manuscript

Supermodularity in Two-Stage Distributionally Robust Optimization

Daniel Zhuoyu Long
Department of Systems Engineering and Engineering Management, Chinese University of Hong Kong, zylong@se.cuhk.edu.hk

Jin Qi
Department of Industrial Engineering and Decision Analytics, Hong Kong University of Science and Technology, jinqi@ust.hk

Aiqi Zhang
Rotman School of Management, University of Toronto, aq.zhang@rotman.utoronto.ca

In this paper, we solve a class of two-stage distributionally robust optimization problems which have the property of supermodularity. We exploit the explicit worst-case expectation of supermodular functions and derive the worst-case distribution for the robust counterpart. This enables us to develop an efficient method to obtain an exact optimal solution to these two-stage problems. Further, we provide a necessary and sufficient condition for checking whether any given two-stage optimization problem has the supermodularity property. We also investigate the optimality of the segregated affine decision rules when problems have the property of supermodularity. We apply this framework to several classic problems, including the multi-item newsvendor problem, the facility location problem, the lot-sizing problem on a network, the appointment scheduling problem, and the assemble-to-order problem. While these problems are typically computationally challenging, they can be solved efficiently under our assumptions. Finally, numerical examples are conducted to illustrate the effectiveness of our approach.

Key words: distributionally robust optimization; two-stage optimization; supermodularity; assemble-to-order

1. Introduction

Many real-world optimization problems with uncertainties can be formulated as two-stage optimization models. In such problems, we first make a "here-and-now" decision. In the second stage, after the uncertainties are realized, we choose an optimal action, which we call the "wait-and-see" decision.

This two-stage optimization formulation has drawn extensive attention from both the operations management and optimization communities as it can model a wide range of operational problems. For instance, in an assemble-to-order (ATO) system, the here-and-now decision is the ordering quantities of the components while the wait-and-see decision is the assembly plan which determines the amount of each type of component to be used to assemble each type of product on demand. In appointment scheduling problems, the here-and-now decision is the scheduled appointment time while we introduce auxiliary second-stage decisions to evaluate the nonlinear objective. Other examples include multi-item newsvendor, facility location, unit commitment problems, etc.

One classic solution approach to two-stage optimization problems is stochastic programming (e.g., Shapiro et al. 2009, Birge and Louveaux 2011), in which uncertainties are assumed to follow some given probability distributions. To incorporate ambiguity, robust optimization is adopted to solve the two-stage optimization problems. Using robust optimization, instead of optimizing the expectation of objective functions, we seek solutions that are immune to a distribution-free uncertainty set. However, this type of problem is still hard to solve in general because of its two-stage nature. Some approximation methods have been proposed to address the intractable nature of the problem, such as the linear decision rule (Ben-Tal et al. 2004), and more complex methods including the polynomial (Bertsimas et al. 2011), segregated affine (Chen et al. 2008) and piecewise linear (Ben-Tal et al. 2009) decision rules. These approaches restrict solutions to specific functions of the uncertainty realizations (such as affine functions). The functions are parameterized by a finite number of coefficients and lead to computational tractability.

In addition, if problems have some special structures, the approximated solutions can be proved to be near-optimal or even optimal. Bertsimas and Goyal (2010) show that for a two-stage stochastic problem, the static solutions derived from the corresponding robust version give a 2-approximation to the original stochastic problem if both the uncertainty set and the probability measure are symmetric. For the linear decision rule, Bertsimas et al. (2010b) prove its optimality in multi-period robust optimization problems when the problem is one-dimensional with convex costs. Bertsimas and Goyal (2012) further give the result that linear decision rules can be optimal in a two-stage setting if the uncertainty set is a simplex. Kuhn et al. (2011) apply the linear decision rule approximation to the primal and dual problems separately, in both stochastic programming and robust optimization problems, where the gap between the two approximated values is used to estimate the loss of optimality. The numerical example shows that in the specific setting they adopt, the relative gap between the bounds can be consistently low.

However, since classic robust optimization does not use any frequency information, the solution can be overly conservative and therefore too extreme for practical applications. To overcome this, by incorporating an ambiguity set of probability distributions, distributionally robust optimization (DRO) has been developed to seek solutions which protect against the worst-case distribution over all admissible ones (Delage and Ye 2010, Goh and Sim 2010, Wiesemann et al. 2014). The distributional ambiguity set containing all possible probability distributions is characterized by certain distributional information, such as moment information or statistical measures. Chen et al. (2020) recently propose a scenario-based distributional ambiguity set, which can model a broader class of uncertainty sets, e.g., uncertainty sets with both moment and Wasserstein distance information. For the two-stage DRO, while solutions can be derived by many parametric decision rules as in robust optimization, very few results have been reported to theoretically evaluate the performance of these approximations. Ardestani-Jaafari and Delage (2016) show the optimality of segregated affine decision rules to a distributionally robust multi-item newsvendor problem with budgeted support, means and first order partial moments as distributional information. Incorporating auxiliary random variables and considering a class of lifted linear decision rules, Bertsimas et al. (2019) prove the optimality of such approximations when the problem has complete recourse and only one second-stage decision variable. Georghiou et al. (2021) identify five conditions for the objective function and feasible region of the second-stage problem, and show that if all the five conditions hold, linear decision rules can be optimal regardless of the structure of the ambiguity set. On the other hand, examples are also given for the case where linear decision rules can be infeasible even for problems with complete recourse (Bertsimas et al. 2019).

Further, few studies have been conducted to examine the equivalent reformulations and tractability conditions required to solve for exact analytical solutions. Bertsimas et al. (2010a) investigate the cases with ambiguity sets constructed using first and second moments and objective functions being nondecreasing piecewise linear convex disutility functions of the second-stage costs. They show that, if uncertainties only appear in the objective function of the second stage, then the original problems can be equivalently reformulated as semidefinite programs. Bansal et al. (2018) propose decomposition algorithms for two-stage distributionally robust linear problems with discrete distributions, as well as conditions under which the algorithms are finitely convergent. Hanasusanto and Kuhn (2018) show that for problems with complete recourse and the ambiguity sets being 2-Wasserstein balls centered on a discrete distribution, if the uncertainty appears only in constraints of the second-stage problem, then there exists a co-positive cone reformulation.

We extend the previous literature by exploiting the property of supermodularity for a broad class of two-stage DRO problems. Hence, besides the DRO, supermodularity is another stream of studies that are closely related to our work. The concept of supermodularity has proved its importance in the areas of economics and operations research. In particular, it has economic implications in terms of complementarity between resources. A widely studied problem is to explore supermodularity in parametric optimization problems in order to derive certain monotone comparative statics. However, the results are rather scattered and the proof is usually problem-specific. For the general case, Topkis (1998) first introduces lattice conditions on the feasible set to derive the property of supermodularity. While the lattice condition is quite restrictive, Chen et al. (2013) extend it and study the sufficient condition for a class of two-dimensional parametric optimization problems. A recent work by Chen et al. (2021) has provided a systematic study of the conditions both necessary and sufficient to identify the property of supermodularity. Because of the essential implication of complementarity, in a few studies, supermodularity is incorporated within robust optimization to analyze the worst-case performance. Specifically, Agrawal et al. (2010) prove that when the marginal distributions are two-point distributions and the cost function is convex and supermodular, there exists a polynomial-time algorithm for the optimization problem under uncertainties. In multi-stage robust optimization, Iancu et al. (2013) show that the linear decision rule gives an optimal solution when the objective function is supermodular and the uncertainty set has a certain lattice structure.

In this paper, we solve a class of two-stage DRO problems in which the second-stage optimal value is supermodular in the realization of uncertainties. Under the setting of scenario-based ambiguity sets with supports, means and upper bounds of mean absolute deviations (MADs), we exploit the explicit worst-case expectation of supermodular functions and derive the worst-case distribution in the robust counterpart. This can make the two-stage DRO problem tractable. Further, we provide a necessary and sufficient condition to check whether any given two-stage optimization problem has this property. We also discuss the optimality of the segregated affine decision rules when problems have the property of supermodularity. We then illustrate the applicability of our theoretical results by identifying a class of two-stage optimization problems with supermodularity. These include several classic problems, e.g., multi-item newsvendor, facility location, lot-sizing on a network, appointment scheduling with random no-shows, and general ATO systems. While these problems are typically computationally challenging, they can be solved efficiently under our assumptions.

Our key contributions are summarized as follows.

1. In a two-stage optimization problem with mean and MAD as the distributional information, whenever the second-stage problem has the property of supermodularity, we can explicitly find its worst-case distribution in polynomial time. With this distribution, we obtain the worst-case expectation of the second-stage cost, and the original two-stage problem can be reduced to a deterministic optimization problem of polynomial size.
2. When the second-stage problem has a linear programming formulation, we provide a necessary and sufficient condition to check its supermodularity. An algorithm is proposed to determine whether the condition is satisfied.
3. Leveraging the special structure of the worst-case distribution, we show that when the property of supermodularity holds, the segregated affine decision rules can return the same optimal solution either when the constraints are relaxed to only on a given subset of uncertainty realizations, or when the feasible region satisfies some given conditions.
4. We provide three extensions to generalize the results and further apply them to several important problems, including multi-item newsvendor, facility location, lot-sizing, appointment

scheduling, and the ATO problems. For the first four applications, the objective is supermodular, and we can reduce them to tractable formulations. For ATO systems, we explore several special structures in which supermodularity holds.

The rest of this paper is organized as follows. In Section 2, we define the model and illustrate the requirement of supermodularity for tractability. In Section 3, we demonstrate the equivalent conditions for checking the supermodularity. In Section 4, we investigate the relationship between the optimality of affine decision rules and the supermodularity property. We then discuss applications in Section 5, provide numerical studies in Section 6, and finally conclude the paper in Section 7. For the sake of readability, we provide several extensions in Appendix B and all proofs are relegated to Appendix C with a table of contents in the beginning.

Notation and convention: For any integer $K \geq 1$, we define $[K] = \{1, \cdots, K\}$, which is the set of positive running indices to K. We use $|\cdot|$ to represent the cardinality of a set. We represent column vectors and matrices by lower- and upper-case boldface characters, respectively. An n-dimensional column vector $\boldsymbol{x}$ is equivalently denoted by $(x_1, \ldots, x_n)$, where we put all elements $x_i, i \in [n]$ in parenthesis and separate each element with a comma. For several matrices (or vectors) with compatible sizes, we use square brackets to join them together, e.g. $[\boldsymbol{A}\ \boldsymbol{B}]$ or $\begin{bmatrix} \boldsymbol{A} \\ \boldsymbol{B} \end{bmatrix}$. Given any matrix $\boldsymbol{A} = (a_{ij})_{i\in[m],j\in[n]} \in \Re^{m\times n}$, we let $\boldsymbol{a}_i^\top$ and $\boldsymbol{A}_j$ be its i-th row vector and j-th column vector, respectively. Further, we use $\boldsymbol{A}_{\mathcal{I}}$ to represent its submatrix $(a_{ij})_{i\in\mathcal{I},j\in[n]} \in \Re^{|\mathcal{I}|\times n}$ for any $\mathcal{I} \subseteq [m]$. We denote span$(\boldsymbol{A})$ to be the column space of $\boldsymbol{A}$. For any two vectors $\boldsymbol{x}', \boldsymbol{x}'' \in \Re^n$, we denote by $\boldsymbol{x}' \leq \boldsymbol{x}''$ if $x_i' \leq x_i''$ for all $i \in [n]$; moreover, we say $\boldsymbol{x}', \boldsymbol{x}''$ are ordered if either $\boldsymbol{x}' \leq \boldsymbol{x}''$ or $\boldsymbol{x}'' \leq \boldsymbol{x}'$, and they are unordered otherwise. We also define two operations join ("$\vee$") and meet ("$\wedge$") such that $\boldsymbol{x}' \vee \boldsymbol{x}'' = (\max\{x_i', x_i''\})_{i=1,\ldots,n}$ and $\boldsymbol{x}' \wedge \boldsymbol{x}'' = (\min\{x_i', x_i''\})_{i=1,\ldots,n}$ for any vectors $\boldsymbol{x}', \boldsymbol{x}'' \in \Re^n$. We let $\boldsymbol{e}_i$ be the vector with only the i-th entry being 1 and all others being 0, and $\mathbf{1}$ be the vector with all the entries being 1. Random variables are represented by characters with the tilde sign, for example, $\tilde{z}$ with z being its realization.

2. Tractability of Two-stage Problems with Supermodularity

In this section, we explore computational tractability in a special class of two-stage DRO problems which exhibit the property of supermodularity.

2.1. Model

The decision maker faces a two-stage problem. In the first stage, the decision maker must make the *here-and-now* decisions $\boldsymbol{x} \in \Re^l$ before the uncertainty $\tilde{\boldsymbol{z}}$, an n-dimensional random vector, is realized. After that, the uncertainty is revealed and observed by the decision maker, who then moves to the second stage and makes the *wait-and-see* decisions $\boldsymbol{y} \in \Re^m$. For a given first-stage decision $\boldsymbol{x}$ and an uncertainty realization $\boldsymbol{z}$, we denote the second-stage cost by $g(\boldsymbol{x},\boldsymbol{z})$. It can be evaluated by the following linear program,

$$\begin{aligned} g(\boldsymbol{x},\boldsymbol{z}) = \min \ & \boldsymbol{b}^\top \boldsymbol{y} \\ \text{s.t.} \ & \boldsymbol{W}\boldsymbol{x} + \boldsymbol{U}\boldsymbol{y} \geq \boldsymbol{V}\boldsymbol{z} + \boldsymbol{v}^0, \end{aligned} \tag{1}$$

where $\boldsymbol{b} \in \Re^m, \boldsymbol{W} \in \Re^{r\times l}, \boldsymbol{U} \in \Re^{r\times m}, \boldsymbol{V} \in \Re^{r\times n}$ and $\boldsymbol{v}^0 \in \Re^r$ are given constants, and $g(\boldsymbol{x},\boldsymbol{z}) = \infty$ if Problem (1) is infeasible. In our current setting, the uncertainties only appear on the right-hand side. This formulation has received extensive attention in the literature (see for instance, Zeng and Zhao 2013, Gupta et al. 2014, Bertsimas and Shtern 2018, El Housni and Goyal 2021) and is intractable in general (Feige et al. 2007, Bertsimas and Goyal 2012). Though it has uncertainties on the right-hand side only, this model can cover a broad range of practical two-stage problems, which will be introduced in Section 5. We will later generalize our results to include left-hand-side uncertainty in Appendix B.1.

We consider the distributionally robust setting such that the true distribution of $\tilde{\boldsymbol{z}}$ is only known to belong to an ambiguity set $\mathcal{F}$. Therefore, for a given first-stage decision $\boldsymbol{x}$, the expected second-stage cost is evaluated under the worst-case distribution and hence is

$$\sup_{\mathbb{P}\in\mathcal{F}} \ \mathbb{E}_{\mathbb{P}}\left[g(\boldsymbol{x},\tilde{\boldsymbol{z}})\right].$$

By choosing the first-stage decision $\boldsymbol{x}$, the decision maker aims to minimize the sum of the deterministic first-stage cost and the worst-case expected second-stage cost. It can be formulated as

$$\min_{\boldsymbol{x}\in\mathcal{X}} \left\{ \boldsymbol{a}^\top \boldsymbol{x} + \sup_{\mathbb{P}\in\mathcal{F}} \mathbb{E}_{\mathbb{P}}\left[g(\boldsymbol{x},\tilde{\boldsymbol{z}})\right] \right\}, \tag{2}$$

where $\boldsymbol{a} \in \Re^l$ is a given constant vector, $\mathcal{X} \subseteq \Re^l$ is the set of all feasible first-stage decisions. We assume that Problem (2) has a finite optimal value. We also consider risk averse objectives such as an expected disutility or risk measure of the second-stage cost. Interested readers are referred to Appendix B.2.

In order to capture the distributional information of $\tilde{\boldsymbol{z}}$, we adopt a special case of the scenario-wise ambiguity set which is recently proposed by Chen et al. (2020). Specifically, we assume

$$\mathcal{F} = \left\{ \mathbb{P} \ \left| \ \begin{array}{ll} \mathbb{E}_{\mathbb{P}}[\tilde{\boldsymbol{z}} \,|\, \tilde{k} = k] = \boldsymbol{\mu}^k, & \forall k \in [K] \\ \mathbb{E}_{\mathbb{P}}\left[|\tilde{z}_i - \mu_i^k| \,\big|\, \tilde{k} = k\right] \leq \delta_i^k, & \forall k \in [K], \ \forall i \in [n] \\ \mathbb{P}\big(\underline{\boldsymbol{z}}^k \leq \tilde{\boldsymbol{z}} \leq \overline{\boldsymbol{z}}^k \,\big|\, \tilde{k} = k\big) = 1, & \forall k \in [K] \\ \mathbb{P}(\tilde{k} = k) = q_k, & \forall k \in [K] \\ \boldsymbol{q} \in \mathcal{Q} & \end{array} \right. \right\}. \tag{3}$$

Here a random scenario $\tilde{k}$ is introduced and its realization affects the distributional information of $\tilde{\boldsymbol{z}}$. In particular, if the random scenario is realized as $k \in [K]$, we have corresponding distributional information for $\tilde{\boldsymbol{z}}$: mean being $\boldsymbol{\mu}^k$, MAD of $\tilde{z}_i$ being bounded by δ_i^k for all $i \in [n]$, and support being $[\underline{\boldsymbol{z}}^k, \overline{\boldsymbol{z}}^k]$. The probability that $\tilde{k}$ is realized as k is denoted by q_k. We also allow ambiguity in $\boldsymbol{q} = (q_k)_{k\in[K]}$ and only know that $\boldsymbol{q}$ is in a given polyhedron $\mathcal{Q} = \{\boldsymbol{q} \mid \boldsymbol{R}\boldsymbol{q} \le \boldsymbol{\nu}, \boldsymbol{q} \ge \boldsymbol{0}\} \subseteq \{\boldsymbol{q} \in \Re_+^K \mid \boldsymbol{1}^\top \boldsymbol{q} = 1\}$. Without loss of generality (WLOG), we make the following assumptions about $\mathcal{F}$ to avoid trivial cases. If there are i,k such that $\delta_i^k = 0$, by the constraint on MAD, $\tilde{z}_i$ realizes at μ_i^k almost surely when the random scenario $\tilde{k}$ takes value at k, and hence we can let $\overline{z}_i^k = \mu_i^k = \underline{z}_i^k$ for notational simplification. Similarly, for any i,k with $\mu_i^k \in \{\overline{z}_i^k, \underline{z}_i^k\}$, by the constraint on mean and MAD, we can also let $\overline{z}_i^k = \mu_i^k = \underline{z}_i^k$ and $\delta_i^k = 0$ for notational simplification. Moreover, the polyhedron $\mathcal{Q}$ is such that for all $k \in [K]$, there exists $\boldsymbol{q} \in \mathcal{Q}$ with $q_k > 0$, otherwise the scenario k almost surely does not happen and we can ignore it. Note that the distributional information for each scenario can be generalized to other expected piece-wise linear forms of the uncertainty. In Appendix B.3 we characterize the most general case of ambiguity sets that our methods can handle.

We first consider the case of $K = 1$. The distributional ambiguity set $\mathcal{F}$ is reduced to a conventional one with means, supports and MADs information, which has been studied in the literature. Examples can be seen in Ben-Tal and Hochman (1972), Qi (2017), Postek et al. (2019), Conejo et al. (2021), van Eekelen et al. (2022). In practice, the MAD information is also easy to estimate (Postek et al. 2018). Comparing with the general moment information, the MAD information allows us to derive a tractable formulation for the two-stage optimization problem and calculate exact solutions, as we will show later.

The incorporation of random scenarios brings modeling flexibility and can capture a broad class of information in a more intuitive way, e.g., multi-modal distribution or covariate information. It can also result in less conservative solutions than the case with a fixed scenario. When the set $\mathcal{Q}$ is a singleton and $\delta_i^k = 0$ for any $k \in [K], i \in [n]$, the information set $\mathcal{F}$ reduces to the case with a known discrete distribution. More illustration on the modeling power of $\mathcal{F}$ can be seen in Ghosal et al. (2021).

To explore the solvability of Problem (2), we will first investigate the worst-case distribution of $\tilde{\boldsymbol{z}}$ conditioning on a given scenario. After that, we provide a computationally tractable reformulation for Problem (2) with a random scenario, i.e., with $\mathcal{F}$ defined in Equation (3).

2.2. The case with a fixed scenario

When the scenario $\tilde{k}$ is realized as k for some $k \in [K]$, we define $\mathcal{F}^k$ to be a set of probability distributions in this specific scenario. That is,

$$\mathcal{F}^k = \left\{ \mathbb{P}^k \;\middle|\; \begin{array}{ll} \mathbb{E}_{\mathbb{P}^k}[\tilde{\boldsymbol{z}}] = \boldsymbol{\mu}^k, & \\ \mathbb{E}_{\mathbb{P}^k}[|\tilde{z}_i - \mu_i^k|] \le \delta_i^k, & \forall i \in [n] \\ \mathbb{P}^k\left(\underline{\boldsymbol{z}}^k \le \tilde{\boldsymbol{z}} \le \overline{\boldsymbol{z}}^k\right) = 1 & \end{array} \right\}. \tag{4}$$

We show that the worst-case distribution in the case of $\tilde{k} = k$ has the following characteristics.

Proposition 1 *For any $\boldsymbol{x}$, there exists $\mathbb{P}^{k*} \in \arg\sup_{\mathbb{P}^k \in \mathcal{F}^k} \mathbb{E}_{\mathbb{P}^k}[g(\boldsymbol{x},\tilde{\boldsymbol{z}})]$ such that for all $i \in [n]$, the marginal distribution is independent of $\boldsymbol{x}$ and can be calculated as*

$$\mathbb{P}^{k*}(\tilde{z}_i = w) = \begin{cases} \frac{\hat{\delta}_i^k}{2(\mu_i^k - \underline{z}_i^k)} & \text{if } w = \underline{z}_i^k \\ 1 - \frac{\hat{\delta}_i^k(\overline{z}_i^k - \underline{z}_i^k)}{2(\overline{z}_i^k - \mu_i^k)(\mu_i^k - \underline{z}_i^k)} & \text{if } w = \mu_i^k \\ \frac{\hat{\delta}_i^k}{2(\overline{z}_i^k - \mu_i^k)} & \text{if } w = \overline{z}_i^k \\ 0 & \text{otherwise,} \end{cases} \tag{5}$$

where $\hat{\delta}_i^k = \min\left\{\delta_i^k, \frac{2(\overline{z}_i^k - \mu_i^k)(\mu_i^k - \underline{z}_i^k)}{\overline{z}_i^k - \underline{z}_i^k}\right\}$ for all $i \in [n]$ with $\overline{z}_i^k > \underline{z}_i^k$.

According to Proposition 1, there exists a worst-case distribution such that at each dimension i, $i \in [n]$, the marginal distribution of $\tilde{z}_i$ has non-zero probability mass at only three points: the lower bound, mean and upper bound (for i with $\overline{z}_i^k = \underline{z}_i^k$, obviously $\mathbb{P}^{k*}(\tilde{z}_i = \overline{z}_i^k) = \mathbb{P}^{k*}(\tilde{z}_i = \mu_i^k) = \mathbb{P}^{k*}(\tilde{z}_i = \underline{z}_i^k) = 1$). Hence, to evaluate $\sup_{\mathbb{P}^k \in \mathcal{F}^k} \mathbb{E}_{\mathbb{P}^k}[g(\boldsymbol{x},\tilde{\boldsymbol{z}})]$, it suffices to focus on the distributions with support $\left\{\boldsymbol{z} \mid z_i \in \{\underline{z}_i^k, \mu_i^k, \overline{z}_i^k\},\ i \in [n]\right\}$. Unfortunately, since the number of these points grows exponentially in n, the two-stage problem is still computationally challenging to solve.

Postek et al. (2018) also consider the support-mean-MAD information and derive 3-point "worst-case marginals". There are two main differences between our ambiguity set and theirs. First, we consider the inequality form for the MAD information, while they use equality correspondingly. This inequality form incurs additional proof of monotonicity to show the optimality of 3-point marginals. Second, Postek et al. (2018) additionally restrict the uncertainties to be independent of each other. Hence, their worst-case joint distribution can be uniquely determined as the Cartesian product of the marginals (with a 3^n-point support). However, as we do not impose mutual independence between uncertain factors, we can only show the existence of a worst-case distribution with such marginals. Still, the exact joint distribution remains to be unknown in general.

We next show that if the function $g(\boldsymbol{x},\boldsymbol{z})$ is supermodular in $\boldsymbol{z}$, the joint distribution can be characterized efficiently and hence the computational burden can be eased. We first define supermodularity as follows.

Definition 1 *A function $f : \Re^n \to \Re$ is supermodular if $f(\boldsymbol{w}') + f(\boldsymbol{w}'') \le f(\boldsymbol{w}' \wedge \boldsymbol{w}'') + f(\boldsymbol{w}' \vee \boldsymbol{w}'')$ for all $\boldsymbol{w}', \boldsymbol{w}'' \in \Re^n$.*

In transportation and copula theory, it is well-known that when the uncertainty is two-dimensional, supermodularity leads to an explicit dependence structure of the worst-case distribution as follows.

Lemma 1 *(Rachev and Rüschendorf 1998) Consider any supermodular function $f : \Re^2 \to \Re$, and any two-dimensional random vector $\tilde{\boldsymbol{w}}$ with the marginal cumulative distribution function for $\tilde{w}_1, \tilde{w}_2$*

being F_1, F_2, respectively. Let $\mathcal{P} = \{\mathbb{P} \mid \mathbb{P}(\tilde{w}_i \le x) = F_i(x)\ \forall x \in \Re, i = 1, 2\}$ be the set of all possible distributions for $\tilde{\boldsymbol{w}}$. Then

$$\mathbb{E}_{\mathbb{P}}\left[f(\tilde{w}_1, \tilde{w}_2)\right] \le \int_0^1 f(F_1^{-1}(u), F_2^{-1}(u))du \quad \forall \mathbb{P} \in \mathcal{P}.$$

Clearly, the upper bound in Lemma 1 is achieved when $(\tilde{w}_1, \tilde{w}_2) \stackrel{d}{=} (F_1^{-1}(\tilde{u}), F_2^{-1}(\tilde{u}))$ with $\tilde{u}$ being uniformly distributed on $[0,1]$. In this worst-case distribution, considering any two realizations $\boldsymbol{w}', \boldsymbol{w}''$, we then have $u', u'' \in [0,1]$ such that $\boldsymbol{w}' = (F_1^{-1}(u'), F_2^{-1}(u'))$ and $\boldsymbol{w}'' = (F_1^{-1}(u''), F_2^{-1}(u''))$. This implies $\boldsymbol{w}', \boldsymbol{w}''$, and hence all pairs of realizations are ordered. Intuitively, this is because we can move the probability mass of any unordered pair to the corresponding join and meet, such that the marginal distribution is unchanged and the expectation of $f(\tilde{\boldsymbol{w}})$ increases due to the supermodularity of f. Interestingly, this result can be extended to the case with general dimensions and significantly reduces the number of possible realizations for the worst-case distribution.[1]

Proposition 2 *Consider any function $f : \Re^n \to \Re$. The following statements are equivalent.*

1) *f is supermodular.*
2) *Consider any given strictly positive integers m_i and $p_{ij} > 0, x_{ij}, j \in [m_i]$ such that $x_{i1} < \cdots < x_{im_i}$ and $\sum_{j \in [m_i]} p_{ij} = 1$, for all $i \in [n]$. Define $\mathcal{P} = \{\mathbb{P} \mid \mathbb{P}(\tilde{w}_i = x_{ij}) = p_{ij},\ j \in [m_i], i \in [n]\}$. Then there exists $\mathbb{P}^* \in \arg\sup_{\mathbb{P} \in \mathcal{P}} \mathbb{E}_{\mathbb{P}}\left[f(\tilde{\boldsymbol{w}})\right]$ such that the set $\mathcal{W}_{\mathbb{P}^*} = \{\boldsymbol{w} \in \Re^n \mid \mathbb{P}^*(\tilde{\boldsymbol{w}} = \boldsymbol{w}) > 0\}$ forms a chain of at most $(\sum_{i \in [n]}(m_i - 1) + 1)$ points.*

A chain is a partially ordered set which does not contain an unordered pair of elements. Moreover, Proposition 2 also shows that the chained structure is embedded in the worst-case distribution only if the function is supermodular. Figure 1 illustrates the intuition behind Proposition 2.

Intuitively, when moving the same amount of probability mass from any two points $\boldsymbol{w}', \boldsymbol{w}''$ to $\boldsymbol{w}' \wedge \boldsymbol{w}'', \boldsymbol{w}' \vee \boldsymbol{w}''$, the marginal distribution does not change but the expectation of $f(\tilde{\boldsymbol{w}})$ is higher because of the supermodularity of f. Hence, a worst-case distribution is to move all probability mass from the unordered pair to their join and meet. This seemingly leads to a worst-case distribution that is highly positively correlated and hence not always realistic in some applications. However, we will show later that our adoption of the scenario-wise ambiguity set addresses this issue and the worst-case distribution in our model can be correlated in any way.

According to Proposition 2, if $g(\boldsymbol{x}, \boldsymbol{z})$ is supermodular in $\boldsymbol{z}$, then the worst-case distribution for $\sup_{\mathbb{P}^k \in \mathcal{F}^k} \mathbb{E}[g(\boldsymbol{x}, \tilde{\boldsymbol{z}})]$ has a chained support. Nevertheless, the number of possible chains within the support can be exponentially large. Interestingly, with Proposition 1, which shows that the

[1] A concurrent work (Chen et al. 2022) makes similar extensions, but only for the case of continuous functions. Our work differs with theirs in two aspects. First, we do not restrict f to be a continuous function. Second, we show in Proposition 2 the necessity of supermodularity for such a chained structure of the worst-case distribution.

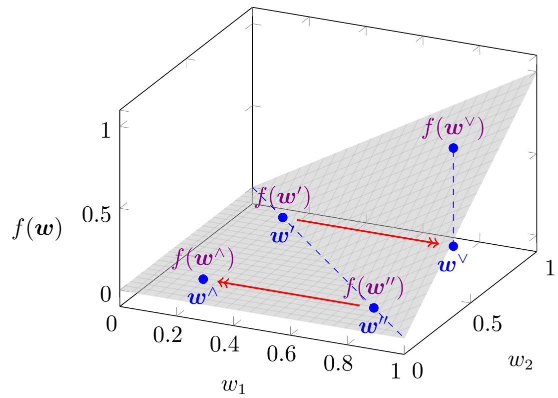


Figure 1 **Consider a distribution $\mathbb{P}$ placing positive probability masses at $\boldsymbol{w}', \boldsymbol{w}'', \boldsymbol{w}^{\wedge} = \boldsymbol{w}' \wedge \boldsymbol{w}'', \boldsymbol{w}^{\vee} = \boldsymbol{w}' \vee \boldsymbol{w}''$, where $\boldsymbol{w}', \boldsymbol{w}''$ are unordered. Denote $p^{\circ} = \min\{\mathbb{P}(\tilde{\boldsymbol{z}} = \boldsymbol{w}'), \mathbb{P}(\tilde{\boldsymbol{z}} = \boldsymbol{w}'')\}$. Moving the mass p° from $\boldsymbol{w}'$ ($\boldsymbol{w}''$) to $\boldsymbol{w}^{\vee}$ ($\boldsymbol{w}^{\wedge}$) does not change the marginal distributions, but we obtain a new probability distribution with higher expectation and one less unordered pair in the support.**

worst-case distribution for $\sup_{\mathbb{P}^k \in \mathcal{F}^k} \mathbb{E}_{\mathbb{P}^k}[g(\boldsymbol{x}, \tilde{\boldsymbol{z}})]$ has an explicit three-point distribution in each dimension, we can find this chained support efficiently. We formalize the results, by the following Algorithm 1 and Proposition 3, to explore this worst-case distribution.

Algorithm 1 algorithm for worst-case distribution

Input: $\mathcal{F}^k$ in Equation (4) with given $\boldsymbol{\mu}^k$, $\boldsymbol{\delta}^k$, $\underline{\boldsymbol{z}}^k$, $\overline{\boldsymbol{z}}^k$

Initialization:

- denote $\mathbb{P}^{k*}$ as the worst-case distribution in Proposition 1 and calculate $\mathbb{P}^{k*}(\tilde{z}_i^k = w)$ for $w \in \{\underline{z}_i^k, \mu_i^k, \overline{z}_i^k\}$, $i \in [n]$ using Equation (5)
- $\boldsymbol{z}^1 = \underline{\boldsymbol{z}}^k$, $\boldsymbol{q}^1 = (\mathbb{P}^{k*}(\tilde{z}_1^k = \underline{z}_1^k), \mathbb{P}^{k*}(\tilde{z}_2^k = \underline{z}_2^k) \ldots, \mathbb{P}^{k*}(\tilde{z}_n^k = \underline{z}_n^k))$, $p_1 = \min\{q_1^1, \ldots, q_n^1\}$ and $j = 1$

while $j \leq 2n$ **do**

    choose r_j as the minimal index in $[n]$ such that $q_{r_j}^j = p_j$

    $\boldsymbol{z}^{j+1} = \boldsymbol{z}^j$, $\boldsymbol{q}^{j+1} = \boldsymbol{q}^j - p_j \mathbf{1}$

    update $z_{r_j}^{j+1} = \mu_{r_j}^k$ if its existing value is $\underline{z}_{r_j}^k$, and $z_{r_j}^{j+1} = \overline{z}_{r_j}^k$ if its existing value is $\mu_{r_j}^k$

    update $q_{r_j}^{j+1} = \mathbb{P}^{k*}(\tilde{z}_{r_j} = z_{r_j}^{j+1})$

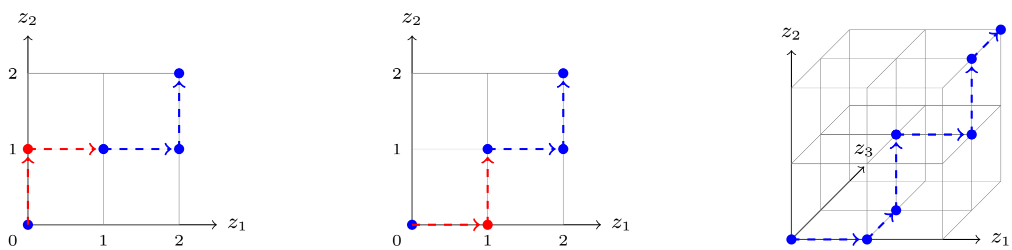
    $p_{j+1} = \min\{q_1^{j+1}, q_2^{j+1}, \ldots, q_n^{j+1}\}$

    update $j = j + 1$

return $\boldsymbol{z}^1, \boldsymbol{z}^2, \ldots, \boldsymbol{z}^{2n+1}$ and $\boldsymbol{p} = (p_1, p_2, \ldots, p_{2n+1})$

Proposition 3 *For any $\boldsymbol{x}$, if $g(\boldsymbol{x}, \boldsymbol{z})$ is supermodular in $\boldsymbol{z}$, we have $\sup_{\mathbb{P}^k \in \mathcal{F}^k} \mathbb{E}[g(\boldsymbol{x}, \tilde{\boldsymbol{z}})] = \sum_{i \in [2n+1]} p_i g(\boldsymbol{x}, \boldsymbol{z}^i)$. Here $\boldsymbol{p}, \boldsymbol{z}^i, i \in [2n+1]$ are output by Algorithm 1 whose time complexity is $O(n)$.*

By moving from $\underline{\boldsymbol{z}}^k$ to $\overline{\boldsymbol{z}}^k$, Algorithm 1 identifies a feasible chain, subject to the marginal distribution provided by Proposition 1. Then Proposition 3 shows that such a feasible chain must constitute the support of the worst-case distribution. The intuition is that there is only one feasible chain satisfying the given marginals. Since the support of the worst-case distribution is a chain (by Proposition 2), the chain identified by Algorithm 1 must be the right one and corresponds to the worst-case distribution. Consequently, Proposition 3 provides an explicit formulation of the worst-case joint distribution. Figure 2 provides examples when the dimension n is 2 or 3.



(a) dimension $n=2$ with $\mathbb{P}(\tilde{z}_1=0)>\mathbb{P}(\tilde{z}_2=0)$

(b) dimension $n=2$ with $\mathbb{P}(\tilde{z}_1=0)<\mathbb{P}(\tilde{z}_2=0)$

(c) dimension $n=3$

Figure 2 **The support of worst-case distributions. For the case of $n=2$, Figures (a) and (b) demonstrate how the chained support can be uniquely determined. They both start from the origin, $(0,0)$. For the case of $\mathbb{P}(\tilde{z}_1=0)>\mathbb{P}(\tilde{z}_2=0)$, as in (a), the next point cannot be $(1,0)$. Similarly, for the case of $\mathbb{P}(\tilde{z}_1=0)<\mathbb{P}(\tilde{z}_2=0)$, as in (b), the next point has to be $(1,0)$ and cannot be $(0,1)$. Figure (c) gives an example of 3-dimensional chain.**

Since the worst-case distribution returned by Algorithm 1 has support on only $(2n+1)$ points and is also independent of the first-stage decision $\boldsymbol{x}$, we can simplify the two-stage optimization problem. While one might criticize that it is rather extreme to have a worst-case distribution independent of the first-stage decision, we remark that such independence is only true when the scenario probabilities $q_k, k\in[K]$ are pre-determined. Indeed, the overall worst-case distribution depends on the first-stage decision since it affects the worst-case probability distribution of the uncertain scenario. This will be demonstrated in the next subsection.

2.3. Incorporating the uncertain scenario

In solving the general two-stage optimization problem (2), Proposition 3 shows how to evaluate the second-stage expected cost efficiently under the worst-case distribution when the uncertain scenario realizes as k. We now incorporate the uncertainty in the scenario $\tilde{k}$.

Based on the definition of $\mathcal{F}$ and $\mathcal{F}^k$ in Equations (3) and (4), we have

$$\sup_{\mathbb{P}\in\mathcal{F}} \mathbb{E}_{\mathbb{P}}\left[g(\boldsymbol{x},\tilde{\boldsymbol{z}})\right] = \max_{\boldsymbol{q}\in\mathcal{Q}} \sup_{\mathbb{P}^k\in\mathcal{F}^k, k\in[K]} \sum_{k\in[K]} q_k \mathbb{E}_{\mathbb{P}^k}\left[g(\boldsymbol{x},\tilde{\boldsymbol{z}})\right] = \max_{\boldsymbol{q}\in\mathcal{Q}} \sum_{k\in[K]} q_k \sup_{\mathbb{P}^k\in\mathcal{F}^k} \mathbb{E}_{\mathbb{P}^k}\left[g(\boldsymbol{x},\tilde{\boldsymbol{z}})\right].$$

We denote by $\boldsymbol{z}^{k,1},\ldots,\boldsymbol{z}^{k,2n+1},\boldsymbol{p}^k$ the output of Algorithm 1 with input $\mathcal{F}^k$ for all $k\in[K]$. Since $\mathcal{Q}$ is a polyhedron, we then have the following reformulation.

Theorem 1 *If $g(\boldsymbol{x},\boldsymbol{z})$ is supermodular in $\boldsymbol{z}$ for any $\boldsymbol{x}$, Problem (2) is equivalent to the following linear program,*

$$
\begin{aligned}
\min\quad & \boldsymbol{a}^\top\boldsymbol{x}+\boldsymbol{\nu}^\top\boldsymbol{l} \\
\text{s.t.}\quad & \boldsymbol{R}_k^\top\boldsymbol{l}\geq\sum_{i\in[2n+1]}p_i^k\boldsymbol{b}^\top\boldsymbol{y}^{k,i}, && k\in[K] \\
& \boldsymbol{W}\boldsymbol{x}+\boldsymbol{U}\boldsymbol{y}^{k,i}\geq\boldsymbol{V}\boldsymbol{z}^{k,i}+\boldsymbol{v}^0, && k\in[K],i\in[2n+1] \\
& \boldsymbol{l}\geq\boldsymbol{0},\ \boldsymbol{x}\in\mathcal{X}.
\end{aligned}
\tag{6}
$$

Intuitively, the reformulation in Theorem 1 incorporates all possible realizations in the worst-case distribution, and assigns a corresponding second-stage decision to each of those realizations. Therefore, the two-stage problem can be formulated as a static linear optimization problem. Nevertheless, the classic approach using this idea has to handle an exponential number of realizations, leading to computational intractability. Here, by exploring the potential property of supermodularity in the uncertainties, we reduce the number of realizations to $K(2n+1)$, which is of polynomial size and makes the problem tractable.

Moreover, our approach works without requiring relatively complete recourse. This is because Problem (6) is an equivalent reformulation of the original problem, and hence Problem (6) maintains the same feasibility for any given first-stage decision $\boldsymbol{x}$. Indeed, the feasibility issue, which is the essential focus of the relatively complete recourse requirement in typical two-stage problems, is addressed by the assumption of supermodularity of $g(\boldsymbol{x},\boldsymbol{z})$ already. In particular, if Problem (6) has a finite optimal value, then at the optimal $\boldsymbol{x}$, the second-stage problem is feasible when $\tilde{\boldsymbol{z}}$ takes any pre-determined realizations ($\boldsymbol{z}^{k,i}$ in Problem (6)). By supermodularity, these pre-determined realizations constitute the worst-case distribution. It implies that when $\tilde{\boldsymbol{z}}$ takes other realizations, the second-stage cost should also be finite, i.e., the second-stage problem is feasible. We further elaborate this by the following corollary.

Corollary 1 *If $\boldsymbol{x}_{opt}$ is optimal to Problem (6), then for all $\boldsymbol{z}\in\bigcup_{k\in[K]}\left[\underline{\boldsymbol{z}}^k,\overline{\boldsymbol{z}}^k\right]$, $g(\boldsymbol{x}_{opt},\boldsymbol{z})$ is finite, i.e., the second-stage problem is feasible when $\boldsymbol{x}=\boldsymbol{x}_{opt}$.*

We also remark that, given a scenario realization k, the worst-case distribution may be positively correlated. However, by incorporating the random scenario, the correlation between any pair of uncertain factors can be negative. For example, when $\delta_i^k=0$, for all $i\in[n],k\in[K]$, we have that $\mathcal{F}^k=\{\mathbb{P}^k\mid\mathbb{P}^k(\tilde{\boldsymbol{z}}=\boldsymbol{\mu}^k)=1\}$. The distributional uncertainty set $\mathcal{F}$ reduces to a set of discrete distributions that can be positively or negatively correlated.

3. Conditions for supermodularity of the second-stage problems

If the second-stage cost, $g(\boldsymbol{x},\boldsymbol{z})$, is supermodular in $\boldsymbol{z}$, Section 2 has shown that a tractable formulation can be achieved. Unfortunately, supermodularity in $\boldsymbol{z}$ is not a feature embedded in all two-stage problems. It depends on the structure of the two-stage problem. In this section, we aim to identify a broad class of two-stage problems where the second-stage cost is supermodular in the uncertain factors.

We reformulate the second-stage cost $g(\boldsymbol{x},\boldsymbol{z})$ as

$$g(\boldsymbol{x},\boldsymbol{z}) = \min \quad \boldsymbol{b}^\top \boldsymbol{y} \qquad \text{s.t.} \quad \boldsymbol{U}\boldsymbol{y} - \boldsymbol{V}\boldsymbol{z} \geq -\boldsymbol{W}\boldsymbol{x} + \boldsymbol{v}^0. \tag{7}$$

It is the optimal value of a parametric optimization problem which is parametrized by $\boldsymbol{z}$, and we need to explore the supermodularity in this parameter. Note that since we only focus on the supermodularity in $\boldsymbol{z}$ but not in $\boldsymbol{x}$, we do not consider $\boldsymbol{x}$ as a parameter in this parametric optimization problem. Hence, we move $\boldsymbol{x}$ to the right-hand-side of the constraint. In the parametric optimization literature, the supermodularity of the optimal value in parameters has been studied systematically for maximization problems (Chen et al. 2013, 2021). However, in Equation (7), we have a minimization problem, which leads to an essential difference from previous studies. It is worth mentioning that while a minimization problem can be formulated as an equivalent maximization problem, inevitably, that reformulation exchanges supermodularity for submodularity. In particular, if we equivalently represent $g(\boldsymbol{x},\boldsymbol{z}) = -\max\left\{-\boldsymbol{b}^\top\boldsymbol{y} \;\middle|\; \boldsymbol{U}\boldsymbol{y} - \boldsymbol{V}\boldsymbol{z} \geq -\boldsymbol{W}\boldsymbol{x} + \boldsymbol{v}^0\right\}$, then the supermodularity condition for g is equivalent to the submodularity condition for the inner maximization problem. It is then again different from the literature which is on supermodularity for maximization problems. Therefore, we cannot rely on the literature of maximization problems to resolve the challenge of the minimization problem. For further illustration, some operations management literature has indicated the significant difference between the supermodularity in maximization (which implies complementarity) and in minimization (which implies substitutability) problems. For example, in perishable inventory control (Chen 2017) and ATO problems (e.g., Lu and Song 2005, Nadar et al. 2014), a monotone structure can be shown in the optimal policies when components are complementary, but similar structural analysis cannot be obtained when components are substitutable.

