldsymbol{x}\in\mathcal{X}} \max_{\boldsymbol{\lambda}\geq\boldsymbol{0}} \left\{ \sum_{k\in[K]} \lambda_k \left[{\boldsymbol{\xi}_k^0}^\top \bar{\boldsymbol{a}}_k(\boldsymbol{x}) \right] + \sum_{k\in[K]} \lambda_k b_k(\boldsymbol{x}) - \theta \sum_{k\in[K]} \lambda_k \left\| \bar{\boldsymbol{a}}_k(\boldsymbol{x}) \right\|_1 : \sum_{k\in[K]} \lambda_k = 1 \right\}.$$

Taking the dual of the inner maximization problem and using strong duality from linear programming, we have

$$\min_{\boldsymbol{x}\in\mathcal{X},\eta} \left\{ \eta \colon \eta \geq \left[{\boldsymbol{\xi}_k^0}^\top \bar{\boldsymbol{a}}_k(\boldsymbol{x}) \right] + b_k(\boldsymbol{x}) - \theta \left\| \bar{\boldsymbol{a}}_k(\boldsymbol{x}) \right\|_1, \forall k \in [K] \right\}. \tag{42}$$

According to Part (i) in Theorem 9, the rDFO (42) is MICP-R. □

The following Corollary 7 shows the MICP-R formulations of the two cases discussed in Theorem 11.

Corollary 7 *When $p = \infty$, suppose that $\mathcal{X} \subseteq [\boldsymbol{l}, \boldsymbol{u}]$.*

(i) If $\boldsymbol{a}_k(\boldsymbol{x}) = \boldsymbol{a}_1(\boldsymbol{x})$ for each $k \in [K]$, the rDFO (38) *can be reformulated as the following MICP*

$$\begin{aligned} v^* = \min_{\boldsymbol{x}\in\mathcal{X},\eta} \quad & \eta, \\ \text{s.t.} \quad & \eta \geq {\boldsymbol{\xi}^0}^\top \boldsymbol{a}_1(\boldsymbol{x}) + b_k(\boldsymbol{x}) - \theta \sum_{i\in[m]} s_{1i}, \forall k \in [K], \\ & (s_{1i}, z_{1i}, a_{1i}(\boldsymbol{x})) \in \mathcal{MI}\left(-1, 1, \widehat{l}_{1i}, \widehat{u}_{1i}\right), \forall i \in [m], \end{aligned}$$

where we let $\widehat{l}_{ki} = \sum_{j\in[n]} \min\{\widehat{A}_{kij} l_j, \widehat{A}_{kij} u_j\} + \widehat{a}_{ki}$ and $\widehat{u}_{ki} = \sum_{j\in[n]} \max\{\widehat{A}_{kij} l_j, \widehat{A}_{kij} u_j\} + \widehat{a}_{ki}$ for each $i \in [m]$ such that $\boldsymbol{a}_k(\boldsymbol{x}) \in [\widehat{\boldsymbol{l}}_k, \widehat{\boldsymbol{u}}_k]$ for each $k \in [K]$; and

(ii) Suppose $\boldsymbol{\xi} := [\boldsymbol{\xi}_1, \ldots, \boldsymbol{\xi}_K]$ such that $\boldsymbol{\xi}_i$ and $\boldsymbol{\xi}_j$ do not overlap for each $i \neq j$, and $\boldsymbol{\xi}^\top \boldsymbol{a}_k(\boldsymbol{x}) = \boldsymbol{\xi}_k^\top \bar{\boldsymbol{a}}_k(\boldsymbol{x})$ for each $k \in [K]$, where $\boldsymbol{\xi}_k \in \mathbb{R}^{m_k}$, $\bar{\boldsymbol{a}}_k(\boldsymbol{x}) = \bar{\boldsymbol{A}}_k \boldsymbol{x} + \bar{\boldsymbol{a}}_k \in \mathbb{R}^{m_k}$ with $\bar{\boldsymbol{A}}_k \in \mathbb{R}^{m_k \times n}, \bar{\boldsymbol{a}}_k \in \mathbb{R}^{m_k}$ such that $\bar{\boldsymbol{A}}_i$ and $\bar{\boldsymbol{A}}_j$, $\bar{\boldsymbol{a}}_i$ and $\bar{\boldsymbol{a}}_j$ do not overlap each $i \neq j$ with $\sum_{k\in[K]} m_k = m$ and each m_k is nonnegative. Then, the rDFO (38) *can be reformulated as the following MICP*