Typically, the lattice structure of the feasible set is a key for supermodularity in the parametric maximization problem. By contrast, to investigate the parametric minimization problem, we introduce the following concept called the *inverse additive lattice*.

Definition 2 *Given two positive integers m,n, a set $\mathcal{S} \subseteq \Re^m \times \Re^n$ is an inverse additive lattice if for any $\boldsymbol{p},\boldsymbol{q} \in \Re^m$, $\boldsymbol{z}',\boldsymbol{z}'' \in \Re^n$ with $(\boldsymbol{p},\boldsymbol{z}' \wedge \boldsymbol{z}''),(\boldsymbol{q},\boldsymbol{z}' \vee \boldsymbol{z}'') \in \mathcal{S}$, there exist $\boldsymbol{y}',\boldsymbol{y}'' \in \Re^m$ such that $(\boldsymbol{y}',\boldsymbol{z}'),(\boldsymbol{y}'',\boldsymbol{z}'') \in \mathcal{S}$ and $\boldsymbol{y}' + \boldsymbol{y}'' = \boldsymbol{p} + \boldsymbol{q}$.*

We now show that the inverse additive lattice is a necessary and sufficient condition for supermodularity in the parametric minimization problem. Given any first-stage decision $\boldsymbol{x}$, we denote the set of all feasible pairs of $(\boldsymbol{y},\boldsymbol{z})$ as $\mathcal{S}(\boldsymbol{x})$, i.e.,

$$\mathcal{S}(\boldsymbol{x}) = \left\{(\boldsymbol{y},\boldsymbol{z}) \mid \boldsymbol{U}\boldsymbol{y} - \boldsymbol{V}\boldsymbol{z} \geq -\boldsymbol{W}\boldsymbol{x} + \boldsymbol{v}^0\right\}. \tag{8}$$

Proposition 4 *Given any $\boldsymbol{x}$, $g(\boldsymbol{x},\boldsymbol{z})$ is supermodular in $\boldsymbol{z}$ for any $\boldsymbol{b}$ if and only if $\mathcal{S}(\boldsymbol{x})$ is an inverse additive lattice.*

Proposition 4 presents a necessary and sufficient condition for the second-stage cost being supermodular in the uncertainty $\boldsymbol{z}$ for a given first-stage decision $\boldsymbol{x}$. Now it remains to characterize the structure of the second-stage problem such that the condition can always be satisfied for any $\boldsymbol{x}$.

Theorem 2 *$g(\boldsymbol{x},\boldsymbol{z})$ is supermodular in $\boldsymbol{z}$ for any $\boldsymbol{x},\boldsymbol{b}$ and $\boldsymbol{v}^0$ if and only if $\boldsymbol{U}\in\Re^{r\times m}$ and $\boldsymbol{V}\in\Re^{r\times n}$ satisfy one of the following conditions:*

1) $\operatorname{rank}(\boldsymbol{U}) = r$,

2) for all $\mathcal{I}\subseteq[r]$, $\boldsymbol{\beta}\in\Re^n_+$ satisfying $|\mathcal{I}| = \operatorname{rank}(\boldsymbol{U})+1$, $\operatorname{rank}(\boldsymbol{U}_\mathcal{I}) = \operatorname{rank}(\boldsymbol{U})$ and $\boldsymbol{V}_\mathcal{I}\boldsymbol{\beta}\in\operatorname{span}(\boldsymbol{U}_\mathcal{I})$, we must have $\beta_i(\boldsymbol{V}_\mathcal{I})_i\in\operatorname{span}(\boldsymbol{U}_\mathcal{I})$ holds for every $i\in[n]$.

For any matrices $\boldsymbol{U},\boldsymbol{V}$, we introduce an algorithm in Appendix A.1 to check explicitly whether the condition in Theorem 2 is met. We next provide the following examples for illustration.

- $\boldsymbol{U} = \boldsymbol{I}_{r\times r}$ or $\boldsymbol{U} = [\boldsymbol{I}_{r\times r}\ \boldsymbol{U}^\circ]$ for some $\boldsymbol{U}^\circ\in\Re^{r\times(m-r)}$. The first condition of Theorem 2 is satisfied, hence $g(\boldsymbol{x},\boldsymbol{z})$ is supermodular in $\boldsymbol{z}$ for any arbitrary $\boldsymbol{V}\in\Re^{r\times n}$.
- $\boldsymbol{U} = \begin{bmatrix}\boldsymbol{I}_{m\times m}\\ \boldsymbol{u}_r^\top\end{bmatrix}$ for some $\boldsymbol{u}_r\in\Re^m$. In this case, $\operatorname{span}(\boldsymbol{U}) = \left\{\boldsymbol{\xi}\in\Re^r \;\middle|\; \sum_{i\in[m]} u_{ri}\xi_i = \xi_r\right\}$. Correspondingly, based on the second condition of Theorem 2, we can prove $g(\boldsymbol{x},\boldsymbol{z})$ is supermodular in $\boldsymbol{z}$ if and only if $(\boldsymbol{u}_r,-1)^\top\boldsymbol{V}_1,\ldots,(\boldsymbol{u}_r,-1)^\top\boldsymbol{V}_n$ have the same sign.
- Given any $g(\boldsymbol{x},\boldsymbol{z}) = \min\left\{\boldsymbol{b}^\top\boldsymbol{y} \mid \boldsymbol{U}\boldsymbol{y} - \boldsymbol{V}\boldsymbol{z} \geq -\boldsymbol{W}\boldsymbol{x} + \boldsymbol{v}^0\right\}$, we consider the problem with partial constraints, i.e., $g^\mathcal{I}(\boldsymbol{x},\boldsymbol{z}) = \min\left\{\boldsymbol{b}^\top\boldsymbol{y} \mid \boldsymbol{U}_\mathcal{I}\boldsymbol{y} - \boldsymbol{V}_\mathcal{I}\boldsymbol{z} \geq -\boldsymbol{W}_\mathcal{I}\boldsymbol{x} + \boldsymbol{v}^0_\mathcal{I}\right\}$ for some $\mathcal{I}\subseteq[r]$. If $g(\boldsymbol{x},\boldsymbol{z})$ is supermodular in $\boldsymbol{z}$, so is $g^\mathcal{I}(\boldsymbol{x},\boldsymbol{z})$.
- $\boldsymbol{U} = \begin{bmatrix}\boldsymbol{I}_{m\times m}\\ \boldsymbol{U}^\circ\end{bmatrix}\in\Re^{r\times m}, \boldsymbol{V} = \begin{bmatrix}\boldsymbol{V}^1\\ \boldsymbol{0}_{(r-m)\times n}\end{bmatrix}\in\Re^{r\times n}$. This choice of $\boldsymbol{U}$ and $\boldsymbol{V}$ includes the ATO system, the detail of which will be discussed later, as a special case. The corresponding result can be formalized as follows.

Proposition 5 *Assume $\boldsymbol{U} = \begin{bmatrix}\boldsymbol{I}_{m\times m}\\ \boldsymbol{U}^\circ\end{bmatrix}\in\Re^{r\times m}, \boldsymbol{V} = \begin{bmatrix}\boldsymbol{V}^1\\ \boldsymbol{0}_{(r-m)\times n}\end{bmatrix}\in\Re^{r\times m}_+$ and for each row of $\boldsymbol{U}^\circ$, all components have the same sign. The function $g(\boldsymbol{x},\boldsymbol{z})$ is supermodular in $\boldsymbol{z}$ for any $\boldsymbol{x},\boldsymbol{b},\boldsymbol{v}^0$ and $\boldsymbol{V}^1$, if and only if every 2×3 submatrix of $\boldsymbol{U}^\circ$ contains at least one pair of column vectors which are linearly dependent.*

By now, given any two-stage optimization problem (2), we can use the conditions in Theorem 2 (or Algorithm 2 in Appendix A.1) to verify whether the second-stage cost function is supermodular in $\boldsymbol{z}$. If the answer is positive, we can use the result in Theorem 1 to obtain an equivalent formulation as Problem (6), and derive the optimal solution efficiently.

4. Optimality of segregated affine decision rules

Though leading to sub-optimal solutions in general, affine decision rules have been widely applied in solving two-stage problems due to their computational efficiency. Interestingly, leveraging the benefits of the $K(2n+1)$-point worst-case distribution, which is derived from the ambiguity set $\mathcal{F}$ and supermodularity, we show that a scenario-wise segregated affine decision rule, which generalizes the classic one proposed by Chen et al. (2008), Goh and Sim (2010) to be scenario dependent, can return the optimal solution for Problem (2) subject to some conditions.

We observe that in the two-stage problem, the second-stage decision $\boldsymbol{y}$ is indeed a function of the uncertainty realization $(\tilde{k}, \tilde{\boldsymbol{z}})$. With a slight abuse of notation, we denote the second-stage decision as a function $\boldsymbol{y}(k, \boldsymbol{z})$, and hence our main problem (2) can be formulated equivalently as

$$\begin{aligned}
\min_{\boldsymbol{x}, \boldsymbol{y}(k, \tilde{\boldsymbol{z}})} \quad & \boldsymbol{a}^\top \boldsymbol{x} + \sup_{\mathbb{P} \in \mathcal{F}} \ \mathbb{E}_{\mathbb{P}}[\boldsymbol{b}^\top \boldsymbol{y}(\tilde{k}, \tilde{\boldsymbol{z}})] \\
\text{s.t.} \quad & \boldsymbol{W}\boldsymbol{x} + \boldsymbol{U}\boldsymbol{y}(k, \boldsymbol{z}) \geq \boldsymbol{V}\boldsymbol{z} + \boldsymbol{v}^0 \qquad \forall \boldsymbol{z} \in [\underline{\boldsymbol{z}}^k, \bar{\boldsymbol{z}}^k], k \in [K], \\
& \boldsymbol{x} \in \mathcal{X}.
\end{aligned} \tag{9}$$

In general, the above formulation involves a functional decision $\boldsymbol{y}(\tilde{k}, \tilde{\boldsymbol{z}})$ and hence induces computational complexity. We now prove that in our setting, it might suffice to consider the class of segregated affine functions for the optimal decision.

To this end, we start by considering the case that the uncertain scenario $\tilde{k}$ realizes at a given $k \in [K]$. Proposition 3 has shown that the worst-case distribution is a $(2n+1)$-point distribution. We follow the notation in Section 2.3 and denote the corresponding support, which is the output of Algorithm 1 with input $\mathcal{F}^k$, as $\boldsymbol{z}^{k,1}, \ldots, \boldsymbol{z}^{k,2n+1} \in \Re^n$. We first lift the support to $\Re^{2n}$ by defining

$$\boldsymbol{\zeta}^{k,i} = \begin{bmatrix} \boldsymbol{\omega}^{k,i} \\ \boldsymbol{v}^{k,i} \end{bmatrix} \tag{10}$$

where $\boldsymbol{\omega}^{k,i} = (\boldsymbol{\mu}^k - \boldsymbol{z}^{k,i})^+, \boldsymbol{v}^{k,i} = (\boldsymbol{z}^{k,i} - \boldsymbol{\mu}^k)^+, i \in [2n+1]$. The following result presents a geometric property of $\boldsymbol{\zeta}^{k,i}, i \in [2n+1]$.

Lemma 2 *For any given $k \in [K]$, the convex hull of $\{\boldsymbol{\zeta}^{k,1}, \ldots, \boldsymbol{\zeta}^{k,2n+1}\}$ is a $2n$-simplex.*

Using the above property, we first show the optimality of a segregated affine decision rule in a revised formulation. Specifically, we restrain the recourse decision $\boldsymbol{y}(k,\boldsymbol{z})$ to be an affine function in $\begin{bmatrix}(\boldsymbol{\mu}^k-\boldsymbol{z})^+\\(\boldsymbol{z}-\boldsymbol{\mu}^k)^+\end{bmatrix}$, and obtain the following problem based on Problem (9),

$$\begin{aligned}\min_{\boldsymbol{x},\boldsymbol{\Theta}^k,\boldsymbol{\phi}^k,k\in[K]}\quad & \boldsymbol{a}^\top\boldsymbol{x}+\sup_{\mathbb{P}\in\mathcal{F}}\ \mathbb{E}_{\mathbb{P}}\left[\boldsymbol{b}^\top\left(\boldsymbol{\Theta}^{\tilde{k}}\begin{bmatrix}(\boldsymbol{\mu}^{\tilde{k}}-\tilde{\boldsymbol{z}})^+\\(\tilde{\boldsymbol{z}}-\boldsymbol{\mu}^{\tilde{k}})^+\end{bmatrix}+\boldsymbol{\phi}^{\tilde{k}}\right)\right]\\ \text{s.t.}\quad & \boldsymbol{W}\boldsymbol{x}+\boldsymbol{U}\left(\boldsymbol{\Theta}^k\begin{bmatrix}(\boldsymbol{\mu}^k-\boldsymbol{z})^+\\(\boldsymbol{z}-\boldsymbol{\mu}^k)^+\end{bmatrix}+\boldsymbol{\phi}^k\right)\geq\boldsymbol{V}\boldsymbol{z}+\boldsymbol{v}^0\qquad\forall\boldsymbol{z}\in\{\boldsymbol{z}^{k,1},\ldots,\boldsymbol{z}^{k,2n+1}\},k\in[K],\\ & \boldsymbol{x}\in\mathcal{X}.\end{aligned}\tag{11}$$

Denote the optimal solution for $(\boldsymbol{x},\boldsymbol{y}^{k,i})$ to Problem (6) by $(\boldsymbol{x}_{opt},\boldsymbol{y}^{k,i}_{opt})$, $k\in[K],i\in[2n+1]$, which can be considered as given constants. Further, for any $k\in[K]$, we define a matrix $\boldsymbol{D}^k\in\Re^{2n\times 2n}$, a matrix $\boldsymbol{\Theta}^k_{opt}\in\Re^{m\times 2n}$ and a vector $\boldsymbol{\phi}^k_{opt}\in\Re^m$ as follows,

$$\begin{aligned}\boldsymbol{D}^k&=\left[\boldsymbol{\zeta}^{k,1}-\boldsymbol{\zeta}^{k,2n+1}\ \cdots\ \boldsymbol{\zeta}^{k,2n}-\boldsymbol{\zeta}^{k,2n+1}\right],\\ \boldsymbol{\Theta}^k_{opt}&=\left[\boldsymbol{y}^{k,1}_{opt}-\boldsymbol{y}^{k,2n+1}_{opt}\ \cdots\ \boldsymbol{y}^{k,2n}_{opt}-\boldsymbol{y}^{k,2n+1}_{opt}\right](\boldsymbol{D}^k)^{-1},\\ \boldsymbol{\phi}^k_{opt}&=\boldsymbol{y}^{k,2n+1}_{opt}-\boldsymbol{\Theta}^k_{opt}\boldsymbol{\zeta}^{k,2n+1},\end{aligned}$$

where $\boldsymbol{\zeta}^{k,i}$ is the lifted uncertainty realization defined in Equation (10). Note that $\boldsymbol{D}^k$ is invertible since by Lemma 2, $\boldsymbol{\zeta}^{k,1},\ldots,\boldsymbol{\zeta}^{k,2n+1}$ are affinely independent, and hence $\boldsymbol{\Theta}^k_{opt}$ is well defined. We then have the following result.

Proposition 6 *If $g(\boldsymbol{x},\boldsymbol{z})$ is supermodular in $\boldsymbol{z}$ for any $\boldsymbol{x}$, then Problem (11) and Problem (9) have the same optimal value. Specifically, $\boldsymbol{x}=\boldsymbol{x}_{opt}$, $\boldsymbol{\Theta}^k=\boldsymbol{\Theta}^k_{opt}$ and $\boldsymbol{\phi}^k=\boldsymbol{\phi}^k_{opt}$, $k\in[K]$ is an optimal solution for Problem (11).*

By Proposition 6, with the supermodularity of $g(\boldsymbol{x},\boldsymbol{z})$ in $\boldsymbol{z}$, the optimal value and optimal first-stage decision can be solved by restricting the second-stage decision as affinely dependent on the lifted uncertainty realization $\begin{bmatrix}(\boldsymbol{\mu}^k-\boldsymbol{z})^+\\(\boldsymbol{z}-\boldsymbol{\mu}^k)^+\end{bmatrix}$.

It is worth mentioning that when we change the original optimization problem (9) to the affine decision rule formulation (11), we do not enforce the constraint to be feasible for all possible $\boldsymbol{z}$. Instead, we only enforce the constraint for the $(2n+1)$ realizations of $\tilde{\boldsymbol{z}}$ at each scenario. This is for two reasons. First, if we enforce the constraint for all possible $\boldsymbol{z}$, the affine decision rule formulation of the original problem becomes

$$\begin{aligned}\min_{\boldsymbol{x},\boldsymbol{\Theta}^k,\boldsymbol{\phi}^k,k\in[K]}\quad & \boldsymbol{a}^\top\boldsymbol{x}+\sup_{\mathbb{P}\in\mathcal{F}}\ \mathbb{E}_{\mathbb{P}}\left[\boldsymbol{b}^\top\left(\boldsymbol{\Theta}^{\tilde{k}}\begin{bmatrix}(\boldsymbol{\mu}^{\tilde{k}}-\tilde{\boldsymbol{z}})^+\\(\tilde{\boldsymbol{z}}-\boldsymbol{\mu}^{\tilde{k}})^+\end{bmatrix}+\boldsymbol{\phi}^{\tilde{k}}\right)\right]\\ \text{s.t.}\quad & \boldsymbol{W}\boldsymbol{x}+\boldsymbol{U}\left(\boldsymbol{\Theta}^k\begin{bmatrix}(\boldsymbol{\mu}^k-\boldsymbol{z})^+\\(\boldsymbol{z}-\boldsymbol{\mu}^k)^+\end{bmatrix}+\boldsymbol{\phi}^k\right)\geq\boldsymbol{V}\boldsymbol{z}+\boldsymbol{v}^0\qquad\forall\boldsymbol{z}\in[\underline{\boldsymbol{z}}^k,\overline{\boldsymbol{z}}^k],k\in[K],\\ & \boldsymbol{x}\in\mathcal{X}.\end{aligned}\tag{12}$$

The lifted set $\left\{(\boldsymbol{\omega},\boldsymbol{v}) \mid \boldsymbol{\omega}=(\boldsymbol{\mu}^k-\boldsymbol{z})^+,\boldsymbol{v}=(\boldsymbol{z}-\boldsymbol{\mu}^k)^+,\boldsymbol{z}\in[\underline{\boldsymbol{z}}^k,\overline{\boldsymbol{z}}^k]\right\}$ is not guaranteed to be convex. As a result, despite having an affine structure, Problem (12) is still computationally intractable. We refer the interested readers to Sections 4.5, 4.6 of Goh and Sim (2010) for a detailed illustration of lifted sets of such kind. Without computational tractability, it is meaningless to investigate the corresponding affine decision rule formulation. The second reason is that the first-stage decision is usually the essential focus in two-stage optimization problems. Proposition 6 shows that solving Problem (11) can provide the optimal value as well as the optimal first-stage decision.

Indeed, in the segregated affine decision rule formulation, if enforcing the constraint to all possible $\boldsymbol{z}$, i.e., adopting the formulation (12) instead of (11), in general, we may not get the same optimal solution as Problem (9). A counterexample is provided in Appendix A.2 for illustration. This concludes that the supermodularity property in the second-stage function cannot guarantee the optimality of the segregated affine decision rule in Problem (12). Interestingly, by adding slightly more conditions, we can still obtain the optimal solution from the affine decision rule formulation (12). Recall that $\mathcal{S}(\boldsymbol{x})$, as defined in Equation (8), is the feasible region of $(\boldsymbol{y},\boldsymbol{z})$ given first-stage decision $\boldsymbol{x}$.

Theorem 3 *Suppose $\mathcal{S}(\boldsymbol{x})$ satisfies that for any given $\boldsymbol{x}\in\mathcal{X}$, $\boldsymbol{p},\boldsymbol{q},\boldsymbol{y}'\in\Re^m$, $\boldsymbol{z}',\boldsymbol{z}''\in\Re^n$ with $(\boldsymbol{p},\boldsymbol{z}'\wedge\boldsymbol{z}''),(\boldsymbol{q},\boldsymbol{z}'\vee\boldsymbol{z}''),(\boldsymbol{y}',\boldsymbol{z}')\in\mathcal{S}(\boldsymbol{x})$, we must have $(\boldsymbol{p}+\boldsymbol{q}-\boldsymbol{y}',\boldsymbol{z}'')\in\mathcal{S}(\boldsymbol{x})$. Then Problem (12) has the same optimal value with Problem (9). Specifically, $\boldsymbol{x}=\boldsymbol{x}_{opt}$, $\boldsymbol{\Theta}^k=\boldsymbol{\Theta}^k_{opt}$ and $\boldsymbol{\phi}^k=\boldsymbol{\phi}^k_{opt}$, $k\in[K]$ is an optimal solution for Problem (12).*

The condition in Theorem 3 is more restrictive than that for supermodularity given in Proposition 4. Given any $\boldsymbol{x}\in\mathcal{X}$, $\boldsymbol{p},\boldsymbol{q}\in\Re^m$, $\boldsymbol{z}',\boldsymbol{z}''\in\Re^n$ with $(\boldsymbol{p},\boldsymbol{z}'\wedge\boldsymbol{z}''),(\boldsymbol{q},\boldsymbol{z}'\vee\boldsymbol{z}'')\in\mathcal{S}(\boldsymbol{x})$, it requires $(\boldsymbol{p}+\boldsymbol{q}-\boldsymbol{y}',\boldsymbol{z}'')\in\mathcal{S}(\boldsymbol{x})$ for all $\boldsymbol{y}'$ such that $(\boldsymbol{y}',\boldsymbol{z}')\in\mathcal{S}(\boldsymbol{x})$, while in Proposition 4 the requirement is only for one such $\boldsymbol{y}'$. Hence, compared with the condition in Proposition 4, the condition in Theorem 3 is indeed a sufficient condition for preserving supermodularity. We would also like to remark that the obtained segregated affine decision rule, from the formulation of either (11) or (12), is not implementable since the realization of the uncertain scenario $\tilde{k}$ might not always be observable. Nevertheless, in most two-stage problems, it suffices to obtain an appealing first-stage decision, to which the segregated affine decision rule approach can serve. After observing the uncertainty realization $\boldsymbol{z}$, the second-stage decision should be determined by solving the second-stage problem, rather than simply by the affine function. Please see Bertsimas et al. (2019) for a related discussion.

For robust optimization which does not use any distributional information except the support, Bertsimas and Goyal (2012) have shown the optimality of affine decision rules when the support is a simplex. However, the optimality of affine decision rules is not true in general if we extend to DRO problems even when the support is a simplex. In Proposition 6 and Theorem 3, we show that by lifting the uncertainty realization $\boldsymbol{z} \in \Re^n$ to $\begin{bmatrix} (\boldsymbol{\mu}^k - \boldsymbol{z})^+ \\ (\boldsymbol{z} - \boldsymbol{\mu}^k)^+ \end{bmatrix} \in \Re^{2n}$, we can construct an affine decision rule which turns out to be an optimal solution. This optimality relies on the chained structure of the support of the worst-case distribution, which is due to the supermodularity of $g(\boldsymbol{x}, \boldsymbol{z})$.

Extending the result of Bertsimas and Goyal (2012), Iancu et al. (2013) show the optimality of the affine decision rule for the unconstrained multi-stage problem when the objective function is convex and supermodular in uncertain parameters and the uncertainty set is a union of simplices that forms a sublattice of the unit hypercube. Our result differs in the sense that we focus on a constrained DRO problem; moreover, the union of all supports from each given scenario realization is not necessarily a lattice within our setting.

5. Applications

In this section, we apply the above theoretical results to several classic operational problems, which are difficult to solve in general. Section 5.1 considers a single-period multi-item newsvendor problem, where the objective is to optimize the retailer's expected disutility or CVaR. In Sections 5.2 and 5.3, we revisit the facility location problem and the lot-sizing problem, respectively. By proving the property of supermodularity, we provide new perspectives and simpler reformulations. Section 5.4 presents an appointment scheduling problem with random no-shows. Finally, a general formulation of ATO systems is discussed in Section 5.5, where we identify a class of systems which are tractable under our assumption. In the following applications, some common notations may have different meanings in different applications.

5.1. Multi-item newsvendor problems

Multi-item newsvendor problems seek the optimal inventory levels of multiple goods with fixed prices and uncertain demands (Hadley and Whitin 1963). Since these items are correlated with each other either through some budget constraint or by a particular utility function, the problem may become much harder to solve. In the distributionally robust setting, Hanasusanto et al. (2015) assume a risk-averse decision maker who minimizes a linear combination of CVaR and expectation of the profit function, and the demand distribution to be multi-modal. They show that the resulting problem is NP-hard and solve it approximately with a semidefinite program by applying the quadratic decision rule. Natarajan et al. (2017) use semi-variance to capture the asymmetry of demand distributions, and also develop a semidefinite program to derive a lower bound for the original problem. We next use our reformulation technique to show that the multi-item newsvendor problem can be solved efficiently within our setting.

Consider a single-period multi-item newsvendor problem with n different items. The selling price, ordering cost and salvage value of item i are denoted by r_i, t_i and s_i ($s_i < r_i$), respectively. Before the random demand $\tilde{\boldsymbol{z}}$ is resolved, we need to decide the ordering quantity $\boldsymbol{x}$, which is subject to a budget Γ. Our goal is to minimize the worst-case expected disutility of cost. This yields the following optimization problem

$$\min_{\boldsymbol{x}\in\mathcal{X}^{news}} \sup_{\mathbb{P}\in\mathcal{F}} \mathbb{E}_{\mathbb{P}}\left[u\left(-\boldsymbol{r}^\top\boldsymbol{x}+(\boldsymbol{r}-\boldsymbol{s})^\top(\boldsymbol{x}-\tilde{\boldsymbol{z}})^+\right)\right], \tag{13}$$

where $\mathcal{X}^{news}=\left\{\boldsymbol{x}\in\Re^n \mid \boldsymbol{t}^\top\boldsymbol{x}\le\Gamma,\boldsymbol{x}\ge\boldsymbol{0}\right\}$ and $u:\Re\to\Re$ is a piecewise linear convex and non-decreasing disutility function defined as $u(w)=\max_{j\in[J]}\ \{c_jw+d_j\}, w\in\Re$, for some constants $c_j\ge0$ and $d_j,\ j\in[J]$. We then have the following result.

Proposition 7 *The function $u\left(-\boldsymbol{r}^\top\boldsymbol{x}+(\boldsymbol{r}-\boldsymbol{s})^\top(\boldsymbol{x}-\boldsymbol{z})^+\right)$ is supermodular in $\boldsymbol{z}$, hence Problem (13) can be reformulated as a linear optimization problem.*

Alternatively, when minimizing the CVaR as Hanasusanto et al. (2015) do, for any $\rho\in(0,1)$, the problem of CVaR minimization is

$$\min_{\boldsymbol{x}\in\mathcal{X}^{news}}\ \sup_{\mathbb{P}\in\mathcal{F}}\ \inf_{\theta\in\Re}\ \left\{\theta+\mathbb{E}_{\mathbb{P}}\left[\frac{1}{\rho}\cdot\left(-\boldsymbol{r}^\top\boldsymbol{x}+(\boldsymbol{r}-\boldsymbol{s})^\top(\boldsymbol{x}-\boldsymbol{z})^+-\theta\right)^+\right]\right\}. \tag{14}$$

This problem can also be reformulated as a linear optimization problem.

Proposition 8 *Problem (14) has a polynomial size linear programming reformulation.*

With the objective of minimizing CVaR and the multi-modal demand assumption, our work differs from Hanasusanto et al. (2015) in the scenario-based distributional information. While their work considers the first two moments and derive an approximate solution by solving a semidefinite programming problem, we focus on partial marginal information and obtain an exact linear programming reformulation of the original problem. Further, with the stockout costs considered in Hanasusanto et al. (2015), we can show that the total cost function is still supermodular in the demand $\boldsymbol{z}$. Hence, the problem can be easily solved if only considering the expected cost. However, when a general convex disutility is incorporated, due to the presence of the stockout cost, the total cost no longer decreases with $\boldsymbol{z}$ and hence the supermodularity can not hold (the details can be referred to Appendix B.2). We illustrate this with the following example. Consider a 2-item problem with the selling price $\boldsymbol{r}=(6,3)$, salvage value $\boldsymbol{s}=(2,2)$, stockout cost $\boldsymbol{b}=(2,2)$ and disutility function $u(w)=(w+5)^+$. Then the total cost is

$$-\boldsymbol{r}^\top\min\{\boldsymbol{x},\boldsymbol{z}\}-\boldsymbol{s}^\top(\boldsymbol{x}-\boldsymbol{z})^++\boldsymbol{b}^\top(\boldsymbol{z}-\boldsymbol{x})^+=-(8,5)^\top\boldsymbol{x}+(2,2)^\top\boldsymbol{z}+(6,3)^\top(\boldsymbol{x}-\boldsymbol{z})^+.$$

Let $g(\boldsymbol{x},\boldsymbol{z})=(2,2)^\top\boldsymbol{z}+(6,3)^\top(\boldsymbol{x}-\boldsymbol{z})^+$ and consider $\boldsymbol{x}=(1,1),\ \boldsymbol{z}'=(2,0),\boldsymbol{z}''=(0,2)$. By simple calculation, we have $u(-(\boldsymbol{r}+\boldsymbol{b})^\top\boldsymbol{x}+g(\boldsymbol{x},\boldsymbol{z}'\wedge\boldsymbol{z}''))+u(-(\boldsymbol{r}+\boldsymbol{b})^\top\boldsymbol{x}+g(\boldsymbol{x},\boldsymbol{z}'\vee\boldsymbol{z}''))=1+0<0+2=u(-(\boldsymbol{r}+\boldsymbol{b})^\top\boldsymbol{x}+g(\boldsymbol{x},\boldsymbol{z}'))+u(-(\boldsymbol{r}+\boldsymbol{b})^\top\boldsymbol{x}+g(\boldsymbol{x},\boldsymbol{z}''))$, which violates supermodularity.

5.2. Reliable facility location

Consider the problem of locating facilities at a set of candidate locations $i \in [n]$, to serve a set of customers $j \in [m]$. In the first stage, the facility location decision $\boldsymbol{x} = (x_1, \ldots, x_n)$ is made, where $x_i = 1$ if facility is opened at location i, and $x_i = 0$ otherwise. Let a_i be the fixed cost of opening a facility at location $i \in [n]$. In the second stage, customers are allocated to the facilities. Denote the transportation cost of location i serving customer j by c_{ij}. The facilities are subject to random disruptions, captured by $\tilde{\boldsymbol{z}} = (\tilde{z}_1, \ldots, \tilde{z}_n)$, which are realized after the first-stage decision $\boldsymbol{x}$ is made. We denote by $z_i = 0$ if location i is disrupted, and by $z_i = 1$ otherwise. The disruption happens at location i with probability M_i. The cost minimization problem is formulated as

$$\min_{\boldsymbol{x} \in \mathcal{X}^{fac}} \left\{ \boldsymbol{a}^\top \boldsymbol{x} + \sup_{\mathbb{P} \in \mathcal{F}^{fac}} \mathbb{E}_{\mathbb{P}} \left[g(\boldsymbol{x}, \tilde{\boldsymbol{z}}) \right] \right\},$$

where $\mathcal{X}^{fac} = \{0,1\}^n$, $\mathcal{F}^{fac} = \{\mathbb{P} \mid \mathbb{P}(\tilde{z}_i = 0) = M_i, \mathbb{P}(\tilde{z}_i = 1) = 1 - M_i, i \in [n]\}$, and

$$\begin{aligned} g(\boldsymbol{x}, \boldsymbol{z}) = \ \min \quad & \sum_{i \in [n], j \in [m]} c_{ij} y_{ij} \\ \text{s.t.} \quad & \sum_{i \in [n]} y_{ij} = 1, \qquad j \in [m], \\ & 0 \le y_{ij} \le x_i z_i, \quad i \in [n], j \in [m]. \end{aligned} \tag{15}$$

We remark that the above second-stage problem always has an optimal solution $y_{ij} \in \{0,1\}$ for all $i \in [n], j \in [m]$, and hence we do not include the binary constraints on y_{ij} in the problem explicitly. Lu et al. (2015) prove the supermodularity of $g(\boldsymbol{x}, \boldsymbol{z})$ by verifying the definition. We show that the same result can be obtained by a direct application of Theorem 2.

Proposition 9 *The function $g(\boldsymbol{x}, \boldsymbol{z})$ defined by Equation (15) is supermodular in $\boldsymbol{z}$ for all $\boldsymbol{x} \in \mathcal{X}^{fac}$.*

5.3. Lot-sizing on a network

Lot sizing is one of the most important and difficult problems in production planning. We adopt the model setting from Bertsimas and de Ruiter (2016) and investigate the lot-sizing problem on a network. Consider n stores in total, each corresponding to a random demand $\tilde{z}_i$, $i \in [n]$. In the first stage, we determine an allocation x_i for the i-th store. The feasible set $\mathcal{X}^{lot}$ describes the capacity of the stores, i.e. $0 \le x_i \le K_i$ for some $(K_1, \ldots, K_n) \in \Re^n_+$. The unit storage cost at store i is denoted as a_i. In the second stage, after the demands are observed, we transport stock y_{ij} from store i to store j at unit cost b_{ij} such that all the demands are met. The goal is to minimize the worst-case expected total cost. We express the model as a two-stage linear optimization problem,

$$\min_{\boldsymbol{x} \in \mathcal{X}^{lot}} \left\{ \boldsymbol{a}^\top \boldsymbol{x} + \sup_{\mathbb{P} \in \mathcal{F}} \mathbb{E}_{\mathbb{P}} \left[\min \left\{ \sum_{s,j \in [n]} b_{sj} y_{sj} \;\middle|\; \begin{array}{ll} \sum_{j \in [n]} y_{js} - \sum_{j \in [n]} y_{sj} \ge z_s - x_s, & s \in [n] \\ y_{sj} \ge 0, & s, j \in [n] \end{array} \right\} \right] \right\}. \tag{16}$$

While Bertsimas and de Ruiter (2016) derive an approximation for the robust version of Problem (16) when the uncertainty set is a polyhedron, we next show that the problem can be solved exactly in polynomial time within our setting. Let $g(\boldsymbol{x},\boldsymbol{z})$ be the second-stage cost for a given allocation $\boldsymbol{x}$ and realized demand $\boldsymbol{z}$.

Proposition 10 *The function $g(\boldsymbol{x},\boldsymbol{z})$ defined by the inner minimization problem in (16) is supermodular in $\boldsymbol{z}$ for all $\boldsymbol{x}$. Hence, Problem (16) can be reformulated as a linear program.*

When the transported amount y_{sj} is bounded by a capacity c_{sj}, as in Bertsimas and Shtern (2018), the second-stage cost becomes

$$\begin{aligned}\hat{g}(\boldsymbol{x},\boldsymbol{z}) = \ \min\ & \sum_{s,j\in[n]} b_{sj}y_{sj} \\ \text{s.t.}\ & \sum_{j\in[n]} y_{js} - \sum_{j\in[n]} y_{sj} \geq z_s - x_s, \quad s\in[n] \\ & 0 \leq y_{sj} \leq c_{sj}, \quad s\in[n], j\in[n].\end{aligned} \tag{17}$$

By a similar analysis, we can verify that $\hat{g}(\boldsymbol{x},\boldsymbol{z})$ defined by Equation (17) is also supermodular. Hence, our method can be applied to obtain an exact solution.

5.4. Appointment scheduling with random no-shows

The appointment scheduling problem, which schedules the arrival times of customers, has wide applications in service delivery systems (Gupta and Denton 2008). In this section, we focus on robust appointment scheduling problems where no-shows are possible. Jiang et al. (2017) assume the means of no-shows and means, supports of the uncertain service times, and propose an integer programming-based decomposition algorithm to minimize the worst-case expected sum of waiting time and overtime. Further, Jiang et al. (2019) provide a copositive reformulation when the ambiguity set is a Wasserstein ball. When no-shows are not considered, Kong et al. (2013) and Mak et al. (2014) conduct thorough studies with the same objective function. Specifically, Kong et al. (2013) propose a tractable semidefinite approximation when the mean and covariance information are known. Mak et al. (2014) provide an exact conic programming reformulation when marginal moments are given. Although Qi (2017) also uses the mean and MAD information, and provides a linear formulation, this linearity arises from the use of a different objective function. Given the scenario-based ambiguity set with MAD information, we next show that the problem can be reduced to a polynomial sized linear program, which is simpler than the formulations derived in previous studies.

We schedule n appointments within a given time period Γ. For all $i\in[n]$, we assume customer i shows up with probability θ_i and use $\tilde{\xi}_i\in\{0,1\}$ to characterize this event, i.e., $\mathbb{P}(\tilde{\xi}_i=1)=\theta_i, \mathbb{P}(\tilde{\xi}_i=0)=1-\theta_i$. Let $\tilde{z}_i$ be the actual duration of the i-th service. We decide the scheduled duration x_i for each appointment i to minimize the worst-case expected sum of waiting time and overtime. For the i-th waiting time $\tilde{w}_i$ and the system overtime $\tilde{w}_{n+1}$, we have $\tilde{w}_1 = 0$ and $\tilde{w}_{i+1} = \max\left\{\tilde{w}_i + \tilde{\xi}_i\tilde{z}_i - x_i, 0\right\}$, for all $i \in [n-1]$. We follow Jiang et al. (2017) and formulate the problem using a two-stage optimization structure as

$$\min_{\boldsymbol{x}\in\mathcal{X}^{app}} \sup_{\mathbb{P}\in\mathcal{G}} \mathbb{E}_{\mathbb{P}}\left[g(\boldsymbol{x}, \tilde{\boldsymbol{\xi}}, \tilde{\boldsymbol{z}})\right], \tag{18}$$

Here the feasible set is defined as $\mathcal{X}^{app} = \left\{\boldsymbol{x} \in \Re^n_+ \mid \mathbf{1}^\top \boldsymbol{x} \leq \Gamma\right\}$, and the second-stage problem can be written as

$$g(\boldsymbol{x}, \boldsymbol{\xi}, \boldsymbol{z}) = \min\left\{\mathbf{1}^\top \boldsymbol{y} \;\middle|\; \begin{array}{ll} y_t \geq \sum_{s=j}^{t}(\xi_s z_s - x_s), & j \in [t], t \in [n] \\ y_t \geq 0, & t \in [n] \end{array}\right\},$$

where the optimal y_t is indeed the realization of $\tilde{w}_{t+1}$. The ambiguity set for $(\tilde{\boldsymbol{\xi}}, \tilde{\boldsymbol{z}})$ is specified as $\mathcal{G} = \{\mathbb{P} \mid \boldsymbol{\Pi}_{\boldsymbol{\xi}}\mathbb{P} \in \mathcal{F}_{\boldsymbol{\xi}}, \boldsymbol{\Pi}_{\boldsymbol{z}}\mathbb{P} \in \mathcal{F}\}$, where $\boldsymbol{\Pi}_{\boldsymbol{\xi}}\mathbb{P}$, $\boldsymbol{\Pi}_{\boldsymbol{z}}\mathbb{P}$ denotes the marginal distribution of $\tilde{\boldsymbol{\xi}}$ and $\tilde{\boldsymbol{z}}$, respectively under $\mathbb{P}$. The distributional uncertainty set $\mathcal{F}_{\boldsymbol{\xi}}$ is defined as

$$\mathcal{F}_{\boldsymbol{\xi}} = \left\{\mathbb{P}_{\boldsymbol{\xi}} \;\middle|\; \begin{array}{l} \mathbb{E}_{\mathbb{P}_{\boldsymbol{\xi}}}\left[\tilde{\boldsymbol{\xi}} \;\middle|\; \tilde{k} = k\right] = \boldsymbol{\theta}^k, \mathbb{P}_{\boldsymbol{\xi}}\left(\tilde{k} = k\right) = q_k, \boldsymbol{q} \in \mathcal{Q} \\ \mathbb{P}_{\boldsymbol{\xi}}\left(\tilde{\boldsymbol{\xi}} \in \boldsymbol{\Xi}^k \;\middle|\; \tilde{k} = k\right) = 1 \end{array}\right\} \tag{19}$$

with $\boldsymbol{\Xi}^k = \{0,1\}^n$ and $\mathcal{F}$ is defined by Equation (3). Though the information set $\mathcal{G}$ differs slightly from that in Equation (3), the key idea and process of our approach are still applicable. Using the condition in Theorem 2, we next demonstrate the supermodularity of the function $g(\boldsymbol{x}, \boldsymbol{\xi}, \boldsymbol{z})$.

Proposition 11 *Function $g(\boldsymbol{x}, \boldsymbol{\xi}, \boldsymbol{z})$ is supermodular in $(\boldsymbol{\xi}, \boldsymbol{z})$ for all $\boldsymbol{x}$. Hence, Problem (18) has a polynomial size linear programming reformulation.*

Therefore, given the information set $\mathcal{G}$, we can reformulate Problem (18) in a computationally tractable manner. Our linear programming reformulation provides an exact solution and the computational complexity is reduced significantly compared to the literature.

To rule out unlikely scenarios such as consecutive no-shows, we can modify our scenario-based support set $\boldsymbol{\Xi}^k$ in Equation (19). For example, if we want to exclude the scenarios in which all patients are absent (i.e. $\boldsymbol{\xi} = \mathbf{0}$), we can let $K = n$ and consider $\boldsymbol{\Xi}^k = \{\xi \in \{0,1\}^n \mid \xi_k = 1\} = \{0,1\} \times \cdots \times \{1\} \times \cdots \times \{0,1\}$ for all $k \in [n]$. The problem remains tractable, since the support of $\boldsymbol{\xi}$ is still a Cartesian product with $\{0,1\}$ modified to $\{1\}$ at the k-th dimension; in other words, the uncertainty at that dimension is reduced to be deterministic.

Chen et al. (2022) prove that when no-shows are not considered, the objective function is supermodular in the uncertain appointment durations $\boldsymbol{z}$. Our setting is more general since we consider no-shows and scenario-based uncertainty set. Further, our proof is based on a systematic tool, which verifies the general conditions in Theorem 2.

5.5. Assemble-to-order systems

The ATO system is an important operational problem (Song and Zipkin 2003). Although this problem has attracted substantial attention, it is still not clear how to derive the optimal decision in general. We now apply our theoretical result and identify a class of systems where the optimal decision can be obtained efficiently.

Here we formally describe the problem using the formulation of Song and Zipkin (2003). For any component i, $i \in [l]$, first we decide the order-up-to inventory level x_i. Then the uncertain demand $\tilde{z}_j$ for end product j is realized as z_j, $j \in [n]$. After that, we make the second-stage decision y_j, which is the quantity of product j to be assembled. To minimize the worst-case expected cost, we have the following formulation,

$$\begin{aligned} \min \quad & \boldsymbol{c}^\top(\boldsymbol{x} - \boldsymbol{x}_{int}) + \sup_{\mathbb{P} \in \mathcal{F}} \mathbb{E}_{\mathbb{P}}\big[g(\boldsymbol{x}, \tilde{\boldsymbol{z}})\big] \\ \text{s.t.} \quad & \boldsymbol{x} \geq \boldsymbol{x}_{int}, \end{aligned}$$

where

$$\begin{aligned} g(\boldsymbol{x}, \boldsymbol{z}) = \min \quad & \boldsymbol{h}^\top(\boldsymbol{x} - \boldsymbol{A}\boldsymbol{y}) + \boldsymbol{p}^\top(\boldsymbol{z} - \boldsymbol{y}) - \boldsymbol{r}^\top \boldsymbol{y} \\ \text{s.t.} \quad & \boldsymbol{A}\boldsymbol{y} \leq \boldsymbol{x}, \ \boldsymbol{y} \leq \boldsymbol{z}, \ \boldsymbol{y} \geq \boldsymbol{0}. \end{aligned}$$

Here $\boldsymbol{c}$ and $\boldsymbol{x}_{int}$ are the per-unit ordering cost and initial inventory level of the components, respectively; $\boldsymbol{h}$ is the per-unit inventory holding cost of the leftover components; $\boldsymbol{p}$ and $\boldsymbol{r}$ are the per-unit penalty cost of the shortage and per-unit selling price of the end products, respectively. The elements in matrix $\boldsymbol{A}$, i.e., $a_{ij} \geq 0$, represent the number of units of component i required to assemble one unit of end product j. Different ATO systems are characterized by different matrices $\boldsymbol{A} \in \Re_+^{l \times n}$. We next provide a condition on $\boldsymbol{A}$ such that the function $g(\boldsymbol{x}, \boldsymbol{z})$ is supermodular in $\boldsymbol{z}$, hence the optimal order-up-to level for each component can be derived based on Theorem 1.

Theorem 4 *The function $g(\boldsymbol{x}, \boldsymbol{z})$ is supermodular in $\boldsymbol{z}$ for any $\boldsymbol{x}, \boldsymbol{h}, \boldsymbol{p}, \boldsymbol{r}$ if and only if every 2×3 submatrix of the matrix $\boldsymbol{A}$ contains at least one pair of column vectors which are linearly dependent.*

We next test the condition of Theorem 4 on practical ATO systems. Consider the *Tree Family* of systems proposed by Zipkin (2016). For any $i \in [l]$, denote $S_i = \{j \in [n] \mid a_{ij} \neq 0\}$ being the index set of products which require component i. A system belongs to the Tree Family if for any two components i, i' with $S_i \cap S_{i'} \neq \emptyset$, either $S_i \subseteq S_{i'}$ or $S_{i'} \subseteq S_i$ holds. That is, if a product uses two distinct components i, i', then the set of products using component i (or i') must contain that of component i' (or i). Observing that general Tree Family systems do not guarantee the supermodularity of $g(\boldsymbol{x}, \boldsymbol{z})$, we define the *Proportional Tree Family* as follows.

Definition 3 *An ATO system belongs to the Proportional Tree Family, if it belongs to the Tree Family and for any two components i, i' with the set of common products $S_i \cap S_{i'} \neq \emptyset$, $a_{ij}/a_{i'j}$ takes the same value for all $j \in S_i \cap S_{i'}$.*

With Theorem 4, Proportional Tree Family has the property of supermodularity.

Corollary 2 *The function $g(\boldsymbol{x},\boldsymbol{z})$ is supermodular in $\boldsymbol{z}$ for any $\boldsymbol{x},\boldsymbol{h},\boldsymbol{p},\boldsymbol{r}$ if the system belongs to Proportional Tree Family.*

We next discuss several typical ATO systems and check whether supermodularity holds or not.

- $\boldsymbol{A}\in\Re_+^{l\times 1}$ or $\boldsymbol{A}\in\Re_+^{l\times 2}$, i.e. there are at most two products in the system. This does not necessarily belong to the Proportional Tree Family but satisfies the condition in Theorem 4.
- Binary Tree Family (Zipkin 2016): a system belonging to Tree Family and with all elements in $\boldsymbol{A}$ being binary. We can show it is in the Proportional Tree Family.
- The generalized W System (Zipkin 2016, Chen et al. 2021): $\boldsymbol{A}=\begin{bmatrix}\boldsymbol{D}\\ \boldsymbol{c}^\top\end{bmatrix}\in\Re_+^{(n+1)\times n}$ with $\boldsymbol{D}$ being a diagonal matrix. This system has $(n+1)$ components and n products. The last component is a common component and used in all products; for all other components, each is specific to a single product. Obviously, this belongs to Proportional Tree Family.
- The generalized M System (Lu and Song 2005, Nadar et al. 2014): $\boldsymbol{A}=[\boldsymbol{D}\ \boldsymbol{c}]\in\Re_+^{n\times(n+1)}$ with $\boldsymbol{D}$ being a diagonal matrix. This system has n components and $(n+1)$ products. The last product uses all components; for all other products, each is specific to a single component. The generalized M system violates the condition of Theorem 4, and hence $g(\boldsymbol{x},\boldsymbol{z})$ is not supermodular in $\boldsymbol{z}$.

Zipkin (2016) considers the Binary Tree Family systems with known demand distribution and shows that an optimal inventory level can be solved approximately, while the computational complexity is indeed not guaranteed. Recognizing that the Binary Tree Family is a special case of the Proportional Tree Family, our method suggests that an exact linear programming reformulation of polynomial size can be obtained in the distributionally robust setting.

6. Numerical studies

In this section, we investigate and compare the computational performances of our method against other solution methods for two-stage DRO models. Specifically, we consider the column-and-constraint generation (CCG) algorithm (see for instance, Zeng and Zhao 2013, Saif and Delage 2021) and the segregated linear decision rules. We use the three methods to solve the ATO problems. The program is coded in Python and run on an 12-Core Intel PC with a 2.7 GHz CPU.

6.1. Settings

To provide a brief and clear computational results, we consider the distributional uncertainty set $\mathcal{F}$ given as Equation (4) with only one fixed scenario. We relegate the formulations of the CCG algorithm and the segregated linear decision rule in Appendix A.3. We use the ATO system with n products and n components as an example. The assemble matrix $\boldsymbol{A}\in\Re_+^{n\times n}$ is specified as

$\boldsymbol{A} = \begin{bmatrix} \boldsymbol{I}_{(n-1)\times(n-1)} & \sigma\boldsymbol{1} \\ \boldsymbol{\rho}^\top & \sigma \end{bmatrix}$, where $\boldsymbol{\rho} = (1, 2, \ldots, n-1)$. In this case, when $\sigma = 0$, the system reduces to the generalized W system and the function $g(\boldsymbol{x}, \boldsymbol{z})$ in the ATO system is supermodular (see Corollary 2). For all other parameters, we randomly generate the values from uniform distributions (denoted by $U(\underline{u}, \overline{u})$, where $\underline{u}, \overline{u}$ are respectively the lower and upper bounds). For each product $i, i \in [n]$, we generate the demand information $\mu_i \sim U(45, 55), \underline{z}_i \sim U(40, \mu_i - 2), \overline{z}_i \sim U(\mu_i + 2, 60), \delta_i \in U\left(0, \frac{2(\mu_i - \underline{z}_i)(\overline{z}_i - \mu_i)}{\overline{z}_i - \underline{z}_i}\right)$, with penalty cost $p_i \sim U(10, 20)$ and retail price $r_i \sim U(10, 20)$. For each component $j, j \in [n]$, we generate ordering cost $c_j \sim U(0, 10)$ and holding cost $h_j \sim U(0, 10)$.

We conduct experiments for $n = 5$ and $n = 20$. In both cases, we randomly generate 200 instances using the procedures mentioned above. For each instance, we vary the parameter σ in the assemble matrix $\boldsymbol{A}$. Specifically, for instances with $n = 5$, we vary σ from 0 to 15 with the step size of 0.5; for instances with $n = 20$, we vary σ from 0 to 60 with the step size of 2.

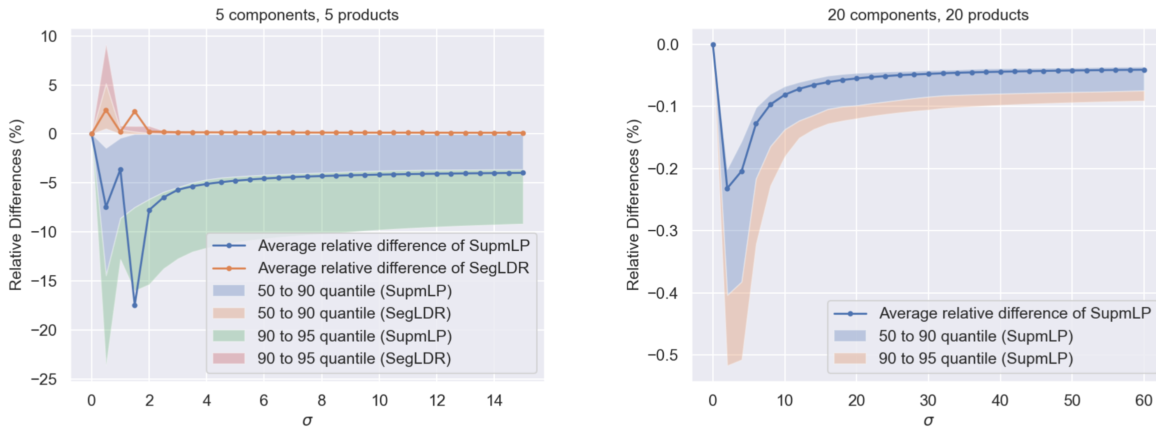
6.2. Quality of the solutions

The CCG algorithm derives the optimal value for Problem (2), while the segregated linear decision rule and our method provide the upper and lower bounds, respectively. We normalize the optimal value derived by CCG algorithm as 1 and calculate the relative difference with the other two methods as $\frac{OPT_{SupmLP} - OPT_{CCG}}{OPT_{CCG}}$ and $\frac{OPT_{SegLDR} - OPT_{CCG}}{OPT_{CCG}}$. Here we abbreviate our method as "SupmLP" and the segregated linear decision rule as "SegLDR". We show the performance of these three methods with 5×5 ATO systems in Figure 3(a). For each value of σ, we consider all the 200 instances and plot the average and quantile about the relative difference. When $\sigma = 0$, these three methods provide the same optimal value, since the property of supermodularity holds. When $\sigma \neq 0$, the property of supermodularity cannot be guaranteed. We observe that the relative differences exist but are not large. Although the segregated linear decision rule performs closer to the CCG algorithm than our method, the two approaches play different roles (upper and lower bounds, respectively) in quantifying the optimal value of the problem.

Figure 3(b) shows the performance for the 20×20 ATO systems. In this case, unfortunately, the segregated linear decision rule formulation includes more than 3^{20} constraints and cannot be solved efficiently. Therefore, we only compare the performance of the CCG algorithm and our method. We observe that our method performs quite well in general even though the property of supermodularity is not guaranteed when $\sigma \neq 0$.

6.3. Computational time

We also compare the computational time for the three methods based on the ATO systems mentioned above. Within two hours, the segregated linear decision rule and the CCG algorithm can solve the ATO problems when $n = 11$ and $n = 120$ respectively. For our method, the average computational time to solve the ATO system when $n = 120$ is around 4 seconds.



(a) 5×5 ATO system

(b) 20×20 ATO system

Figure 3 Performance comparison in ATO systems.

7. Conclusion

This paper identifies a tractable class of two-stage DRO problems and derives exact optimal solutions when the scenario-based ambiguity set is considered. Given any realization of the uncertain scenario, we know the information of supports, means and MADs for the underlying uncertainties. Our results show that any two-stage problem has a computationally tractable reformulation whenever its second-stage cost function is supermodular in the uncertainty realization. This reformulation relies on the common worst-case distribution, which can be pre-calculated via an efficient algorithm. As a result, our reformulation preserves the original structure of the problem and retains the same computational complexity as the nominal problem. While the reformulation is based on the requirement for supermodularity in the second-stage problem, we provide a necessary and sufficient condition to check whether this requirement is met for any given two-stage problem. We also discuss the optimality of scenario-wise segregated affine decision rules in our setting.

Subsequently, it can be verified that a wide range of practical problems fit within our framework of two-stage DRO with supermodularity. Instances include multi-item newsvendor, reliable facility location, lot-sizing, appointment scheduling with random no-shows, and general ATO systems. While these problems are considered computationally challenging in general, under our assumption, they can be solved exactly and efficiently.

There are several promising directions to explore. Theoretically, it is interesting to investigate whether our efficient solution method can be applied to a broader class of two-stage problems, such as those with integer recourse decisions where the second-stage cost function might not even be continuous. Moreover, whether our framework can be extended to multi-periods optimization problems would also be an important research question. Last but not least, since there are numerous two-stage problems with significant practical impact, we can verify if supermodularity property exists in those problems with our results and further validate the applicability of our approach.

References

Agrawal, Shipra, Yichuan Ding, Amin Saberi, Yinyu Ye. 2010. Correlation robust stochastic optimization. *Proceedings of the twenty-first annual ACM-SIAM symposium on Discrete Algorithms*. SIAM, 1087–1096.

Ardestani-Jaafari, Amir, Erick Delage. 2016. Robust optimization of sums of piecewise linear functions with application to inventory problems. *Operations Research* **64**(2) 474–494.

Bansal, Manish, Kuo-Ling Huang, Sanjay Mehrotra. 2018. Decomposition algorithms for two-stage distributionally robust mixed binary programs. *SIAM Journal on Optimization* **28**(3) 2360–2383.

Ben-Tal, Aharon, Laurent El Ghaoui, Arkadi Nemirovski. 2009. *Robust Optimization*, vol. 28. Princeton University Press.

Ben-Tal, Aharon, Alexander Goryashko, Elana Guslitzer, Arkadi Nemirovski. 2004. Adjustable robust solutions of uncertain linear programs. *Mathematical Programming* **99**(2) 351–376.

Ben-Tal, Aharon, Eithan Hochman. 1972. More bounds on the expectation of a convex function of a random variable. *Journal of Applied Probability* **9**(4) 803–812.

Bertsimas, Dimitris, Frans JCT de Ruiter. 2016. Duality in two-stage adaptive linear optimization: Faster computation and stronger bounds. *INFORMS Journal on Computing* **28**(3) 500–511.

Bertsimas, Dimitris, Xuan Vinh Doan, Karthik Natarajan, Chung-Piaw Teo. 2010a. Models for minimax stochastic linear optimization problems with risk aversion. *Mathematics of Operations Research* **35**(3) 580–602.

Bertsimas, Dimitris, Vineet Goyal. 2010. On the power of robust solutions in two-stage stochastic and adaptive optimization problems. *Mathematics of Operations Research* **35**(2) 284–305.

Bertsimas, Dimitris, Vineet Goyal. 2012. On the power and limitations of affine policies in two-stage adaptive optimization. *Mathematical Programming* **134**(2) 491–531.

Bertsimas, Dimitris, Dan A Iancu, Pablo A Parrilo. 2010b. Optimality of affine policies in multistage robust optimization. *Mathematics of Operations Research* **35**(2) 363–394.

Bertsimas, Dimitris, Dan A Iancu, Pablo A Parrilo. 2011. A hierarchy of near-optimal policies for multistage adaptive optimization. *IEEE Transactions Auto. Control* **56**(12) 2809–2824.

Bertsimas, Dimitris, Shimrit Shtern. 2018. A scalable algorithm for two-stage adaptive linear optimization. *arXiv preprint arXiv:1807.02812* .

Bertsimas, Dimitris, Melvyn Sim, Meilin Zhang. 2019. Adaptive distributionally robust optimization. *Management Science* **65**(2) 604–618.

Birge, John R, Francois Louveaux. 2011. *Introduction to Stochastic Programming*. Springer Science & Business Media.

Chen, Louis, Will Ma, Karthik Natarajan, David Simchi-Levi, Zhenzhen Yan. 2022. Distributionally robust linear and discrete optimization with marginals. *Operations Research* .

Chen, Xin. 2017. $L^\natural$-convexity and its applications in operations. *Frontiers of Engineering Management* **4**(3) 283–294.

Chen, Xin, Peng Hu, Simai He. 2013. Preservation of supermodularity in parametric optimization problems with nonlattice structures. *Operations Research* **61**(5) 1166–1173.

Chen, Xin, Daniel Zhuoyu Long, Jin Qi. 2021. Preservation of supermodularity in parametric optimization: Necessary and sufficient conditions on constraint structures. *Operations Research* **69**(1) 1–12.

Chen, Xin, Melvyn Sim, Peng Sun, Jiawei Zhang. 2008. A linear decision-based approximation approach to stochastic programming. *Operations Research* **56**(2) 344–357.

Chen, Zhi, Melvyn Sim, Peng Xiong. 2020. Robust stochastic optimization made easy with RSOME. *Management Science* **66**(8) 3329–3339.

Conejo, Antonio J, Nicholas G Hall, Daniel Zhuoyu Long, Runhao Zhang. 2021. Robust capacity planning for project management. *INFORMS Journal on Computing* **33**(4) 1533–1550.

Delage, Erick, Yinyu Ye. 2010. Distributionally robust optimization under moment uncertainty with application to data-driven problems. *Operations Research* **58**(3) 595–612.

El Housni, Omar, Vineet Goyal. 2021. On the optimality of affine policies for budgeted uncertainty sets. *Mathematics of Operations Research* **46**(2) 674–711.

Feige, Uriel, Kamal Jain, Mohammad Mahdian, Vahab Mirrokni. 2007. Robust combinatorial optimization with exponential scenarios. *International Conference on Integer Programming and Combinatorial Optimization*. Springer, 439–453.

Georghiou, Angelos, Angelos Tsoukalas, Wolfram Wiesemann. 2021. On the optimality of affine decision rules in robust and distributionally robust optimization. *Available at Optimization Online* .

Ghosal, Shubhechyya, Chin Pang Ho, Wolfram Wiesemann. 2021. A unifying framework for the capacitated vehicle routing problem under risk and ambiguity. *Available at Optimization Online* .

Goh, Joel, Melvyn Sim. 2010. Distributionally robust optimization and its tractable approximations. *Operations Research* **58**(4-1) 902–917.

Gupta, Anupam, Viswanath Nagarajan, R Ravi. 2014. Thresholded covering algorithms for robust and max–min optimization. *Mathematical Programming* **146**(1) 583–615.

Gupta, Diwakar, Brian Denton. 2008. Appointment scheduling in health care: Challenges and opportunities. *IIE Transactions* **40**(9) 800–819.

Hadley, George, Thomson M Whitin. 1963. Analysis of inventory systems. Tech. rep.

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

Hanasusanto, Grani A, Daniel Kuhn, Stein W Wallace, Steve Zymler. 2015. Distributionally robust multi-item newsvendor problems with multimodal demand distributions. *Mathematical Programming* **152**(1-2) 1–32.

Iancu, Dan A, Mayank Sharma, Maxim Sviridenko. 2013. Supermodularity and affine policies in dynamic robust optimization. *Operations Research* **61**(4) 941–956.

Jiang, Ruiwei, Minseok Ryu, Guanglin Xu. 2019. Data-driven distributionally robust appointment scheduling over Wasserstein balls. *arXiv preprint arXiv:1907.03219* .

Jiang, Ruiwei, Siqian Shen, Yiling Zhang. 2017. Integer programming approaches for appointment scheduling with random no-shows and service durations. *Operations Research* **65**(6) 1638–1656.

Kong, Qingxia, Chung-Yee Lee, Chung-Piaw Teo, Zhichao Zheng. 2013. Scheduling arrivals to a stochastic service delivery system using copositive cones. *Operations Research* **61**(3) 711–726.

Kuhn, Daniel, Wolfram Wiesemann, Angelos Georghiou. 2011. Primal and dual linear decision rules in stochastic and robust optimization. *Mathematical Programming* **130**(1) 177–209.

Lu, Mengshi, Lun Ran, Zuo-Jun Max Shen. 2015. Reliable facility location design under uncertain correlated disruptions. *Manufacturing & Service Operations Management* **17**(4) 445–455.

Lu, Yingdong, Jing-Sheng Song. 2005. Order-based cost optimization in assemble-to-order systems. *Operations Research* **53**(1) 151–169.

Mak, Ho-Yin, Ying Rong, Jiawei Zhang. 2014. Appointment scheduling with limited distributional information. *Management Science* **61**(2) 316–334.

Nadar, Emre, Mustafa Akan, Alan Scheller-Wolf. 2014. Optimal structural results for assemble-to-order generalized m-systems. *Operations Research* **62**(3) 571–579.

Natarajan, Karthik, Melvyn Sim, Joline Uichanco. 2017. Asymmetry and ambiguity in newsvendor models. *Management Science* **64**(7) 3146–3167.

Postek, Krzysztof, Aharon Ben-Tal, Dick Den Hertog, Bertrand Melenberg. 2018. Robust optimization with ambiguous stochastic constraints under mean and dispersion information. *Operations Research* **66**(3) 814–833.

Postek, Krzysztof, Ward Romeijnders, Dick Den Hertog, Maarten H van der Vlerk. 2019. An approximation framework for two-stage ambiguous stochastic integer programs under mean-MAD information. *European Journal of Operational Research* **274**(2) 432–444.

Qi, Jin. 2017. Mitigating delays and unfairness in appointment systems. *Management Science* **63**(2) 566–583.

Rachev, Svetlozar T, Ludger Rüschendorf. 1998. *Mass Transportation Problems: Volume I: Theory*, vol. 1. Springer Science & Business Media.

Saif, Ahmed, Erick Delage. 2021. Data-driven distributionally robust capacitated facility location problem. *European Journal of Operational Research* **291**(3) 995–1007.

Shapiro, Alexander, Darinka Dentcheva, Andrzej Ruszczyński. 2009. *Lectures on Stochastic Programming: Modeling and Theory*. SIAM.

Song, Jing-Sheng, Paul Zipkin. 2003. Supply chain operations: Assemble-to-order systems. *Handbooks in Operations Research and Management Science* **11** 561–596.

Topkis, Donald M. 1998. *Supermodularity and Complementarity*. Princeton university press.

van Eekelen, Wouter, Dick den Hertog, Johan SH van Leeuwaarden. 2022. MAD dispersion measure makes extremal queue analysis simple. *INFORMS Journal on Computing* .

Wiesemann, Wolfram, Daniel Kuhn, Melvyn Sim. 2014. Distributionally robust convex optimization. *Operations Research* **62**(6) 1358–1376.

Zeng, Bo, Long Zhao. 2013. Solving two-stage robust optimization problems using a column-and-constraint generation method. *Operations Research Letters* **41**(5) 457–461.

Zipkin, Paul. 2016. Some specially structured assemble-to-order systems. *Operations Research Letters* **44**(1) 136–142.

Appendices

A. Supplement

A.1. Algorithm for checking supermodularity in Section 3

For any given matrices $\boldsymbol{U}, \boldsymbol{V}$, we provide the following algorithm to check explicitly whether the condition in Theorem 2 is met.

Algorithm 2 algorithm for checking supermodularity

1: **Input:** $\boldsymbol{U} \in \Re^{r\times m}, \boldsymbol{V} \in \Re^{r\times n}$

2: **Initialization:** $r_0 = \mathrm{rank}(\boldsymbol{U}), s = 1$

3: **if** $r_0 < r$ **then**

4: arbitrarily remove columns in $\boldsymbol{U}$, if any, until $\boldsymbol{U}$ has only r_0 linearly independent columns

5: **for all** $\mathcal{I} \subseteq [r]$ with $|\mathcal{I}| = r_0$ and $\boldsymbol{U}_{\mathcal{I}}$ invertible, **do**

6: **for** $i \in [r]\backslash\mathcal{I}$ **do**

7: $\boldsymbol{d}_i^\top = \boldsymbol{v}_i^\top - \boldsymbol{u}_i^\top \boldsymbol{U}_{\mathcal{I}}^{-1} \boldsymbol{V}_{\mathcal{I}}$

8: **if** there exist components d_{ia}, d_{ib} such that $d_{ia}d_{ib} < 0$ **then**

9: $s = 0$, **go to** line 10

10: **return** s

Theorem 5 *The condition in Theorem 2 is satisfied if and only if Algorithm 2 returns* $s = 1$.

We note that Algorithm 2 may take exponential number of steps. Specifically, the complexity is reflected in line 5, where we search for all row index sets subject to conditions on the number of rows and rank. The high complexity is essentially because this algorithm is for the necessary and sufficient condition. Indeed, if we aim only for necessary conditions, then it can be simplified by reducing the range of search. For example, only checking for submatrices containing consecutive rows of $\boldsymbol{U}$ and $\boldsymbol{V}$ can also be a necessary condition. If the condition is violated for any tested index set $\mathcal{I}$, then the function $g(\boldsymbol{x}, \boldsymbol{z})$ must not be supermodular for all $\boldsymbol{x}, \boldsymbol{b}, \boldsymbol{v}^0$. On the other hand, if we aim at sufficient conditions only, some matrices with simple structures can be easily shown to satisfy the conditions.

A.2. A counterexample where segregated affine decision rules are suboptimal in Section 4

For simplicity, we drop the first-stage decision and consider the problem

$$\sup_{\mathbb{P}\in\mathcal{F}} \mathbb{E}_{\mathbb{P}}\left[g(\tilde{z}_1, \tilde{z}_2)\right]$$

as an example, where we define

$$\mathcal{F} = \left\{ \mathbb{P} \;\middle|\; \begin{array}{l} \mathbb{E}_{\mathbb{P}}[\tilde{z}_1] = 3,\ \mathbb{E}_{\mathbb{P}}[\tilde{z}_2] = 1 \\ \mathbb{E}_{\mathbb{P}}[|\tilde{z}_1 - 3|] \le 1.2,\ \mathbb{E}_{\mathbb{P}}[|\tilde{z}_2 - 1|] \le 1.2 \\ \mathbb{P}(\tilde{z}_1 \in [0,4]) = \mathbb{P}(\tilde{z}_2 \in [0,4]) = 1 \end{array} \right\}, \qquad g(z_1, z_2) = \min \left\{ y \;\middle|\; \begin{array}{l} y \ge z_1 + z_2 - 4 \\ y \ge -z_1 - z_2 + 4 \\ y \le 4 \end{array} \right\}$$

for the problem setting. Easily we can check that $g(z_1, z_2)$ satisfies the condition in Theorem 2, hence is supermodular in $\boldsymbol{z}$. Applying Algorithm 1 we obtain a worst-case distribution $\mathbb{P}^\star$ with $\mathbb{P}^\star(\tilde{\boldsymbol{z}} = (0,0)) = \mathbb{P}^\star(\tilde{\boldsymbol{z}} = (3,0)) = \mathbb{P}^\star(\tilde{\boldsymbol{z}} = (4,0)) = \mathbb{P}^\star(\tilde{\boldsymbol{z}} = (4,1)) = \mathbb{P}^\star(\tilde{\boldsymbol{z}} = (4,4)) = 0.2$, hence the problem has an optimal value

$$\begin{aligned} \sup_{\mathbb{P} \in \mathcal{F}} \mathbb{E}_{\mathbb{P}}[g(\tilde{z}_1, \tilde{z}_2)] &= \mathbb{E}_{\mathbb{P}^\star}[g(\tilde{z}_1, \tilde{z}_2)] \\ &= 0.2\,(g(0,0) + g(3,0) + g(4,0) + g(4,1) + g(4,4)) \\ &= 0.2\,(4 + 1 + 0 + 1 + 4) \\ &= 2. \end{aligned}$$

On the other hand, we consider the decision following the segregated linear decision rule, i.e., $y(z_1, z_2) = \boldsymbol{\theta}^\top ((3 - z_1)^+, (1 - z_2)^+, (z_1 - 3)^+, (z_2 - 1)^+) + \phi$. The problem with segregated linear decision rule (The detailed reformulation can be referred to Problem (26) in Appendix A.3) then becomes

$$\begin{aligned} \min \quad & \boldsymbol{\theta}^\top (0.6, 0.6, 0.6, 0.6) + \phi \\ \text{s.t.} \quad & \boldsymbol{\theta}^\top \left((3 - z_1)^+, (1 - z_2)^+, (z_1 - 3)^+, (z_2 - 1)^+\right) \ge z_1 + z_2 - 4, && \forall (z_1, z_2) \in \{0,3,4\} \times \{0,1,4\} \\ & \boldsymbol{\theta}^\top \left((3 - z_1)^+, (1 - z_2)^+, (z_1 - 3)^+, (z_2 - 1)^+\right) \ge -z_1 - z_2 + 4, && \forall (z_1, z_2) \in \{0,3,4\} \times \{0,1,4\} \\ & \boldsymbol{\theta}^\top \left((3 - z_1)^+, (1 - z_2)^+, (z_1 - 3)^+, (z_2 - 1)^+\right) \le 4, && \forall (z_1, z_2) \in \{0,3,4\} \times \{0,1,4\}. \end{aligned}$$

This is a linear program with 27 constraints and 5 decision variables, hence can be handled by ordinary solvers. The optimal segregated linear decision rule turns out to be $y(z_1, z_2) = \frac{1}{3}(3 - z_1)^+ + (1 - z_2)^+ + (z_1 - 3)^+ + \frac{1}{3}(z_2 - 1)^+ + 2$, yielding the optimal value $0.6(\frac{1}{3} + 1 + 1 + \frac{1}{3}) + 2 = 3.6 > 2$. This implies that segregated linear decision rules can be suboptimal even if the second-stage cost is supermodular in $\boldsymbol{z}$.

A.3. The CCG algorithm and segregated affine decision rules in Section 6

CCG algorithm

The CCG algorithm we use in this paper mostly resembles Saif and Delage (2021), who study a two-stage distributionally robust facility location problem. For Problem (2), the inner supremum problem $\sup_{\mathbb{P} \in \mathcal{F}}$ can be equivalently written as an infimum problem and the strong duality holds. Then, we reformulate Problem (2) as

$$\begin{aligned} \min \quad & \boldsymbol{a}^\top \boldsymbol{x} + \alpha + \boldsymbol{\mu}^\top \boldsymbol{\beta} + \boldsymbol{\delta}^\top \boldsymbol{\gamma} \\ \text{s.t.} \quad & \alpha + \boldsymbol{\beta}^\top \boldsymbol{z} + \boldsymbol{\gamma}^\top (|\boldsymbol{z} - \boldsymbol{\mu}|) \ge g(\boldsymbol{x}, \boldsymbol{z}), \quad \forall \boldsymbol{z} \in [\underline{\boldsymbol{z}}, \overline{\boldsymbol{z}}] \\ & \boldsymbol{\gamma} \ge \boldsymbol{0},\ \boldsymbol{x} \in \mathcal{X}. \end{aligned} \tag{20}$$

For the CCG algorithm, we assume that the relatively complete recourse holds and consider multiple iterations. In iteration τ ($\tau = 0, 1, 2, \ldots$), we solve a relaxation of Problem (20) that imposes the first constraint to hold only on a subset $\{\boldsymbol{z}^0, \boldsymbol{z}^1, \ldots, \boldsymbol{z}^\tau\} \subseteq [\underline{\boldsymbol{z}}, \overline{\boldsymbol{z}}]$. Formally, let

$$\begin{aligned}
\hat{L}_\tau := \min \quad & \boldsymbol{a}^\top \boldsymbol{x} + \alpha + \boldsymbol{\mu}^\top \boldsymbol{\beta} + \boldsymbol{\delta}^\top \boldsymbol{\gamma} \\
\text{s.t.} \quad & \alpha + \boldsymbol{\beta}^\top \boldsymbol{z}^i + \boldsymbol{\gamma}^\top (|\boldsymbol{z}^i - \boldsymbol{\mu}|) \geq \boldsymbol{b}^\top \boldsymbol{y}^i, \quad i \in [\tau] \\
& \boldsymbol{W}\boldsymbol{x} + \boldsymbol{U}\boldsymbol{y}^i \geq \boldsymbol{V}\boldsymbol{z}^i + \boldsymbol{v}^0, \qquad\qquad i \in [\tau] \\
& \boldsymbol{\gamma} \geq \boldsymbol{0},\ \boldsymbol{x} \in \mathcal{X}.
\end{aligned} \tag{21}$$

This problem provides a lower bound to the optimal value of the original problem (2). Denoting by $\left(\hat{\boldsymbol{x}}, \hat{\alpha}, \hat{\boldsymbol{\beta}}, \hat{\boldsymbol{\gamma}}, \{\hat{\boldsymbol{y}}^i\}_{i\in[\tau]}\right)$ a solution to Problem (21), we then examine the violation of the constraint in Problem (20) by evaluating

$$\begin{aligned}
h\left(\hat{\boldsymbol{x}}, \hat{\boldsymbol{\beta}}, \hat{\gamma}\right) &= \max\left\{ g(\hat{\boldsymbol{x}}, \boldsymbol{z}) - \hat{\boldsymbol{\beta}}^\top \boldsymbol{z} - \hat{\gamma}^\top (|\boldsymbol{z} - \boldsymbol{\mu}|) \;\middle|\; \boldsymbol{z} \in [\underline{\boldsymbol{z}}, \overline{\boldsymbol{z}}] \right\} \\
&= \max\left\{ g(\hat{\boldsymbol{x}}, \boldsymbol{z}) - \hat{\boldsymbol{\beta}}^\top \boldsymbol{z} - \hat{\gamma}^\top \boldsymbol{\theta} \;\middle|\; \begin{array}{l} \boldsymbol{\theta} \geq \boldsymbol{z} - \boldsymbol{\mu},\ \boldsymbol{\theta} \geq \boldsymbol{\mu} - \boldsymbol{z},\ \boldsymbol{\theta} \leq M_0 \boldsymbol{1} \\ \boldsymbol{z} \in [\underline{\boldsymbol{z}}, \overline{\boldsymbol{z}}] \end{array} \right\} \\
&= \max\left\{ (\boldsymbol{V}\boldsymbol{z} + \boldsymbol{v}^0 - \boldsymbol{W}\hat{\boldsymbol{x}})^\top \boldsymbol{\eta} - \hat{\boldsymbol{\beta}}^\top \boldsymbol{z} - \hat{\gamma}^\top \boldsymbol{\theta} \;\middle|\; \begin{array}{l} \boldsymbol{U}^\top \boldsymbol{\eta} = \boldsymbol{b},\ \boldsymbol{\eta} \geq \boldsymbol{0} \\ \boldsymbol{\theta} \geq \boldsymbol{z} - \boldsymbol{\mu},\ \boldsymbol{\theta} \geq \boldsymbol{\mu} - \boldsymbol{z},\ \boldsymbol{\theta} \leq M_0 \boldsymbol{1} \\ \boldsymbol{z} \in [\underline{\boldsymbol{z}}, \overline{\boldsymbol{z}}] \end{array} \right\} \\
&= \max_{\substack{\boldsymbol{\eta}: \boldsymbol{U}^\top \boldsymbol{\eta} = \boldsymbol{b} \\ \boldsymbol{\eta} \geq \boldsymbol{0}}} \left\{ (\boldsymbol{v}^0 - \boldsymbol{W}\hat{\boldsymbol{x}})^\top \boldsymbol{\eta} + \max\left\{ (\boldsymbol{V}^\top \boldsymbol{\eta} - \hat{\boldsymbol{\beta}})^\top \boldsymbol{z} - (\hat{\gamma})^\top \boldsymbol{\theta} \;\middle|\; \begin{array}{l} \boldsymbol{z} - \boldsymbol{\theta} \leq \boldsymbol{\mu} \\ -\boldsymbol{z} - \boldsymbol{\theta} \leq -\boldsymbol{\mu} \\ \boldsymbol{\theta} \leq M_0 \boldsymbol{1} \\ \boldsymbol{z} \leq \overline{\boldsymbol{z}} \\ -\boldsymbol{z} \leq -\underline{\boldsymbol{z}} \end{array} \right\} \right\}.
\end{aligned} \tag{22}$$

Here the second equality follows from $\hat{\boldsymbol{\gamma}} \geq \boldsymbol{0}$, and the constraint $\boldsymbol{\theta} \leq M_0 \boldsymbol{1}$ with M_0 being a sufficiently large real number guarantees the boundedness of the feasible set while keeping the optimal value unchanged. The third equality follows from the dual form of $g(\hat{\boldsymbol{x}}, \boldsymbol{z})$, since the relatively complete course assumption guarantees that the strong duality holds. In the last line we seperate the maximization of $\boldsymbol{\eta}$ and $\boldsymbol{z}, \boldsymbol{\theta}$. Further, replacing the constraints of the inner maximization by its KKT conditions, we get

$$
\begin{aligned}
h\left(\hat{\boldsymbol{x}},\hat{\boldsymbol{\beta}},\hat{\boldsymbol{\gamma}}\right) = \max\quad & (\boldsymbol{v}^0-\boldsymbol{W}\hat{\boldsymbol{x}})^\top\boldsymbol{\eta}+\boldsymbol{\mu}^\top(\overline{\boldsymbol{\phi}}-\underline{\boldsymbol{\phi}})+M_0\mathbf{1}^\top\boldsymbol{\xi}+\overline{\boldsymbol{z}}^\top\overline{\boldsymbol{\rho}}-\underline{\boldsymbol{z}}^\top\underline{\boldsymbol{\rho}} \\
\text{s.t.}\quad & \boldsymbol{U}^\top\boldsymbol{\eta}=\boldsymbol{b} \\
& \boldsymbol{z}-\boldsymbol{\theta}\le\boldsymbol{\mu} & (\overline{\boldsymbol{\phi}}) \\
& -\boldsymbol{z}-\boldsymbol{\theta}\le-\boldsymbol{\mu} & (\underline{\boldsymbol{\phi}}) \\
& \boldsymbol{\theta}\le M_0\mathbf{1} & (\boldsymbol{\xi}) \\
& \boldsymbol{z}\le\overline{\boldsymbol{z}} & (\overline{\boldsymbol{\rho}}) \\
& -\boldsymbol{z}\le-\underline{\boldsymbol{z}} & (\underline{\boldsymbol{\rho}}) \\
& \overline{\boldsymbol{\phi}}-\underline{\boldsymbol{\phi}}+\overline{\boldsymbol{\rho}}-\underline{\boldsymbol{\rho}}=\boldsymbol{V}^\top\boldsymbol{\eta}-\hat{\boldsymbol{\beta}} & (\boldsymbol{z}) \\
& \overline{\boldsymbol{\phi}}-\underline{\boldsymbol{\phi}}-\boldsymbol{\xi}=\hat{\boldsymbol{\gamma}} & (\boldsymbol{\theta}) \\
& \boldsymbol{\mu}-\boldsymbol{z}+\boldsymbol{\theta}\le M_0\boldsymbol{\lambda}^{\overline{\boldsymbol{\phi}}},\quad \overline{\boldsymbol{\phi}}\le M_0(\mathbf{1}-\boldsymbol{\lambda}^{\overline{\boldsymbol{\phi}}}) \\
& -\boldsymbol{\mu}+\boldsymbol{z}+\boldsymbol{\theta}\le M_0\boldsymbol{\lambda}^{\underline{\boldsymbol{\phi}}},\quad \underline{\boldsymbol{\phi}}\le M_0(\mathbf{1}-\boldsymbol{\lambda}^{\underline{\boldsymbol{\phi}}}) \\
& M_0\mathbf{1}-\boldsymbol{\theta}\le M_0\boldsymbol{\lambda}^{\boldsymbol{\xi}},\quad \boldsymbol{\xi}\le M_0(\mathbf{1}-\boldsymbol{\lambda}^{\boldsymbol{\xi}}) \\
& \overline{\boldsymbol{z}}-\boldsymbol{z}\le M_0\boldsymbol{\lambda}^{\overline{\boldsymbol{\rho}}},\quad \overline{\boldsymbol{\rho}}\le M_0(\mathbf{1}-\boldsymbol{\lambda}^{\overline{\boldsymbol{\rho}}}) \\
& \boldsymbol{z}-\underline{\boldsymbol{z}}\le M_0\boldsymbol{\lambda}^{\underline{\boldsymbol{\rho}}},\quad \underline{\boldsymbol{\rho}}\le M_0(\mathbf{1}-\boldsymbol{\lambda}^{\underline{\boldsymbol{\rho}}}) \\
& \overline{\boldsymbol{\phi}},\underline{\boldsymbol{\phi}},\boldsymbol{\xi},\overline{\boldsymbol{\rho}},\underline{\boldsymbol{\rho}},\boldsymbol{\eta}\ge\mathbf{0} \\
& \boldsymbol{\lambda}^{\overline{\boldsymbol{\phi}}},\boldsymbol{\lambda}^{\underline{\boldsymbol{\phi}}},\boldsymbol{\lambda}^{\boldsymbol{\xi}},\boldsymbol{\lambda}^{\overline{\boldsymbol{\rho}}},\boldsymbol{\lambda}^{\underline{\boldsymbol{\rho}}}\in\{0,1\}^n.
\end{aligned}
\tag{23}
$$

Here the variables in parentheses specify the associated dual variable for each constraint, and the binary variables $\boldsymbol{\lambda}^{\overline{\boldsymbol{\phi}}},\boldsymbol{\lambda}^{\underline{\boldsymbol{\phi}}},\boldsymbol{\lambda}^{\boldsymbol{\xi}},\boldsymbol{\lambda}^{\overline{\boldsymbol{\rho}}},\boldsymbol{\lambda}^{\underline{\boldsymbol{\rho}}}$ are introduced such that $\boldsymbol{\mu}-\boldsymbol{z}+\boldsymbol{\theta}\le M\boldsymbol{\lambda}^{\overline{\boldsymbol{\phi}}},\overline{\boldsymbol{\phi}}\le M(\mathbf{1}-\boldsymbol{\lambda}^{\overline{\boldsymbol{\phi}}})$ is equivalent to the complementary slackness condition $(\mu_i-z_i+\theta_i)\overline{\boldsymbol{\phi}}_i=0\ (i\in[n])$ for $\boldsymbol{\lambda}^{\overline{\boldsymbol{\phi}}}$ and so forth for the other binary variables. We hence reformulate the bilinear problem in the third line of Equation (22) into a mixed integer program.

By solving Problem (23), we derive the optimal solution of $\boldsymbol{z}$ and denote as $\boldsymbol{z}^{\tau+1}$. Further, we obtain an upper bound of Problem (2) as

$$
\hat{U}_\tau := \boldsymbol{a}^\top\hat{\boldsymbol{x}}+h\left(\hat{\boldsymbol{x}},\hat{\boldsymbol{\beta}},\hat{\boldsymbol{\gamma}}\right)+\boldsymbol{\mu}^\top\hat{\boldsymbol{\beta}}+\boldsymbol{\delta}^\top\hat{\boldsymbol{\gamma}} \tag{24}
$$

This is an upper bound because the fixed $\left(\hat{\boldsymbol{x}},\hat{\boldsymbol{\beta}},\hat{\boldsymbol{\gamma}}\right)$ and $\alpha=h\left(\hat{\boldsymbol{x}},\hat{\boldsymbol{\beta}},\hat{\boldsymbol{\gamma}}\right)$ form a feasible solution to Problem (20).

We terminate the algorithm when the upper and lower bounds converge. If the algorithm does not stop in iteration τ, we then add a decision variable $\boldsymbol{y}^{\tau+1}$ and the following constraints

$$
\begin{cases}
\alpha+\boldsymbol{\beta}^\top\boldsymbol{z}^{\tau+1}+\boldsymbol{\gamma}^\top(|\boldsymbol{z}^{\tau+1}-\boldsymbol{\mu}|)\ge\boldsymbol{b}^\top\boldsymbol{y}^{\tau+1}, \\
\boldsymbol{W}\boldsymbol{x}+\boldsymbol{U}\boldsymbol{y}^{\tau+1}\ge\boldsymbol{V}\boldsymbol{z}^{\tau+1}+\boldsymbol{v}^0
\end{cases}
$$

to the lower bound problem (21) and solve for $\hat{L}_{\tau+1}$ in iteration $\tau+1$. The full procedure is formalized in Algorithm 3.