$$\begin{aligned} v^* = \min_{\boldsymbol{x}\in\mathcal{X},\eta} \quad & \eta, \\ \text{s.t.} \quad & \eta \geq {\boldsymbol{\xi}^0}^\top \boldsymbol{a}_k(\boldsymbol{x}) + b_k(\boldsymbol{x}) - \theta \sum_{i\in[m]} s_{ki}, \forall k \in [K], \\ & (s_{ki}, z_{ki}, \bar{a}_{ki}(\boldsymbol{x})) \in \mathcal{MI}\left(-1, 1, \bar{l}_{ki}, \bar{u}_{ki}\right), \forall k \in [K], i \in [m], \end{aligned}$$

where we let $\bar{l}_{ki} = \sum_{j\in[n]} \min\{\bar{A}_{kij} l_j, \bar{A}_{kij} u_j\} + \bar{a}_{ki}$ and $\bar{u}_{ki} = \sum_{j\in[n]} \max\{\bar{A}_{kij} l_j, \bar{A}_{kij} u_j\} + \bar{a}_{ki}$ for each $i \in [m]$ such that $\bar{\boldsymbol{a}}_k(\boldsymbol{x}) \in [\bar{\boldsymbol{l}}_k, \bar{\boldsymbol{u}}_k]$ for each $k \in [K]$.

D.3 DFO (2) with Type$-\infty$ Wasserstein Ambiguity Set

In this subsection, we extend the tractability and complexity results to DFO (2). In particular, we focus on the type$-\infty$ Wasserstein ambiguity set, which reads $\mathcal{P}_\infty^W = \{\mathbb{P}\colon \mathbb{P}\{\tilde{\boldsymbol{\xi}} \in \mathcal{U}\} = 1, W_\infty(\mathbb{P}, \mathbb{P}_{\tilde{\boldsymbol{\zeta}}}) \leq \theta\}$, where $\mathbb{P}_{\tilde{\boldsymbol{\zeta}}}$ is a discrete empirical reference distribution of random parameters $\tilde{\boldsymbol{\zeta}}$ generated by N i.i.d. samples such that $\mathbb{P}_{\tilde{\boldsymbol{\zeta}}}\{\tilde{\boldsymbol{\zeta}} = \boldsymbol{\zeta}^i\} = 1/N$, i.e., $\mathbb{P}_{\tilde{\boldsymbol{\zeta}}} = 1/N \sum_{i\in[N]} \delta_{\boldsymbol{\zeta}^i}$ and $\delta_{\boldsymbol{\zeta}^i}$ is the Dirac function that places unit mass on the realization $\tilde{\boldsymbol{\zeta}} = \boldsymbol{\zeta}^i$ for each $i \in [N]$, $\theta \geq 0$ is the Wasserstein radius, and the $\infty-$Wasserstein distance between two probability distributions $\mathbb{P}_1, \mathbb{P}_2$ is defined as

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

Under this setting, DFO (2) admits the following representation (see, e.g., [8, 71]):

$$v^* = \min_{\boldsymbol{x}\in\mathcal{X}} \left\{ \boldsymbol{c}^\top \boldsymbol{x} + \inf_{\mathbb{P}\in\mathcal{P}_\infty^W} \mathbb{E}_{\mathbb{P}} \left[Q(\boldsymbol{x}, \tilde{\boldsymbol{\xi}}) \right] \right\} = \min_{\boldsymbol{x}\in\mathcal{X}} \left\{ \boldsymbol{c}^\top \boldsymbol{x} + \frac{1}{N} \sum_{i\in[N]} \left[\inf_{\boldsymbol{\xi}} \left\{ Q(\boldsymbol{x}, \boldsymbol{\xi})\colon \|\boldsymbol{\xi} - \boldsymbol{\zeta}^i\|_p \leq \theta \right\} \right] \right\}. \tag{43}$$

Similar to the discussions in Section D.1 and Section D.2, we then consider the function $\boldsymbol{c}^\top \boldsymbol{x} + Q(\boldsymbol{x}, \boldsymbol{\xi})$ to be convex and concave piecewise affine, respectively.