Algorithm 3 A CCG algorithm for two-stage DRO

1: **Input:** two-stage formulation (2), second-stage formulation of $g(\boldsymbol{x},\boldsymbol{z})$, $\epsilon \geq 0$
2: **Initialization:** $LB=-\infty, UB=\infty$, $\tau=0$, $\boldsymbol{z}^0=\boldsymbol{\mu}$
3: **while** $UB-LB>\epsilon$ **do**
4: solve Problem (21) with $\{\boldsymbol{z}^i\}_{i\in[\tau]}$
5: denote by $(\hat{\boldsymbol{x}},\hat{\boldsymbol{\beta}},\hat{\boldsymbol{\gamma}})$ the optimal solution, by $\hat{L}_\tau$ the optimal value
6: update $LB=\hat{L}_\tau$
7: solve Problem (23) with $(\hat{\boldsymbol{x}},\hat{\boldsymbol{\beta}},\hat{\boldsymbol{\gamma}})$
8: denote by $\boldsymbol{z}^{\tau+1}$ the optimal solution for $\boldsymbol{z}$, by $h(\hat{\boldsymbol{x}},\hat{\boldsymbol{\beta}},\hat{\boldsymbol{\gamma}})$ the optimal value
9: calculate $\hat{U}_\tau$ based on Equation (24) with $h(\hat{\boldsymbol{x}},\hat{\boldsymbol{\beta}},\hat{\boldsymbol{\gamma}})$
10: update $UB=\min\left\{UB,\hat{U}_\tau\right\}$
11: update $\tau=\tau+1$
12: **Output:** optimal solution $\hat{\boldsymbol{x}}$

Segregated affine decision rule

For Problem (2), when we use the segregated decision rule method, the problem is reformulated as

$$\begin{aligned}
\min_{\boldsymbol{x},\boldsymbol{\Theta},\boldsymbol{\phi}} \quad & \boldsymbol{a}^\top\boldsymbol{x}+\sup_{\mathbb{P}\in\mathcal{F}}\ \mathbb{E}_\mathbb{P}\left[\boldsymbol{b}^\top\left(\boldsymbol{\Theta}\begin{bmatrix}(\boldsymbol{\mu}-\tilde{\boldsymbol{z}})^+\\(\tilde{\boldsymbol{z}}-\boldsymbol{\mu})^+\end{bmatrix}+\boldsymbol{\phi}\right)\right]\\
\text{s.t.} \quad & \boldsymbol{W}\boldsymbol{x}+\boldsymbol{U}\left(\boldsymbol{\Theta}\begin{bmatrix}(\boldsymbol{\mu}-\boldsymbol{z})^+\\(\boldsymbol{z}-\boldsymbol{\mu})^+\end{bmatrix}+\boldsymbol{\phi}\right)\geq\boldsymbol{V}\boldsymbol{z}+\boldsymbol{v}^0, \qquad \forall\boldsymbol{z}\in[\underline{\boldsymbol{z}},\overline{\boldsymbol{z}}]\\
& \boldsymbol{x}\in\mathcal{X}.
\end{aligned}$$

By the property of segregated linearity, we can equivalently impose the constraints only on the breakpoints, hence obtain the following reformulation.

$$\begin{aligned}
\min \quad & \boldsymbol{a}^\top\boldsymbol{x}+\sup_{\mathbb{P}\in\mathcal{F}}\mathbb{E}_\mathbb{P}\left[\boldsymbol{b}^\top\left(\boldsymbol{\Theta}\begin{bmatrix}(\boldsymbol{\mu}-\tilde{\boldsymbol{z}})^+\\(\tilde{\boldsymbol{z}}-\boldsymbol{\mu})^+\end{bmatrix}+\boldsymbol{\phi}\right)\right]\\
\text{s.t.} \quad & \boldsymbol{W}\boldsymbol{x}+\boldsymbol{U}\left(\boldsymbol{\Theta}\begin{bmatrix}(\boldsymbol{\mu}-\boldsymbol{z})^+\\(\boldsymbol{z}-\boldsymbol{\mu})^+\end{bmatrix}+\boldsymbol{\phi}\right)\geq\boldsymbol{V}\boldsymbol{z}+\boldsymbol{v}^0, \quad \forall\boldsymbol{z}\in\prod_{i\in[n]}\{\underline{z}_i,\mu_i,\overline{z}_i\},\\
& \boldsymbol{x}\in\mathcal{X}.
\end{aligned} \tag{25}$$

Similar to the proof of Proposition 1, we can derive a worst-case distribution $\mathbb{P}^\star$ and show that the MAD under $\mathbb{P}^\star$ is $\mathbb{E}_{\mathbb{P}^\star}[|\tilde{\boldsymbol{z}}-\boldsymbol{\mu}|]=\hat{\boldsymbol{\delta}}$. Observing that $\mathbb{E}_{\mathbb{P}^\star}[(\boldsymbol{\mu}-\tilde{\boldsymbol{z}})^++(\tilde{\boldsymbol{z}}-\boldsymbol{\mu})^+]=\mathbb{E}_{\mathbb{P}^\star}[|\tilde{\boldsymbol{z}}-\boldsymbol{\mu}|]=\hat{\boldsymbol{\delta}}$,

$\mathbb{E}_{\mathbb{P}^\star}[(\boldsymbol{\mu}-\tilde{\boldsymbol{z}})^+ - (\tilde{\boldsymbol{z}}-\boldsymbol{\mu})^+] = \mathbb{E}_{\mathbb{P}^\star}[\boldsymbol{\mu}-\tilde{\boldsymbol{z}}] = \boldsymbol{0}$, we have $\mathbb{E}_{\mathbb{P}^\star}[(\boldsymbol{\mu}-\tilde{\boldsymbol{z}})^+] = \mathbb{E}_{\mathbb{P}^\star}[(\tilde{\boldsymbol{z}}-\boldsymbol{\mu})^+] = \hat{\boldsymbol{\delta}}/2$. Hence, Problem (25) can be further reformulated as

$$\begin{aligned} \min \quad & \boldsymbol{a}^\top \boldsymbol{x} + \boldsymbol{b}^\top \boldsymbol{\Theta} \begin{bmatrix} \hat{\boldsymbol{\delta}}/2 \\ \hat{\boldsymbol{\delta}}/2 \end{bmatrix} + \boldsymbol{b}^\top \boldsymbol{\phi} \\ \text{s.t.} \quad & \boldsymbol{W}\boldsymbol{x} + \boldsymbol{U}\left(\boldsymbol{\Theta}\begin{bmatrix} (\boldsymbol{\mu}-\boldsymbol{z})^+ \\ (\boldsymbol{z}-\boldsymbol{\mu})^+ \end{bmatrix} + \boldsymbol{\phi}\right) \geq \boldsymbol{V}\boldsymbol{z} + \boldsymbol{v}^0, \quad \forall \boldsymbol{z} \in \prod_{i\in[n]} \{\underline{z}_i, \mu_i, \overline{z}_i\}, \\ & \boldsymbol{x} \in \mathcal{X}. \end{aligned} \tag{26}$$

This problem is essentially a linear program with an exponential size of constraints.

B. Extensions

In this section, we introduce three possible extensions and show that, when the property of supermodularity holds, the exact tractable reformulation can be applied to more general settings.

B.1. Left-hand-side uncertainties in the constraints

We consider that the matrix $\boldsymbol{W}$ on the left-hand side of the constraints is an affine function of the uncertain vector $\tilde{\boldsymbol{z}}$ as $\boldsymbol{W}(\tilde{\boldsymbol{z}}) = \boldsymbol{W}^0 + \sum_{i\in[n]} \boldsymbol{W}^i \tilde{z}_i$. In this case, the second-stage problem becomes

$$\begin{aligned} g^W(\boldsymbol{x},\boldsymbol{z}) = \min \quad & \boldsymbol{b}^\top \boldsymbol{y} \\ \text{s.t.} \quad & \left(\boldsymbol{W}^0 + \sum_{i\in[n]} \boldsymbol{W}^i z_i\right)\boldsymbol{x} + \boldsymbol{U}\boldsymbol{y} \geq \boldsymbol{V}\boldsymbol{z} + \boldsymbol{v}^0, \end{aligned}$$

where $\boldsymbol{W}^i, i \in \{0,1,\ldots,n\}$ are given constant matrices in $\Re^{r\times l}$. We next establish an equivalent condition for the supermodularity of g^W.

Theorem 6 *$g^W(\boldsymbol{x},\boldsymbol{z})$ is supermodular in $\boldsymbol{z}$ for any $\boldsymbol{x},\boldsymbol{b}$ and $\boldsymbol{v}^0$ if and only if $\boldsymbol{U}\in\Re^{r\times m}$, $\boldsymbol{V}\in\Re^{r\times n}$ and $\boldsymbol{W}_i \in \Re^{r\times l}, i\in[n]$ satisfy one of the following conditions:*

1) $\operatorname{rank}(\boldsymbol{U}) = r$,

2) for all $\mathcal{I}\subseteq[r], \boldsymbol{\eta}\in\Re^{|\mathcal{I}|}$ with $|\mathcal{I}| = \operatorname{rank}(\boldsymbol{U})+1, \operatorname{rank}(\boldsymbol{U}_\mathcal{I}) = \operatorname{rank}(\boldsymbol{U})$ and $\boldsymbol{U}_\mathcal{I}^\top\boldsymbol{\eta} = \boldsymbol{0}$, we have

2a) $(\boldsymbol{\eta}^\top(\boldsymbol{V}_\mathcal{I})_i)\cdot(\boldsymbol{\eta}^\top(\boldsymbol{V}_\mathcal{I})_j) \geq 0$, $(\boldsymbol{W}_\mathcal{I}^i)^\top\boldsymbol{\eta}\boldsymbol{\eta}^\top\boldsymbol{W}_\mathcal{I}^j$ is positive semidefinite, for all $i,j\in[n]$;

2b) $(\boldsymbol{\eta}^\top(\boldsymbol{V}_\mathcal{I})_i)\cdot\left(\boldsymbol{\eta}^\top\boldsymbol{W}_\mathcal{I}^j\right) = (\boldsymbol{\eta}^\top(\boldsymbol{V}_\mathcal{I})_j)\cdot(\boldsymbol{\eta}^\top\boldsymbol{W}_\mathcal{I}^i)$, for all $i,j\in[n]$.

For Condition 2) in Theorem 6, considering any concerned $\mathcal{I}$, i.e., $|\mathcal{I}| = \operatorname{rank}(\boldsymbol{U})+1$ and $\operatorname{rank}(\boldsymbol{U}_\mathcal{I}) = \operatorname{rank}(\boldsymbol{U})$, the null space of $\boldsymbol{U}_\mathcal{I}$ is of dimension 1. That is, there exists $\boldsymbol{\eta}^o$ such that for all $\boldsymbol{\eta}$ with $\boldsymbol{U}_\mathcal{I}^\top\boldsymbol{\eta} = \boldsymbol{0}$ we have $\boldsymbol{\eta} = k\boldsymbol{\eta}^o$ for some $k\in\Re$. We can easily observe that both Conditions 2a) and 2b) hold for all such $\boldsymbol{\eta}$ if and only if they hold for $\boldsymbol{\eta}^o$. Therefore, to verify whether Conditions 2a) and 2b) hold, it suffices to check for $\boldsymbol{\eta}^o$ only. Hence, as in Theorem 5, we can similarly build a corresponding algorithm to check the supermodularity of g^W.

B.2. Non-linearity in the objective function

We extend our results by considering the objective as a more general function, which is nonlinear of the second-stage cost. For example, the objective can be either an expected disutility or a risk measure. Specifically, when the second-stage cost itself is supermodular in the uncertainty, the following lemma identifies mild conditions which are sufficient to preserve supermodularity. We subsequently show how our method can help us obtain tractable reformulations.

Lemma 3 *Given any convex and non-decreasing function $u:\Re\to\Re$ and any monotone supermodular function $h:\Re^n\to\Re$, the function $\phi:\Re^n\to\Re$ defined as $\phi(\boldsymbol{z})=u(h(\boldsymbol{z}))$ is supermodular.*

This result can be applied when maximizing the decision maker's expected utility, or equivalently, minimizing the expected disutility. Consider the following problem

$$\min_{\boldsymbol{x}\in\mathcal{X}}\ \sup_{\mathbb{P}\in\mathcal{F}}\ \mathbb{E}_{\mathbb{P}}\left[u\left(\boldsymbol{a}^\top\boldsymbol{x}+g(\boldsymbol{x},\tilde{\boldsymbol{z}})\right)\right], \tag{27}$$

where $g(\boldsymbol{x},\boldsymbol{z})$ is the second-stage cost function defined by (1), and $u:\Re\to\Re$ is a piecewise linear convex and non-decreasing disutility function defined as

$$u(w)=\max_{j\in[J]}\ \{c_jw+d_j\},\quad \forall w\in\Re, \tag{28}$$

for some constants $c_j\geq 0$ and $d_j,\ j\in[J]$.

Proposition 12 *If $g(\boldsymbol{x},\boldsymbol{z})$ is monotone and supermodular in $\boldsymbol{z}$ for all $\boldsymbol{x}\in\mathcal{X}$, then Problem (27) is equivalent to the following problem*

$$\begin{array}{lll}\min & \boldsymbol{\nu}^\top\boldsymbol{l} & \\ \text{s.t.} & \boldsymbol{R}_k^\top\boldsymbol{l}\geq\sum\limits_{i\in[2n+1]}p_i^kf^{k,i}, & k\in[K]\\ & \boldsymbol{W}\boldsymbol{x}+\boldsymbol{U}\boldsymbol{y}^{k,i}\geq\boldsymbol{V}\boldsymbol{z}^{k,i}+\boldsymbol{v}^0, & k\in[K],i\in[2n+1]\\ & f^{k,i}\geq c_j\left(\boldsymbol{a}^\top\boldsymbol{x}+\boldsymbol{b}^\top\boldsymbol{y}^{k,i}\right)+d_j, & k\in[K],i\in[2n+1],j\in[J]\\ & \boldsymbol{l}\geq\boldsymbol{0},\ \boldsymbol{x}\in\mathcal{X}, & \end{array} \tag{29}$$

where $p_i^k,\boldsymbol{z}^{k,i},\ i\in[2n+1]$ are obtained from Algorithm 1 given the ambiguity sets $\mathcal{F}^k,k\in[K]$ defined by (4).

We can also apply Lemma 3 when some risk measures are included in the objective. In particular, we study the case where the objective function is based on Optimized Certainty Equivalent (OCE) (Ben-Tal and Teboulle 1986). It is shown that the OCE models a broad range of risk measures

(Ben-Tal and Teboulle 2007), and includes the Conditional-Value-at-Risk (CVaR) as a special case. When evaluating the total cost by OCE, the two-stage problem is as follows,

$$\min_{\boldsymbol{x}\in\mathcal{X}} \sup_{\mathbb{P}\in\mathcal{F}} \mathrm{OCE}_u\big(\boldsymbol{a}^\top\boldsymbol{x}+g(\boldsymbol{x},\tilde{\boldsymbol{z}})\big)=\min_{\boldsymbol{x}\in\mathcal{X}} \sup_{\mathbb{P}\in\mathcal{F}} \inf_{\theta\in\Re} \{\theta+\mathbb{E}_{\mathbb{P}}\left[u\big(\boldsymbol{a}^\top\boldsymbol{x}+g(\boldsymbol{x},\tilde{\boldsymbol{z}})-\theta\big)\right]\}, \tag{30}$$

where $u(\cdot)$ is a piecewise linear convex and non-decreasing disutility function taken the form of (28). We now show that our method is applicable to Problem (30).

Corollary 3 *If $g(\boldsymbol{x},\boldsymbol{z})$ is monotone and supermodular in $\boldsymbol{z}$ for all $\boldsymbol{x}\in\mathcal{X}$, then the OCE minimization problem (30) is equivalent to the following linear program*

$$\begin{array}{lll} \min & \theta+\boldsymbol{\nu}^\top\boldsymbol{l} & \\ \text{s.t.} & \boldsymbol{R}_k^\top\boldsymbol{l}\geq \sum\limits_{i\in[2n+1]} p_i^k f^{k,i}, & k\in[K] \\ & \boldsymbol{W}\boldsymbol{x}+\boldsymbol{U}\boldsymbol{y}^{k,i}\geq \boldsymbol{V}\boldsymbol{z}^{k,i}+\boldsymbol{v}^0, & k\in[K], i\in[2n+1] \\ & f^{k,i}\geq c_j\left(\boldsymbol{a}^\top\boldsymbol{x}+\boldsymbol{b}^\top\boldsymbol{y}^{k,i}-\theta\right)+d_j, & k\in[K], i\in[2n+1], j\in[J] \\ & \boldsymbol{l}\geq\boldsymbol{0},\ \boldsymbol{x}\in\mathcal{X}, & \end{array} \tag{31}$$

where $p_i^k,\boldsymbol{z}^{k,i},\ i\in[2n+1]$ are obtained from Algorithm 1 given the ambiguity sets $\mathcal{F}^k, k\in[K]$ defined by (4).

B.3. General ambiguity set

While the previous results are based on the ambiguity set that is constructed by mean, support and MAD in each scenario (see Equation (3)), now we extend that ambiguity set to a more general one and show that it is the most general case in which our results are still applicable. We define the ambiguity set based on piecewise linear convex functions, which are rather general and still maintain the linear structure in the reformulation. For notational simplicity, we do not incorporate the random scenario in this subsection. Specifically, we consider the ambiguity set defined as follows,

$$\mathcal{F}^G=\left\{\mathbb{P}\ \middle|\ \begin{array}{l} \mathbb{P}\big(\underline{\boldsymbol{z}}\leq\tilde{\boldsymbol{z}}\leq\overline{\boldsymbol{z}}\big)=1 \\ \mathbb{E}_{\mathbb{P}}[\tilde{\boldsymbol{z}}]=\boldsymbol{\mu} \\ \mathbb{E}_{\mathbb{P}}\left[h_i^j(\tilde{z}_i)\right]\leq\delta_i^j, \quad i\in[n], j\in[J_i] \end{array}\right\}, \tag{32}$$

where $J_i\geq 1$ is an integer, h_i^j is a given piecewise linear convex function, $i\in[n],\ j\in[J_i]$. We assume h_i^j has at least two pieces in $[\underline{z}_i,\overline{z}_i]$ to avoid the trivial case. The ambiguity set $\mathcal{F}^G$ generalizes $\mathcal{F}^k$, defined in Equation (4), as it replaces the MAD information in $\mathcal{F}^k$ by the expectations of several piecewise linear convex functions. Obviously, $\mathcal{F}^G$ includes $\mathcal{F}^k$ as a special case by choosing $J_i=1$ and $h_i^j(z)=|z-\mu_i|$ for all $z\in\Re$.

Unfortunately, as we will show later in this subsection, not all ambiguity sets $\mathcal{F}^G$ defined by Equation (32) can lead to a tractable reformulation using the procedures we discussed in Section 2.

Here we aim to identify the conditions for $\mathcal{F}^G$ such that the corresponding two-stage optimization problem, whenever the property of supermodularity holds for the second-stage cost function, can be solved with the methods in Section 2.

For any $i \in [n]$, we let $z_i^1 = \underline{z}_i, z_i^2 = \overline{z}_i$ and denote $z_i^3, z_i^4, \ldots, z_i^{S_i} \in (\underline{z}_i, \overline{z}_i)$ as the breakpoints of the piecewise linear functions $h_i^1, \ldots, h_i^{J_i}$. We now have the following result, which is essential for using the procedures in Section 2.

Theorem 7 *The following two statements are equivalent.*

1. *Given any* $\delta_i^j,\ i \in [n],\ j \in [J_i]$ *satisfying* $\mathcal{F}^G \neq \emptyset$, *there exists* $\boldsymbol{p}_i = (p_{i1}, \ldots, p_{iS_i}) \in \Re_+^{S_i}, i \in [n]$ *such that for all convex function* $f : \Re^n \to \Re$, *we have* $\mathbb{P}^* \in \arg\sup_{\mathbb{P} \in \mathcal{F}^G} \mathbb{E}_{\mathbb{P}}[f(\tilde{\boldsymbol{z}})]$ *and for any* $i \in [n]$,
$$\mathbb{P}^*(\tilde{z}_i = w) = \begin{cases} p_{is} & \text{if } w = z_i^s,\ s \in [S_i], \\ 0 & \text{otherwise.} \end{cases}$$
2. *For all* $i \in [n],\ j \in [J_i]$, h_i^j *has exactly two pieces on* $[\underline{z}_i, \overline{z}_i]$.

We observe that the worst-case distribution $\mathbb{P}^*$ provided in Theorem 7 has the same structure as that in Proposition 1. Essentially, we can characterize their marginal distributions for both settings. Moreover, the marginal distribution depends only on the ambiguity set itself, and is independent of the objective function f (in Theorem 7) or the first-stage decision $\boldsymbol{x}$ (in Proposition 1). Therefore, if $\mathcal{F}^G$ satisfies the condition in Theorem 7, we can adopt a similar procedure to that in Section 2 to solve the two-stage optimization problem. In particular, we first obtain the marginal distribution, and then find the worst-case distribution based on the chained support, after which we can reformulate the two-stage problem as a linear program with low dimension. By contrast, if $\mathcal{F}^G$ violates the condition in Theorem 7, there are two-stage problems such that the worst-case distribution would depend on the first-stage decision $\boldsymbol{x}$, and hence our method cannot work.

C. Proofs

C.1. Proof of Proposition 1

For a one-dimensional random variable, when its MAD is known exactly, Ben-Tal and Hochman (1972, Theorem 3) have derived the worst-case distribution. Here since in $\mathcal{F}^k$, it involves multiple dimensions and only the upper bound of MADs is known, we need to prove differently as follows. For notational brevity, we drop the superscript k.

We now consider any $i \in [n]$ such that $\overline{z}_i > \underline{z}_i$. For the i-th marginal, given the support $[\underline{z}_i, \overline{z}_i]$ and mean μ_i, the maximum possible value of MAD is $\frac{2(\overline{z}_i-\mu_i)(\mu_i-\underline{z}_i)}{\overline{z}_i-\underline{z}_i}$ (see Ben-Tal and Hochman (1972, Lemma 1)). Hence, we let $\hat{\delta}_i = \min\left\{\delta_i, \frac{2(\overline{z}_i-\mu_i)(\mu_i-\underline{z}_i)}{\overline{z}_i-\underline{z}_i}\right\}$. Then, the worst-case expectation of $g(\boldsymbol{x}, \tilde{\boldsymbol{z}})$ under the $\mathcal{F}^k$ defined in Equation (4) can be reformulated as

$$\sup_{\mathbb{P}\in\mathcal{F}} \mathbb{E}_{\mathbb{P}}\left[g(\boldsymbol{x}, \tilde{\boldsymbol{z}})\right] = \sup\left\{ \mathbb{E}_{\mathbb{P}}\left[g(\boldsymbol{x}, \tilde{\boldsymbol{z}})\right] \;\middle|\; \begin{array}{l} \mathbb{E}_{\mathbb{P}}[\tilde{\boldsymbol{z}}] = \boldsymbol{\mu}, \\ \mathbb{E}_{\mathbb{P}}[|\tilde{\boldsymbol{z}} - \boldsymbol{\mu}|] \leq \boldsymbol{\delta}, \\ \mathbb{P}(\underline{\boldsymbol{z}} \leq \boldsymbol{z} \leq \bar{\boldsymbol{z}}) = 1 \end{array} \right\} = \sup\left\{ \mathbb{E}_{\mathbb{P}}\left[g(\boldsymbol{x}, \tilde{\boldsymbol{z}})\right] \;\middle|\; \begin{array}{l} \mathbb{E}_{\mathbb{P}}[\tilde{\boldsymbol{z}}] = \boldsymbol{\mu}, \\ \mathbb{E}_{\mathbb{P}}[|\tilde{\boldsymbol{z}} - \boldsymbol{\mu}|] \leq \hat{\boldsymbol{\delta}}, \\ \mathbb{P}(\underline{\boldsymbol{z}} \leq \boldsymbol{z} \leq \bar{\boldsymbol{z}}) = 1 \end{array} \right\}$$
$$= \sup_{\boldsymbol{0}\leq \boldsymbol{d} \leq \hat{\boldsymbol{\delta}}} V(\boldsymbol{d}),$$

where

$$V(\boldsymbol{d}) = \sup\left\{ \mathbb{E}_{\mathbb{P}}\left[g(\boldsymbol{x}, \tilde{\boldsymbol{z}})\right] \;\middle|\; \begin{array}{l} \mathbb{E}_{\mathbb{P}}[\tilde{\boldsymbol{z}}] = \boldsymbol{\mu}, \\ \mathbb{E}_{\mathbb{P}}[|\tilde{\boldsymbol{z}} - \boldsymbol{\mu}|] = \boldsymbol{d}, \\ \mathbb{P}(\underline{\boldsymbol{z}} \leq \boldsymbol{z} \leq \bar{\boldsymbol{z}}) = 1 \end{array} \right\}.$$

We prove our proposition by two steps.

Step 1. Considering any given $\boldsymbol{d} \in [\boldsymbol{0}, \hat{\boldsymbol{\delta}}]$, we will show that there must exist an optimal probability distribution, $\mathbb{P}^*$, for the problem in defining $V(\boldsymbol{d})$ such that the marginal distribution is as follows,

$$\mathbb{P}^*\left(\tilde{z}_i = w\right) = \begin{cases} \frac{d_i}{2(\mu_i - \underline{z}_i)} & \text{if } w = \underline{z}_i \\ 1 - \frac{d_i(\overline{z}_i - \underline{z}_i)}{2(\overline{z}_i - \mu_i)(\mu_i - \underline{z}_i)} & \text{if } w = \mu_i \\ \frac{d_i}{2(\overline{z}_i - \mu_i)} & \text{if } w = \overline{z}_i \\ 0 & \text{otherwise.} \end{cases} \tag{33}$$

We prove this by discussing two scenarios, depending on whether $V(\boldsymbol{d})$ is finite or not.

Consider the first case where $V(\boldsymbol{d}) = \infty$. In this case, there exists $\mathbb{P}'$ such that $\mathbb{E}_{\mathbb{P}'}[g(\boldsymbol{x}', \tilde{\boldsymbol{z}})] = \infty, \mathbb{E}_{\mathbb{P}'}[\tilde{\boldsymbol{z}}] = \boldsymbol{\mu}, \mathbb{E}_{\mathbb{P}'}[|\tilde{\boldsymbol{z}} - \boldsymbol{\mu}|] = \boldsymbol{d}$ and $\mathbb{P}'(\underline{\boldsymbol{z}} \leq \boldsymbol{z} \leq \bar{\boldsymbol{z}}) = 1$. We denote by supp($\mathbb{P}$) the support of any probability distribution $\mathbb{P}$. Observing that the feasible set of $V(\boldsymbol{d})$ is nonempty (any distribution with marginal distribution as in (33) is feasible), $\mathbb{E}_{\mathbb{P}'}[g(\boldsymbol{x}', \tilde{\boldsymbol{z}})] = \infty$ implies that there must exist $\boldsymbol{z}' \in \text{supp}(\mathbb{P}') \subseteq [\underline{\boldsymbol{z}}, \overline{\boldsymbol{z}}]$ such that $g(\boldsymbol{x}, \boldsymbol{z}') = \infty$. We let $\mathbb{P}''$ be any probability distribution with marginal distribution as defined in (33), then $\text{supp}(\mathbb{P}'') = \prod_{i\in[n]} S_i$ where $S_i = \{\mu_i\}$ if $d_i = 0$, $S_i = \{\underline{z}_i, \mu_i, \overline{z}_i\}$ if $d_i \in (0, \hat{\delta}_i)$ and $S_i = \{\underline{z}_i, \overline{z}_i\}$ if $d_i = \hat{\delta}_i$, $i \in [n]$. Now, consider any $i \in [n]$. If $d_i = 0$, we must have $\mathbb{P}'(\tilde{z}_i = \mu_i) = 1$; hence, $z_i' = \mu_i \in \text{conv}(S_i)$; if $d_i > 0$, $z_i' \in \text{conv}(S_i) = [\underline{z}_i, \overline{z}_i]$ since $\boldsymbol{z} \in [\underline{\boldsymbol{z}}, \overline{\boldsymbol{z}}]$. Hence, in any case, $z_i' \in \text{conv}(S_i)$. Consequently, we have $\boldsymbol{z}' \in \text{conv}(\text{supp}(\mathbb{P}''))$. Since function $g(\boldsymbol{x}, \boldsymbol{z})$ is convex in $\boldsymbol{z}$ (see Theorem 2, Section 3.1 in Birge and Louveaux (2011)), there must exist $\boldsymbol{z}'' \in \text{supp}(\mathbb{P}'')$ such that $g(\boldsymbol{x}', \boldsymbol{z}'') = \infty$. Hence, $\mathbb{P}''$ is also a worst-case distribution.

For the second case where $V(\boldsymbol{d})$ is finite, by strong duality (e.g., Shapiro 2001),

$$V(\boldsymbol{d}) = \min\left\{s + \boldsymbol{\mu}^\top \boldsymbol{t} + \boldsymbol{d}^\top \boldsymbol{r} \;\middle|\; s + \boldsymbol{z}^\top \boldsymbol{t} + (|\boldsymbol{z} - \boldsymbol{\mu}|)^\top \boldsymbol{r} \geq g(\boldsymbol{x}, \boldsymbol{z}),\ \forall \underline{\boldsymbol{z}} \leq \boldsymbol{z} \leq \overline{\boldsymbol{z}}\right\}. \tag{34}$$

For any given $\boldsymbol{t}, \boldsymbol{\mu}, \boldsymbol{r}$, since $g(\boldsymbol{x}, \boldsymbol{z})$ is convex in $\boldsymbol{z}$, the function $g(\boldsymbol{x}, \boldsymbol{z}) - \boldsymbol{z}^\top \boldsymbol{t} - (|\boldsymbol{z} - \boldsymbol{\mu}|)^\top \boldsymbol{r}$ is convex in $\boldsymbol{z}$ if $\boldsymbol{z} \in [\boldsymbol{a}, \boldsymbol{b}]$ where for all $i \in [n]$, (a_i, b_i) takes value of $(\underline{z}_i, \mu_i)$ or $(\mu_i, \overline{z}_i)$. Hence, the constraint in (34) is equivalent to

$$s \geq g(\boldsymbol{x}, \boldsymbol{z}) - \boldsymbol{z}^\top \boldsymbol{t} - (|\boldsymbol{z} - \boldsymbol{\mu}|)^\top \boldsymbol{r}, \quad \forall \boldsymbol{z} \in \prod_{i\in[n]} \{\underline{z}_i, \mu_i, \overline{z}_i\}. \tag{35}$$

Substituting the constraints in Problem (34) by (35) and writing its dual form again, we obtain

$$V(\boldsymbol{d}) = \sup\left\{ \sum_{\tau=1}^{3^n} p_\tau g(\boldsymbol{x}, \boldsymbol{z}^\tau) \;\middle|\; \begin{array}{ll} \sum_{\tau=1}^{3^n} p_\tau z_i^\tau = \mu_i, & i \in [n] \\ \sum_{\tau=1}^{3^n} p_\tau |z_i^\tau - \mu_i| = d_i, & i \in [n] \\ \sum_{\tau=1}^{3^n} p_\tau = 1 & \\ p_\tau \geq 0, & \tau \in [3^n] \end{array} \right\}, \tag{36}$$

where $\boldsymbol{z}^1, \ldots, \boldsymbol{z}^{3^n}$ represent all $\boldsymbol{z} \in \prod_{i\in[n]} \{\underline{z}_i, \mu_i, \overline{z}_i\}$, and $p_1, \ldots, p_{3^n}$ are the associated decision variables. Therefore, we can find a distribution $\mathbb{P}^*$ which is optimal to $V(\boldsymbol{d})$, with its support being $\boldsymbol{z}^1,\ldots,\boldsymbol{z}^{3^n}$. This implies $\mathbb{P}^*(\tilde{z}_i=w)=0$ whenever $w\notin\{\underline{z}_i,\mu_i,\overline{z}_i\}$, for all $i\in[n]$. Given any three-point support $\{\underline{z}_i,\mu_i,\overline{z}_i\}$, mean μ_i and MAD value $d_i\in[0,\hat{\delta}_i]$, we observe that a distribution which places $\frac{d_i}{2(\mu_i-\underline{z}_i)}$ amount of mass at $\underline{z}_i$, $1-\frac{d_i(\overline{z}_i-\underline{z}_i)}{2(\overline{z}_i-\mu_i)(\mu_i-\underline{z}_i)}$ at μ_i, and $\frac{d_i}{2(\overline{z}_i-\mu_i)}$ at $\overline{z}_i$, is uniquely determined. Hence, $\mathbb{P}^*$ must have a marginal distribution as in (33).

Step 2. We next show that function $V(\boldsymbol{d})$ is non-decreasing in $\boldsymbol{d}$.

Consider any $\boldsymbol{0}\le\boldsymbol{d}'\le\boldsymbol{d}''\le\hat{\boldsymbol{\delta}}$ with $\boldsymbol{d}''-\boldsymbol{d}'=\theta\boldsymbol{e}_{i^\circ}$ for some $\theta>0, i^\circ\in[n]$, and the probability distribution $\mathbb{P}'$ with $\mathbb{P}'(\tilde{\boldsymbol{z}}=\boldsymbol{z}^\tau)=p'_\tau,\tau\in[3^n]$ such that $p'_1,\ldots,p'_{3^n}$ is the worst-case distribution in Problem (36) when $\boldsymbol{d}=\boldsymbol{d}'$. WLOG, we let $i^\circ=1$. We define another distribution $\mathbb{P}''$ with $\mathbb{P}''(\tilde{\boldsymbol{z}}=\boldsymbol{z}^\tau)=p''_\tau,\tau\in[3^n]$ as

$$\mathbb{P}''(\tilde{\boldsymbol{z}}=\boldsymbol{z})=\begin{cases}\mathbb{P}'(\tilde{\boldsymbol{z}}=\boldsymbol{z})+\frac{\epsilon(\boldsymbol{z})}{\mu_1-\underline{z}_1} & \text{if } z_1=\underline{z}_1,\\ \mathbb{P}'(\tilde{\boldsymbol{z}}=\boldsymbol{z})-\frac{\epsilon(\boldsymbol{z})}{\mu_1-\underline{z}_1}-\frac{\epsilon(\boldsymbol{z})}{\overline{z}_1-\mu_1} & \text{if } z_1=\mu_1,\\ \mathbb{P}'(\tilde{\boldsymbol{z}}=\boldsymbol{z})+\frac{\epsilon(\boldsymbol{z})}{\overline{z}_1-\mu_1} & \text{if } z_1=\overline{z}_1,\end{cases}$$

where $\epsilon:\prod_{i\in[n]}\{\underline{z}_i,\mu_i,\overline{z}_i\}\to[0,1]$ is a mapping defined by

$$\epsilon(\boldsymbol{z})=\epsilon(\boldsymbol{z}+(\underline{z}_1-\mu_1)\boldsymbol{e}_1)=\epsilon(\boldsymbol{z}+(\overline{z}_1-\mu_1)\boldsymbol{e}_1)=\frac{\theta/2}{1-\frac{d'_1}{2}\left(\frac{1}{\mu_1-\underline{z}_1}+\frac{1}{\overline{z}_1-\mu_1}\right)}\mathbb{P}'(\tilde{\boldsymbol{z}}=\boldsymbol{z}) \tag{37}$$

for all $\boldsymbol{z}\in\prod_{i\in[n]}\{\underline{z}_i,\mu_i,\overline{z}_i\}$ with $z_1=\mu_1$. We next verify that $p''_1,\ldots,p''_{3^n}$ satisfy the constraints in (36) when replacing $\boldsymbol{d}$ by $\boldsymbol{d}''$.

From the definition of θ, we observe that for all $\boldsymbol{z}$ such that $z_1=\mu_1$,

$$\frac{\epsilon(\boldsymbol{z})}{\mu_1-\underline{z}_1}+\frac{\epsilon(\boldsymbol{z})}{\overline{z}_1-\mu_1}=\frac{\frac{d''_1-d'_1}{2}\left(\frac{1}{\mu_1-\underline{z}_1}+\frac{1}{\overline{z}_1-\mu_1}\right)}{1-\frac{d'_1}{2}\left(\frac{1}{\mu_1-\underline{z}_1}+\frac{1}{\overline{z}_1-\mu_1}\right)}\mathbb{P}'(\tilde{\boldsymbol{z}}=\boldsymbol{z})=\left(1-\frac{1-\frac{d''_1}{2}\left(\frac{1}{\mu_1-\underline{z}_1}+\frac{1}{\overline{z}_1-\mu_1}\right)}{1-\frac{d'_1}{2}\left(\frac{1}{\mu_1-\underline{z}_1}+\frac{1}{\overline{z}_1-\mu_1}\right)}\right)\mathbb{P}'(\tilde{\boldsymbol{z}}=\boldsymbol{z}).$$

Because $0<d'_1<d''_1\le\hat{\delta}_1$ and the three-point distribution is uniquely determined, we have $1-\frac{d'_1}{2}\left(\frac{1}{\mu_1-\underline{z}_1}+\frac{1}{\overline{z}_1-\mu_1}\right)\ge1-\frac{d''_1}{2}\left(\frac{1}{\mu_1-\underline{z}_1}+\frac{1}{\overline{z}_1-\mu_1}\right)\ge0$. Hence $\mathbb{P}''(\tilde{\boldsymbol{z}}=\boldsymbol{z})\in[0,\mathbb{P}'(\tilde{\boldsymbol{z}}=\boldsymbol{z})]$ for all $\boldsymbol{z}$ with $z_1=\mu_1$. By the definition of $\mathbb{P}''$ we notice that $\sum_{\boldsymbol{z}\in\prod_{i\in[n]}\{\underline{z}_i,\mu_i,\overline{z}_i\}}\mathbb{P}''(\tilde{\boldsymbol{z}}=\boldsymbol{z})=\sum_{\boldsymbol{z}\in\prod_{i\in[n]}\{\underline{z}_i,\mu_i,\overline{z}_i\}}\mathbb{P}'(\tilde{\boldsymbol{z}}=\boldsymbol{z})=1$, we have $\mathbb{P}''(\tilde{\boldsymbol{z}}=\boldsymbol{z})\in[0,1]$ for all $\boldsymbol{z}\in\prod_{i\in[n]}\{\underline{z}_i,\mu_i,\overline{z}_i\}$.

To see $\mathbb{E}_{\mathbb{P}''}[\tilde{\boldsymbol{z}}]=\boldsymbol{\mu}$, we observe that for the first dimension,

$$\begin{aligned}\sum_{\tau=1}^{3^n}p''_\tau z_1^\tau&=\sum_{\boldsymbol{z}:z_1=\mu_1}\mathbb{P}''(\tilde{\boldsymbol{z}}=\boldsymbol{z})\mu_1+\sum_{\boldsymbol{z}:z_1=\underline{z}_1}\mathbb{P}''(\tilde{\boldsymbol{z}}=\boldsymbol{z})\underline{z}_1+\sum_{\boldsymbol{z}:z_1=\overline{z}_1}\mathbb{P}''(\tilde{\boldsymbol{z}}=\boldsymbol{z})\overline{z}_1\\&=\sum_{\boldsymbol{z}:z_1=\mu_1}\left(\mathbb{P}''(\tilde{\boldsymbol{z}}=\boldsymbol{z})\mu_1+\mathbb{P}''(\tilde{\boldsymbol{z}}=\boldsymbol{z}+(\underline{z}_1-\mu_1)\boldsymbol{e}_1)\underline{z}_1+\mathbb{P}''(\tilde{\boldsymbol{z}}=\boldsymbol{z}+(\overline{z}_1-\mu_1)\boldsymbol{e}_1)\overline{z}_1\right)\\&=\sum_{\tau=1}^{3^n}p'_\tau z_1^\tau+\sum_{\boldsymbol{z}:z_1=\mu_1}\left(\left(-\frac{\epsilon(\boldsymbol{z})}{\mu_1-\underline{z}_1}-\frac{\epsilon(\boldsymbol{z})}{\overline{z}_1-\mu_1}\right)\mu_1+\frac{\epsilon(\boldsymbol{z})}{\mu_1-\underline{z}_1}\underline{z}_1+\frac{\epsilon(\boldsymbol{z})}{\overline{z}_1-\mu_1}\overline{z}_1\right)\\&=\mu_1+\sum_{\boldsymbol{z}:z_1=\mu_1}\left(\frac{\epsilon(\boldsymbol{z})}{\mu_1-\underline{z}_1}(\underline{z}_1-\mu_1)+\frac{\epsilon(\boldsymbol{z})}{\overline{z}_1-\mu_1}(\overline{z}_1-\mu_1)\right)\\&=\mu_1,\end{aligned}$$

where the third equality is due to the property of ϵ stated in (37). For any other dimension i with $i \neq 1$, by the construction of $\mathbb{P}''$ we can observe that the marginal masses on the i-th dimension remain identical with $\mathbb{P}'$. Hence, $\sum_{\tau=1}^{3^n} p''_\tau \boldsymbol{z}^\tau = \boldsymbol{\mu}$.

For the MAD information, we start from the first dimension and notice that

$$
\begin{aligned}
\sum_{\boldsymbol{z}:z_1=\mu_1} \mathbb{P}''(\tilde{\boldsymbol{z}}=\boldsymbol{z}) &= \sum_{\boldsymbol{z}:z_1=\mu_1} \mathbb{P}'(\tilde{\boldsymbol{z}}=\boldsymbol{z}) - \left(\frac{1}{\mu_1-\underline{z}_1}+\frac{1}{\overline{z}_1-\mu_1}\right)\sum_{\boldsymbol{z}:z_1=\mu_1}\epsilon(\boldsymbol{z}) \\
&= \left(1-\left(\frac{1}{\mu_1-\underline{z}_1}+\frac{1}{\overline{z}_1-\mu_1}\right)\frac{\theta/2}{1-\frac{d'_1}{2}\left(\frac{1}{\mu_1-\underline{z}_1}+\frac{1}{\overline{z}_1-\mu_1}\right)}\right)\sum_{\boldsymbol{z}:z_1=\mu_1}\mathbb{P}'(\tilde{\boldsymbol{z}}=\boldsymbol{z}) \\
&= \left(1-\left(\frac{1}{\mu_1-\underline{z}_1}+\frac{1}{\overline{z}_1-\mu_1}\right)\frac{\theta/2}{1-\frac{d'_1}{2}\left(\frac{1}{\mu_1-\underline{z}_1}+\frac{1}{\overline{z}_1-\mu_1}\right)}\right)\left(1-\frac{d'_1}{2}\left(\frac{1}{\mu_1-\underline{z}_1}+\frac{1}{\overline{z}_1-\mu_1}\right)\right) \\
&= 1-\frac{d''_1}{2}\left(\frac{1}{\mu_1-\underline{z}_1}+\frac{1}{\overline{z}_1-\mu_1}\right),
\end{aligned}
$$

where the second equality is due to (37), the third equality holds since the marginal distribution is uniquely determined in $\mathbb{P}'$. Similarly, we have $\sum_{\boldsymbol{z}:z_1=\underline{z}_1}\mathbb{P}''(\tilde{\boldsymbol{z}}=\boldsymbol{z}) = \frac{d''_1}{2(\mu_1-\underline{z}_1)}$ and $\sum_{\boldsymbol{z}:z_1=\overline{z}_1}\mathbb{P}''(\tilde{\boldsymbol{z}}=\boldsymbol{z}) = \frac{d''_1}{2(\overline{z}_1-\mu_1)}$. It then follows that $\sum_{\tau=1}^{3^n} p''_\tau |z_1^\tau-\mu_1| = d''_1$. Since the marginal masses at all remaining dimensions are unchanged, we have $\sum_{\tau=1}^{3^n} p''_\tau |\boldsymbol{z}^\tau-\boldsymbol{\mu}| = \boldsymbol{d}''$. Hence $\mathbb{P}''$ is a feasible solution to the set (36) when we replace $\boldsymbol{d}$ by $\boldsymbol{d}''$.

Consequently,

$$
\begin{aligned}
V(\boldsymbol{d}'') &\geq \sum_{\tau=1}^{3^n} p''_\tau g(\boldsymbol{x},\boldsymbol{z}^\tau) \\
&= \sum_{\tau=1}^{3^n} p'_\tau g(\boldsymbol{x},\boldsymbol{z}^\tau) + \sum_{\boldsymbol{z}:z_1=\mu_1}\left(-\frac{\epsilon(\boldsymbol{z})}{\mu_1-\underline{z}_1}-\frac{\epsilon(\boldsymbol{z})}{\overline{z}_1-\mu_1}\right)g(\boldsymbol{x},\boldsymbol{z}) + \sum_{\boldsymbol{z}:z_1=\underline{z}_1}\frac{\epsilon(\boldsymbol{z})}{\mu_1-\underline{z}_1}g(\boldsymbol{x},\boldsymbol{z}) + \sum_{\boldsymbol{z}:z_1=\overline{z}_1}\frac{\epsilon(\boldsymbol{z})}{\overline{z}_1-\mu_1}g(\boldsymbol{x},\boldsymbol{z}) \\
&= \sum_{\tau=1}^{3^n} p'_\tau g(\boldsymbol{x},\boldsymbol{z}^\tau) + \sum_{\boldsymbol{z}:z_1=\mu_1}\Bigg(\left(-\frac{\epsilon(\boldsymbol{z})}{\mu_1-\underline{z}_1}-\frac{\epsilon(\boldsymbol{z})}{\overline{z}_1-\mu_1}\right)g(\boldsymbol{x},\boldsymbol{z}) \\
&\qquad + \frac{\epsilon(\boldsymbol{z})}{\mu_1-\underline{z}_1}g(\boldsymbol{x},\boldsymbol{z}+(\underline{z}_1-\mu_1)\boldsymbol{e}_1) + \frac{\epsilon(\boldsymbol{z})}{\overline{z}_1-\mu_1}g(\boldsymbol{x},\boldsymbol{z}+(\overline{z}_1-\mu_1)\boldsymbol{e}_1)\Bigg) \\
&= \sum_{\tau=1}^{3^n} p'_\tau g(\boldsymbol{x},\boldsymbol{z}^\tau) + \sum_{\boldsymbol{z}:z_1=\mu_1}\epsilon(\boldsymbol{z})\Bigg(\frac{1}{\mu_1-\underline{z}_1}\left(g(\boldsymbol{x},\boldsymbol{z}+(\underline{z}_1-\mu_1)\boldsymbol{e}_1)-g(\boldsymbol{x},\boldsymbol{z})\right) \\
&\qquad + \frac{1}{\overline{z}_1-\mu_1}\left(g(\boldsymbol{x},\boldsymbol{z}+(\overline{z}_1-\mu_1)\boldsymbol{e}_1)-g(\boldsymbol{x},\boldsymbol{z})\right)\Bigg) \\
&\geq \sum_{\tau=1}^{3^n} p'_\tau g(\boldsymbol{x},\boldsymbol{z}^\tau) = V(\boldsymbol{d}'),
\end{aligned}
$$

where the first inequality holds because $p''_1,\ldots,p''_{3^n}$ is a feasible solution, the second equality is based on (37), and the second inequality follows from the convexity of g. Hence, $V(\boldsymbol{d})$ is non-decreasing on $[\mathbf{0}, \hat{\boldsymbol{\delta}}]$. Therefore, $\sup_{\mathbb{P}\in\mathcal{F}} \mathbb{E}_{\mathbb{P}}[g(\boldsymbol{x}, \tilde{\boldsymbol{z}})] = \sup_{\mathbf{0}\le \boldsymbol{d}\le\hat{\boldsymbol{\delta}}} V(\boldsymbol{d}) = V(\hat{\boldsymbol{\delta}})$. The worst-case is with the form of (33) when $\boldsymbol{d} = \hat{\boldsymbol{\delta}}$, which is as proposed in our Proposition. □

C.2. Proof of Proposition 2

1) $\Longrightarrow$ 2). Consider any $\mathbb{P}\in\mathcal{P}$ such that there exists an unordered pair $\boldsymbol{w}', \boldsymbol{w}''$ with $p' = \mathbb{P}(\tilde{\boldsymbol{w}} = \boldsymbol{w}') > 0$, $p'' = \mathbb{P}(\tilde{\boldsymbol{w}} = \boldsymbol{w}'') > 0$. WLOG, assume $p' \le p''$. We construct a new probability distribution $\mathbb{P}^o$, such that

$$\mathbb{P}^o(\tilde{\boldsymbol{w}} = \boldsymbol{w}) = \begin{cases} 0 & \text{if } \boldsymbol{w} = \boldsymbol{w}' \\ p'' - p' & \text{if } \boldsymbol{w} = \boldsymbol{w}'' \\ \mathbb{P}(\tilde{\boldsymbol{w}} = \boldsymbol{w}' \wedge \boldsymbol{w}'') + p' & \text{if } \boldsymbol{w} = \boldsymbol{w}' \wedge \boldsymbol{w}'' \\ \mathbb{P}(\tilde{\boldsymbol{w}} = \boldsymbol{w}' \vee \boldsymbol{w}'') + p' & \text{if } \boldsymbol{w} = \boldsymbol{w}' \vee \boldsymbol{w}'' \\ \mathbb{P}(\tilde{\boldsymbol{w}} = \boldsymbol{w}) & \text{otherwise.} \end{cases}$$

In particular, based on $\mathbb{P}$, we move the probability mass p' from the realization of $\boldsymbol{w}', \boldsymbol{w}''$ to $\boldsymbol{w}' \wedge \boldsymbol{w}'', \boldsymbol{w}' \vee \boldsymbol{w}''$. That does not change the marginal distribution and hence $\mathbb{P}^o \in \mathcal{P}$. Moreover, compared with the support of $\mathbb{P}$, that of $\mathbb{P}^o$ has one less unordered pair. We also observe that

$$\mathbb{E}_{\mathbb{P}^o}[f(\tilde{\boldsymbol{w}})] - \mathbb{E}_{\mathbb{P}}[f(\tilde{\boldsymbol{w}})] = p'\left(f(\boldsymbol{w}' \wedge \boldsymbol{w}'') + f(\boldsymbol{w}' \vee \boldsymbol{w}'') - f(\boldsymbol{w}') - f(\boldsymbol{w}'')\right) \ge 0,$$

where the last inequality is due to the supermodularity of f. Therefore, we can always reduce the number of unordered pairs (if there is any) in the support while the value of expectation on $f(\tilde{\boldsymbol{w}})$ either increases or remains unchanged. Since any $\mathbb{P}\in\mathcal{P}$ has nonzero probability mass only at a finite number of discrete points (by the definition of $\mathcal{P}$), the number of unordered pairs must be finite and hence will be decreased to zero after a finite number of such steps. Therefore, finally we obtain $\mathbb{P}^* \in \arg\sup_{\mathbb{P}\in\mathcal{P}} \mathbb{E}_{\mathbb{P}}[f(\tilde{\boldsymbol{w}})]$ such that the support of $\mathbb{P}^*$ has no unordered pair. Since there are m_i points in the support of the i-th marginal, moving along the chain in ascending order from $(x_{11}, \ldots, x_{n1})$ to $(x_{1m_1}, \ldots, x_{nm_n})$ takes $m_i - 1$ steps on the i-th dimension. Hence, the chain has its maximum length being 1+the total number of steps, i.e., $\sum_{i\in[n]}(m_i - 1) + 1$.

2) $\Longrightarrow$ 1). Assuming the contrary of 1), i.e., f is not supermodular, then there exists a pair of unordered $\boldsymbol{w}', \boldsymbol{w}'' \in \Re^n$ such that $f(\boldsymbol{w}') + f(\boldsymbol{w}'') > f(\boldsymbol{w}' \wedge \boldsymbol{w}'') + f(\boldsymbol{w}' \vee \boldsymbol{w}'')$. We denote $\mathcal{I}' = \{i \in [n] \mid w_i' < w_i''\}$, $\mathcal{I}'' = \{i \in [n] \mid w_i' > w_i''\}$ and $\mathcal{I}_e = \{i \in [n] \mid w_i' = w_i''\}$. As $\boldsymbol{w}', \boldsymbol{w}''$ are unordered, we know $\mathcal{I}'$, $\mathcal{I}''$ are both nonempty. For all $i \in \mathcal{I}' \cup \mathcal{I}''$, we let $m_i = 2$, $x_{i1} = w_i' \wedge w_i''$, $x_{i2} = w_i' \vee w_i''$, $p_{i1} = p_{i2} = 0.5$; for all $i \in \mathcal{I}_e$, we let $m_i = 1$, $x_{i1} = w_i' = w_i''$ and $p_{i1} = 1$. Correspondingly, $\mathcal{P} = \{\mathbb{P} \mid \mathbb{P}(\tilde{w}_i = x_{ij}) = p_{ij}, j \in [m_i], i \in [n]\}$, and any $\mathbb{P}\in\mathcal{P}$ must has its support in $\mathcal{W} = \Pi_{i\in\mathcal{I}'\cup\mathcal{I}''}\{x_{i1}, x_{x2}\} \times \Pi_{i\in\mathcal{I}_e}\{x_{i1}\}$. Consider any $\mathbb{P}^o \in \mathcal{P}$ such that its support $\mathcal{W}_{\mathbb{P}^o} = \{\boldsymbol{w} \in \Re^n \mid \mathbb{P}^o(\tilde{\boldsymbol{w}} = \boldsymbol{w}) > 0\}$ forms a chain. We now show that $\mathbb{P}^o \notin \arg\sup_{\mathbb{P}\in\mathcal{P}} \mathbb{E}_{\mathbb{P}}[f(\tilde{\boldsymbol{w}})]$ and hence statement 2) in the proposition is false, then the proof can be completed. To this end, we notice since $\mathcal{W}_{\mathbb{P}^o}$ forms a chain, we can label the elements in $\mathcal{W}_{\mathbb{P}^o}$ in ascending order, i.e., $\boldsymbol{w}^1 \le \boldsymbol{w}^2 \le \ldots$.

We first show $\boldsymbol{w}^1 = \boldsymbol{w}' \wedge \boldsymbol{w}''$. Consider any $i \in [n]$. If $w_i^1 < x_{i1}$, then $w_i^1 \notin \{x_{ij} \mid j \in [m_i]\}$, contradicts with $\mathbb{P}^o \in \mathcal{P}$. If $w_i^1 > x_{i1}$, then $w_i > x_{i1}$ for all $\boldsymbol{w} \in \mathcal{W}_{\mathbb{P}^o}$, $\mathbb{P}^o(\tilde{w}_i = x_{i1}) = 0$, which also contradicts with $\mathbb{P}^o \in \mathcal{P}$. Therefore, we must have $w_i^1 = x_{i1}$ for all $i \in [n]$, i.e., $\boldsymbol{w}^1 = (x_{11}, \ldots, x_{n1}) = \boldsymbol{w}' \wedge \boldsymbol{w}''$.

We next show $\boldsymbol{w}^2 = \boldsymbol{w}' \vee \boldsymbol{w}''$. Assume that there exists $i \in \mathcal{I}' \cup \mathcal{I}''$ with $w_i^2 = w_i^1$. Since $\boldsymbol{w}^2 \geq \boldsymbol{w}^1$ and $\boldsymbol{w}^2 \neq \boldsymbol{w}^1$, we know that there exists $j \in \mathcal{I}' \cup \mathcal{I}''$ with $w_j^2 > w_j^1$. By $w_i^2 = w_i^1 = x_{i1}$, $\mathbb{P}^o(\tilde{\boldsymbol{w}} = \boldsymbol{w}^1) + \mathbb{P}^o(\tilde{\boldsymbol{w}} = \boldsymbol{w}^2) \leq \mathbb{P}^o(\tilde{w}_i = x_{i1}) = p_{i1} = 0.5$, we know $\mathbb{P}^o(\tilde{\boldsymbol{w}} = \boldsymbol{w}^1) < 0.5$; by $w_j^2 > w_j^1 = x_{j1}$, we know $\mathbb{P}^o(\tilde{\boldsymbol{w}} = \boldsymbol{w}^1) = p_{j1} = 0.5$. Hence, we have contradiction. It follows that $w_i^2 > w_i^1$ for all $i \in \mathcal{I}' \cup \mathcal{I}''$, i.e., $\boldsymbol{w}^2 = \boldsymbol{w}' \vee \boldsymbol{w}''$. Moreover, we have $|\mathcal{W}| = 2$ since $\boldsymbol{w}^2$ is the maximum element of $\mathcal{W}$.

Therefore, $\mathbb{P}^o$ is such that $\mathbb{P}^o(\tilde{\boldsymbol{w}} = \boldsymbol{w}' \wedge \boldsymbol{w}'') = \mathbb{P}^o(\tilde{\boldsymbol{w}} = \boldsymbol{w}' \vee \boldsymbol{w}'') = 0.5$. Consider another distribution $\mathbb{P}^*$ such that $\mathbb{P}^*(\tilde{\boldsymbol{w}} = \boldsymbol{w}') = \mathbb{P}^*(\tilde{\boldsymbol{w}} = \boldsymbol{w}'') = 0.5$. We can easily have $\mathbb{P}^* \in \mathcal{P}$, and

$$\mathbb{E}_{\mathbb{P}^o}[f(\tilde{\boldsymbol{w}})] = 0.5 \times (f(\boldsymbol{w}' \wedge \boldsymbol{w}'') + f(\boldsymbol{w}' \vee \boldsymbol{w}'')) < 0.5 \times (f(\boldsymbol{w}') + f(\boldsymbol{w}'')) = \mathbb{E}_{\mathbb{P}^*}[f(\tilde{\boldsymbol{w}})].$$

□

C.3. Proof of Proposition 3

For notational simplicity, we drop the superscript k which represents the scenario k; we also assume $\underline{z}_i = -1, \mu_i = 0, \bar{z}_i = 1$ for all $i \in [n]$, since the general case can be proved in the same way.

During the progress of this algorithm, for each $j \in [2n+1]$, we define $\mathbf{mp}^{j,i}$, which stands for the remaining marginal probability for iteration j at dimension i, as

$$\mathbf{mp}^{j,i} = \begin{cases} q_i^j & \text{if } z_i^j = 1 \quad (\mathbf{mp}^{j,i} \in \Re \text{ in this case}) \\ (q_i^j,\ \mathbb{P}^*(\tilde{z}_i = 1)) & \text{if } z_i^j = 0 \quad (\mathbf{mp}^{j,i} \in \Re^2 \text{ in this case}) \\ (q_i^j,\ \mathbb{P}^*(\tilde{z}_i = 0),\ \mathbb{P}^*(\tilde{z}_i = 1)) & \text{if } z_i^j = -1 \ (\mathbf{mp}^{j,i} \in \Re^3 \text{ in this case}) \end{cases}.$$

We also define $c_j = \mathbf{1}^\top \mathbf{mp}^{j,1}$ which represents the remaining total probability mass. Correspondingly, we denote the set of information $\mathcal{I}^j = \{\boldsymbol{z}^j, \mathbf{mp}^{j,1}, \ldots, \mathbf{mp}^{j,n}, c_j\}$.

Given a set of information $\mathcal{I}^j$, we say it is valid if it satisfies the following four conditions: 1) $\boldsymbol{z}^j \in \{-1,0,1\}^n$; 2) $\mathbf{mp}^{j,i} \in [0,1]^{2-z_i^j}$ for all $i \in [n]$; 3) $\mathrm{mp}_{\mathrm{end}}^{j,i} > 0$ for all $i \in [n]$, where we denote $\mathrm{mp}_{\mathrm{end}}^{j,i}$ as the last element of the vector $\mathbf{mp}^{j,i}$; and 4) $\mathbf{1}^\top \mathbf{mp}^{j,i} = c_j$ for all $i \in [n]$.

By induction, we now show that $\mathcal{I}^j$ is valid for all $j \in [2n+1]$.

First, when $j = 1$, the conditions 1), 2) and 3) are obviously satisfied. The condition 4) is also satisfied since $\mathbf{1}^\top \mathbf{mp}^{1,i} = \mathbb{P}^*(\tilde{z}_i = -1) + \mathbb{P}^*(\tilde{z}_i = 0) + \mathbb{P}^*(\tilde{z}_i = 1) = 1$ for all $i \in [n]$, and $c_1 = 1$.

Suppose $\mathcal{I}^j$ is valid for some $j \in [2n]$. Based on the algorithm, the elements in $\mathcal{I}^{j+1}$ are obtained as follows. First, $p_j = \min\{\mathrm{mp}_1^{j,1}, \ldots, \mathrm{mp}_1^{j,n}\}$, $r_j = \min\{i \in [n] \mid \mathrm{mp}_1^{j,i} = p_j\}$. After that, $\boldsymbol{z}^{j+1} = \boldsymbol{z}^j + \boldsymbol{e}_{r_j}$. We now prove that $z_{r_j}^j \neq 1$ by contradiction. Assume to the contrary, i.e., $z_{r_j}^j = 1$, then $\mathbf{mp}^{j,r_j} \in \Re$, we have $c_j = \mathbf{1}^\top \mathbf{mp}^{j,r_j} = \mathrm{mp}_1^{j,r_j} = p_j$. For any $i \in [n] \setminus \{r_j\}$, we observe i) $\mathrm{mp}_1^{j,i} \geq p_j = c_j$ (the inequality is because of our choice of p_j); ii) $\mathrm{mp}_{\mathrm{end}}^{j,i} > 0$; and iii) $\mathbf{1}^\top \mathbf{mp}^{j,i} = c_j$ and $\mathbf{mp}^{j,i} \geq \mathbf{0}$.

The last two observations are because $\mathcal{I}^j$ is valid and hence satisfies conditions 2), 3) and 4). Hence, we have $\mathbf{mp}^{j,i} \in \Re$ and then $z_i^j = 1$. That implies $\boldsymbol{z}^j = \mathbf{1}$. We notice that for any $t \in [j-1]$, $\boldsymbol{z}^{t+1} = \boldsymbol{z}^t + \boldsymbol{e}_i$ for some $i \in [n]$. So moving from $\boldsymbol{z}^1 = -\mathbf{1}$ to $\boldsymbol{z}^j = \mathbf{1}$ requires $2n$ steps, i.e., $j = 2n+1$, which contradicts $j \in [2n]$. Hence, $z_{r_j}^j = 1$ is false, and we must have $z_{r_j}^j \in \{-1, 0\}$. We can conclude that $\boldsymbol{z}^{j+1} = \boldsymbol{z}^j + \boldsymbol{e}_{r_j} \in \{-1,0,1\}^n$, the condition 1) is satisfied for $\mathcal{I}^{j+1}$. As a result, condition 2) is obviously satisfied by the way $\mathbf{mp}^{j,i}$ is calculated.

With the algorithm, we know $\mathbf{mp}^{j+1,r_j}$ can be obtained from the vector of $\mathbf{mp}^{j,r_j}$ by removing the first component. Therefore, $\mathrm{mp}_{\mathrm{end}}^{j+1,r_j} = \mathrm{mp}_{\mathrm{end}}^{j,r_j} > 0$, the condition 3) is satisfied when $i = r_j$. Moreover, $\mathbf{1}^\top \mathbf{mp}^{j+1,r_j} = \mathbf{1}^\top \mathbf{mp}^{j,r_j} - \mathrm{mp}_1^{j,r_j} = c_j - p_j$. We also observe $c_j - p_j = \mathbf{1}^\top \mathbf{mp}^{j+1,r_j} \geq \mathrm{mp}_{\mathrm{end}}^{j+1,r_j} > 0$ and hence $c_j > p_j$.

For any $i \in [n] \setminus \{r_j\}$, since $z_i^{j+1} = z_i^j$, $\mathbf{mp}^{j+1,i}$ and $\mathbf{mp}^{j,i}$ are both of dimension $(2 - z^{j+1,i})$, they differ only at the first dimension; in particular,

$$\mathrm{mp}_s^{j+1,i} = \begin{cases} \mathrm{mp}_1^{j,i} - p_j & \text{if } s = 1 \\ \mathrm{mp}_s^{j,i} & \text{if } z_i^{j+1} \in \{-1, 0\} \text{ and } s \neq 1 \end{cases}. \tag{38}$$

We note that if $z_i^{j+1} = z_i^j = 1$, then $\mathbf{mp}^{j,i}, \mathbf{mp}^{j+1,i} \in \Re_+$, and $\mathrm{mp}_1^{j,i} = \mathbf{1}^\top \mathbf{mp}^{j,i} = c_j > p_j$, $\mathrm{mp}_{\mathrm{end}}^{j+1,i} = \mathrm{mp}_1^{j+1,i} = \mathrm{mp}_1^{j,i} - p_j > 0$. If $z_i^{j+1} = z_i^j \in \{-1, 0\}$, obviously $\mathrm{mp}_{\mathrm{end}}^{j+1,i} = \mathrm{mp}_{\mathrm{end}}^{j,i} > 0$. Therefore, condition 3) is satisfied for i. Moreover, by Equation (38) we also know $\mathbf{1}^\top \mathbf{mp}^{j+1,i} = \mathbf{1}^\top \mathbf{mp}^{j,i} - p_j = c_j - p_j$. Since we have previously obtained $\mathbf{1}^\top \mathbf{mp}^{j+1,r_j} = c_j - p_j$, condition 4) is also satisfied. We conclude $\mathcal{I}^{j+1}$ is also valid and it finishes the induction, i.e., $\mathcal{I}^j$ is valid for all $j \in [2n+1]$.

Now, for any $j \in [2n+1]$, we define $\mathcal{Q}^j$ as the set of all mass functions with the marginal mass given by $\mathbf{mp}^{j,1}, \ldots, \mathbf{mp}^{j,n}$ and the possible realizations forming a chain. More specifically, define $\boldsymbol{w}^{j,i} \in \{-1,0,1\}^{2-z_i^j}$ by

$$\boldsymbol{w}^{j,i} = \begin{cases} (-1,0,1) & \text{if } z_i^j = -1 \\ (0,1) & \text{if } z_i^j = 0 \\ 1 & \text{if } z_i^j = 1 \end{cases},$$

which is the vector of all possible realizations at dimension i, $i \in [n]$, and $\mathcal{W}^j = \{\boldsymbol{z} \mid \boldsymbol{z}^j \leq \boldsymbol{z} \leq \mathbf{1}\} \cap \{-1,0,1\}^n$ which is the set of all possible realizations of vector $\boldsymbol{z}$; then

$$\mathcal{Q}^j = \left\{ f^j : \mathcal{W}^j \to [0,1] \;\middle|\; \begin{array}{l} \sum_{\boldsymbol{z} \in \mathcal{W}^j : z_i = w_s^{j,i}} f^j(\boldsymbol{z}) = \mathrm{mp}_s^{j,i}, \quad i \in [n], s \in [2 - z_i^j] \\ \{\boldsymbol{z} \mid f^j(\boldsymbol{z}) > 0\} \text{ forms a chain} \end{array} \right\}.$$

Noticing that $\mathcal{W}^{j+1} = \left\{ \boldsymbol{z} \in \mathcal{W}^j \;\middle|\; z_{r_j} \neq z_{r_j}^j \right\}$, we define another set $\hat{\mathcal{Q}}^j$ by

$$\hat{\mathcal{Q}}^j = \left\{ f^j : \mathcal{W}^j \to [0,1] \;\middle|\; \begin{array}{ll} f^j(\boldsymbol{z}^j) = p_j & \\ f^j(\boldsymbol{z}) = 0, & \forall \boldsymbol{z} \in \mathcal{W}^j \text{ such that } z_{r_j} = z_{r_j}^j, \boldsymbol{z} \neq \boldsymbol{z}^j \\ f^j(\boldsymbol{z}) = f^{j+1}(\boldsymbol{z}), & \forall \boldsymbol{z} \in \mathcal{W}^{j+1} \\ f^{j+1} \in \mathcal{Q}^{j+1} & \end{array} \right\}.$$

We next prove $\mathcal{Q}^j = \hat{\mathcal{Q}}^j$.

First, consider any $f^j \in \mathcal{Q}^j$. Suppose there exist $\boldsymbol{z}^o \in \mathcal{W}^j$ with $z^o_{r_j} = z^j_{r_j}$ and $\boldsymbol{z}^o \neq \boldsymbol{z}^j$ such that $f^j(\boldsymbol{z}^o) > 0$. That implies the existence of $s \in [n] \setminus \{r_j\}$ such that $z^o_s > z^j_s$. Hence,

$$\sum_{\substack{\boldsymbol{z}\in\mathcal{W}^j: z_{r_j}=z^j_{r_j}\\ z_s=z^j_s}} f^j(\boldsymbol{z}) = \sum_{\boldsymbol{z}\in\mathcal{W}^j: z_{r_j}=z^j_{r_j}} f^j(\boldsymbol{z}) - \sum_{\substack{\boldsymbol{z}\in\mathcal{W}^j: z_{r_j}=z^j_{r_j}\\ z_s\neq z^j_s}} f^j(\boldsymbol{z}) = p_j - \sum_{\substack{\boldsymbol{z}\in\mathcal{W}^j: z_{r_j}=z^j_{r_j}\\ z_s\neq z^j_s}} f^j(\boldsymbol{z}) \leq p_j - f^j(\boldsymbol{z}^o) < p_j,$$

$$\sum_{\substack{\boldsymbol{z}\in\mathcal{W}^j: z_{r_j}>z^j_{r_j}\\ z_s=z^j_s}} f^j(\boldsymbol{z}) = \sum_{\boldsymbol{z}\in\mathcal{W}^j: z_s=z^j_s} f^j(\boldsymbol{z}) - \sum_{\substack{\boldsymbol{z}\in\mathcal{W}^j: z_{r_j}=z^j_{r_j}\\ z_s=z^j_s}} f^j(\boldsymbol{z}) \geq p_j - \sum_{\substack{\boldsymbol{z}\in\mathcal{W}^j: z_{r_j}=z^j_{r_j}\\ z_s=z^j_s}} f^j(\boldsymbol{z}) > 0.$$

Therefore, we have $\boldsymbol{z}^* \in \mathcal{W}^j$ such that $z^*_{r_j} > z^j_{r_j} = z^o_{r_j}$, $z^*_s = z^j_s < z^0_s$ and $f^j(\boldsymbol{z}^*) > 0$, contradicting that $\{\boldsymbol{z} \mid f^j(\boldsymbol{z}) > 0\}$ forms a chain. Therefore, $f^j(\boldsymbol{z}) = 0$ whenever $\boldsymbol{z} \in \mathcal{W}^j$ has $z_{r_j} = z^j_{r_j}, \boldsymbol{z} \neq \boldsymbol{z}^j$, and $f^j(\boldsymbol{z}^j) = \mathrm{mp}_1^{j,r_j} - \sum_{\boldsymbol{z}\in\mathcal{W}^j, z_{r_j}=z^j_{r_j}, \boldsymbol{z}\neq\boldsymbol{z}^j} f^j(\boldsymbol{z}) = p_j - 0 = p_j$. Therefore, f^j satisfies the first two conditions in $\hat{Q}^j$. The corresponding f^{j+1} is in $\mathcal{Q}^{j+1}$ can be easily verified by showing the chain structure and checking the equality constraints on the marginal mass. Hence, we have $f^j \in \hat{\mathcal{Q}}^j$.

We now prove the reverse. Consider any $f^j \in \hat{\mathcal{Q}}^j$ and we check whether it satisfies the two conditions in $\mathcal{Q}^j$. The first condition, which is on the marginal mass, can be verified by standard algebra. The second condition, which is on the chain structure, is straightforward. Therefore, we have $f^j \in \mathcal{Q}^j$. We can conclude that $\mathcal{Q}^j = \hat{\mathcal{Q}}^j$ for all $j \in [2n+1]$.

Finally, by representing $\mathcal{Q}^j$ in the form of $\hat{Q}^j$, with recursion we can easily get

$$\mathcal{Q}^1 = \left\{ f: \mathcal{W}^j \to [0,1] \;\middle|\; \begin{array}{ll} f(\boldsymbol{z}^i) = p_i, & i \in [2n] \\ f(\boldsymbol{z}) = 0, & \forall \boldsymbol{z} \in \mathcal{W}^1 \setminus \{\boldsymbol{z}^i, i \in [2n]\} \setminus \mathcal{W}^{2n+1} \\ f(\boldsymbol{z}) = \hat{f}(\boldsymbol{z}), & \forall \boldsymbol{z} \in \mathcal{W}^{2n+1} \\ \hat{f} \in \mathcal{Q}^{2n+1} & \end{array} \right\} \tag{39}$$

We note that since $\boldsymbol{z}^j \in \{-1,0,1\}^n$, $\boldsymbol{z}^1 = -\mathbf{1}$, and any time the movement from $\boldsymbol{z}^j$ to $\boldsymbol{z}^{j+1}$ is to increase one dimension by 1 while maintaining other dimensions unchanged, and hence we have $\boldsymbol{z}^{2n+1} = \mathbf{1}$. Therefore, $\mathcal{W}^{2n+1} = \{\boldsymbol{z}^{2n+1}\}$. Then by Equation (39), we have

$$\mathcal{Q}^1 = \left\{ f: \mathcal{W}^j \to [0,1] \;\middle|\; \begin{array}{l} f(\boldsymbol{z}^i) = p_i,\ i \in [2n+1] \\ f(\boldsymbol{z}) = 0, \quad \forall \boldsymbol{z} \in \mathcal{W}^1 \setminus \{\boldsymbol{z}^i, i \in [2n+1]\} \end{array} \right\}$$

Hence, the result is proved. □

C.4. Proof of Theorem 1

We define the function $f(\boldsymbol{x})$ as

$$\begin{aligned} f(\boldsymbol{x}) = \min \quad & \boldsymbol{\nu}^\top \boldsymbol{l} \\ \text{s.t.} \quad & \boldsymbol{R}_k^\top \boldsymbol{l} \geq \sum_{i\in[2n+1]} p_i^k \boldsymbol{b}^\top \boldsymbol{y}^{k,i}, && k \in [K] \\ & \boldsymbol{W}\boldsymbol{x} + \boldsymbol{U}\boldsymbol{y}^{k,i} \geq \boldsymbol{V}\boldsymbol{z}^{k,i} + \boldsymbol{v}^0, && k \in [K], i \in [2n+1] \\ & \boldsymbol{l} \geq \mathbf{0}, \end{aligned}$$

then Problem (6) is equivalent with $\min_{\boldsymbol{x}\in\mathcal{X}} f(\boldsymbol{x})$.

We further denote $\mathcal{X}_{fea} = \{\boldsymbol{x}\in\mathcal{X} \mid \sup_{\mathbb{P}\in\mathcal{F}} \mathbb{E}_{\mathbb{P}}[g(\boldsymbol{x},\tilde{\boldsymbol{z}})] < \infty\}$. Recall that we assume Problem (2) has finite optimal value, so $\mathcal{X}_{fea} \neq \emptyset$.

Consider any $\boldsymbol{x}\in\mathcal{X}\setminus\mathcal{X}_{fea}$, we have

$$\infty = \sup_{\mathbb{P}\in\mathcal{F}} \mathbb{E}_{\mathbb{P}}[g(\boldsymbol{x},\tilde{\boldsymbol{z}})] = \max\left\{ \sum_{k\in[K]} q_k \sup_{\mathbb{P}^k\in\mathcal{F}^k} \mathbb{E}_{\mathbb{P}^k}[g(\boldsymbol{x},\tilde{\boldsymbol{z}})] \;\middle|\; \boldsymbol{q}\in\mathcal{Q} \right\}.$$

Since any feasible $\boldsymbol{q}\in\mathcal{Q}\subseteq\{\boldsymbol{q}\in\Re^K_+ \mid \sum_{k\in[K]} q_k = 1\}$ is bounded, there must be $k\in[K]$ such that

$$\infty = \sup_{\mathbb{P}^k\in\mathcal{F}^k} \mathbb{E}_{\mathbb{P}^k}[g(\boldsymbol{x},\tilde{\boldsymbol{z}})] = \sum_{i\in[2n+1]} p_i^k g(\boldsymbol{x},\boldsymbol{z}^{k,i}),$$

where the last equality follows from Proposition 3. Again, since $p_i^k \in [0,1]$ for all $i\in[2n+1]$, there exists a specific $i\in[2n+1]$ such that $g(\boldsymbol{x},\boldsymbol{z}^{k,i}) = \infty$. It is equivalent to the infeasibility of the constraint $\boldsymbol{W}\boldsymbol{x} + \boldsymbol{U}\boldsymbol{y}^{k,i} \geq \boldsymbol{V}\boldsymbol{z}^{k,i} + \boldsymbol{v}^0$, which is involved in the problem defining $f(\boldsymbol{x})$. Hence, $f(\boldsymbol{x}) = \infty$.

Therefore, Problem (6) is equivalent with $\min_{\boldsymbol{x}\in\mathcal{X}_{fea}} f(\boldsymbol{x})$. We notice that Problem (2) is equivalent to $\min_{\boldsymbol{x}\in\mathcal{X}_{fea}} \sup_{\mathbb{P}\in\mathcal{F}} \mathbb{E}_{\mathbb{P}}[g(\boldsymbol{x},\tilde{\boldsymbol{z}})]$. Hence, for proving this theorem, now it suffices to show that for all $\boldsymbol{x}\in\mathcal{X}_{fea}$, we have $\sup_{\mathbb{P}\in\mathcal{F}} \mathbb{E}_{\mathbb{P}}[g(\boldsymbol{x},\tilde{\boldsymbol{z}})] = f(\boldsymbol{x})$. To this end, consider any $\boldsymbol{x}\in\mathcal{X}_{fea}$, we then know $\sup_{\mathbb{P}\in\mathcal{F}} \mathbb{E}_{\mathbb{P}}[g(\boldsymbol{x},\tilde{\boldsymbol{z}})]$ is finite. Notice that 1) $\sup_{\mathbb{P}\in\mathcal{F}} \mathbb{E}_{\mathbb{P}}[g(\boldsymbol{x},\tilde{\boldsymbol{z}})] = \max\left\{\sum_{k\in[K]} q_k \sup_{\mathbb{P}^k\in\mathcal{F}^k} \mathbb{E}_{\mathbb{P}^k}[g(\boldsymbol{x},\tilde{\boldsymbol{z}})] \;\middle|\; \boldsymbol{q}\in\mathcal{Q}\right\}$ and 2) by the assumption on $\mathcal{Q}$, for any $k\in[K]$ there exists $\boldsymbol{q}\in\mathcal{Q}$ with $q_k > 0$. Hence, for all $k\in[K]$, $\sup_{\mathbb{P}^k\in\mathcal{F}^k} \mathbb{E}_{\mathbb{P}^k}[g(\boldsymbol{x},\tilde{\boldsymbol{z}})]$ must be finite. It implies that $g(\boldsymbol{x},\boldsymbol{z})$ is finite for all $\boldsymbol{z}\in[\underline{\boldsymbol{z}}^k, \overline{\boldsymbol{z}}^k]$. Moreover,

$$\begin{aligned}
\sup_{\mathbb{P}\in\mathcal{F}} \mathbb{E}_{\mathbb{P}}[g(\boldsymbol{x},\tilde{\boldsymbol{z}})] &= \max\left\{ \sum_{k\in[K]} q_k \sup_{\mathbb{P}^k\in\mathcal{F}^k} \mathbb{E}_{\mathbb{P}^k}[g(\boldsymbol{x},\tilde{\boldsymbol{z}})] \;\middle|\; \boldsymbol{R}\boldsymbol{q}\leq\boldsymbol{\nu}, \boldsymbol{q}\geq\boldsymbol{0} \right\} \\
&= \min\left\{ \boldsymbol{\nu}^\top\boldsymbol{l} \;\middle|\; \begin{array}{l} \boldsymbol{R}_k^\top\boldsymbol{l} \geq \sup_{\mathbb{P}^k\in\mathcal{F}^k} \mathbb{E}_{\mathbb{P}^k}[g(\boldsymbol{x},\tilde{\boldsymbol{z}})],\ k\in[K] \\ \boldsymbol{l}\geq\boldsymbol{0} \end{array} \right\} \\
&= f(\boldsymbol{x}),
\end{aligned}$$

where the second equality is due to strong duality. □

C.5. Proof of Corollary 1

It has been proved in the proof for Theorem 1. □

C.6. Proof of Proposition 4

We first prove the "if" part. Suppose $\mathcal{S}(\boldsymbol{x})$ is an inverse additive lattice, then given any $\boldsymbol{z}', \boldsymbol{z}'', \boldsymbol{p}, \boldsymbol{q}$ with $(\boldsymbol{p}, \boldsymbol{z}'\wedge\boldsymbol{z}''), (\boldsymbol{q}, \boldsymbol{z}'\vee\boldsymbol{z}'')\in\mathcal{S}(\boldsymbol{x})$, there exist $\boldsymbol{y}', \boldsymbol{y}''$ such that $(\boldsymbol{y}', \boldsymbol{z}'), (\boldsymbol{y}'', \boldsymbol{z}'')\in\mathcal{S}(\boldsymbol{x})$ and $\boldsymbol{y}' + \boldsymbol{y}'' = \boldsymbol{p} + \boldsymbol{q}$. We then have

$$g(\boldsymbol{x},\boldsymbol{z}') + g(\boldsymbol{x},\boldsymbol{z}'') \leq \boldsymbol{b}^\top\boldsymbol{y}' + \boldsymbol{b}^\top\boldsymbol{y}'' = \boldsymbol{b}^\top\boldsymbol{p} + \boldsymbol{b}^\top\boldsymbol{q}.$$

Taking the minimum on the right-hand-side over all $\boldsymbol{p},\boldsymbol{q}$ with $(\boldsymbol{p},\boldsymbol{z}'\wedge\boldsymbol{z}''),(\boldsymbol{q},\boldsymbol{z}'\vee\boldsymbol{z}'')\in\mathcal{S}(\boldsymbol{x})$, we obtain $g(\boldsymbol{x},\boldsymbol{z}')+g(\boldsymbol{x},\boldsymbol{z}'')\le g(\boldsymbol{x},\boldsymbol{z}'\wedge\boldsymbol{z}'')+g(\boldsymbol{x},\boldsymbol{z}'\vee\boldsymbol{z}'')$.

Next we prove the "only if" part by contradiction. Suppose $\mathcal{S}(\boldsymbol{x})$ is not an inverse additive lattice, then there exist $\boldsymbol{z}',\boldsymbol{z}'',\boldsymbol{p},\boldsymbol{q}$ with $(\boldsymbol{p},\boldsymbol{z}'\wedge\boldsymbol{z}''),(\boldsymbol{q},\boldsymbol{z}'\vee\boldsymbol{z}'')\in\mathcal{S}(\boldsymbol{x})$ but $\boldsymbol{p}+\boldsymbol{q}\notin\mathcal{W}$, where the set $\mathcal{W}$ is defined as $\mathcal{W}=\{\boldsymbol{r}+\boldsymbol{s}\mid(\boldsymbol{r},\boldsymbol{z}'),(\boldsymbol{s},\boldsymbol{z}'')\in\mathcal{S}(\boldsymbol{x})\}$. According to the definition of $\mathcal{S}(\boldsymbol{x})$, we can easily see that $\mathcal{W}$ is convex and closed. By the Hyperplane Separation Theorem, there exist a vector $\boldsymbol{\eta}$ and a real number λ such that,

$$\boldsymbol{\eta}^\top(\boldsymbol{p}+\boldsymbol{q})<\lambda<\boldsymbol{\eta}^\top\boldsymbol{w}\qquad\forall\boldsymbol{w}\in\mathcal{W}.$$

Consider the second-stage cost function $g(\boldsymbol{x},\boldsymbol{z})$ (defined in Equation (1)) with coefficient $\boldsymbol{b}=\boldsymbol{\eta}$. We have

$$\begin{aligned}
g(\boldsymbol{x},\boldsymbol{z}')+g(\boldsymbol{x},\boldsymbol{z}'')&=\min\left\{\boldsymbol{\eta}^\top\boldsymbol{y}\;\middle|\;(\boldsymbol{y},\boldsymbol{z}')\in\mathcal{S}(\boldsymbol{x})\right\}+\min\left\{\boldsymbol{\eta}^\top\boldsymbol{y}\;\middle|\;(\boldsymbol{y},\boldsymbol{z}'')\in\mathcal{S}(\boldsymbol{x})\right\}\\
&=\min\left\{\boldsymbol{\eta}^\top(\boldsymbol{r}+\boldsymbol{s})\;\middle|\;(\boldsymbol{r},\boldsymbol{z}'),(\boldsymbol{s},\boldsymbol{z}'')\in\mathcal{S}(\boldsymbol{x})\right\}\\
&=\min\left\{\boldsymbol{\eta}^\top\boldsymbol{w}\;\middle|\;\boldsymbol{w}\in\mathcal{W}\right\}>\lambda,\\
g(\boldsymbol{x},\boldsymbol{z}'\wedge\boldsymbol{z}'')+g(\boldsymbol{x},\boldsymbol{z}'\vee\boldsymbol{z}'')&=\min\left\{\boldsymbol{\eta}^\top\boldsymbol{y}\;\middle|\;(\boldsymbol{y},\boldsymbol{z}'\wedge\boldsymbol{z}'')\in\mathcal{S}(\boldsymbol{x})\right\}+\min\left\{\boldsymbol{\eta}^\top\boldsymbol{y}\;\middle|\;(\boldsymbol{y},\boldsymbol{z}''\vee\boldsymbol{z}'')\in\mathcal{S}(\boldsymbol{x})\right\}\\
&\le\boldsymbol{\eta}^\top(\boldsymbol{p}+\boldsymbol{q})<\lambda.
\end{aligned}$$

Therefore, $g(\boldsymbol{x},\boldsymbol{z}')+g(\boldsymbol{x},\boldsymbol{z}'')>g(\boldsymbol{x},\boldsymbol{z}'\wedge\boldsymbol{z}'')+g(\boldsymbol{x},\boldsymbol{z}'\vee\boldsymbol{z}'')$, which contradicts the supermodularity. The "only if" part is completed. □

C.7. Proof of Theorem 2

Based on Proposition 4, the above theorem is equivalent to this statement: $\mathcal{S}(\boldsymbol{x})$ is an additive inverse lattice for all $\boldsymbol{x}$ and $\boldsymbol{v}^0$ if and only if $\boldsymbol{U}$ and $\boldsymbol{V}$ satisfy one of the two conditions in the above theorem. We prove the equivalent statement as follows.

First we prove the "if" direction by contradiction. Suppose there exist $\boldsymbol{x}$ and $\boldsymbol{v}^0$ such that $\mathcal{S}(\boldsymbol{x})$ is not an additive inverse lattice, i.e., we have $\boldsymbol{z}',\boldsymbol{z}'',\boldsymbol{p},\boldsymbol{q}$ with $\boldsymbol{z}^\wedge=\boldsymbol{z}'\wedge\boldsymbol{z}'',\boldsymbol{z}^\vee=\boldsymbol{z}'\vee\boldsymbol{z}''$ and $(\boldsymbol{p},\boldsymbol{z}^\wedge),(\boldsymbol{q},\boldsymbol{z}^\vee)\in\mathcal{S}(\boldsymbol{x})$, such that $\boldsymbol{y}'+\boldsymbol{y}''\ne\boldsymbol{p}+\boldsymbol{q}$ holds for all $\boldsymbol{y}',\boldsymbol{y}''$ with $(\boldsymbol{y}',\boldsymbol{z}'),(\boldsymbol{y}'',\boldsymbol{z}'')\in\mathcal{S}(\boldsymbol{x})$.

We denote $\boldsymbol{c}=-\boldsymbol{W}\boldsymbol{x}+\boldsymbol{v}^0$, $\boldsymbol{t}^1=\boldsymbol{U}\boldsymbol{p}-\boldsymbol{V}\boldsymbol{z}^\wedge\ge\boldsymbol{c}$, $\boldsymbol{t}^2=\boldsymbol{U}\boldsymbol{q}-\boldsymbol{V}\boldsymbol{z}^\vee\ge\boldsymbol{c}$. Here the two inequalities are due to $(\boldsymbol{p},\boldsymbol{z}^\wedge),(\boldsymbol{q},\boldsymbol{z}^\vee)\in\mathcal{S}(\boldsymbol{x})$ and the definition of $\mathcal{S}(\boldsymbol{x})$. We define a set $\mathcal{W}$ as

$$\mathcal{W}=\left\{\boldsymbol{y}\in\Re^m\;\middle|\;(\boldsymbol{t}^1\wedge\boldsymbol{t}^2)+\boldsymbol{V}\boldsymbol{z}'\le\boldsymbol{U}\boldsymbol{y}\le(\boldsymbol{t}^1\vee\boldsymbol{t}^2)+\boldsymbol{V}\boldsymbol{z}'\right\}.$$

Note that $\mathcal{W}$ should be an empty set, otherwise there exists a $\boldsymbol{y}^0\in\mathcal{W}$ and hence

$$\begin{gathered}
\boldsymbol{U}\boldsymbol{y}^0-\boldsymbol{V}\boldsymbol{z}'\ge(\boldsymbol{t}^1\wedge\boldsymbol{t}^2)\ge\boldsymbol{c},\\
\boldsymbol{U}(\boldsymbol{p}+\boldsymbol{q}-\boldsymbol{y}^0)-\boldsymbol{V}\boldsymbol{z}''=\boldsymbol{U}\boldsymbol{p}-\boldsymbol{V}\boldsymbol{z}^\wedge+\boldsymbol{U}\boldsymbol{q}-\boldsymbol{V}\boldsymbol{z}^\vee-(\boldsymbol{U}\boldsymbol{y}^0-\boldsymbol{V}\boldsymbol{z}')\ge\boldsymbol{t}^1+\boldsymbol{t}^2-(\boldsymbol{t}^1\vee\boldsymbol{t}^2)=\boldsymbol{t}^1\wedge\boldsymbol{t}^2\ge\boldsymbol{c},
\end{gathered}$$

which implies both $(\boldsymbol{y}^0, \boldsymbol{z}'), (\boldsymbol{p}+\boldsymbol{q}-\boldsymbol{y}^0, \boldsymbol{z}'') \in \mathcal{S}(\boldsymbol{x})$, and contradicts the previous statement on $\boldsymbol{y}', \boldsymbol{y}''$ resulting from the assumption.

We now show that the first part of the condition in our theorem is not true. If $\text{rank}(\boldsymbol{U}) = r$, we can solve $\boldsymbol{y}$ with $\boldsymbol{U}\boldsymbol{y} = (\boldsymbol{t}^1 \wedge \boldsymbol{t}^2) + \boldsymbol{V}\boldsymbol{z}' \le (\boldsymbol{t}^1 \vee \boldsymbol{t}^2) + \boldsymbol{V}\boldsymbol{z}'$, which contradicts the emptiness of $\mathcal{W}$. Therefore, $\text{rank}(\boldsymbol{U}) < r$.