Special Case I. Concave Piecewise Recourse Function. We first consider the concave piecewise affine recourse function. Following the same notation as Section D.1, i.e., function $\boldsymbol{c}^\top \boldsymbol{x} + Q(\boldsymbol{x}, \boldsymbol{\xi})$ is the minimum of piecewise affine functions, DFO (43) can be written as

$$v^* = \min_{\boldsymbol{x}\in\mathcal{X}} \frac{1}{N} \sum_{i\in[N]} \left[\min_{k\in[K]} {\boldsymbol{\zeta}^i}^\top \boldsymbol{a}_k(\boldsymbol{x}) + b_k(\boldsymbol{x}) - \theta \left\| \boldsymbol{a}_k(\boldsymbol{x}) \right\|_{p^*} \right]. \tag{44}$$

We notice that if there is only one sample available in the empirical distribution $\mathbb{P}_{\tilde{\boldsymbol{\zeta}}}$ (i.e., $N = 1$), then DFO (44) reduces to the favorable optimization (36). Thus, according to Theorem 8, if $p \in (1, \infty]$, solving DFO (44) is, in general, NP-hard. It turns out that even with $p = 1$, solving DFO (44) is also NP-hard.

Proposition 11 *For any $p \in [1, \infty]$, solving DFO* (44) *is, in general, NP-hard.*

Proof: For any $p \in (1, \infty]$, DFO (44) reduces to the favorable optimization (38) if there is only one sample available for the empirical distribution $\mathbb{P}_{\tilde{\boldsymbol{\zeta}}}$ (i.e., $N = 1$). Thus, according to Theorem 8, solving DFO (44) is, in general, NP-hard.

It remains to show that solving DFO (44) is also NP-hard when $p = 1$. Recall the NP-complete problem - feasibility problem of a general binary program, which asks

> ***Feasibility of a binary program.*** *Given an integer matrix $\boldsymbol{D} \in \mathbb{Z}^{\tau\times n}$, and integer vector $\boldsymbol{d} \in \mathbb{Z}^\tau$, is there a vector $\boldsymbol{x} \in \{-1, 1\}^n$ such that $\boldsymbol{D}\boldsymbol{x} \leq \boldsymbol{d}$?*

Let us consider the following special case of DFO (44). Let set $\mathcal{X} = \{\boldsymbol{x} : \boldsymbol{D}\boldsymbol{x} \leq \boldsymbol{d}, -1 \leq x_i \leq 1, \forall i \in [n]\}$ and suppose $\theta = 0$, $b_k(\boldsymbol{x}) = 1$ for each $k \in [K]$, $\boldsymbol{\zeta}^i = \boldsymbol{e}_i$ for each $i \in [N]$, $N = n$, $K = 2n, m = n$, and

$$\boldsymbol{a}_k(\boldsymbol{x}) = \begin{cases} x_k \boldsymbol{e}_k, & k \leq n, \\ -x_k \boldsymbol{e}_k, & n+1 \leq k \leq 2n. \end{cases}$$

Then, the inner minimum $\min_{k\in[K]} {\boldsymbol{\zeta}^i}^\top \boldsymbol{a}_k(\boldsymbol{x}) + b_k(\boldsymbol{x})$ reduces to

$$\min_{k\in[K]} {\boldsymbol{\zeta}^i}^\top \boldsymbol{a}_k(\boldsymbol{x}) + b_k(\boldsymbol{x}) = \min\{1, 1 - x_i, 1 + x_i\} = \min\{1, 1 - |x_i|\} = 1 - |x_i|.$$

Thus, DFO (44) can be written as

$$v^* = \min_{\substack{\boldsymbol{D}\boldsymbol{x}\leq\boldsymbol{d},\\ \boldsymbol{x}\in[-1,1]^n}} \frac{1}{n}\sum_{i\in[n]}\left[1-|x_i|\right]. \tag{45}$$