We then focus on the second part of the condition in our theorem. The emptiness of $\mathcal{W}$ leads to the infeasibility of the following optimization problem:

$$\begin{aligned} \max \quad & 0 \\ \text{s.t.} \quad & \begin{bmatrix} \boldsymbol{U} \\ -\boldsymbol{U} \end{bmatrix} \boldsymbol{y} \le \begin{bmatrix} (\boldsymbol{t}^1 \vee \boldsymbol{t}^2) + \boldsymbol{V}\boldsymbol{z}' \\ -(\boldsymbol{t}^1 \wedge \boldsymbol{t}^2) - \boldsymbol{V}\boldsymbol{z}' \end{bmatrix}. \end{aligned}$$

Furthermore, by Lemma 4 we know that there exists $\mathcal{I} \subseteq [r], |\mathcal{I}| = \text{rank}(\boldsymbol{U}) + 1$ with $\text{rank}(\boldsymbol{U}_{\mathcal{I}}) = \text{rank}(\boldsymbol{U})$ such that the problem

$$\begin{aligned} \max \quad & 0 \\ \text{s.t.} \quad & \begin{bmatrix} \boldsymbol{U}_{\mathcal{I}} \\ -\boldsymbol{U}_{\mathcal{I}} \end{bmatrix} \boldsymbol{y} \le \begin{bmatrix} (\boldsymbol{t}^1_{\mathcal{I}} \vee \boldsymbol{t}^2_{\mathcal{I}}) + \boldsymbol{V}_{\mathcal{I}}\boldsymbol{z}' \\ -(\boldsymbol{t}^1_{\mathcal{I}} \wedge \boldsymbol{t}^2_{\mathcal{I}}) - \boldsymbol{V}_{\mathcal{I}}\boldsymbol{z}' \end{bmatrix} \end{aligned} \tag{40}$$

is also infeasible. We write the dual of (40) as follows,

$$\begin{aligned} \min \quad & \boldsymbol{r}^\top \left((\boldsymbol{t}^1_{\mathcal{I}} \vee \boldsymbol{t}^2_{\mathcal{I}}) + \boldsymbol{V}_{\mathcal{I}}\boldsymbol{z}' \right) - \boldsymbol{s}^\top \left((\boldsymbol{t}^1_{\mathcal{I}} \wedge \boldsymbol{t}^2_{\mathcal{I}}) + \boldsymbol{V}_{\mathcal{I}}\boldsymbol{z}' \right) \\ \text{s.t.} \quad & \boldsymbol{U}_{\mathcal{I}}^\top (\boldsymbol{r} - \boldsymbol{s}) = \boldsymbol{0} \\ & \boldsymbol{r}, \boldsymbol{s} \ge \boldsymbol{0}. \end{aligned} \tag{41}$$

Observing that $\boldsymbol{r} = \boldsymbol{s} = \boldsymbol{0}$ gives a feasible solution of (41), the infeasibility of the primal problem implies the unboundedness of the above dual problem. Therefore, there exist $\boldsymbol{r}, \boldsymbol{s} \ge \boldsymbol{0}$ with $\boldsymbol{U}_{\mathcal{I}}^\top (\boldsymbol{r} - \boldsymbol{s}) = \boldsymbol{0}$ such that the following inequalities holds,

$$\begin{aligned} 0 &> \boldsymbol{r}^\top \left((\boldsymbol{t}^1_{\mathcal{I}} \vee \boldsymbol{t}^2_{\mathcal{I}}) + \boldsymbol{V}_{\mathcal{I}}\boldsymbol{z}' \right) - \boldsymbol{s}^\top \left((\boldsymbol{t}^1_{\mathcal{I}} \wedge \boldsymbol{t}^2_{\mathcal{I}}) + \boldsymbol{V}_{\mathcal{I}}\boldsymbol{z}' \right) \\ &= \boldsymbol{r}^\top \left((\boldsymbol{t}^1_{\mathcal{I}} \vee \boldsymbol{t}^2_{\mathcal{I}}) + \boldsymbol{V}_{\mathcal{I}}\boldsymbol{z}' - \boldsymbol{U}_{\mathcal{I}}\boldsymbol{q} \right) - \boldsymbol{s}^\top \left((\boldsymbol{t}^1_{\mathcal{I}} \wedge \boldsymbol{t}^2_{\mathcal{I}}) + \boldsymbol{V}_{\mathcal{I}}\boldsymbol{z}' - \boldsymbol{U}_{\mathcal{I}}\boldsymbol{q} \right) \\ &\ge \boldsymbol{r}^\top \left(\boldsymbol{t}^2_{\mathcal{I}} + \boldsymbol{V}_{\mathcal{I}}\boldsymbol{z}' - \boldsymbol{U}_{\mathcal{I}}\boldsymbol{q} \right) - \boldsymbol{s}^\top \left(\boldsymbol{t}^2_{\mathcal{I}} + \boldsymbol{V}_{\mathcal{I}}\boldsymbol{z}' - \boldsymbol{U}_{\mathcal{I}}\boldsymbol{q} \right) \\ &= (\boldsymbol{r} - \boldsymbol{s})^\top \boldsymbol{V}_{\mathcal{I}}(\boldsymbol{z}' - \boldsymbol{z}^\vee), \end{aligned}$$

where the first inequality is obtained from the unboundedness of (41), the first equality is due to $\boldsymbol{U}_{\mathcal{I}}^\top (\boldsymbol{r} - \boldsymbol{s}) = \boldsymbol{0}$, the second inequality follows from $\boldsymbol{t}^1_{\mathcal{I}} \wedge \boldsymbol{t}^2_{\mathcal{I}} \le \boldsymbol{t}^1_{\mathcal{I}} \le \boldsymbol{t}^1_{\mathcal{I}} \vee \boldsymbol{t}^2_{\mathcal{I}}$, and the second equality comes from $\boldsymbol{t}^2_{\mathcal{I}} = \boldsymbol{U}_{\mathcal{I}}\boldsymbol{q} - \boldsymbol{V}_{\mathcal{I}}\boldsymbol{z}^\vee$. We remark that in the above equation, if we use $\boldsymbol{U}_{\mathcal{I}}\boldsymbol{p}$ instead of $\boldsymbol{U}_{\mathcal{I}}\boldsymbol{q}$ in the first equality, and $\boldsymbol{t}^1_{\mathcal{I}}$ instead of $\boldsymbol{t}^2_{\mathcal{I}}$ in the second inequality, then $0 > (\boldsymbol{r} - \boldsymbol{s})^\top \boldsymbol{V}_{\mathcal{I}}(\boldsymbol{z}' - \boldsymbol{z}^\wedge)$ can be obtained similarly.

We define $\Delta_1 = (\boldsymbol{r} - \boldsymbol{s})^\top \boldsymbol{V}_{\mathcal{I}}(\boldsymbol{z}' - \boldsymbol{z}^\vee)$, $\Delta_2 = (\boldsymbol{r} - \boldsymbol{s})^\top \boldsymbol{V}_{\mathcal{I}}(\boldsymbol{z}' - \boldsymbol{z}^\wedge)$, and $\boldsymbol{\beta} = \dfrac{\boldsymbol{z}' - \boldsymbol{z}^\vee}{\Delta_1} - \dfrac{\boldsymbol{z}' - \boldsymbol{z}^\wedge}{\Delta_2}$.

We have three observations on $\boldsymbol{\beta}$. First, $\boldsymbol{\beta} \ge \boldsymbol{0}$ since $\Delta_1, \Delta_2 < 0$ and $\boldsymbol{z}^\wedge \le \boldsymbol{z}' \le \boldsymbol{z}^\vee$.

Second, $\boldsymbol{V}_{\mathcal{I}}\boldsymbol{\beta} \in \text{span}(\boldsymbol{U}_{\mathcal{I}})$. To see this, recall that for any matrix, its column space is the orthogonal complement of the null space of its transpose; therefore, we can equivalently show that $\text{null}\,(\boldsymbol{U}_{\mathcal{I}}^\top) \subseteq \text{null}\,((\boldsymbol{V}_{\mathcal{I}}\boldsymbol{\beta})^\top)$, where $\text{null}\,(\cdot)$ is the null space of a given matrix. Since $|\mathcal{I}| = \text{rank}(\boldsymbol{U}) + 1 = \text{rank}\,(\boldsymbol{U}_{\mathcal{I}}) + 1 = \text{rank}\,(\boldsymbol{U}_{\mathcal{I}}^\top) + 1$, $\text{null}\,(\boldsymbol{U}_{\mathcal{I}}^\top)$ is of dimension 1. That implies for any $\boldsymbol{w} \in \text{null}\,(\boldsymbol{U}_{\mathcal{I}}^\top)$, we have $\boldsymbol{w} = k(\boldsymbol{r} - \boldsymbol{s})$ for some $k \in \Re$. Therefore,

$$(\boldsymbol{V}_{\mathcal{I}}\boldsymbol{\beta})^\top \boldsymbol{w} = k(\boldsymbol{r}-\boldsymbol{s})^\top \boldsymbol{V}_{\mathcal{I}}\boldsymbol{\beta} = k\left(\frac{(\boldsymbol{r}-\boldsymbol{s})^\top \boldsymbol{V}_{\mathcal{I}}(\boldsymbol{z}'-\boldsymbol{z}^\vee)}{\Delta_1} - \frac{(\boldsymbol{r}-\boldsymbol{s})^\top \boldsymbol{V}_{\mathcal{I}}(\boldsymbol{z}'-\boldsymbol{z}^\wedge)}{\Delta_2}\right) = k\,(1-1) = 0.$$

That is, $\boldsymbol{w} \in \text{null}\,((\boldsymbol{V}_{\mathcal{I}}\boldsymbol{\beta})^\top)$. Hence, $\text{null}\,(\boldsymbol{U}_{\mathcal{I}}^\top) \subseteq \text{null}\,((\boldsymbol{V}_{\mathcal{I}}\boldsymbol{\beta})^\top)$ and then we have $\boldsymbol{V}_{\mathcal{I}}\boldsymbol{\beta} \in \text{span}(\boldsymbol{U}_{\mathcal{I}})$.

The third observation is that there exists some $i \in [n]$ such that $(\boldsymbol{V}_{\mathcal{I}})_i\beta_i \notin \text{span}(\boldsymbol{U}_{\mathcal{I}})$. To show this, we denote $\mathcal{H} = \{i \in [n] \mid z_i' \le z_i''\}$. We then have for every $i \in \mathcal{H}$, $z_i^\wedge = z_i'$, $z_i^\vee = z_i''$ and hence $\beta_i = \dfrac{z_i' - z_i^\vee}{\Delta_1}$. In addition, since for every $i \in [n] \setminus \mathcal{H}$, $z_i' > z_i''$, $\dfrac{z_i' - z_i^\vee}{\Delta_1} = \dfrac{z_i' - z_i'}{\Delta_1} = 0$, we have

$$(\boldsymbol{r}-\boldsymbol{s})^\top \sum_{i\in\mathcal{H}} \beta_i (\boldsymbol{V}_{\mathcal{I}})_i = (\boldsymbol{r}-\boldsymbol{s})^\top \left(\sum_{i\in\mathcal{H}} \beta_i(\boldsymbol{V}_{\mathcal{I}})_i + \sum_{i\in[n]\setminus\mathcal{H}} 0\cdot(\boldsymbol{V}_{\mathcal{I}})_i\right) = (\boldsymbol{r}-\boldsymbol{s})^\top \boldsymbol{V}_{\mathcal{I}} \frac{\boldsymbol{z}'-\boldsymbol{z}^\vee}{\Delta_1} = 1.$$

Hence, $(\boldsymbol{r} - \boldsymbol{s}) \notin \text{null}\left(\left(\sum_{i\in\mathcal{H}} \beta_i(\boldsymbol{V}_{\mathcal{I}})_i\right)^\top\right)$, which implies that $\text{null}\,(\boldsymbol{U}_{\mathcal{I}}^\top)$ is not a subset of $\text{null}\left(\left(\sum_{i\in\mathcal{H}} \beta_i(\boldsymbol{V}_{\mathcal{I}})_i\right)^\top\right)$. Consequently we have $\sum_{i\in\mathcal{H}} \beta_i(\boldsymbol{V}_{\mathcal{I}})_i \notin \text{span}(\boldsymbol{U}_{\mathcal{I}})$, implying that there exists some $i \in \mathcal{H}$ such that $(\boldsymbol{V}_{\mathcal{I}})_i\beta_i \notin \text{span}(\boldsymbol{U}_{\mathcal{I}})$.

With the three observations, we have a contradiction of the second condition in Theorem 2.

We next prove the "only if" direction by contradiction. Assume the condition in the theorem is not satisfied. That is, $\text{rank}(\boldsymbol{U}) < r$ and there exist some $\mathcal{I} \subseteq [r]$, $\boldsymbol{\beta} \in \Re_+^n$ satisfying $|\mathcal{I}| = \text{rank}(\boldsymbol{U}) + 1$, $\text{rank}(\boldsymbol{U}_{\mathcal{I}}) = \text{rank}(\boldsymbol{U})$ and $\boldsymbol{V}_{\mathcal{I}}\boldsymbol{\beta} \in \text{span}(\boldsymbol{U}_{\mathcal{I}})$, such that $\beta_i(\boldsymbol{V}_{\mathcal{I}})_i \notin \text{span}(\boldsymbol{U}_{\mathcal{I}})$ for some $i \in [n]$. Note that in this case, we can find a vector $\boldsymbol{\alpha} \in \Re^m$ such that $\boldsymbol{U}_{\mathcal{I}}\boldsymbol{\alpha} = \boldsymbol{V}_{\mathcal{I}}\boldsymbol{\beta}$.

We arbitrarily choose $\boldsymbol{z}^\wedge \in \Re^n, \boldsymbol{p} \in \Re^m$ and let $\boldsymbol{z}^\vee = \boldsymbol{z}^\wedge + \boldsymbol{\beta} \ge \boldsymbol{z}^\wedge, \boldsymbol{q} = \boldsymbol{p} + \boldsymbol{\alpha}$, then $\boldsymbol{U}_{\mathcal{I}}\boldsymbol{p} - \boldsymbol{V}_{\mathcal{I}}\boldsymbol{z}^\wedge = \boldsymbol{U}_{\mathcal{I}}(\boldsymbol{q}-\boldsymbol{\alpha}) - \boldsymbol{V}_{\mathcal{I}}(\boldsymbol{z}^\vee - \boldsymbol{\beta}) = \boldsymbol{U}_{\mathcal{I}}\boldsymbol{q} - \boldsymbol{V}_{\mathcal{I}}\boldsymbol{z}^\vee$. We also arbitrarily choose $\boldsymbol{x}$, and then choose $\boldsymbol{v}^0$ such that $\boldsymbol{c} = -\boldsymbol{W}\boldsymbol{x} + \boldsymbol{v}^0$ is with $\boldsymbol{c}_{\mathcal{I}} = \boldsymbol{U}_{\mathcal{I}}\boldsymbol{p} - \boldsymbol{V}_{\mathcal{I}}\boldsymbol{z}^\wedge$ and c_j being sufficiently small for all $j \notin \mathcal{I}$. Then we have $(\boldsymbol{p}, \boldsymbol{z}^\wedge), (\boldsymbol{q}, \boldsymbol{z}^\vee) \in \mathcal{S}(\boldsymbol{x})$. We further define $\boldsymbol{z}' = \boldsymbol{z}^\wedge + \beta_i\boldsymbol{e}_i, \boldsymbol{z}'' = \boldsymbol{z}^\vee - \beta_i\boldsymbol{e}_i$ so that $\boldsymbol{z}' \wedge \boldsymbol{z}'' = \boldsymbol{z}^\wedge, \boldsymbol{z}' \vee \boldsymbol{z}'' = \boldsymbol{z}^\vee$. Then we have

$$\boldsymbol{c}_{\mathcal{I}} + \boldsymbol{V}_{\mathcal{I}}\boldsymbol{z}' = \boldsymbol{c}_{\mathcal{I}} + \boldsymbol{V}_{\mathcal{I}}(\boldsymbol{z}^\wedge + \beta_i\boldsymbol{e}_i) = \boldsymbol{U}_{\mathcal{I}}\boldsymbol{p} - \boldsymbol{V}_{\mathcal{I}}\boldsymbol{z}^\wedge + \boldsymbol{V}_{\mathcal{I}}(\boldsymbol{z}^\wedge + \beta_i\boldsymbol{e}_i) = \boldsymbol{U}_{\mathcal{I}}\boldsymbol{p} + \beta_i(\boldsymbol{V}_{\mathcal{I}})_i \notin \text{span}(\boldsymbol{U}_{\mathcal{I}}),$$

where the last relationship holds since $\boldsymbol{U}_{\mathcal{I}}\boldsymbol{p} \in \text{span}(\boldsymbol{U}_{\mathcal{I}})$ but $\beta_i(\boldsymbol{V}_{\mathcal{I}})_i \notin \text{span}(\boldsymbol{U}_{\mathcal{I}})$.

Hence, $\{\boldsymbol{y} \in \Re^m \mid \boldsymbol{U}_{\mathcal{I}}\boldsymbol{y} = \boldsymbol{c}_{\mathcal{I}} + \boldsymbol{V}_{\mathcal{I}}\boldsymbol{z}'\} = \emptyset$, i.e. for any $\boldsymbol{y}'$ satisfying $\boldsymbol{U}\boldsymbol{y}' - \boldsymbol{V}\boldsymbol{z}' \ge \boldsymbol{c}$, there exists $j \in \mathcal{I}$ such that $\boldsymbol{u}_j^\top \boldsymbol{y}' - \boldsymbol{v}_j^\top \boldsymbol{z}' > c_j$. If there exists some $\boldsymbol{y}''$ with $\boldsymbol{U}\boldsymbol{y}'' - \boldsymbol{V}\boldsymbol{z}'' \ge \boldsymbol{c}$ satisfies $\boldsymbol{y}' + \boldsymbol{y}'' = \boldsymbol{p} + \boldsymbol{q}$,

$$\begin{aligned}
\boldsymbol{U}_{\mathcal{I}}\boldsymbol{y}'' - \boldsymbol{V}_{\mathcal{I}}\boldsymbol{z}'' &= \boldsymbol{U}_{\mathcal{I}}\,(\boldsymbol{p} + \boldsymbol{q} - \boldsymbol{y}') - \boldsymbol{V}_{\mathcal{I}}\,(\boldsymbol{z}^\wedge + \boldsymbol{z}^\vee - \boldsymbol{z}') \\
&= \boldsymbol{U}_{\mathcal{I}}\boldsymbol{p} - \boldsymbol{V}_{\mathcal{I}}\boldsymbol{z}^\wedge + \boldsymbol{U}_{\mathcal{I}}\boldsymbol{q} - \boldsymbol{V}_{\mathcal{I}}\boldsymbol{z}^\vee - (\boldsymbol{U}_{\mathcal{I}}\boldsymbol{y}' - \boldsymbol{V}_{\mathcal{I}}\boldsymbol{z}') \\
&= 2\boldsymbol{c}_{\mathcal{I}} - (\boldsymbol{U}_{\mathcal{I}}\boldsymbol{y}' - \boldsymbol{V}_{\mathcal{I}}\boldsymbol{z}'),
\end{aligned}$$

then we should have $2c_j - (\boldsymbol{u}_j^\top \boldsymbol{y}' - \boldsymbol{v}_j^\top \boldsymbol{z}') < c_j$ for the above mentioned j, which contradicts the assumption $(\boldsymbol{y}'', \boldsymbol{z}'') \in \mathcal{S}(\boldsymbol{x})$. Hence we prove the necessity of the conditions on $\boldsymbol{U}, \boldsymbol{V}$. □

Lemma 4 (Chen et al. 2021) *Consider any matrix* $\boldsymbol{U} \in \Re^{r\times m}$ *with* $\operatorname{rank}(\boldsymbol{U}) < r$. *Suppose that system* $\begin{cases} \boldsymbol{U}\boldsymbol{x} \le \overline{\boldsymbol{c}} \\ -\boldsymbol{U}\boldsymbol{x} \le -\underline{\boldsymbol{c}} \end{cases}$ *is infeasible. Then there exists* $\mathcal{I} \subseteq [r]$ *with* $|\mathcal{I}| = \operatorname{rank}(\boldsymbol{U}) + 1$ *and* $\operatorname{rank}(\boldsymbol{U}_\mathcal{I}) = \operatorname{rank}(\boldsymbol{U})$ *such that system* $\begin{cases} \boldsymbol{U}_\mathcal{I}\boldsymbol{x} \le \overline{\boldsymbol{c}}_\mathcal{I} \\ -\boldsymbol{U}_\mathcal{I}\boldsymbol{x} \le -\underline{\boldsymbol{c}}_\mathcal{I} \end{cases}$ *is also infeasible.*

C.8. Proof of Proposition 5

"⇐" Assume there exists a 2×3 submatrix of $\boldsymbol{U}^\circ$ such that any pair of columns in it are linearly independent. WLOG, let $\boldsymbol{U}^\circ_{\{1,2\},\{1,2,3\}}$ be such matrix and we denote it by $\boldsymbol{A} = [\boldsymbol{A}_1 \ \boldsymbol{A}_2 \ \boldsymbol{A}_3] \in \Re^{2\times 3}$. WLOG, assume $\boldsymbol{A}_3 = t_1\boldsymbol{A}_1 + t_2\boldsymbol{A}_2$ with $t_1, t_2 > 0$. Choose $\boldsymbol{V}^1 = \boldsymbol{I}_{m\times m}$, $\mathcal{I} = [m+2] \setminus \{3\}$, $\boldsymbol{\beta} = t_1\boldsymbol{e}_1 + t_2\boldsymbol{e}_2 \ge \boldsymbol{0}$, $\boldsymbol{\alpha} = t_1\boldsymbol{e}_1 + t_2\boldsymbol{e}_2 - \boldsymbol{e}_3$. We then have $\boldsymbol{V}_\mathcal{I}\boldsymbol{\beta} = t_1\boldsymbol{e}_1 + t_2\boldsymbol{e}_2$; at the same time, $\boldsymbol{U}_\mathcal{I}\boldsymbol{\alpha} = t_1\boldsymbol{e}_1 + t_2\boldsymbol{e}_2$ since $\boldsymbol{A}_3 = t_1\boldsymbol{A}_1 + t_2\boldsymbol{A}_2$. Hence, $\boldsymbol{V}_\mathcal{I}\boldsymbol{\beta} = \boldsymbol{U}_\mathcal{I}\boldsymbol{\alpha} \in \operatorname{span}(\boldsymbol{U}_\mathcal{I})$. However, $(\boldsymbol{V}_\mathcal{I})_1\beta_1 = \beta_1\boldsymbol{e}_1 \notin \operatorname{span}(\boldsymbol{U}_\mathcal{I})$. Therefore, the second condition in Theorem 2 is violated, there exists an instance of $g(\boldsymbol{x}, \boldsymbol{z})$ which is not supermodular in $\boldsymbol{z}$.

"⇒" Assume that every 2×3 submatrix of $\boldsymbol{U}^\circ$ contains at least one pair of column vectors which are linearly dependent. We prove the result by showing the second condition in Theorem 2 is always satisfied. To see this, consider any $\mathcal{I} \subseteq [r]$ such that $|\mathcal{I}| = m+1$, $\operatorname{rank}(\boldsymbol{U}_\mathcal{I}) = m$. Let $\mathcal{I}_1 = \mathcal{I} \cap [m]$ and $\mathcal{I}_0 = \mathcal{I} \cap \{m+1, \ldots, r\}$ be a partition of $\mathcal{I}$, hence the submatrix $\boldsymbol{U}_{\mathcal{I}_1}$ is extracted from $\boldsymbol{I}_{m\times m}$ and $\boldsymbol{U}_{\mathcal{I}_0}$ is from $\boldsymbol{U}^\circ$. We further let $\mathcal{J}_1$, $\mathcal{J}_0$ be a partition of $[m]$ such that $U_{\mathcal{I}_1,\mathcal{J}_1}$ contains all nonzero columns in $U_{\mathcal{I}_1}$ and hence $U_{\mathcal{I}_1,\mathcal{J}_0} = \boldsymbol{0}$. Noting that $U_{\mathcal{I}_1}$ contains rows extracted from $\boldsymbol{I}_{m\times m}$, we know $\mathcal{I}_1 = \mathcal{J}_1$. Hence, $|\mathcal{I}_0| = m+1-|\mathcal{I}_1| = m+1-|\mathcal{J}_1| = m+1-(m-|\mathcal{J}_0|) = |\mathcal{J}_0|+1$. We illustrate the partition of $U_\mathcal{I}$ as follows,

$$\boldsymbol{U}_\mathcal{I} = \begin{bmatrix} \boldsymbol{U}_{\mathcal{I}_1} \\ \boldsymbol{U}_{\mathcal{I}_0} \end{bmatrix} = \begin{bmatrix} \boldsymbol{I}_{|\mathcal{I}_1|\times|\mathcal{I}_1|} & \boldsymbol{0}_{|\mathcal{I}_1|\times|\mathcal{J}_0|} \\ \boldsymbol{U}_{\mathcal{I}_0,\mathcal{J}_1} & \boldsymbol{U}_{\mathcal{I}_0,\mathcal{J}_0} \end{bmatrix}.$$

We first prove that there exists a unit vector $\boldsymbol{p} \in \Re^{|\mathcal{I}_0|}$, such that it is orthogonal to $\operatorname{span}(\boldsymbol{U}_{\mathcal{I}_0,\mathcal{J}_0})$ and $\operatorname{span}[\boldsymbol{U}_{\mathcal{I}_0,\mathcal{J}_0} \ \boldsymbol{p}] = \Re^{|\mathcal{I}_0|}$. Notice that $\boldsymbol{U}_\mathcal{I}$ is of full column rank, and hence so does its submatrix $\boldsymbol{U}_{\mathcal{I},\mathcal{J}_0} = \begin{bmatrix} \boldsymbol{0}_{|\mathcal{I}_1|\times|\mathcal{J}_0|} \\ \boldsymbol{U}_{\mathcal{I}_0,\mathcal{J}_0} \end{bmatrix}$, which implies $\boldsymbol{U}_{\mathcal{I}_0,\mathcal{J}_0} \in \Re^{|\mathcal{I}_0|\times|\mathcal{J}_0|}$ is also of full column rank. Therefore, $\operatorname{span}(\boldsymbol{U}_{\mathcal{I}_0,\mathcal{J}_0})$ is of dimension $|\mathcal{J}_0| = |\mathcal{I}_0| - 1$, the existence of $\boldsymbol{p}$ can be proved.

We now show that the orthogonal unit vector $\boldsymbol{p}$ can be chosen such that for all $i \in \mathcal{J}_1$, there exist some $\boldsymbol{s}_i \in \operatorname{span}(\boldsymbol{U}_{\mathcal{I}_0,\mathcal{J}_0})$ and $\gamma_i \ge 0$ such that $(\boldsymbol{U}_{\mathcal{I}_0})_i = \boldsymbol{s}_i + \gamma_i\boldsymbol{p}$. For those $i \in \mathcal{J}_1$ with $(\boldsymbol{U}_{\mathcal{I}_0})_i \in \operatorname{span}(\boldsymbol{U}_{\mathcal{I}_0,\mathcal{J}_0})$, we always have $\gamma_i = 0$ regardless of the choice of orthogonal vector $\boldsymbol{p}$. Now we consider any given $j \in \mathcal{J}_1$ with $(\boldsymbol{U}_{\mathcal{I}_0})_j \notin \operatorname{span}(\boldsymbol{U}_{\mathcal{I}_0,\mathcal{J}_0})$. Since $\operatorname{span}[\boldsymbol{U}_{\mathcal{I}_0,\mathcal{J}_0} \ \boldsymbol{p}] = \Re^{|\mathcal{I}_0|}$, we can surely represent $(\boldsymbol{U}_{\mathcal{I}_0})_j = \boldsymbol{s}_j + \gamma_j\boldsymbol{p}$ for some $\boldsymbol{s}_j \in \operatorname{span}(\boldsymbol{U}_{\mathcal{I}_0,\mathcal{J}_0})$ and $\gamma_j \ne 0$. Moreover, the unit vector $\boldsymbol{p}$ can be chosen (as $-\boldsymbol{p}$, if necessary) to make $\gamma_j > 0$. Consider any $k \in \mathcal{J}_1 \setminus \{j\}$ with $(\boldsymbol{U}_{\mathcal{I}_0})_j \notin \operatorname{span}(\boldsymbol{U}_{\mathcal{I}_0,\mathcal{J}_0})$. Denote

$\boldsymbol{Q} = [(\boldsymbol{U}_{\mathcal{I}_0})_j \ (\boldsymbol{U}_{\mathcal{I}_0})_k \ \boldsymbol{U}_{\mathcal{I}_0,\mathcal{J}_0}]$. Notice that every 2×3 submatrix of $\boldsymbol{U}^\circ$, and hence that of $\boldsymbol{Q}$, contains at least one pair of column vectors which are linearly dependent. By Lemma 5, there are at least one pair of columns in $\boldsymbol{Q}$ which are linearly dependent. Since $\boldsymbol{U}_{\mathcal{I}_0,\mathcal{J}_0}$ is of full column rank and $(\boldsymbol{U}_{\mathcal{I}_0})_j, (\boldsymbol{U}_{\mathcal{I}_0})_k \notin \operatorname{span}(\boldsymbol{U}_{\mathcal{I}_0,\mathcal{J}_0})$, the two linearly dependent columns can only be $(\boldsymbol{U}_{\mathcal{I}_0})_j, (\boldsymbol{U}_{\mathcal{I}_0})_k$, i.e., $(\boldsymbol{U}_{\mathcal{I}_0})_k = \zeta(\boldsymbol{U}_{\mathcal{I}_0})_j$ for some $\zeta \neq 0$ (recall that both $(\boldsymbol{U}_{\mathcal{I}_0})_k$ and $(\boldsymbol{U}_{\mathcal{I}_0})_j$ are nonzero vector since they are not in $\operatorname{span}(\boldsymbol{U}_{\mathcal{I}_0,\mathcal{J}_0})$). As all components in the same row of $\boldsymbol{U}$ are with the same sign, we know $\zeta > 0$. Therefore, $(\boldsymbol{U}_{\mathcal{I}_0})_k = \zeta\,(\boldsymbol{s}_j + \gamma_j \boldsymbol{p}) = \zeta \boldsymbol{s}_j + \zeta\gamma_j \boldsymbol{p}$ where $\zeta\gamma_j > 0$.

We are now ready to prove the second condition in Theorem 2 holds. Consider any $\boldsymbol{\beta} \geq \boldsymbol{0}$ and $\boldsymbol{\alpha}$ such that $\boldsymbol{V}_{\mathcal{I}}\boldsymbol{\beta} = \boldsymbol{U}_{\mathcal{I}}\boldsymbol{\alpha}$. Observing the first block, characterized by $\mathcal{I}_1$, we have $\boldsymbol{V}_{\mathcal{I}_1}\boldsymbol{\beta} = \boldsymbol{U}_{\mathcal{I}_1}\boldsymbol{\alpha} = (\boldsymbol{I}_{m\times m})_{\mathcal{I}_1}\boldsymbol{\alpha} = \boldsymbol{\alpha}_{\mathcal{I}_1}$; since $\boldsymbol{V}, \boldsymbol{\beta}$ are both nonnegative, we have $\boldsymbol{\alpha}_{\mathcal{I}_1} \geq \boldsymbol{0}$. Observing the second block, characterized by $\mathcal{I}_0$, by $\boldsymbol{V}_{\mathcal{I}_0} = \boldsymbol{0}$, we have

$$\boldsymbol{0} = \boldsymbol{V}_{\mathcal{I}_0}\boldsymbol{\beta} = \boldsymbol{U}_{\mathcal{I}_0}\boldsymbol{\alpha} = [\boldsymbol{U}_{\mathcal{I}_0,\mathcal{J}_1}\ \boldsymbol{U}_{\mathcal{I}_0,\mathcal{J}_0}]\begin{bmatrix}\boldsymbol{\alpha}_{\mathcal{J}_1}\\ \boldsymbol{\alpha}_{\mathcal{J}_0}\end{bmatrix} = \boldsymbol{U}_{\mathcal{I}_0,\mathcal{J}_1}\boldsymbol{\alpha}_{\mathcal{J}_1} + \boldsymbol{U}_{\mathcal{I}_0,\mathcal{J}_0}\boldsymbol{\alpha}_{\mathcal{J}_0} = \boldsymbol{s} + \boldsymbol{p}\sum_{i\in\mathcal{N}}\alpha_i\gamma_i \tag{42}$$

for some $\boldsymbol{s} \in \operatorname{span}(\boldsymbol{U}_{\mathcal{I}_0,\mathcal{J}_0})$. Here we denote the index set $\mathcal{N} = \{i \in \mathcal{J}_1 \mid (\boldsymbol{U}_{\mathcal{I}_0})_i \notin \operatorname{span}(\boldsymbol{U}_{\mathcal{I}_0,\mathcal{J}_0})\}$ and hence the last equality above holds due to the argument proved in the last paragraph. Moreover, since $\boldsymbol{s}$ and $\boldsymbol{p}$ are orthogonal, by (42) we have $\sum_{i\in\mathcal{N}}\alpha_i\gamma_i = 0$, which implies $\alpha_i = 0$ for all $i \in \mathcal{N}$, as we have already known $\gamma_i > 0, \alpha_i \geq 0$ holds for all $i \in \mathcal{N}$ (recall that $\mathcal{N} \subseteq \mathcal{J}_1 = \mathcal{I}_1$, and $\boldsymbol{\alpha}_{\mathcal{I}_1} \geq \boldsymbol{0}$). Therefore, the equation $0 = \alpha_i = \boldsymbol{u}_i^\top\boldsymbol{\alpha} = \boldsymbol{v}_i^\top\boldsymbol{\beta}$ holds for any $i \in \mathcal{N}$, where the last equality is due to $\boldsymbol{V}_{\mathcal{I}}\boldsymbol{\beta} = \boldsymbol{U}_{\mathcal{I}}\boldsymbol{\alpha}$. As $\boldsymbol{V}, \boldsymbol{\beta} \geq \boldsymbol{0}$ for all $i \in \mathcal{N}$, $\boldsymbol{v}_i^\top\boldsymbol{\beta} = 0$ implies $v_{ik}\beta_k = 0$ for all $k \in [m]$. We now consider any $j \in [m]$ and it remains to show $(\boldsymbol{V}_{\mathcal{I}})_j\beta_j = \boldsymbol{U}_{\mathcal{I}}\boldsymbol{\eta}$ for some $\boldsymbol{\eta} \in \Re^m$. To this end, we choose $\boldsymbol{\eta} \in \Re^m$ with $\eta_i = v_{ij}\beta_j$ for all $i \in \mathcal{J}_1 = \mathcal{I}_1$ and we determine $\boldsymbol{\eta}_{\mathcal{J}_0}$ later. Then $\boldsymbol{u}_i^\top\boldsymbol{\eta} = \eta_i = v_{ij}\beta_j$ for all $i \in \mathcal{J}_1 = \mathcal{I}_1$. We additionally observe that $\eta_i = 0$ for all $i \in \mathcal{N}$, following from $v_{ij}\beta_j = 0$. We now move on to $\mathcal{I}_0$, and have

$$\boldsymbol{U}_{\mathcal{I}_0}\boldsymbol{\eta} = [\boldsymbol{U}_{\mathcal{I}_0,\mathcal{J}_1}\ \boldsymbol{U}_{\mathcal{I}_0,\mathcal{J}_0}]\begin{bmatrix}\boldsymbol{\eta}_{\mathcal{J}_1}\\ \boldsymbol{\eta}_{\mathcal{J}_0}\end{bmatrix} = \boldsymbol{U}_{\mathcal{I}_0,\mathcal{J}_1}\boldsymbol{\eta}_{\mathcal{J}_1} + \boldsymbol{U}_{\mathcal{I}_0,\mathcal{J}_0}\boldsymbol{\eta}_{\mathcal{J}_0} = \sum_{i\in\mathcal{J}_1\setminus\mathcal{N}}\boldsymbol{s}_i\eta_i + \boldsymbol{U}_{\mathcal{I}_0,\mathcal{J}_0}\boldsymbol{\eta}_{\mathcal{J}_0},$$

where the last equality is due to that when $i \in \mathcal{N}$, $\eta_i = 0$ and when $j \in \mathcal{J}_1 \setminus \mathcal{N}$, $(\boldsymbol{U}_{\mathcal{I}_0})_j = \boldsymbol{s}_j + \gamma_j\boldsymbol{u}$ with $\gamma_j = 0$. Since $\boldsymbol{s}_i \in \operatorname{span}(\boldsymbol{U}_{\mathcal{I}_0,\mathcal{J}_0})$, we can choose $\boldsymbol{\eta}_{\mathcal{J}_0}$ such that $\sum_{i\in\mathcal{J}_1\setminus\mathcal{N}}\boldsymbol{s}_i\eta_i + \boldsymbol{U}_{\mathcal{I}_0,\mathcal{J}_0}\boldsymbol{\eta}_{\mathcal{J}_0} = \boldsymbol{0}$. In this case, $\boldsymbol{U}_{\mathcal{I}_0}\boldsymbol{\eta} = \boldsymbol{0} = (\boldsymbol{V}_{\mathcal{I}_0})_j\beta_j$. Hence, we conclude $(\boldsymbol{V}_{\mathcal{I}})_j\beta_j = \boldsymbol{U}_{\mathcal{I}}\boldsymbol{\eta} \in \operatorname{span}(\boldsymbol{U}_{\mathcal{I}})$. □

Lemma 5 (Chen et al. 2021) *Consider any matrix* $\boldsymbol{Q} \in \Re^{s\times(s+1)}$ *with* $\operatorname{rank}(\boldsymbol{Q}) = s, s \geq 2$. *If every* 2×3 *submatrix of* $\boldsymbol{Q}$ *contains at least one pair of column vectors which are linearly dependent, then* $\boldsymbol{Q}$ *has at least one pair of column vectors which are linearly dependent.*

C.9. Proof of Lemma 2

For notational simplicity, we remove the superscript k throughout this proof. To see $\boldsymbol{\zeta}^1,\ldots,\boldsymbol{\zeta}^{2n+1}$ are vertices of a $2n$-simplex, it suffices to show these $2n+1$ points are affinely independent. That is, we need to prove that $\boldsymbol{\zeta}^2-\boldsymbol{\zeta}^1,\ldots,\boldsymbol{\zeta}^{2n+1}-\boldsymbol{\zeta}^1$ are linearly independent. First, we scale each elements in $\boldsymbol{\omega}^i,\boldsymbol{v}^i,\boldsymbol{\zeta}^i$ such that all nonzero elements become 1 and denote the corresponding vectors as $\hat{\boldsymbol{\omega}}^i,\hat{\boldsymbol{v}}^i,\hat{\boldsymbol{\zeta}}^i$. Notice that we still have $\hat{\boldsymbol{\zeta}}^i=\begin{bmatrix}\hat{\boldsymbol{\omega}}^i\\ \hat{\boldsymbol{v}}^i\end{bmatrix}$. In this case, $\hat{\boldsymbol{\omega}}^1=\mathbf{1},\hat{\boldsymbol{v}}^1=\mathbf{0}$ since $\boldsymbol{z}^1=\underline{\boldsymbol{z}}$, $\hat{\boldsymbol{\omega}}^{2n+1}=\mathbf{0},\hat{\boldsymbol{v}}^{2n+1}=\mathbf{1}$ since $\boldsymbol{z}^{2n+1}=\overline{\boldsymbol{z}}$. Moreover, we have

$$\{\hat{\boldsymbol{\omega}}^i-\hat{\boldsymbol{\omega}}^{i+1},\hat{\boldsymbol{v}}^{i+1}-\hat{\boldsymbol{v}}^i\}=\{\mathbf{0},\boldsymbol{e}_{\kappa_i}\}$$

for some $\kappa_i\in[n]$, $i\in[2n]$. This follows from that $\boldsymbol{z}^{i+1}-\boldsymbol{z}^i$ has exactly one nonzero entry, the index of which is denoted as κ_i. Specifically, for the κ_i-th entry where $\boldsymbol{z}^i$ moves to $\boldsymbol{z}^{i+1}$, 1) if the move is from the lower bound to the mean, then $\hat{\boldsymbol{\omega}}^{i+1}=\hat{\boldsymbol{\omega}}^i-\boldsymbol{e}_{\kappa_i},\hat{\boldsymbol{v}}^{i+1}=\hat{\boldsymbol{v}}^i$ and hence $\hat{\boldsymbol{\zeta}}^{i+1}-\hat{\boldsymbol{\zeta}}^i=\begin{bmatrix}-\boldsymbol{e}_{\kappa_i}\\ \mathbf{0}\end{bmatrix}$; 2) if the move is from the mean to the upper bound, then $\hat{\boldsymbol{\omega}}^{i+1}=\hat{\boldsymbol{\omega}}^i,\hat{\boldsymbol{v}}^{i+1}=\hat{\boldsymbol{v}}^i+\boldsymbol{e}_{\kappa_i}$ and hence $\hat{\boldsymbol{\zeta}}^{i+1}-\hat{\boldsymbol{\zeta}}^i=\begin{bmatrix}\mathbf{0}\\ \boldsymbol{e}_{\kappa_i}\end{bmatrix}$. We also notice that for each dimension, there is exactly one move from the lower bound to the mean, and one from the mean to the upper bound. Therefore, the matrix $\begin{bmatrix}\hat{\boldsymbol{\zeta}}^2-\hat{\boldsymbol{\zeta}}^1 & \cdots & \hat{\boldsymbol{\zeta}}^{2n+1}-\hat{\boldsymbol{\zeta}}^{2n}\end{bmatrix}=\begin{bmatrix}-\boldsymbol{I}_{n\times n} & \mathbf{0}\\ \mathbf{0} & \boldsymbol{I}_{n\times n}\end{bmatrix}\boldsymbol{P}$ for a $2n\times 2n$ permutation matrix $\boldsymbol{P}$. Notice that $\boldsymbol{\zeta}^i=\begin{bmatrix}((\mu_j-\underline{z}_j)\cdot\hat{\omega}^i_j)_{j\in[n]}\\ ((\overline{z}_j-\mu_j)\cdot\hat{v}^i_j)_{j\in[2n]}\end{bmatrix}=\begin{bmatrix}\operatorname{diag}(\boldsymbol{\mu}-\underline{\boldsymbol{z}}) & \mathbf{0}\\ \mathbf{0} & \operatorname{diag}(\overline{\boldsymbol{z}}-\boldsymbol{\mu})\end{bmatrix}\hat{\boldsymbol{\zeta}}^i$ for all $i\in[2n+1]$. We then have

$$\begin{aligned}[\boldsymbol{\zeta}^2-\boldsymbol{\zeta}^1 \ \cdots\ \boldsymbol{\zeta}^{2n+1}-\boldsymbol{\zeta}^{2n}] &= \begin{bmatrix}\operatorname{diag}(\boldsymbol{\mu}-\underline{\boldsymbol{z}}) & \mathbf{0}\\ \mathbf{0} & \operatorname{diag}(\overline{\boldsymbol{z}}-\boldsymbol{\mu})\end{bmatrix}\begin{bmatrix}\hat{\boldsymbol{\zeta}}^2-\hat{\boldsymbol{\zeta}}^1 & \cdots & \hat{\boldsymbol{\zeta}}^{2n+1}-\hat{\boldsymbol{\zeta}}^{2n}\end{bmatrix}\\ &= \begin{bmatrix}\operatorname{diag}(\boldsymbol{\mu}-\underline{\boldsymbol{z}}) & \mathbf{0}\\ \mathbf{0} & \operatorname{diag}(\overline{\boldsymbol{z}}-\boldsymbol{\mu})\end{bmatrix}\begin{bmatrix}-\boldsymbol{I}_{n\times n} & \mathbf{0}\\ \mathbf{0} & \boldsymbol{I}_{n\times n}\end{bmatrix}\boldsymbol{P}\\ &= \begin{bmatrix}\operatorname{diag}(\underline{\boldsymbol{z}}-\boldsymbol{\mu}) & \mathbf{0}\\ \mathbf{0} & \operatorname{diag}(\overline{\boldsymbol{z}}-\boldsymbol{\mu})\end{bmatrix}\boldsymbol{P}.\end{aligned}$$

This implies that the matrix $[\boldsymbol{\zeta}^2-\boldsymbol{\zeta}^1\ \cdots\ \boldsymbol{\zeta}^{2n+1}-\boldsymbol{\zeta}^1]$ are also invertible. □

C.10. Proof of Proposition 6

We first let V_{adapt} and V_{ldr} represent the optimal values for Problems (9) and (11), respectively. Our aim is to show that $V_{adapt}=V_{ldr}$.