We observe that the optimal value of DFO (45) $v^* = 0$ if and only if $|x_i| = 1$ for all $i \in [n]$, i.e., if and only if there exists a binary feasible solution $\boldsymbol{x} \in \{-1,1\}^n$ such that $\boldsymbol{D}\boldsymbol{x} \leq \boldsymbol{d}$. Thus, solving problem (45) is NP-hard, so is DFO (44). □

Albeit being NP-hard, when $p \in \{1,\infty\}$, we are able to provide MICP-R formulations for the DFO (44).

Theorem 12 *Suppose $\mathcal{X} \subseteq [\boldsymbol{l},\boldsymbol{u}]$, let $\widehat{l}^b_k = \sum_{j\in[n]}\min\{\widehat{B}_{kj}l_j, \widehat{B}_{kj}u_j\}+\widehat{b}_k$ and $\widehat{u}^b_k = \sum_{j\in[n]}\max\{\widehat{B}_{kj}l_j, \widehat{B}_{kj}u_j\}+\widehat{b}_k$ such that $\boldsymbol{b}_k(\boldsymbol{x}) \in [\widehat{l}^b_k, \widehat{u}^b_k]$ for each $k \in [K]$, and let $\widehat{l}^a_{ki} = \sum_{j\in[n]}\min\{\widehat{A}_{kij}l_j, \widehat{A}_{kij}u_j\} + \widehat{a}_{ki}$ and $\widehat{u}^a_{ki} = \sum_{j\in[n]}\max\{\widehat{A}_{kij}l_j, \widehat{A}_{kij}u_j\} + \widehat{a}_{ki}$ for each $i \in [m]$ such that $\boldsymbol{a}_k(\boldsymbol{x}) \in [\widehat{\boldsymbol{l}}^a_k, \widehat{\boldsymbol{u}}^a_k]$ for each $k \in [K]$. When $p \in \{1,\infty\}$, DFO (44) is MICP-R.*

Proof: We first introduce binary variables $\boldsymbol{\lambda}$ to reformulate the inner minimum, that is,

$$\begin{aligned} v^* = \min_{\boldsymbol{x}\in\mathcal{X},\boldsymbol{\lambda}} \quad & \frac{1}{N}\sum_{i\in[N]}\sum_{k\in[K]}\lambda_{ki}\left[{\boldsymbol{\zeta}^i}^\top\boldsymbol{a}_k(\boldsymbol{x}) + b_k(\boldsymbol{x}) - \theta\left\|\boldsymbol{a}_k(\boldsymbol{x})\right\|_{p^*}\right], \\ \text{s.t.} \quad & \sum_{k\in[K]}\lambda_{ki} = 1, \forall i\in[N], \\ & \lambda_{ki}\in\{0,1\}, \forall k\in[K], i\in[N]. \end{aligned}$$

Since set $\mathcal{X}$ is compact, we can apply McCormick inequalities [49] to linearize the terms $\{\lambda_{ki}\boldsymbol{a}_k(\boldsymbol{x})\}_{i\in[N],k\in[K]}$ and $\{\lambda_{ki}b_k(\boldsymbol{x})\}_{i\in[N],k\in[K]}$. It remains to provide the MICP-R formulation for the term $\{\lambda_{ki}\|\boldsymbol{a}_k(\boldsymbol{x})\|_{p^*}\}_{i\in[N],k\in[K]}$.

When $p=\infty$, i.e., the dual norm is L_1, the term $\{\lambda_{ki}\|\boldsymbol{a}_k(\boldsymbol{x})\|_1\}$ can be linearized by applying McCormick inequalities twice for each $i \in [N]$ and $k \in [K]$.

When $p=1$, i.e., the dual norm is L_∞, we can apply disjunctive programming [5] to the term $\{\lambda_{ki}\|\boldsymbol{a}_k(\boldsymbol{x})\|_\infty\}$ and then apply McCormick inequalities to linearize $\{\lambda_{ki}\boldsymbol{a}_{kj}(\boldsymbol{x})\}_{i\in[N],k\in[K],j\in[m]}$.

Therefore, according to the result in Lemma 2, DFO (44) is MICP-R with $p \in \{1,\infty\}$. □

On the other hand, since the rDFO (36) is a special case of DFO (44) with $N = 1$. According to Theorem 9, when $p \in (1,\infty)$, DFO (44) may not be MICP-R.

Corollary 8 *When $p \in (1,\infty)$, DFO (44) may not be MICP-R.*

Special Case II. Convex Piecewise Recourse Function. Following the same notation introduced in Section D.2, i.e., function $Q(\boldsymbol{x},\boldsymbol{\xi})$ is the maximum of piecewise affine function, DFO (43) can be written as

$$v^* = \min_{\boldsymbol{x}\in\mathcal{X}}\frac{1}{N}\sum_{i\in[N]}\left[\inf_{\boldsymbol{\xi}}\left\{\max_{k\in[K]}\boldsymbol{\xi}^\top\boldsymbol{a}_k(\boldsymbol{x}) + b_k(\boldsymbol{x})\colon \|\boldsymbol{\xi}-\boldsymbol{\zeta}^i\|_p \leq \theta\right\}\right]. \tag{46}$$

Under this special case, it turns out that all the results in Section D.2 can be naturally extended to DFO (46), as presented below. The proofs and formulations are omitted for brevity.

Corollary 9 *For DFO (46), the complexity and tractability results in Section D.2 directly follow.*

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

In statistical robustness (see the details in [25, 47]), if the influence curve of a statistic estimator is bounded, then that estimator is called "outlier robust." Let $\mathbb{P}$ denote the reference probability measure of $\tilde{\boldsymbol{\xi}}$ and $\delta_{\boldsymbol{\xi}^o}$ is the Dirac measure for the perturbation data $\boldsymbol{\xi}^o \in \mathcal{U}$. For any decision $\boldsymbol{x} \in \mathcal{X}$ with corresponding function values $Q(\boldsymbol{x}, \tilde{\boldsymbol{\xi}})$, the statistic estimator $T(\mathbb{P}, \cdot)$ is "outlier robust" if the following condition is satisfied:

$$\lim_{\gamma \to 0} \frac{1}{\gamma} \left[T\left((1-\gamma)\mathbb{P} + \gamma \delta_{\delta_{\boldsymbol{\xi}^o}}, Q(\boldsymbol{x}, \tilde{\boldsymbol{\xi}}) \right) - T\left(\mathbb{P}, Q(\boldsymbol{x}, \tilde{\boldsymbol{\xi}}) \right) \right] < \infty. \tag{47}$$

Then, based on condition (47), we first illustrate that $\mathrm{VaR}_{1-\varepsilon}\{Q(\boldsymbol{x}, \tilde{\boldsymbol{\xi}})\}$ (i.e., a quantile) may not be outlier robust.

Example 7 When the reference probability distribution is discrete, that is, suppose $\mathbb{P}\{\tilde{\boldsymbol{\xi}} : \tilde{\boldsymbol{\xi}} = \boldsymbol{\xi}^i\} = 1/N$ for each $i \in [N]$ and the perturbation $Q(\boldsymbol{x}, \boldsymbol{\xi}^o)$, $\mathrm{VaR}_{1-\varepsilon}\{Q(\boldsymbol{x}, \tilde{\boldsymbol{\xi}})\}$ is "outlier robust" if the condition (47) is satisfied. Suppose $\varepsilon = 0.1$, $N = 10\bar{N}$, 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 $\gamma \to 0$, $\mathrm{VaR}_{1-\varepsilon}\{(1-\gamma)\mathbb{P} + \gamma\delta_{\delta_{\boldsymbol{\xi}^o}}, Q(\boldsymbol{x}, \tilde{\boldsymbol{\xi}})\} = \bar{N}$, and $\mathrm{VaR}_{1-\varepsilon}\{\mathbb{P}, Q(\boldsymbol{x}, \tilde{\boldsymbol{\xi}})\} = \bar{N} - 1$. Then, condition (47) is simplified as

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

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

Under the similar setting of Example 7, we can show that $\mathrm{FCVaR}_{1-\varepsilon}\{Q(\boldsymbol{x}, \tilde{\boldsymbol{\xi}})\}$ (i.e., LTS) may not be outlier robust.