We first prove $V_{adapt}\le V_{ldr}$. To show this, we define a new problem by relaxing Problem (9) such that the constraints of second-stage problem apply only to the realizations $\boldsymbol{z}^{k,i},k\in[K],i\in[2n+1]$ and denote the optimal value as V_{relax}, i.e.,

$$\begin{aligned}V_{relax}=\min\quad & \boldsymbol{a}^\top\boldsymbol{x}+\sup_{\mathbb{P}\in\mathcal{F}}\ \mathbb{E}_{\mathbb{P}}\left[\boldsymbol{b}^\top\boldsymbol{y}(\tilde{k},\tilde{\boldsymbol{z}})\right]\\ \text{s.t.}\quad & \boldsymbol{W}\boldsymbol{x}+\boldsymbol{U}\boldsymbol{y}(k,\boldsymbol{z}^{k,i})\ge\boldsymbol{V}\boldsymbol{z}^{k,i}+\boldsymbol{v}^0,\quad k\in[K],i\in[2n+1]\\ & \boldsymbol{x}\in\mathcal{X}.\end{aligned}\tag{43}$$

By the minimax theorem (von Neumann 1928), we can interchange "$\sup_{\mathbb{P}}$" and "$\min_{\boldsymbol{y}(k,\boldsymbol{z})}$" equivalently. Omitting the dependency between $\boldsymbol{y}$ and the uncertainty realizations, we rewrite $V_{relax} = \min_{\boldsymbol{x}\in\mathcal{X}} \boldsymbol{a}^\top\boldsymbol{x} + \sup_{\mathbb{P}\in\mathcal{F}} \mathbb{E}_{\mathbb{P}}[g'(\boldsymbol{x},\boldsymbol{z})]$, where

$$g'(\boldsymbol{x},\boldsymbol{z}) = \begin{cases} \min\left\{\boldsymbol{b}^\top\boldsymbol{y} \;\middle|\; \boldsymbol{W}\boldsymbol{x}+\boldsymbol{U}\boldsymbol{y} \geq \boldsymbol{V}\boldsymbol{z}+\boldsymbol{v}^0\right\} & \text{if } \boldsymbol{z}\in\bigcup_{k\in[K]}\{\boldsymbol{z}^{k,1},\ldots,\boldsymbol{z}^{k,2n+1}\}, \\ -\infty & \text{otherwise.} \end{cases}$$

Fixing any $\boldsymbol{x}\in\mathcal{X}$, we recall that $p_i^k, \boldsymbol{z}^{k,i}, k\in[K], i\in[2n+1]$ returned by Algorithm 1 gives a worst-case distribution to Problem (9), and, at the same time, is an admissible probability distribution to Problem (11) because the two problems share the same ambiguity set. It follows that

$$\begin{aligned} V_{adapt} &= \min_{\boldsymbol{x}\in\mathcal{X}} \ \boldsymbol{a}^\top\boldsymbol{x} + \sup_{\mathbb{P}\in\mathcal{F}} \ \mathbb{E}_{\mathbb{P}}[g(\boldsymbol{x},\boldsymbol{z})] \\ &= \min_{\boldsymbol{x}\in\mathcal{X}} \ \boldsymbol{a}^\top\boldsymbol{x} + \max_{\boldsymbol{q}\in\mathcal{Q}} \sum_{k\in[K],i\in[2n+1]} q_k p_i^k g(\boldsymbol{x},\boldsymbol{z}^{k,i}) \\ &= \min_{\boldsymbol{x}\in\mathcal{X}} \ \boldsymbol{a}^\top\boldsymbol{x} + \max_{\boldsymbol{q}\in\mathcal{Q}} \sum_{k\in[K],i\in[2n+1]} q_k p_i^k g'(\boldsymbol{x},\boldsymbol{z}^{k,i}) \\ &\leq \min_{\boldsymbol{x}\in\mathcal{X}} \ \boldsymbol{a}^\top\boldsymbol{x} + \sup_{\mathbb{P}\in\mathcal{F}} \ \mathbb{E}_{\mathbb{P}}[g'(\boldsymbol{x},\boldsymbol{z})] = V_{relax}, \end{aligned}$$

where the second equality follows from Proposition 3, the third equality holds because g and g' have the same value whenever $\boldsymbol{z}\in\bigcup_{k\in[K]}\{\boldsymbol{z}^{k,1},\ldots,\boldsymbol{z}^{k,2n+1}\}$, and the inequality follows from the feasibility of the distribution characterized by $p_i^k, \boldsymbol{z}^{k,i}, i\in[2n+1]$.

Further, we observe that Problem (11) can be directly obtained from Problem (43) by imposing a restriction of linearity structure on $\boldsymbol{y}(k,\boldsymbol{z})$. This implies any feasible $\boldsymbol{\Theta}^k, \boldsymbol{\phi}^k$ to Problem (11) determines a function $\boldsymbol{y}(k,\boldsymbol{z})$ that is feasible to Problem (43). Hence, $V_{relax}\leq V_{ldr}$. We then conclude that $V_{adapt}\leq V_{relax}\leq V_{ldr}$.

We next show $V_{adapt}\geq V_{ldr}$. To this end, we construct a recourse decision rule that is feasible to Problem (11) and returns the optimal value of Problem (9).

We first consider the case of fixed scenario; for brevity, we remove the notation k (or $\tilde{k}$) that denotes realized (or random) scenarios. The construction is similar to the proof of Bertsimas and Goyal (2012, Theorem 1). Define auxiliary uncertain factors $\tilde{\boldsymbol{\omega}} = (\boldsymbol{\mu}-\tilde{\boldsymbol{z}})^+, \tilde{\boldsymbol{v}} = (\tilde{\boldsymbol{z}}-\boldsymbol{\mu})^+, \tilde{\boldsymbol{\zeta}} = (\tilde{\boldsymbol{\omega}},\tilde{\boldsymbol{v}})$, and let $\boldsymbol{\omega},\boldsymbol{v},\boldsymbol{\zeta}$ be the counterpart when $\tilde{\boldsymbol{z}}$ is realized as $\boldsymbol{z}$. Then $\tilde{\boldsymbol{z}} = \boldsymbol{\mu}-\tilde{\boldsymbol{\omega}}+\tilde{\boldsymbol{v}} = \boldsymbol{\mu}+[-\boldsymbol{I}_{n\times n}\ \boldsymbol{I}_{n\times n}]\,\tilde{\boldsymbol{\zeta}}$, $|\tilde{\boldsymbol{z}}-\boldsymbol{\mu}| = \tilde{\boldsymbol{\omega}}+\tilde{\boldsymbol{v}} = [\boldsymbol{I}_{n\times n}\ \boldsymbol{I}_{n\times n}]\,\tilde{\boldsymbol{\zeta}}$. Define

$$\boldsymbol{y}_{opt}(\boldsymbol{z}) = \boldsymbol{\Theta}_{opt}\begin{bmatrix}(\boldsymbol{\mu}-\boldsymbol{z})^+ \\ (\boldsymbol{z}-\boldsymbol{\mu})^+\end{bmatrix} + \boldsymbol{\phi}_{opt} = \boldsymbol{\Theta}_{opt}\boldsymbol{\zeta}+\boldsymbol{\phi}_{opt}.$$

For all $i\in[2n+1]$,

$$\begin{aligned} \boldsymbol{y}_{opt}(\boldsymbol{z}^i) &= \boldsymbol{y}_{opt}^{2n+1} + \boldsymbol{\Theta}_{opt}(\boldsymbol{\zeta}^i-\boldsymbol{\zeta}^{2n+1}) \\ &= \boldsymbol{y}_{opt}^{2n+1} + \begin{bmatrix}\boldsymbol{y}_{opt}^1-\boldsymbol{y}_{opt}^{2n+1} & \cdots & \boldsymbol{y}_{opt}^{2n}-\boldsymbol{y}_{opt}^{2n+1}\end{bmatrix}\boldsymbol{D}^{-1}(\boldsymbol{\zeta}^i-\boldsymbol{\zeta}^{2n+1}) \\ &= \boldsymbol{y}_{opt}^{2n+1} + \begin{bmatrix}\boldsymbol{y}_{opt}^1-\boldsymbol{y}_{opt}^{2n+1} & \cdots & \boldsymbol{y}_{opt}^{2n}-\boldsymbol{y}_{opt}^{2n+1}\end{bmatrix}\boldsymbol{e}_i \\ &= \boldsymbol{y}_{opt}^i, \end{aligned} \tag{44}$$

where the third last equality holds because $\boldsymbol{\zeta}^i - \boldsymbol{\zeta}^{2n+1} = \boldsymbol{D}\boldsymbol{e}_i$ for all $i \in [2n]$. We notice that $\boldsymbol{b}^\top \boldsymbol{y}_{opt}(\boldsymbol{z})$, as a linear combination of $(\boldsymbol{\mu}-\boldsymbol{z})^+$ and $(\boldsymbol{z}-\boldsymbol{\mu})^+$, is supermodular in $\boldsymbol{z}$ because it is separable. Now, utilizing the worst-case distribution given by Algorithm 1, we get

$$
\begin{aligned}
\sup_{\mathbb{P}\in\mathcal{F}} \mathbb{E}_{\mathbb{P}}\left[\boldsymbol{b}^\top \boldsymbol{y}_{opt}(\tilde{\boldsymbol{z}})\right] &= \sum_{i\in[2n+1]} p_i \boldsymbol{b}^\top \boldsymbol{y}_{opt}(\boldsymbol{z}^i) \\
&= \sum_{i\in[2n+1]} p_i \boldsymbol{b}^\top \boldsymbol{y}^i_{opt} \\
&= \sum_{i\in[2n+1]} p_i \min\left\{\boldsymbol{b}^\top \boldsymbol{y} \;\middle|\; \boldsymbol{W}\boldsymbol{x}_{opt} + \boldsymbol{U}\boldsymbol{y} \geq \boldsymbol{V}\boldsymbol{z}^i + \boldsymbol{v}^0\right\} \\
&= \sup_{\mathbb{P}\in\mathcal{F}} \mathbb{E}_{\mathbb{P}}\left[g(\boldsymbol{x}_{opt}, \tilde{\boldsymbol{z}})\right].
\end{aligned}
$$

The first and last equalities follow from the supermodularity of $\boldsymbol{b}^\top \boldsymbol{y}_{opt}(\boldsymbol{z})$ and $g(\boldsymbol{x}_{opt}, \boldsymbol{z})$ defined by (1), respectively. The second equality holds since $\boldsymbol{y}_{opt}(\boldsymbol{z}^i) = \boldsymbol{y}^i_{opt}, i \in [2n+1]$ (as shown in (44)), while the third one follows from the definition of $\boldsymbol{y}^i_{opt}$. It follows that the worst-case expected cost returned by $\boldsymbol{x}_{opt}, \boldsymbol{y}_{opt}(\boldsymbol{z})$ is the same as the optimal value of Problem (9). Further, we can observe easily that the solution $\boldsymbol{x}_{opt}, \boldsymbol{y}_{opt}(\boldsymbol{z})$ is feasible for Problem (11).

We next consider the case of uncertain scenarios. Following the above proof, we define $\boldsymbol{y}_{opt}(k, \boldsymbol{z})$ as

$$
\boldsymbol{y}_{opt}(k, \boldsymbol{z}) = \boldsymbol{\Theta}^k_{opt} \begin{bmatrix} (\boldsymbol{\mu}^k - \boldsymbol{z})^+ \\ (\boldsymbol{z} - \boldsymbol{\mu}^k)^+ \end{bmatrix} + \boldsymbol{\phi}^k_{opt}. \tag{45}
$$

It is supermodular in $\boldsymbol{z}$ and for any realized scenario k, $\sup_{\mathbb{P}^k\in\mathcal{F}^k} \mathbb{E}_{\mathbb{P}^k}\left[\boldsymbol{b}^\top \boldsymbol{y}_{opt}(k, \tilde{\boldsymbol{z}})\right] = \sup_{\mathbb{P}^k\in\mathcal{F}^k} \mathbb{E}_{\mathbb{P}^k}\left[\min\left\{\boldsymbol{b}^\top \boldsymbol{y} \;\middle|\; \boldsymbol{W}\boldsymbol{x}_{opt} + \boldsymbol{U}\boldsymbol{y} \geq \boldsymbol{V}\tilde{\boldsymbol{z}} + \boldsymbol{v}^0\right\}\right]$. Hence

$$
\begin{aligned}
&\boldsymbol{a}^\top \boldsymbol{x}_{opt} + \sup_{\mathbb{P}\in\mathcal{F}} \mathbb{E}_{\mathbb{P}}\left[\boldsymbol{b}^\top \boldsymbol{y}_{opt}(\tilde{k}, \tilde{\boldsymbol{z}})\right] \\
= \;& \boldsymbol{a}^\top \boldsymbol{x}_{opt} + \max_{\boldsymbol{q}\in\mathcal{Q}} \sum_{k\in[K]} q_k \sup_{\mathbb{P}^k\in\mathcal{F}^k} \mathbb{E}_{\mathbb{P}^k}\left[\boldsymbol{b}^\top \boldsymbol{y}_{opt}(k, \tilde{\boldsymbol{z}})\right] \\
= \;& \boldsymbol{a}^\top \boldsymbol{x}_{opt} + \max_{\boldsymbol{q}\in\mathcal{Q}} \sum_{k\in[K]} q_k \sup_{\mathbb{P}^k\in\mathcal{F}^k} \mathbb{E}_{\mathbb{P}^k}\left[\min\left\{\boldsymbol{b}^\top \boldsymbol{y} \;\middle|\; \boldsymbol{W}\boldsymbol{x}_{opt} + \boldsymbol{U}\boldsymbol{y} \geq \boldsymbol{V}\tilde{\boldsymbol{z}} + \boldsymbol{v}^0\right\}\right] \\
= \;& \boldsymbol{a}^\top \boldsymbol{x}_{opt} + \sup_{\mathbb{P}\in\mathcal{F}} \mathbb{E}_{\mathbb{P}}\left[g(\boldsymbol{x}_{opt}, \tilde{\boldsymbol{z}})\right] \\
= \;& V_{adapt}.
\end{aligned}
$$

Similar to Equation (44), we can check that $\boldsymbol{y}_{opt}(k, \boldsymbol{z}^{k,i}) \in \min\left\{\boldsymbol{b}^\top \boldsymbol{y} \;\middle|\; \boldsymbol{W}\boldsymbol{x}_{opt} + \boldsymbol{U}\boldsymbol{y} \geq \boldsymbol{V}\boldsymbol{z} + \boldsymbol{v}^0\right\}$ for all $k \in [K], i \in [2n+1]$. It follows that $\boldsymbol{x}_{opt}, \boldsymbol{\Theta}^k_{opt}, \boldsymbol{\phi}^k_{opt}, k \in [K]$ is a feasible solution to Problem (11).

Therefore, we can conclude that $V_{ldr} \leq V_{adapt}$. Hence, we have $V_{adapt} = V_{ldr}$ and $\boldsymbol{x}_{opt}, \boldsymbol{\Theta}^k_{opt}, \boldsymbol{\phi}^k_{opt}, k \in [K]$ is an optimal solution to Problem (11). □

C.11. Proof of Theorem 3

Following the proof for Proposition 6, we denote by $\boldsymbol{y}_{opt}(k, \boldsymbol{z})$ the linear decision rule defined by (45). To complete the proof for this theorem, based on Proposition 6, it suffices to show $\boldsymbol{y}_{opt}(k, \boldsymbol{z})$

is feasible, i.e., $(\boldsymbol{y}_{opt}(k,\boldsymbol{z}),\boldsymbol{z}) \in \mathcal{S}(\boldsymbol{x})$, for all $\boldsymbol{z} \in \bigcup_{k\in[K]} \mathcal{Z}^k$, where $\mathcal{Z}^k = \Pi_{i\in[n]}\{\underline{z}_i^k, \mu_i^k, \overline{z}_i^k\}$. We first prove the following claim and then show the feasibility by induction.

Claim. *Fix any scenario k. For all $\boldsymbol{z}', \boldsymbol{z}'' \in \mathcal{Z}^k$ with $\boldsymbol{z}^\wedge = \boldsymbol{z}' \wedge \boldsymbol{z}'', \boldsymbol{z}^\vee = \boldsymbol{z}' \vee \boldsymbol{z}''$, if $(\boldsymbol{y}_{opt}(k,\boldsymbol{z}^\wedge),\boldsymbol{z}^\wedge)$, $(\boldsymbol{y}_{opt}(k,\boldsymbol{z}^\vee),\boldsymbol{z}^\vee)$, $(\boldsymbol{y}_{opt}(k,\boldsymbol{z}'),\boldsymbol{z}') \in \mathcal{S}(\boldsymbol{x})$, then $(\boldsymbol{y}_{opt}(k,\boldsymbol{z}''),\boldsymbol{z}'')$ is also in $\mathcal{S}(\boldsymbol{x})$.*

Proof of Claim. Since the function $\boldsymbol{y}_{opt}(k,\boldsymbol{z})$ is separable in $\boldsymbol{z}$, hence, it is both supermodular and submodular in $\boldsymbol{z}$. Therefore, $\boldsymbol{y}_{opt}(k,\boldsymbol{z}') + \boldsymbol{y}_{opt}(k,\boldsymbol{z}'') = \boldsymbol{y}_{opt}(k,\boldsymbol{z}^\wedge) + \boldsymbol{y}_{opt}(k,\boldsymbol{z}^\vee)$, or equivalently, $\boldsymbol{y}_{opt}(k,\boldsymbol{z}'') = \boldsymbol{y}_{opt}(k,\boldsymbol{z}^\wedge) + \boldsymbol{y}_{opt}(k,\boldsymbol{z}^\vee) - \boldsymbol{y}_{opt}(k,\boldsymbol{z}')$. With the condition in the theorem satisfied, the claim follows directly.

For all $k \in [K], i \in [2n+1]$, recall that by the proof for Proposition 6, $\boldsymbol{y}_{opt}(k,\boldsymbol{z}^{k,i})$ is the optimal second-stage decision when the uncertainty is realized as $\boldsymbol{z}^{k,i}$. This implies $\boldsymbol{y}_{opt}(k,\boldsymbol{z})$ is feasible for all $\boldsymbol{z} \in \{\boldsymbol{z}^{k,i} \mid k \in [K], i \in [2n+1]\}$. Fix any scenario k, define $\mathcal{Z}^{k,i} = \left\{\boldsymbol{z} \in \mathcal{Z}^k \;\middle|\; \boldsymbol{z} \le \boldsymbol{z}^{k,i}\right\}$ for all $i \in [2n+1]$. Observe that $\mathcal{Z}^{k,1} = \{\boldsymbol{z}^{k,1}\}$, we know $\boldsymbol{y}_{opt}(k,\boldsymbol{z})$ is feasible on $\mathcal{Z}^{k,i}$ when $i=1$. Next we inductively show the feasibility on $\mathcal{Z}^{k,i}$ for all $i \in [2n+1]$. Specifically, we assume the statement holds for a given $i \in [2n+1]$, and consider for the set $\mathcal{Z}^{k,i+1}$. By definition, $\boldsymbol{z}^{k,i+1}$ deviates from $\boldsymbol{z}^{k,i}$ in only one dimension, i.e., $z_l^{k,i+1} > z_l^{k,i}$ for some $l \in [n]$ and $z_{l'}^{k,i+1} = z_{l'}^{k,i}$ for all $l' \neq l$. By assumption, it suffices to prove the feasibility for any $\hat{\boldsymbol{z}} \in \mathcal{Z}^{k,i+1} \setminus \mathcal{Z}^{k,i}$. In this case, we have $\hat{z}_l = z_l^{k,i+1}$ and $\hat{z}_{l'} \le z_{l'}^{k,i}$ for all $l' \neq l$, i.e., $\hat{\boldsymbol{z}} = (\hat{z}_1,\ldots,\hat{z}_n) = (\hat{z}_1,\ldots,\hat{z}_{l-1}, z_l^{k,i+1}, \hat{z}_{l+1},\ldots,\hat{z}_n)$. Choosing $\boldsymbol{z}' = \boldsymbol{z}^{k,i} = (z_1^{k,i},\ldots,z_{l-1}^{k,i}, z_l^{k,i}, z_{l+1}^{k,i},\ldots,z_n^{k,i})$ and $\boldsymbol{z}'' = \hat{\boldsymbol{z}}$, we obtain

$$
\begin{aligned}
\boldsymbol{z}^\wedge &= \boldsymbol{z}' \wedge \boldsymbol{z}'' = (\hat{z}_1,\ldots,\hat{z}_{l-1}, z_l^{k,i}, \hat{z}_{l+1},\ldots,\hat{z}_n) \in \mathcal{Z}^{k,i},\\
\boldsymbol{z}^\vee &= \boldsymbol{z}' \vee \boldsymbol{z}'' = (z_1^{k,i},\ldots,z_{l-1}^{k,i}, z_l^{k,i+1}, z_{l+1}^{k,i},\ldots,z_n^{k,i}) = \boldsymbol{z}^{k,i+1}.
\end{aligned}
$$

Since $\boldsymbol{y}_{opt}(k,\boldsymbol{z})$ is feasible when $\boldsymbol{z} = \boldsymbol{z}', \boldsymbol{z}^\wedge, \boldsymbol{z}^\vee$, by the Claim we conclude that $\boldsymbol{y}_{opt}(k,\boldsymbol{z}'') = \boldsymbol{y}_{opt}(k,\hat{\boldsymbol{z}})$ is also feasible. Notice that $\mathcal{Z}^{k,i} = \mathcal{Z}^k$ when $i = 2n+1$, and the same proof goes for any $k \in [K]$, we complete the proof. □

C.12. Proof of Proposition 7

It is obvious that $-\boldsymbol{r}^\top \boldsymbol{x} + (\boldsymbol{r} - \boldsymbol{s})(\boldsymbol{x} - \boldsymbol{z})^+$ is decreasing and supermodular in $\boldsymbol{z}$. Thus, we can apply Proposition 12. Specifically, by substituting $\boldsymbol{a} = -\boldsymbol{r}$ and $g(\boldsymbol{x},\boldsymbol{z}) = (\boldsymbol{r}-\boldsymbol{s})(\boldsymbol{x}-\boldsymbol{z})^+ = \min\left\{(\boldsymbol{r}-\boldsymbol{s})^\top \boldsymbol{y} \;\middle|\; \boldsymbol{y} \ge \boldsymbol{x} - \boldsymbol{z}, \boldsymbol{y} \ge \boldsymbol{0}\right\}$ in the formulation (29), we obtain the following reformulation for Problem (13),

$$\begin{aligned}
\min \quad & \boldsymbol{\nu}^\top \boldsymbol{l} & \\
\text{s.t.} \quad & \boldsymbol{R}_k^\top \boldsymbol{l} \geq \sum_{i\in[2n+1]} p_i^k f^{k,i}, & k \in [K] \\
& \boldsymbol{y}^{k,i} \geq \boldsymbol{x} - \boldsymbol{z}^{k,i}, & k \in [K], i \in [2n+1] \\
& \boldsymbol{y}^{k,i} \geq \boldsymbol{0}, & k \in [K], i \in [2n+1] \\
& f^{k,i} \geq c_j \left(-\boldsymbol{r}^\top \boldsymbol{x} + (\boldsymbol{r}-\boldsymbol{s})^\top \boldsymbol{y}^{k,i}\right) + d_j, & k \in [K], i \in [2n+1], j \in [J], \\
& \boldsymbol{l} \geq \boldsymbol{0},\ \boldsymbol{x} \in \mathcal{X}^{news}, &
\end{aligned}$$

where $p_i^k, \boldsymbol{z}^{k,i}, k \in [K], i \in [2n+1]$ are the output of Algorithm 1 given the ambiguity sets $\mathcal{F}^k, k \in [K]$ defined by Equation (4). □

C.13. Proof of Proposition 8

Notice that (14) is a special case of the formulation (30) with $u(w) = \frac{1}{\rho} w^+$, $\boldsymbol{a} = -\boldsymbol{r}$ and $g(\boldsymbol{x}, \boldsymbol{z}) = \min\left\{(\boldsymbol{r}-\boldsymbol{s})^\top \boldsymbol{y} \;\middle|\; \boldsymbol{y} \geq \boldsymbol{x} - \boldsymbol{z}, \boldsymbol{y} \geq \boldsymbol{0}\right\}$. Hence, a direct application of Corollary 3 gives the following reformulation,

$$\begin{aligned}
\min \quad & \theta + \boldsymbol{\nu}^\top \boldsymbol{l} & \\
\text{s.t.} \quad & \boldsymbol{R}_k^\top \boldsymbol{l} \geq \sum_{i\in[2n+1]} \frac{1}{\rho} \cdot p_i^k f^{k,i}, & k \in [K] \\
& \boldsymbol{y}^{k,i} \geq \boldsymbol{x} - \boldsymbol{z}^{k,i}, & k \in [K], i \in [2n+1] \\
& \boldsymbol{y}^{k,i} \geq \boldsymbol{0}, & k \in [K], i \in [2n+1] \\
& f^{k,i} \geq -\boldsymbol{r}^\top \boldsymbol{x} + (\boldsymbol{r}-\boldsymbol{s})^\top \boldsymbol{y}^{k,i} - \theta, & k \in [K], i \in [2n+1] \\
& f^{k,i} \geq 0, & k \in [K], i \in [2n+1] \\
& \boldsymbol{l} \geq \boldsymbol{0},\ \boldsymbol{x} \in \mathcal{X}^{news}, &
\end{aligned}$$

where $p_i^k, \boldsymbol{z}^{k,i}, k \in [K], i \in [2n+1]$ are the output of Algorithm 1 given the ambiguity sets $\mathcal{F}^k, k \in [K]$ defined by Equation (4). □

C.14. Proof of Proposition 9

We prove the supermodularity by showing that Problem (15) satisfies the conditions in Theorem 2. We first reformulate Problem (15) as the sum of m sub-problems. Denote

$$g^j(\boldsymbol{x}, \boldsymbol{z}) = \min\left\{ \sum_{i\in[n]} c_{ij} y_{ij} \;\middle|\; \begin{array}{l} \sum_{i\in[n]} y_{ij} = 1 \\ 0 \leq y_{ij} \leq x_i z_i,\ i \in [n] \end{array} \right\}.$$

Then it can be verified that $g(\boldsymbol{x}, \boldsymbol{z}) = \sum_{j\in[m]} g^j(\boldsymbol{x}, \boldsymbol{z})$. Hence, it suffices to prove the supermodularity of $g^j(\boldsymbol{x}, \boldsymbol{z})$ for all $j \in [m]$. Observing that $\boldsymbol{x} \in \{0,1\}^n$, we denote $S = \{i \in [n] \mid x_i = 1\}$ and $T =$

$[n]\backslash S$. It follows that the constraints in defining $g^j(\boldsymbol{x},\boldsymbol{z})$ can be reformulated as $\boldsymbol{U}(y_{1j},\ldots,y_{nj}) - \boldsymbol{V}\boldsymbol{z} \geq \boldsymbol{v}^0$, where

$$\boldsymbol{U} = \begin{bmatrix} \boldsymbol{1}^\top \\ -\boldsymbol{1}^\top \\ \boldsymbol{I}_{n\times n} \\ -\boldsymbol{I}_T \\ -\boldsymbol{I}_S \end{bmatrix} \in \Re^{(2+2n)\times n},\ \boldsymbol{V} = \begin{bmatrix} \boldsymbol{0}_{(2+n+|T|)\times n} \\ -\boldsymbol{I}_S \end{bmatrix} \in \Re^{(2+2n)\times n},\ \boldsymbol{v}^0 = \begin{bmatrix} 1 \\ -1 \\ \boldsymbol{0}_{2n\times 1} \end{bmatrix} \in \Re^{2+2n}.$$

Here $\boldsymbol{I}_T, \boldsymbol{I}_S$ are the submatrices of $\boldsymbol{I}_{n\times n}$ consisting of rows which are indexed by elements in T, S, separately. Note that $\text{rank}(\boldsymbol{U}) = n < 2n+2$. We hence can apply Theorem 2 to prove the supermodularity of g^j. To this end, we consider any index set $\mathcal{I}$ such that $|\mathcal{I}| = n+1$ and $\text{rank}(\boldsymbol{U}_\mathcal{I}) = n$, any $\boldsymbol{\beta} \geq \boldsymbol{0} \in \Re^n, \boldsymbol{\alpha} \in \Re^n$ such that

$$\boldsymbol{V}_\mathcal{I}\boldsymbol{\beta} = \boldsymbol{U}_\mathcal{I}\boldsymbol{\alpha}. \tag{46}$$

Consider any $j \in [n]$, we need to show $\beta_j(\boldsymbol{V}_\mathcal{I})_j \in \text{span}(\boldsymbol{U}_\mathcal{I})$. If $\beta_j(\boldsymbol{V}_\mathcal{I})_j = \boldsymbol{0}$, the result is straightforward. We now consider only the case that $\beta_j > 0$ and $(\boldsymbol{V}_\mathcal{I})_j \neq \boldsymbol{0}$.

By $(\boldsymbol{V}_\mathcal{I})_j \neq \boldsymbol{0}$, we have $\boldsymbol{V}_j \neq \boldsymbol{0}$. Based on the structure of $\boldsymbol{V}$, $\boldsymbol{V}_j$ has only one nonzero element and indeed, there exists a unique i such that $\boldsymbol{V}_j = -\boldsymbol{e}_i \in \Re^{2n+2}$. Moreover, $i \in \mathcal{I}$, $i > 2+n+|T|$, $V_{ij} = -1$ is the only nonzero element in the ith row, i.e., $\boldsymbol{v}_i = -\boldsymbol{e}_j \in \Re^n$. Therefore, by (46), we have $\boldsymbol{u}_i^\top\boldsymbol{\alpha} = \boldsymbol{v}_i^\top\boldsymbol{\beta} = -\beta_j$, implying $\alpha_j = \beta_j$ since $U_{ij} = -1$ is also the only nonzero element in $\boldsymbol{u}_i$. We now show that for $\boldsymbol{U}_j$, only zero element from blocks $\boldsymbol{I}_{n\times n}$ and $-\boldsymbol{I}_T$ are included in $(\boldsymbol{U}_\mathcal{I})_j$. Assume to the contrary, i.e., there is a $k \in \{3,\ldots,2+n+|T|\} \cap \mathcal{I}$ with $U_{kj} \neq 0$. Note that U_{kj} is the only nonzero element in $\boldsymbol{u}_k$. Hence, $\boldsymbol{u}_k^\top\boldsymbol{\alpha} = U_{kj}\alpha_j \neq 0$, $\boldsymbol{v}_k^\top\boldsymbol{\beta} = \boldsymbol{0}^\top\boldsymbol{\beta} = 0$, contradicts with (46) and $k \in \mathcal{I}$. Therefore, $U_{kj} = 0$ for all $k \in \{3,\ldots,2+n+|T|\} \cap \mathcal{I}$. Now we consider two scenarios.

In the first scenario, $\{1,2\} \cap \mathcal{I} = \emptyset$, then $(\boldsymbol{U}_\mathcal{I})_j$ has the only one nonzero element which is from $-\boldsymbol{I}_S$, $(\boldsymbol{U}_\mathcal{I})_j = (\boldsymbol{V}_\mathcal{I})_j$, and hence $\beta_j(\boldsymbol{V}_\mathcal{I})_j \in \text{span}(\boldsymbol{U}_\mathcal{I})$.

In the second scenario, $\{1,2\} \cap \mathcal{I} \neq \emptyset$. WLOG, let $1 \in \mathcal{I}$. We then have $\sum_{k\in[n]}\alpha_k = \boldsymbol{1}^\top\boldsymbol{\alpha} = \boldsymbol{u}_1^\top\boldsymbol{\alpha} = \boldsymbol{v}_1^\top\boldsymbol{\beta} = \boldsymbol{0}^\top\boldsymbol{\beta} = 0$, where the third equality is due to (46). From the above analysis, we already have $\alpha_j = \beta_j > 0$, which implies that there exists $k \neq j$ such that $\alpha_k < 0$. We now prove that for $\boldsymbol{U}_k$, only zero elements from blocks $\boldsymbol{I}_{n\times n}$ and $-\boldsymbol{I}_T$ are included in $(\boldsymbol{U}_\mathcal{I})_k$. This can be done with the same logic as that in above when we show only zero elements from blocks $\boldsymbol{I}_{n\times n}$ and $-\boldsymbol{I}_T$ are included in $(\boldsymbol{U}_\mathcal{I})_j$. We next show that for $\boldsymbol{U}_k$, only zero elements from blocks $-\boldsymbol{I}_S$ is included in $(\boldsymbol{U}_\mathcal{I})_k$. Assume to the contrary, i.e., there is an $l \in \{2+n+|T|+1,\ldots,2+2n\} \cap \mathcal{I}$ such that $u_{lk} \neq 0$. Notice that $u_{lk} = -1$ and $v_{lk} = -1$ are the only nonzero elements in $\boldsymbol{u}_l$ and $\boldsymbol{v}_l$, respectively. We have $\boldsymbol{u}_l^\top\boldsymbol{\alpha} = -\alpha_k$, $\boldsymbol{v}_l^\top\boldsymbol{\alpha} = -\beta_k$, and hence $\boldsymbol{u}_l^\top\boldsymbol{\alpha} \neq \boldsymbol{v}_l^\top\boldsymbol{\alpha}$ since $\alpha_k < 0$ and $\boldsymbol{\beta} \geq \boldsymbol{0}$. It contradicts with (46), and we have that only zero elements from blocks $-\boldsymbol{I}_S$ are included in $(\boldsymbol{U}_\mathcal{I})_k$. Therefore, from the two observations above we can conclude that $(\boldsymbol{U}_\mathcal{I})_k$ has all elements as zero from the blocks $\boldsymbol{I}_{n\times n}$, $-\boldsymbol{I}_T$ and $-\boldsymbol{I}_S$. In other words, $(\boldsymbol{U}_\mathcal{I})_k$ and $(\boldsymbol{U}_\mathcal{I})_j$ only differs at $u_{ik} = 0$, $u_{ij} = -1$. We can then easily have $\beta_j(\boldsymbol{V}_\mathcal{I})_j = \beta_j(\boldsymbol{U}_\mathcal{I})_j - \beta_j(\boldsymbol{U}_\mathcal{I})_k \in \text{span}(\boldsymbol{U}_\mathcal{I})$. $\square$

C.15. Proof of Proposition 10

Denote $\boldsymbol{y}=(y_{11},\ldots,y_{n1},\ \ldots,\ y_{1n},\ldots,y_{nn})\in\Re^{n^2}$. Then the second-stage problem can be expressed as

$$\begin{aligned} g(\boldsymbol{x},\boldsymbol{z})=\ \min\ & \sum_{s,j\in[n]} b_{sj}y_{sj} \\ \text{s.t.}\ & \begin{bmatrix} \boldsymbol{U}^1 & \cdots & \boldsymbol{U}^n \\ & \boldsymbol{I}_{n^2\times n^2} & \end{bmatrix}\boldsymbol{y}-\begin{bmatrix}\boldsymbol{I}_{n\times n}\\ \boldsymbol{0}_{n^2\times n}\end{bmatrix}\boldsymbol{z}\geq\begin{bmatrix}-\boldsymbol{x}\\ \boldsymbol{0}_{n^2\times 1}\end{bmatrix}. \end{aligned}$$

For any $s\in[n]$, the matrix $\boldsymbol{U}^s\in\Re^{n\times n}$ has $\boldsymbol{e}_s-\boldsymbol{e}_j$ as its j-th column, $j\in[n]$. Denote $\boldsymbol{U}^0=[\boldsymbol{U}^1\ \cdots\ \boldsymbol{U}^n]$, $\boldsymbol{V}^0=\boldsymbol{I}_{n\times n}$ and $\boldsymbol{U}=\begin{bmatrix}\boldsymbol{U}^0\\ \boldsymbol{I}_{n^2\times n^2}\end{bmatrix}\in\Re^{(n+n^2)\times n^2}$, $\boldsymbol{V}=\begin{bmatrix}\boldsymbol{V}^0\\ \boldsymbol{0}_{n^2\times n}\end{bmatrix}$. Obviously, $\text{rank}(\boldsymbol{U})=n^2$ which is less than the number of rows in $\boldsymbol{U}$. Therefore, to complete the proof, we now show that $\boldsymbol{U},\boldsymbol{V}$ meet the second condition in Theorem 2.

Consider any index set $\mathcal{I}$ such that $|\mathcal{I}|=\text{rank}(\boldsymbol{U})+1=n^2+1,\text{rank}(\boldsymbol{U}_{\mathcal{I}})=n^2$. Denote $\mathcal{I}_0=\mathcal{I}\cap[n],\mathcal{I}_1=\mathcal{I}\backslash\mathcal{I}_0$, then the rows of $\boldsymbol{U}_{\mathcal{I}_0}$ (or $\boldsymbol{V}_{\mathcal{I}_0}$) are extracted from $\boldsymbol{U}^0$ (or $\boldsymbol{V}^0$); the rows of $\boldsymbol{U}_{\mathcal{I}_1}$ (or $\boldsymbol{V}_{\mathcal{I}_1}$) are extracted from $\boldsymbol{I}_{n^2\times n^2}$ (or $\boldsymbol{0}_{n^2\times n}$).

We first let the column index set $\mathcal{J}_0$ be such that the submatrix $\boldsymbol{U}_{\mathcal{I}_0,\mathcal{J}_0}=\boldsymbol{0}$, and let $\mathcal{J}_1=[n^2]\setminus\mathcal{J}_0$. Hence, $\boldsymbol{U}_{\mathcal{I}}$ can be decomposed into four submatrices $\boldsymbol{U}_{\mathcal{I}_0,\mathcal{J}_0},\boldsymbol{U}_{\mathcal{I}_0,\mathcal{J}_1},\boldsymbol{U}_{\mathcal{I}_1,\mathcal{J}_0},\boldsymbol{U}_{\mathcal{I}_1,\mathcal{J}_1}$. Recalling that $\boldsymbol{U}_{\mathcal{I}_1}$ is a submatrix of $\boldsymbol{I}$, there is exactly one entry being one in its each row, and at most one entry being 1 in its each column. Hence, in $\boldsymbol{U}_{\mathcal{I}_1}$, the number of columns being $\boldsymbol{0}$ is $n^2-|\mathcal{I}_1|=n^2-(|\mathcal{I}|-|\mathcal{I}_0|)=|\mathcal{I}_0|-1$. Noticing $\boldsymbol{U}_{\mathcal{I}}$ is full column rank and $\boldsymbol{U}_{\mathcal{I}_0,\mathcal{J}_0}=\boldsymbol{0}$, all of the $|\mathcal{I}_0|-1$ zero columns in $\boldsymbol{U}_{\mathcal{I}_1}$ must be in $\boldsymbol{U}_{\mathcal{I}_1,\mathcal{J}_1}$. Denote the index set $\mathcal{K}_1$ as the set of column index for those zero columns in $\boldsymbol{U}_{\mathcal{I}_1}$, and $\mathcal{K}_2=\mathcal{J}_1\setminus\mathcal{K}_1$. Then $\boldsymbol{U}_{\mathcal{I}_1,\mathcal{J}_1}$ can be further decomposed into two submatrices $\boldsymbol{U}_{\mathcal{I}_1,\mathcal{K}_1},\boldsymbol{U}_{\mathcal{I}_1,\mathcal{K}_2}$ where $\boldsymbol{U}_{\mathcal{I}_1,\mathcal{K}_1}=\boldsymbol{0}_{|\mathcal{I}_1|\times(|\mathcal{I}_0|-1)}$.

Since $\boldsymbol{U}_{\mathcal{I}_0,\mathcal{K}_1}\in\Re^{|\mathcal{I}_0|\times(|\mathcal{I}_0|-1)}$ and it is of full column rank (otherwise it contradicts with $\boldsymbol{U}_{\mathcal{I}}$ being full column rank and $\boldsymbol{U}_{\mathcal{I}_1,\mathcal{K}_1}=\boldsymbol{0}$), we have that $\text{null}(\boldsymbol{U}^\top_{\mathcal{I}_0,\mathcal{K}_1})$ is of dimension 1. Recalling that $\boldsymbol{U}_{\mathcal{I}_0,\mathcal{K}_1}$ is a submatrix of $\boldsymbol{U}^0$, each column can only be either $\pm\boldsymbol{e}_s$ or $\boldsymbol{e}_{s_1}-\boldsymbol{e}_{s_2}$ for some $s,s_1,s_2\in[|\mathcal{I}_0|]$. Let $\mathcal{N}_j\subseteq[|\mathcal{I}_0|]$ be the index set $\left\{s\mid \text{the } s\text{-th entry of } (\boldsymbol{U}_{\mathcal{I}_0})_j \text{ is non-zero}\right\}$ for any $j\in\mathcal{K}_1$. We observe that

$$\text{null}\left(\boldsymbol{U}^\top_{\mathcal{I}_0,\mathcal{K}_1}\right)=\left\{\boldsymbol{\gamma}\ \middle|\ \begin{array}{ll}\forall j\in\mathcal{K}_1 \text{ with } |\mathcal{N}_j|=1:\gamma_s=0 & \text{for } s\in\mathcal{N}_j\\ \forall j\in\mathcal{K}_1 \text{ with } |\mathcal{N}_j|=2:\gamma_{s_1}=\gamma_{s_2} & \text{for } s_1,s_2\in\mathcal{N}_j\end{array}\right\}. \tag{47}$$

Consider any nonzero $\boldsymbol{\gamma}\in\text{null}\left(\boldsymbol{U}^\top_{\mathcal{I}_0,\mathcal{K}_1}\right)$, we now prove that there is no s_1,s_2 such that $\gamma_{s_1},\gamma_{s_2}$ are both nonzero and $\gamma_{s_1}\neq\gamma_{s_2}$. Assume to the contrary, i.e., we can find s_1,s_2 such that $\gamma_{s_1}\gamma_{s_2}\neq0$ and $\gamma_{s_1}\neq\gamma_{s_2}$. We construct a vector $\hat{\boldsymbol{\gamma}}$ such that $\hat{\gamma}_i=0$ for all i such that $\gamma_i=0$, and $\hat{\gamma}_i=1$ for all i such that $\gamma_i\neq0$. As $\boldsymbol{\gamma}$ satisfies the condition in (47), so does $\hat{\boldsymbol{\gamma}}$, and hence $\hat{\boldsymbol{\gamma}}\in\text{null}\left(\boldsymbol{U}^\top_{\mathcal{I}_0,\mathcal{K}_1}\right)$. Nevertheless, $\boldsymbol{\gamma}$ and $\hat{\boldsymbol{\gamma}}$ are obviously linearly independent, and hence we have contradiction to that $\text{null}\left(\boldsymbol{U}^\top_{\mathcal{I}_0,\mathcal{K}_1}\right)$ is of dimension 1. Therefore, we can conclude that all nonzero elements in $\boldsymbol{\gamma}$ have the same value.

Consider any $\boldsymbol{\eta}^0 \in \Re^{|\mathcal{I}_0|}, \boldsymbol{\eta}^1 \in \Re^{|\mathcal{I}_1|}, \boldsymbol{\eta} = (\boldsymbol{\eta}^0, \boldsymbol{\eta}^1)$ such that $\boldsymbol{U}_{\mathcal{I}}^\top \boldsymbol{\eta} = \mathbf{0}$. It implies $\mathbf{0} = \boldsymbol{U}_{\mathcal{I},\mathcal{K}_1}^\top \boldsymbol{\eta} = \boldsymbol{U}_{\mathcal{I}_0,\mathcal{K}_1}^\top \boldsymbol{\eta}^0 + \boldsymbol{U}_{\mathcal{I}_1,\mathcal{K}_1}^\top \boldsymbol{\eta}^1 = \boldsymbol{U}_{\mathcal{I}_0,\mathcal{K}_1}^\top \boldsymbol{\eta}^0$, where the last equality is due to $\boldsymbol{U}_{\mathcal{I}_1,\mathcal{K}_1} = \mathbf{0}$. Hence, $\boldsymbol{\eta}^0 \in \text{null}\left(\boldsymbol{U}_{\mathcal{I}_0,\mathcal{K}_1}^\top\right)$, whose dimension has been shown as 1. Therefore, $\boldsymbol{\eta}^0 = k\boldsymbol{\gamma}$ for some $k \in \Re$. As we have shown above, all nonzero elements in $\boldsymbol{\gamma}$ are equal, WLOG, we can have $\boldsymbol{\eta}^0 \geq \mathbf{0}$. We are now ready to verify the second condition in Theorem 2.