Example 8 When the reference probability distribution is discrete, that is, suppose $\mathbb{P}\{\tilde{\boldsymbol{\xi}} : \tilde{\boldsymbol{\xi}} = \boldsymbol{\xi}^i\} = 1/N$ for each $i \in [N]$ and the perturbation $Q(\boldsymbol{x}, \boldsymbol{\xi}^o)$, $\mathrm{FCVaR}_{1-\varepsilon}\{Q(\boldsymbol{x}, \tilde{\boldsymbol{\xi}})\}$ is "outlier robust" if the condition (47) is satisfied. Suppose $\varepsilon = 0.1$, $N = 10\bar{N}$, 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 (47) is simplified as

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

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

Appendix F. Comparisons of DFO (2) with Main Results in [21]

In this appendix, we show that the framework in this paper is quite different from that in [21]. It turns out that the result of [21] (i.e., theorem 1 [21]) is exactly the same as the worst-case conditional value-at-risk under the Wasserstein ambiguity set, the compact support information $\widetilde{\Xi}$, and an inflated Wasserstein radius. That is, once the support is known, their model is pessimistic; on the other hand, ours is optimistic and can handle the endogenous outliers when the data might not even be contaminated.

Given any $q \in [1, \infty]$, type$-q$ Wasserstein ambiguity set is defined as

$$\mathcal{P}_q^W = \{\mathbb{P}\colon \mathbb{P}\{\tilde{\boldsymbol{\xi}} \in \widetilde{\Xi}\} = 1, W_q(\mathbb{P}, \mathbb{P}_{\tilde{\boldsymbol{\zeta}}}) \leq \theta\},$$

where $W_q(\cdot,\cdot)$ represents the $q-$Wasserstein distance, i.e.,

$$W_q(\mathbb{P}_1, \mathbb{P}_2) = \inf \left\{ \left[\int_{\widetilde{\Xi}\times\widetilde{\Xi}} \left\|\boldsymbol{\xi}^1 - \boldsymbol{\xi}^2\right\|^q \mathbb{Q}(d\boldsymbol{\xi}^1, d\boldsymbol{\xi}^2) \right]^{\frac{1}{q}} : \begin{array}{l} \mathbb{Q} \text{ is a joint distribution of } \tilde{\boldsymbol{\xi}}^1 \text{ and } \tilde{\boldsymbol{\xi}}^2 \\ \text{with marginals } \mathbb{P}_1 \text{ and } \mathbb{P}_2, \text{ respectively} \end{array} \right\},$$

$\theta \geq 0$ is the Wasserstein radius, and $\mathbb{P}_{\tilde{\boldsymbol{\zeta}}}$ denote a discrete empirical reference distribution of random parameters $\tilde{\boldsymbol{\zeta}}$ generated by N i.i.d. samples such that $\mathbb{P}\{\tilde{\boldsymbol{\zeta}} = \boldsymbol{\zeta}^i\} = 1/N$ for each $i \in [N]$.

This result is formally shown below.

Proposition 12 *Given any $\varepsilon \in (0,1)$ and compact support $\widetilde{\Xi}$, then under the type$-q$ Wasserstein ambiguity set with radius θ^q/ε, the formulation in theorem 1 in [21] is equivalent to the worst-case conditional value-at-risk under the type$-q$ Wasserstein ambiguity set with radius θ, i..e,*

$$\sup_{\mathbb{P}\in\mathcal{P}_q^W} \min_{\beta} \left\{ \beta + \frac{1}{\varepsilon}\mathbb{E}_{\mathbb{P}}\left[Q(\boldsymbol{x}, \tilde{\boldsymbol{\zeta}}) - \beta\right]_+ \right\}. \tag{48}$$

Proof: Following the setting as [21] and using theorem 1 in [76], we can switch the infimum and supremum operators in (48), that is,

$$\min_{\beta} \sup_{\mathbb{P}\in\mathcal{P}_q^W} \left\{ \beta + \frac{1}{\varepsilon}\mathbb{E}_{\mathbb{P}}\left[Q(\boldsymbol{x}, \tilde{\boldsymbol{\zeta}}) - \beta\right]_+ \right\}. \tag{49a}$$

Applying the results of theorem 4.2 in [20] and using the definition of function $(\cdot)_+$, the formulation (49a) can be further simplified as

$$\min_{\beta,\lambda\geq 0} \lambda\frac{\theta^q}{\varepsilon} + \beta + \frac{1}{\varepsilon}\mathbb{E}_{\mathbb{P}}\left[\sup_{\boldsymbol{\zeta}\in\widetilde{\Xi}} \left\{ Q(\boldsymbol{x}, \boldsymbol{\zeta}) - \lambda \left\|\boldsymbol{\zeta} - \widehat{\boldsymbol{\zeta}}\right\|^q \right\} - \beta \right]_+. \tag{49b}$$

Denote $\tilde{\theta} = \theta^q/\varepsilon$. Then the formulation (49b) can be written as

$$\begin{aligned} \min_{\beta,\lambda\geq 0,\bar{\boldsymbol{\mu}}} \quad & \lambda\tilde{\theta} + \beta + \frac{1}{N\varepsilon}\sum_{i=1}^N \bar{\mu}_i, \\ \text{s.t.} \quad & \bar{\mu}_i + \beta \geq \sup_{\boldsymbol{\zeta}\in\widetilde{\Xi}} \left\{ Q(\boldsymbol{x}, \boldsymbol{\zeta}) - \lambda \left\|\boldsymbol{\zeta} - \boldsymbol{\zeta}^i\right\|^q \right\}, \forall i \in [N], \\ & \bar{\mu}_i \geq 0, \forall i \in [N], \end{aligned} \tag{49c}$$

which is exactly the same result of theorem 1 in [21] under the type$-q$ Wasserstein ambiguity set with $\boldsymbol{\zeta} = (\boldsymbol{y}, \boldsymbol{z})$. □

The result in Proposition 12 is different from what we propose in DFO, where the latter focuses on the best-case expectation to handle the endogenous outliers. Let us consider the following example for an illustration.

Example 9 Under the same setting as Example 1 except the assumptions that set $\mathcal{Y} = \{y : 0 \leq y \leq 10\}$ and $\tilde{\xi}$ follows the standard Gaussian distribution, i.e., $\tilde{\xi} \sim \mathcal{N}(0, 1)$ (see, e.g., Figure 7). In this setting, due to the lack of relatively complete recourse, the two-stage stochastic program is infeasible. Applying the DFO with interval ambiguity set $\mathcal{P}_I = \{\boldsymbol{\mu} : \boldsymbol{\mu}(\mathcal{U}) = 1, 0 \preceq \boldsymbol{\mu} \preceq \mathbb{P}/(2 - 2\Phi(0.1))\}$, the resulting favorable two-stage problem is feasible and mitigates the effect of endogenous outliers, i.e., region B in Figure 7.

On the other hand, if we apply the results of theorem 1 of [21] to Example 9 by handling the extreme scenarios from region A or C or both from the support $\widetilde{\Xi}$ (see, Figure 7), the remaining optimization problem would be still infeasible. In our DFO, we mitigate the effect of region B, and this information is unknown before solving the optimization model. That is, the endogenous outliers may not be processed based on the results in theorem 1 of [21], while our DFO framework can. ◇

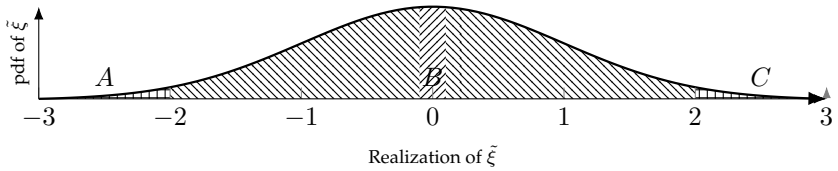


Fig. 7: Illustration of Example 9.