Given any $\boldsymbol{\beta} \in \Re_+^n$ with $\boldsymbol{V}_{\mathcal{I}}\boldsymbol{\beta} \in \text{span}(\boldsymbol{U}_{\mathcal{I}})$, as $\boldsymbol{\eta} \in \text{null}(\boldsymbol{U}_{\mathcal{I}}^\top)$, we have $0 = \boldsymbol{\eta}^\top \boldsymbol{V}_{\mathcal{I}}\boldsymbol{\beta} = (\boldsymbol{\eta}^0)^\top \boldsymbol{V}_{\mathcal{I}_0}\boldsymbol{\beta} + (\boldsymbol{\eta}^1)^\top \boldsymbol{V}_{\mathcal{I}_1}\boldsymbol{\beta} = (\boldsymbol{\eta}^0)^\top \boldsymbol{V}_{\mathcal{I}_0}\boldsymbol{\beta} = \sum_{i\in[n]} \beta_i (\boldsymbol{\eta}^0)^\top (\boldsymbol{V}_{\mathcal{I}_0})_i$, where the third equality is due to $\boldsymbol{V}_{\mathcal{I}_1} = \mathbf{0}$. Since $\boldsymbol{\eta}^0 \geq \mathbf{0}, \boldsymbol{V}_{\mathcal{I}_0} \geq \mathbf{0}, \boldsymbol{\beta} \geq \mathbf{0}$, we have that $\boldsymbol{\eta}^\top \beta_i (\boldsymbol{V}_{\mathcal{I}})_i = \beta_i (\boldsymbol{\eta}^0)^\top (\boldsymbol{V}_{\mathcal{I}_0})_i = 0$ for all $i \in [n]$. Recall that $\text{null}(\boldsymbol{U}_{\mathcal{I}}^\top)$ is of dimension 1, we then have $\beta_i (\boldsymbol{V}_{\mathcal{I}})_i \in \text{span}(\boldsymbol{U}_{\mathcal{I}})$, and the second condition in Theorem 2 is satisfied. Thus $g(\boldsymbol{x}, \boldsymbol{z})$ is supermodular in $\boldsymbol{z}$ for all $\boldsymbol{x}$, and we obtain the following reformulation as a simple corollary of Theorem 1,

$$
\begin{aligned}
\min \quad & \boldsymbol{a}^\top \boldsymbol{x} + \boldsymbol{\nu}^\top \boldsymbol{l} \\
\text{s.t.} \quad & \boldsymbol{R}_k^\top \boldsymbol{l} \geq \sum_{i\in[2n+1]} p_i^k \sum_{s,j\in[n]} b_{sj} y_{sj}^{k,i}, && k \in [K] \\
& \sum_{j\in[n]} y_{js}^{k,i} - \sum_{j\in[n]} y_{sj}^{k,i} \geq z_s^{k,i} - x_s, && s \in [n], k \in [K], i \in [2n+1] \\
& y_{sj}^{k,i} \geq 0, && s \in [n], j \in [n], k \in [K], i \in [2n+1] \\
& \boldsymbol{l} \geq \mathbf{0}, \ \boldsymbol{x} \in \mathcal{X}^{lot},
\end{aligned}
$$

where $p_i^k, \boldsymbol{z}^{k,i}, k \in [K], i \in [2n+1]$ are the output of Algorithm 1 given the ambiguity sets $\mathcal{F}^k, k \in [K]$ defined by Equation (4). □

C.16. Proof of Proposition 11

Let $\hat{\boldsymbol{z}} \in \Re^n$ be such that $\hat{z}_i = \xi_i z_i$ for all $i \in [n]$, and we define $\hat{g}(\boldsymbol{x}, \hat{\boldsymbol{z}}) = \min\left\{ \mathbf{1}^\top \boldsymbol{y} \;\middle|\; \begin{array}{ll} y_t \geq \sum_{s=j}^t (\hat{z}_s - x_s), & j \in [t], t \in [n] \\ y_t \geq 0, & t \in [n] \end{array} \right\}$. Notice that $\hat{g}(\boldsymbol{x}, \hat{\boldsymbol{z}}) = \hat{g}(\boldsymbol{x}, (\xi_1 z_1, \ldots, \xi_n z_n)) = g(\boldsymbol{x}, \boldsymbol{\xi}, \boldsymbol{z})$ defined by Equation

$$
g(\boldsymbol{x}, \boldsymbol{\xi}, \boldsymbol{z}) = \min\left\{ \mathbf{1}^\top \boldsymbol{y} \;\middle|\; \begin{array}{ll} y_t \geq \sum_{s=j}^t (\xi_s z_s - x_s), & j \in [t], t \in [n] \\ y_t \geq 0, & t \in [n] \end{array} \right\}. \tag{48}
$$

To prove the supermodularity of g, we first show $\hat{g}(\boldsymbol{x}, \hat{\boldsymbol{z}})$ is supermodular in $\hat{\boldsymbol{z}}$, and then prove that $g(\boldsymbol{x}, \boldsymbol{\xi}, \boldsymbol{z}) = \hat{g}(\boldsymbol{x}, (\xi_1 z_1, \ldots, \xi_n z_n))$ is supermodular in $(\xi_1, \ldots, \xi_n, z_1, \ldots, z_n)$.

To show the supermodularity of $\hat{g}$, we first rewrite the problem defining $\hat{g}$ in its matrix form, i.e., $\hat{g}(\boldsymbol{x}, \hat{\boldsymbol{z}}) = \min\left\{ \mathbf{1}^\top \boldsymbol{y} \;\middle|\; \boldsymbol{U}\boldsymbol{y} - \boldsymbol{V}\hat{\boldsymbol{z}} \geq -\boldsymbol{W}\boldsymbol{x} \right\}$, where $\boldsymbol{U} = \begin{bmatrix} \bar{\boldsymbol{U}}^1 \\ \vdots \\ \bar{\boldsymbol{U}}^n \\ \bar{\boldsymbol{U}}^{n+1} \end{bmatrix} \in \Re^{\frac{n^2+3n}{2} \times n}$, $\boldsymbol{V} = \boldsymbol{W} = \begin{bmatrix} \bar{\boldsymbol{V}}^1 \\ \vdots \\ \bar{\boldsymbol{V}}^n \\ \bar{\boldsymbol{V}}^{n+1} \end{bmatrix} \in \Re^{\frac{n^2+3n}{2} \times n}$ are such that

$$
\bar{\boldsymbol{U}}^t \in \Re^{t \times n} \text{ are with elements of } \bar{u}_{js}^t = \begin{cases} 1 & \text{if } s = t \\ 0 & \text{otherwise} \end{cases} \text{ for } j \in [t], \ s, t \in [n],
$$

$$\bar{\boldsymbol{V}}^t \in \Re^{t\times n} \text{ are with elements of } \bar{v}^t_{js} = \begin{cases} 1 & \text{if } j \le s \le t \\ 0 & \text{otherwise} \end{cases} \text{ for } j\in[t],\ s,t\in[n],$$

$$\bar{\boldsymbol{U}}^{n+1} = \boldsymbol{I}_{n\times n},\ \bar{\boldsymbol{V}}^{n+1} = \boldsymbol{0}_{n\times n}.$$

We prove $\hat{g}(\boldsymbol{x},\hat{\boldsymbol{z}})$ is supermodular in $\hat{\boldsymbol{z}}$ by verify that $\boldsymbol{U},\boldsymbol{V}$ satisfy the condition in Theorem 2. To this end, consider any $\mathcal{I}\subseteq[(n^2+3n)/2], \boldsymbol{\beta}\in\Re^n_+$ with $|\mathcal{I}|=n+1$, $\text{rank}(\boldsymbol{U}_\mathcal{I})=n$, and $\boldsymbol{V}_\mathcal{I}\boldsymbol{\beta}\in \text{span}(\boldsymbol{U}_\mathcal{I})$. Note that $\text{rank}(\boldsymbol{U})=n$, and each row of $\boldsymbol{U}_\mathcal{I}\in\Re^{(n+1)\times n}$ has only one nonzero element which takes the value of 1. Therefore, there exists $\omega\in[n]$ such that $\boldsymbol{U}$ has two row vectors being $\boldsymbol{e}_\omega$, and exactly one row vector being $\boldsymbol{e}_i$ for each $i\in[n]\setminus\{\omega\}$. WLOG, we let $RI_1,\ldots,RI_{n+1}$ be the distinct row indices such that $\mathcal{I}=\{RI_1,\ldots,RI_{n+1}\}$, $\boldsymbol{u}_{RI_i}=\boldsymbol{e}_i$ for all $i\in[n]$, $\boldsymbol{u}_{RI_{n+1}}=\boldsymbol{e}_\omega$, and $RI_\omega<RI_{n+1}$. Moreover, for notational brevity, we arrange the rows in $\boldsymbol{U}_I$, $\boldsymbol{V}_I$ with the order of $RI_1,\ldots,RI_{n+1}$, which would not change the satisfaction/violation of the condition in Theorem 2. Therefore, $\boldsymbol{U}_\mathcal{I}=\begin{bmatrix}\boldsymbol{I}\\ \boldsymbol{e}_\omega^\top\end{bmatrix}$. In this case, for any $\boldsymbol{\alpha}\in\Re^n$, $\boldsymbol{U}_\mathcal{I}\boldsymbol{\alpha}=\begin{bmatrix}\boldsymbol{\alpha}\\ \alpha_\omega\end{bmatrix}$. This implies that, given any $\boldsymbol{\gamma}\in\Re^{n+1}$, we have $\boldsymbol{\gamma}\in\text{span}(\boldsymbol{U})$ if and only if $\gamma_\omega=\gamma_{n+1}$. Therefore, consider any $\boldsymbol{\beta}$ with $\boldsymbol{V}_\mathcal{I}\boldsymbol{\beta}\in\text{span}(\boldsymbol{U}_\mathcal{I})$, we know $\boldsymbol{v}^\top_{RI_\omega}\boldsymbol{\beta}=\boldsymbol{v}^\top_{RI_{n+1}}\boldsymbol{\beta}$. Our objective is to show $\beta_i(\boldsymbol{V}_\mathcal{I})_i\in\text{span}(\boldsymbol{U}_\mathcal{I})$, i.e, $\beta_i v_{RI_\omega,i}=\beta_i v_{RI_{n+1},i}$, for all $i\in[n]$. To see this, we consider two cases.

- Case 1: both $\boldsymbol{u}_{RI_\omega}$ and $\boldsymbol{u}_{RI_{n+1}}$ are extracted from $\bar{\boldsymbol{U}}^\omega$, i.e., $RI_\omega, RI_{n+1}\in \left\{\frac{\omega(\omega-1)}{2}+1,\ldots,\frac{\omega(\omega-1)}{2}+\omega\right\}$. We denote $j_\omega=RI_\omega-\frac{\omega(\omega-1)}{2}$ and $j_{n+1}=RI_{n+1}-\frac{\omega(\omega-1)}{2}$, i.e., $\boldsymbol{u}^\top_{RI_\omega}$ and $\boldsymbol{u}^\top_{RI_{n+1}}$ are the j_ω-th and j_{n+1}-th rows in $\bar{\boldsymbol{U}}^\omega$, respectively. By the structure of $\bar{\boldsymbol{V}}^\omega$, we know for all $s\in[n]$,

$$v_{RI_\omega,s}=\bar{v}^\omega_{j_\omega,s}=\begin{cases}1 & s=j_\omega,\ldots,\omega\\ 0 & s=1,\ldots,j_\omega-1 \text{ or } s=\omega+1,\ldots,n,\end{cases}$$
$$v_{RI_{n+1},s}=\bar{v}^\omega_{j_{n+1},s}=\begin{cases}1 & s=j_{n+1},\ldots,\omega\\ 0 & s=1,\ldots,j_{n+1}-1 \text{ or } s=\omega+1,\ldots,n.\end{cases}$$

  In this case, $\boldsymbol{v}^\top_{RI_\omega}\boldsymbol{\beta}=\boldsymbol{v}^\top_{RI_{n+1}}\boldsymbol{\beta}$ implies $\sum\limits_{j=j_\omega}^{\omega}\beta_j=\sum\limits_{j=j_{n+1}}^{\omega}\beta_j$; and hence $\beta_j=0$ for all $j\in\{j_\omega,\ldots,j_{n+1}-1\}$ since $\boldsymbol{\beta}\ge\boldsymbol{0}$. Now for any arbitrary $i\in[n]$, the equation $\beta_i v_{RI_\omega,i}=\beta_i v_{RI_{n+1},i}$ always holds since 1) $v_{RI_\omega,i}=v_{RI_{n+1},i}=0$ when $i=1,\ldots j_\omega-1$ or $i=\omega+1,\ldots,n$; 2) $\beta_i=0$ when $i=j_\omega,\ldots,j_{n+1}-1$; 3) $v_{RI_\omega,i}=v_{RI_{n+1},i}=1$ when $i=j_{n+1},\ldots,\omega$.
- Case 2: $\boldsymbol{u}_{RI_\omega}$ is extracted from $\bar{\boldsymbol{U}}^\omega$ while $\boldsymbol{u}_{RI_{n+1}}$ is extracted from $\bar{\boldsymbol{U}}^{n+1}$. The submatrix $\bar{\boldsymbol{V}}^{n+1}=\boldsymbol{0}_{n\times n}$ implies in this case $\boldsymbol{v}_{RI_{n+1}}=\boldsymbol{0}$. Hence, $\boldsymbol{v}^\top_{RI_\omega}\boldsymbol{\beta}=\boldsymbol{v}^\top_{RI_{n+1}}\boldsymbol{\beta}$ implies $0=\boldsymbol{v}^\top_{RI_\omega}\boldsymbol{\beta}=\sum\limits_{i\in[n]}\beta_i v_{RI_\omega,i}$. Since $\boldsymbol{v}_{RI_\omega}\ge\boldsymbol{0}$ and $\boldsymbol{\beta}\ge\boldsymbol{0}$, we then have $\beta_i v_{RI_\omega,i}=0=\beta_i v_{RI_{n+1},i}$ for all $i\in[n]$.

Therefore, $\hat{g}(\boldsymbol{x},\hat{\boldsymbol{z}})$ is supermodular in $\hat{\boldsymbol{z}}$ for all $\boldsymbol{x}$. We next prove that $g(\boldsymbol{x},\boldsymbol{\xi},\boldsymbol{z})$ is supermodular in every two distinct components of $(\boldsymbol{\xi},\boldsymbol{z})$, and hence is jointly supermodular in $(\boldsymbol{\xi},\boldsymbol{z})$.

We first consider argument as the pair (ξ_i,z_i) for some $i\in[n]$ and fix all ξ_s,z_s with $s\in[n]\setminus\{i\}$. As all the remaining elements are fixed, we define $g^i:\Re^2\to\Re$ and $h:\Re\to\Re$ to be such that

$g^i(\xi_i,z_i)=h(\hat{z}_i)=h(\xi_i z_i)=\hat{g}\left(\boldsymbol{x},(\xi_1 z_1,\ldots,\xi_n z_n)\right)=g(\boldsymbol{x},\boldsymbol{\xi},\boldsymbol{z})$. Hence, it is equivalent to show that g^i, as a function of ξ_i, z_i, is supermodular in its arguments. To this end, we first observe that $\xi_i z_i$ is increasing and supermodular in (ξ_i,z_i) (recall that $\xi_i,z_i\geq 0$). Further, as a second-stage cost function, $\hat{g}(\boldsymbol{x},\hat{\boldsymbol{z}})$ has been shown as convex in $\hat{\boldsymbol{z}}$ by literature (e.g., Birge and Louveaux 2011, Theorem 2), and it implies that $h(\hat{z}_i)$ is convex in $\hat{z}_i$. In addition, $h(\hat{z}_i)$ is also increasing in $\hat{z}_i$ by the definition in (48). Therefore, the supermodularity of g^i follows as a corollary of Lemma 3, and $g(\boldsymbol{x},\boldsymbol{\xi},\boldsymbol{z})$ is supermodular in (ξ_i,z_i) for all $i\in[n]$.

Next, if the argument is the pair (ξ_i,z_j) for some distinct $i,j\in[n]$, we prove the supermodularity of $g^{ij}(\xi_i,z_j)=\hat{g}\left(\boldsymbol{x},(\xi_1 z_1,\ldots,\xi_n z_n)\right)$. Consider $\boldsymbol{\xi}',\boldsymbol{\xi}'',\boldsymbol{z}',\boldsymbol{z}''\in\Re^n$ with $\xi_i'<\xi_i'', z_j'>z_j''$, $\xi_s'=\xi_s'', z_s'=z_s''$ and we denote their common values as ξ_s, z_s, respectively, for all $s\in[n]\setminus\{i,j\}$. Since $\boldsymbol{\xi}\in\{0,1\}^n$, by $\xi_i'<\xi_i''$ we know $\xi_i'=0,\xi_i''=1$. Define $\hat{\boldsymbol{z}}'$, $\hat{\boldsymbol{z}}''$, $\hat{\boldsymbol{z}}^{\min}$, $\hat{\boldsymbol{z}}^{\min}\in\Re^n$ such that $\hat{z}_k'=\xi_k'z_k'$, $\hat{z}_k''=\xi_k''z_k''$, $\hat{z}_k^{\min}=(\xi_k'\wedge\xi_k'')(z_k'\wedge z_k'')$, $\hat{z}_k^{\max}=(\xi_k'\vee\xi_k'')(z_k'\vee z_k'')$ for all $k\in[n]$. Then these four vectors differ only in their ith, jth elements. In particular, $(\hat{z}_i',\hat{z}_j')=(0,\xi_j z_j')$, $(\hat{z}_i'',\hat{z}_j'')=(z_i,\xi_j z_j'')$, $(\hat{z}_i^{\min},\hat{z}_j^{\min})=(0,\xi_j z_j'')$, $(\hat{z}_i^{\max},\hat{z}_j^{\max})=(z_i,\xi_j z_j')$. Hence, denoting $\hat{\boldsymbol{z}}^\circ\in\Re^n$ such that $\hat{z}_i^\circ=\hat{z}_j^\circ=0$ and $\hat{z}_s^\circ=\xi_s z_s$ for all $s\in[n]\setminus\{i,j\}$, we have

$$
\begin{aligned}
& g^{ij}(\xi_i'\wedge\xi_i'',z_j'\wedge z_j'')+g^{ij}(\xi_i'\vee\xi_i'',z_j'\vee z_j'')-g^{ij}(\xi_i',z_j')-g^{ij}(\xi_i'',z_j'')\\
=\ & \hat{g}\left(\boldsymbol{x},\hat{\boldsymbol{z}}^{\min}\right)+\hat{g}\left(\boldsymbol{x},\hat{\boldsymbol{z}}^{\max}\right)-\hat{g}\left(\boldsymbol{x},\hat{\boldsymbol{z}}'\right)-\hat{g}\left(\boldsymbol{x},\hat{\boldsymbol{z}}''\right)\\
=\ & \hat{g}\left(\boldsymbol{x},\hat{\boldsymbol{z}}^\circ+\xi_j z_j''\boldsymbol{e}_j\right)+\hat{g}\left(\boldsymbol{x},\hat{\boldsymbol{z}}^\circ+z_i\boldsymbol{e}_i+\xi_j z_j'\boldsymbol{e}_j\right)-\hat{g}\left(\boldsymbol{x},\hat{\boldsymbol{z}}^\circ+\xi_j z_j'\boldsymbol{e}_j\right)-\hat{g}\left(\boldsymbol{x},\hat{\boldsymbol{z}}^\circ+z_i\boldsymbol{e}_i+\xi_j z_j''\boldsymbol{e}_j\right)\\
\geq\ & 0,
\end{aligned}
$$

where the inequality holds because $g\left(\boldsymbol{x},\hat{\boldsymbol{z}}\right)$ is supermodular in $\hat{\boldsymbol{z}}$. Hence, g^{ij} is supermodular and therefore $g(\boldsymbol{x},\boldsymbol{\xi},\boldsymbol{z})$ is supermodular in (ξ_i,z_j).

For (z_i,z_j) or (ξ_i,ξ_j) with $1\leq i<j\leq n$, the proof is similar to the second case. We now conclude that $g(\boldsymbol{x},\boldsymbol{\xi},\boldsymbol{z})$ is supermodular in $(\boldsymbol{\xi},\boldsymbol{z})$.

Noticing that $\mathcal{F}_{\boldsymbol{\xi}}^k$ (or $\mathcal{F}_{\boldsymbol{z}}^k$) determine a set of 0-1 (or three-point) worst-case marginals for $\tilde{\boldsymbol{\xi}}$ (or $\tilde{\boldsymbol{z}}$), we claim that applying Algorithm 1 yields a $(3n+1)$-point joint distribution of $(\tilde{\boldsymbol{\xi}},\tilde{\boldsymbol{z}})$ for each realized scenario. The number of points follows from one plus the number of steps it takes when moving from $(\boldsymbol{0},\underline{\boldsymbol{z}}^k)$ to $(\boldsymbol{1},\overline{\boldsymbol{z}}^k)$ only in the positive directions. The number of steps is $3n$, since there are exactly 3 steps on the i-th dimension—from $\xi_i=0\to 1$, and from $z_i=\underline{z}_i^k\to\mu_i\to\overline{z}_i^k$. We then utilize the results in Theorem 1 and obtain the reformulation as follows.

$$\begin{aligned}
\min \quad & \boldsymbol{\nu}^\top \boldsymbol{l} \\
\text{s.t.} \quad & \boldsymbol{R}_k^\top \boldsymbol{l} \geq \sum_{i\in[3n+1]} p_i^k \mathbf{1}^\top \boldsymbol{y}^{k,i}, && k \in [K] \\
& y_t^{k,i} \geq \sum_{s=j}^{t} (\xi^{k,i} z_s^{k,i} - x_s), && j \in [t], t \in [n], k \in [K], i \in [3n+1] \\
& \boldsymbol{y}^{k,i} \geq \mathbf{0}, && k \in [K], i \in [3n+1] \\
& \boldsymbol{l} \geq \mathbf{0},\ \boldsymbol{x} \in \mathcal{X}^{app},
\end{aligned}$$

where $p_i^k, \boldsymbol{\xi}^{k,i}, \boldsymbol{z}^{k,i}, k \in [K], i \in [3n+1]$ are the output of Algorithm 1, given the ambiguity sets $\mathcal{G}^k, k \in [K]$ defined as $\mathcal{G}^k = \left\{\mathbb{P}^k \;\middle|\; \boldsymbol{\Pi}_{\boldsymbol{\xi}}\mathbb{P}^k \in \mathcal{F}_{\boldsymbol{\xi}}^k, \boldsymbol{\Pi}_{\boldsymbol{z}}\mathbb{P}^k \in \mathcal{F}_{\boldsymbol{z}}^k\right\}$, where $\boldsymbol{\Pi}_{\boldsymbol{\xi}}\mathbb{P}^k$, $\boldsymbol{\Pi}_{\boldsymbol{z}}\mathbb{P}^k$ denotes the marginal distribution of $\tilde{\boldsymbol{\xi}}$ and $\tilde{\boldsymbol{z}}$, respectively under $\mathbb{P}^k$. $\mathcal{F}_{\boldsymbol{\xi}}^k$ is the conditional ambiguity set of $\mathcal{F}_{\boldsymbol{\xi}}$ when $\tilde{k}$ is realized as k, and $\mathcal{F}_{\boldsymbol{z}}^k$ is defined by (4). □

C.17. Proof of Theorem 4

The constraint of the second-stage problem

$$\begin{aligned}
g(\boldsymbol{x}, \boldsymbol{z}) = \min \quad & \boldsymbol{h}^\top (\boldsymbol{x} - \boldsymbol{A}\boldsymbol{y}) + \boldsymbol{p}^\top (\boldsymbol{z} - \boldsymbol{y}) - \boldsymbol{r}^\top \boldsymbol{y} \\
\text{s.t.} \quad & \boldsymbol{A}\boldsymbol{y} \leq \boldsymbol{x},\ \boldsymbol{y} \leq \boldsymbol{z},\ \boldsymbol{y} \geq \mathbf{0}.
\end{aligned}$$

can be represented as $\boldsymbol{U}\boldsymbol{y} - \boldsymbol{V}\boldsymbol{z} \geq -\boldsymbol{W}\boldsymbol{x} + \boldsymbol{v}^0$, where $\boldsymbol{U} = \begin{bmatrix} -\boldsymbol{I} \\ \boldsymbol{I} \\ -\boldsymbol{A} \end{bmatrix}, \boldsymbol{V} = \begin{bmatrix} -\boldsymbol{I} \\ \mathbf{0} \\ \mathbf{0} \end{bmatrix}$ and $\boldsymbol{W} = \begin{bmatrix} \mathbf{0} \\ \mathbf{0} \\ \boldsymbol{I} \end{bmatrix}$.

We first prove the "if" direction. Suppose the condition for $\boldsymbol{A}$ in this theorem is satisfied. By Theorem 2, whether $\boldsymbol{U}, \boldsymbol{V}$ lead to supermodularity of g is equivalent to whether $-\boldsymbol{U}, -\boldsymbol{V}$ do so. Therefore, here we verify the supermodularity based on $-\boldsymbol{U}, -\boldsymbol{V}$. Observing that $-\boldsymbol{U}$ and $-\boldsymbol{V}$ have the structure as in Proposition 5 with $\boldsymbol{U}^\circ = \begin{bmatrix} -\boldsymbol{I} \\ \boldsymbol{A} \end{bmatrix}$, we now show that every 2×3 submatrix of $\boldsymbol{U}^\circ$ contains at least one pair of columns which are linearly dependent. If both rows of the 2×3 submatrix are extracted from $\boldsymbol{A}$, then this submatrix must have two linearly dependent columns by the assumption on $\boldsymbol{A}$. If at least one of the rows are from $-\boldsymbol{I}$, since the rows from $-\boldsymbol{I}$ have at least two zero elements, then this submatrix must have two linearly dependent columns.

We now prove the "only if" direction by contradiction. We first consider the case that $\boldsymbol{A} \in \Re_+^{2\times 3}$. Assume the contrary, i.e., every two columns in $\boldsymbol{A}$ are in different directions. Given that $\boldsymbol{A} \geq \mathbf{0}$, there must be one column in $\boldsymbol{A}$ being a conical combination of the other two columns. WLOG, let $\boldsymbol{A}_3$ be a conical combination of $\boldsymbol{A}_1, \boldsymbol{A}_2$. We remark that multiplying any strictly positive constant by a row/column in $\boldsymbol{A}$, or switching rows, or switching columns does not affect whether the corresponding function g is supermodular. Therefore, we can make the following simplification on $\boldsymbol{A}$. Since $\boldsymbol{A}_1, \boldsymbol{A}_2$ are linearly independent, WLOG, we can let $\boldsymbol{A} = \begin{bmatrix} 1 & a & c \\ b & 1 & d \end{bmatrix}$ with

$ab < 1$. Since $\boldsymbol{A}_3$ is a conical combination of $\boldsymbol{A}_1, \boldsymbol{A}_2$, we have $cd > 0$; WLOG, we can let $d = 1$, i.e., $\boldsymbol{A} = \begin{bmatrix} 1 & a & c \\ b & 1 & 1 \end{bmatrix}$. Multiplying the first row by $1/c$, and then multiplying the first column by c, we have $\boldsymbol{A} = \begin{bmatrix} 1 & a/c & 1 \\ bc & 1 & 1 \end{bmatrix}$. Let a/c, bc be the new a, b, we have $\boldsymbol{A} = \begin{bmatrix} 1 & a & 1 \\ b & 1 & 1 \end{bmatrix}$ with $ab < 1$. Again, since $\boldsymbol{A}_3$ is a conical combination of $\boldsymbol{A}_1, \boldsymbol{A}_2$, we have either $a, b < 1$ or $a, b > 1$. Together with $ab < 1$, we know $a, b < 1$. In summary, WLOG, we let $\boldsymbol{A} = \begin{bmatrix} 1 & a & 1 \\ b & 1 & 1 \end{bmatrix}$ with $a, b \in [0, 1)$.

We define $\bar{g}(\boldsymbol{x}, \boldsymbol{z}) = g(\boldsymbol{x}, \boldsymbol{z}) - \boldsymbol{p}^\top \boldsymbol{z}$, then it is equivalent to prove that $\bar{g}(\boldsymbol{x}, \boldsymbol{z})$ is not supermodular in $\boldsymbol{z}$. We now construct such a counterexample. Let $\boldsymbol{h} = \boldsymbol{0}$, $\boldsymbol{r} = \boldsymbol{0}$, $\boldsymbol{p} = (1, 1, \epsilon)$ with any $\epsilon \in (0, 1)$. We choose $\boldsymbol{x} = (1 - ab)\boldsymbol{1}$, $\boldsymbol{z}' = (1 - a, 0, 1 - ab)$, $\boldsymbol{z}'' = (0, 1 - b, 1 - ab)$. Denote $\boldsymbol{z}^\wedge = \boldsymbol{z}' \wedge \boldsymbol{z}''$, $\boldsymbol{z}^\vee = \boldsymbol{z}' \vee \boldsymbol{z}''$, we have $\boldsymbol{z}^\wedge = (0, 0, 1 - ab)$, $\boldsymbol{z}^\vee = (1 - a, 1 - b, 1 - ab)$. We notice that

$$\begin{aligned}
\bar{g}(\boldsymbol{x}, \boldsymbol{z}) &= \min \left\{ -\boldsymbol{p}^\top \boldsymbol{y} \;\middle|\; \boldsymbol{A}\boldsymbol{y} \leq \boldsymbol{x}, \boldsymbol{0} \leq \boldsymbol{y} \leq \boldsymbol{z} \right\} \\
&= \min \left\{ -y_1 - y_2 - \epsilon y_3 \;\middle|\; \begin{array}{l} y_1 + a y_2 + y_3 \leq 1 - ab \\ b y_1 + y_2 + y_3 \leq 1 - ab \\ (0, 0, 0) \leq (y_1, y_2, y_3) \leq (z_1, z_2, z_3). \end{array} \right\}
\end{aligned} \tag{49}$$

Hence,

$$\begin{aligned}
\bar{g}(\boldsymbol{x}, \boldsymbol{z}') &= \min \left\{ -y_1 - \epsilon y_3 \;\middle|\; \begin{array}{l} y_1 + y_3 \leq 1 - ab, \\ 0 \leq y_1 \leq 1 - a,\; y_2 = 0,\; y_3 \geq 0 \end{array} \right\}, \\
\bar{g}(\boldsymbol{x}, \boldsymbol{z}'') &= \min \left\{ -y_2 - \epsilon y_3 \;\middle|\; \begin{array}{l} y_2 + y_3 \leq 1 - ab, \\ y_1 = 0,\; 0 \leq y_2 \leq 1 - b,\; y_3 \geq 0 \end{array} \right\}, \\
\bar{g}(\boldsymbol{x}, \boldsymbol{z}^\wedge) &= \min \left\{ -\epsilon y_3 \;\middle|\; \begin{array}{l} y_3 \leq 1 - ab, \\ y_1 = y_2 = 0,\; y_3 \geq 0 \end{array} \right\}, \\
\bar{g}(\boldsymbol{x}, \boldsymbol{z}^\vee) &= \min \left\{ -y_1 - y_2 - \epsilon y_3 \;\middle|\; \begin{array}{l} y_1 + a y_2 + y_3 \leq 1 - ab, \\ b y_1 + y_2 + y_3 \leq 1 - ab, \\ 0 \leq y_1 \leq 1 - a,\; 0 \leq y_2 \leq 1 - b,\; y_3 \geq 0 \end{array} \right\}.
\end{aligned}$$

Since $0 < \epsilon < 1$, in the optimization problem for $\bar{g}(\boldsymbol{x}, \boldsymbol{z}')$, the optimal solution should be that y_1 goes to the upper bound, i.e, $y_1 = 1 - a$, $y_2 = 0$ and $y_3 = (1 - ab) - (1 - a) = a(1 - b)$. Similarly, in the optimization problem for $\bar{g}(\boldsymbol{x}, \boldsymbol{z}'')$, the optimal $\boldsymbol{y} = (0, 1 - b, b(1 - a))$; in that for $\bar{g}(\boldsymbol{x}, \boldsymbol{z}^\wedge)$, the optimal $\boldsymbol{y} = (0, 0, 1 - ab)$; in that for $\bar{g}(\boldsymbol{x}, \boldsymbol{z}^\vee)$, the optimal $\boldsymbol{y} = (1 - a, 1 - b, 0)$. We then have

$$\begin{aligned}
& \bar{g}(\boldsymbol{x}, \boldsymbol{z}') + \bar{g}(\boldsymbol{x}, \boldsymbol{z}'') - \bar{g}(\boldsymbol{x}, \boldsymbol{z}^\wedge) - \bar{g}(\boldsymbol{x}, \boldsymbol{z}^\vee) \\
= &\; -\left((1 - a + \epsilon a(1 - b)) + (1 - b + \epsilon b(1 - a)) - \epsilon(1 - ab) - (1 - a + 1 - b) \right) \\
= &\; \epsilon(1 - a)(1 - b) > 0,
\end{aligned}$$

where the last equality holds since $0 < a, b < 1$. Therefore, $\bar{g}(\boldsymbol{x}, \boldsymbol{z}^\wedge) + \bar{g}(\boldsymbol{x}, \boldsymbol{z}^\vee) < \bar{g}(\boldsymbol{x}, \boldsymbol{z}') + \bar{g}(\boldsymbol{x}, \boldsymbol{z}'')$, this function $\bar{g}$ is not supermodular.

For the general case of $\boldsymbol{A} \in \Re_+^{l \times n}$, we can prove the result by the same contradiction. WLOG, we assume the 2×3 submatrix of $\boldsymbol{A}$, which is obtained by deleting all rows except the first two and all columns except the first three, is such that each pair of columns in it are linearly independent. We can then let $z'_i = z''_i = 0$ for $i \in \{4, 5, \ldots, n\}$ and x_i be sufficiently large for $i \in \{3, 4, \ldots, l\}$ such that it would not affect the feasible region of $\boldsymbol{y}$. We then have $\bar{g}(\boldsymbol{x}, \boldsymbol{z})$ with exactly the same expression in Equation (49). Therefore, we still have $\bar{g}(\boldsymbol{x}, \boldsymbol{z}^{\wedge}) + \bar{g}(\boldsymbol{x}, \boldsymbol{z}^{\vee}) < \bar{g}(\boldsymbol{x}, \boldsymbol{z}') + \bar{g}(\boldsymbol{x}, \boldsymbol{z}'')$. □

C.18. Proof of Corollary 2

We first prove the "if" direction using Theorem 4. Consider any 2×3 submatrix of $\boldsymbol{A}$, which, WLOG, is $\boldsymbol{C} = \boldsymbol{A}_{\{1,2\},\{1,2,3\}}$. Let $\hat{S}_i = S_i \cap \{1, 2, 3\},\ i = 1, 2$. If $\hat{S}_1 \cap \hat{S}_2 = \emptyset$, then at least one of the rows in $\boldsymbol{C}$ has two zero elements, and hence $\boldsymbol{C}$ has at least one pair of columns which are linearly dependent. If $\hat{S}_1 \cap \hat{S}_2 \neq \emptyset$, by the definition of Tree Family, we have either $\hat{S}_1 \subseteq \hat{S}_2$ or $\hat{S}_2 \subseteq \hat{S}_1$. WLOG, we let $\hat{S}_1 \subseteq \hat{S}_2$. If $|\hat{S}_1| = 1$, then the first row of $\boldsymbol{C}$ has two zero elements and hence $\boldsymbol{C}$ has at least one pair of columns which are linearly dependent. If $|\hat{S}_1| \geq 2$, WLOG, $\{1, 2\} \subseteq \hat{S}_1$, by the definition of Proportional Tree Family, we have $a_{11}/a_{21} = a_{12}/a_{22}$, hence $\boldsymbol{C}$ has at least one pair of columns which are linearly dependent. In summary, $\boldsymbol{C}$ always have at least one pair of columns which are linearly dependent. □

C.19. Proof of Theorem 5

The case for $\text{rank}(\boldsymbol{U}) = r$ is straightforward, so we only consider the case where $\text{rank}(\boldsymbol{U}) < r$. In that case, we only need to verify whether $\boldsymbol{U}, \boldsymbol{V}$ satisfy the second part of the condition in Theorem 2, which depends solely on the relationship between $\boldsymbol{V}$ and $\text{span}(\boldsymbol{U})$. Thus, removing the dependent columns in $\boldsymbol{U}$ does not change the satisfaction or violation of the conditions. Therefore, the procedure in line 4 of the algorithm does not change the result and WLOG, we can assume $\boldsymbol{U}$ has r_0 columns, i.e., $m = r_0$.

First we look into the case where Algorithm 2 returns $s = 0$. This implies that there exists an index set $\mathcal{I} \subseteq [r]$ and indices $i \in [r] \setminus \mathcal{I}, a, b \in [r_0]$ with $|\mathcal{I}| = r_0$, $\boldsymbol{U}_{\mathcal{I}}$ invertible and $d_{ia} d_{ib} < 0$. WLOG, we let $d_{ia} > 0, d_{ib} < 0$.

Denote $\boldsymbol{\beta} = \dfrac{\boldsymbol{e}_a}{d_{ia}} - \dfrac{\boldsymbol{e}_b}{d_{ib}} \geq \boldsymbol{0}$, $\boldsymbol{\alpha} = \boldsymbol{U}_{\mathcal{I}}^{-1} \left(\dfrac{(\boldsymbol{V}_{\mathcal{I}})_a}{d_{ia}} - \dfrac{(\boldsymbol{V}_{\mathcal{I}})_b}{d_{ib}} \right)$, then

$$\begin{bmatrix} \boldsymbol{V}_{\mathcal{I}} \\ \boldsymbol{v}_i^{\top} \end{bmatrix} \boldsymbol{\beta} = \begin{bmatrix} \dfrac{(\boldsymbol{V}_{\mathcal{I}})_a}{d_{ia}} - \dfrac{(\boldsymbol{V}_{\mathcal{I}})_b}{d_{ib}} \\ \dfrac{v_{ia}}{d_{ia}} - \dfrac{v_{ib}}{d_{ib}} \end{bmatrix} = \begin{bmatrix} \dfrac{(\boldsymbol{V}_{\mathcal{I}})_a}{d_{ia}} - \dfrac{(\boldsymbol{V}_{\mathcal{I}})_b}{d_{ib}} \\ \dfrac{d_{ia} + \boldsymbol{u}_i^{\top} \boldsymbol{U}_{\mathcal{I}}^{-1} (\boldsymbol{V}_{\mathcal{I}})_a}{d_{ia}} - \dfrac{d_{ib} + \boldsymbol{u}_i^{\top} \boldsymbol{U}_{\mathcal{I}}^{-1} (\boldsymbol{V}_{\mathcal{I}})_b}{d_{ib}} \end{bmatrix} = \begin{bmatrix} \boldsymbol{U}_{\mathcal{I}} \\ \boldsymbol{u}_i^{\top} \end{bmatrix} \boldsymbol{\alpha},$$

We let $\hat{\mathcal{I}} = \mathcal{I} \cup \{i\}$. The above equality implies $\boldsymbol{V}_{\hat{\mathcal{I}}} \boldsymbol{\beta} = \boldsymbol{U}_{\hat{\mathcal{I}}} \boldsymbol{\alpha} \in \text{span}(\boldsymbol{U}_{\hat{\mathcal{I}}})$. On the other hand, for $\beta_a (\boldsymbol{V}_{\hat{\mathcal{I}}})_a$ we have

$$\beta_a \begin{bmatrix} (\boldsymbol{V}_{\mathcal{I}})_a \\ v_{ia} \end{bmatrix} = \begin{bmatrix} \dfrac{(\boldsymbol{V}_{\mathcal{I}})_a}{d_{ia}} \\ \dfrac{d_{ia} + \boldsymbol{u}_i^{\top} \boldsymbol{U}_{\mathcal{I}}^{-1} (\boldsymbol{V}_{\mathcal{I}})_a}{d_{ia}} \end{bmatrix} = \begin{bmatrix} \boldsymbol{U}_{\mathcal{I}} \\ \boldsymbol{u}_i^{\top} \end{bmatrix} \frac{\boldsymbol{U}_{\mathcal{I}}^{-1} (\boldsymbol{V}_{\mathcal{I}})_a}{d_{ia}} + \begin{bmatrix} \boldsymbol{0} \\ 1 \end{bmatrix} = \boldsymbol{U}_{\hat{\mathcal{I}}} \frac{\boldsymbol{U}_{\mathcal{I}}^{-1} (\boldsymbol{V}_{\mathcal{I}})_a}{d_{ia}} + \begin{bmatrix} \boldsymbol{0} \\ 1 \end{bmatrix}.$$

Since $\boldsymbol{U}_{\mathcal{I}}$ is invertible, there is no $\boldsymbol{\gamma}\in\Re^{r_0}$ such that $\boldsymbol{U}_{\hat{\mathcal{I}}}\boldsymbol{\gamma}=\begin{bmatrix}\boldsymbol{U}_{\mathcal{I}}\\ \boldsymbol{u}_i^\top\end{bmatrix}\boldsymbol{\gamma}=\begin{bmatrix}\boldsymbol{0}\\ 1\end{bmatrix}$. Hence $\beta_a(\boldsymbol{V}_{\hat{\mathcal{I}}})_a\notin\operatorname{span}(\boldsymbol{U}_{\hat{\mathcal{I}}})$ and the second part of the condition in Theorem 2 is violated.

We now investigate the case where the second part of the condition in Theorem 2 is violated. That means, there exist $\hat{\mathcal{I}}\subseteq[r]$, $\boldsymbol{\beta}\geq\boldsymbol{0}$ and $a\in[r_0]$ such that $|\hat{\mathcal{I}}|=r_0+1,\operatorname{rank}(\boldsymbol{U}_{\hat{\mathcal{I}}})=r_0$, $\boldsymbol{V}_{\hat{\mathcal{I}}}\boldsymbol{\beta}\in\operatorname{span}(\boldsymbol{U}_{\hat{\mathcal{I}}})$ but $\beta_a(\boldsymbol{V}_{\hat{\mathcal{I}}})_a\notin\operatorname{span}(\boldsymbol{U}_{\hat{\mathcal{I}}})$. We choose $\mathcal{I}\subseteq\hat{\mathcal{I}}$ such that $|\mathcal{I}|=r_0$ and $\boldsymbol{U}_{\mathcal{I}}$ is invertible, and denote i as the unique index in $\hat{\mathcal{I}}\backslash\mathcal{I}$. It follows that

$$\boldsymbol{V}_{\hat{\mathcal{I}}}\boldsymbol{\beta}=\begin{bmatrix}\boldsymbol{V}_{\mathcal{I}}\\ \boldsymbol{v}_i^\top\end{bmatrix}\boldsymbol{\beta}=\begin{bmatrix}\boldsymbol{0}\\ \boldsymbol{v}_i^\top-\boldsymbol{u}_i^\top\boldsymbol{U}_{\mathcal{I}}^{-1}\boldsymbol{V}_{\mathcal{I}}\end{bmatrix}\boldsymbol{\beta}+\begin{bmatrix}\boldsymbol{U}_{\mathcal{I}}\\ \boldsymbol{u}_i^\top\end{bmatrix}\boldsymbol{U}_{\mathcal{I}}^{-1}\boldsymbol{V}_{\mathcal{I}}\boldsymbol{\beta}=\begin{bmatrix}\boldsymbol{0}\\ \boldsymbol{d}_i^\top\end{bmatrix}\boldsymbol{\beta}+\boldsymbol{U}_{\hat{\mathcal{I}}}\boldsymbol{U}_{\mathcal{I}}^{-1}\boldsymbol{V}_{\mathcal{I}}\boldsymbol{\beta},$$

$$\beta_a(\boldsymbol{V}_{\hat{\mathcal{I}}})_a=\beta_a\begin{bmatrix}(\boldsymbol{V}_{\mathcal{I}})_a\\ v_{ia}\end{bmatrix}=\beta_a\left(\begin{bmatrix}\boldsymbol{0}\\ v_{ia}-\boldsymbol{u}_i^\top\boldsymbol{U}_{\mathcal{I}}^{-1}(\boldsymbol{V}_{\mathcal{I}})_a\end{bmatrix}+\begin{bmatrix}\boldsymbol{U}_{\mathcal{I}}\\ \boldsymbol{u}_i^\top\end{bmatrix}\boldsymbol{U}_{\mathcal{I}}^{-1}(\boldsymbol{V}_{\mathcal{I}})_a\right)=\begin{bmatrix}\boldsymbol{0}\\ \beta_a d_{ia}\end{bmatrix}+\beta_a\boldsymbol{U}_{\hat{\mathcal{I}}}\boldsymbol{U}_{\mathcal{I}}^{-1}(\boldsymbol{V}_{\mathcal{I}})_a.$$

Since $\boldsymbol{V}_{\hat{\mathcal{I}}}\boldsymbol{\beta},\boldsymbol{U}_{\hat{\mathcal{I}}}\boldsymbol{U}_{\mathcal{I}}^{-1}\boldsymbol{V}_{\mathcal{I}}\boldsymbol{\beta},\beta_a\boldsymbol{U}_{\hat{\mathcal{I}}}\boldsymbol{U}_{\mathcal{I}}^{-1}(\boldsymbol{V}_{\mathcal{I}})_a\in\operatorname{span}(\boldsymbol{U}_{\hat{\mathcal{I}}})$ and $\beta_a(\boldsymbol{V}_{\hat{\mathcal{I}}})_a\notin\operatorname{span}(\boldsymbol{U}_{\hat{\mathcal{I}}})$, the above equations imply $\begin{bmatrix}\boldsymbol{0}\\ \boldsymbol{d}_i^\top\end{bmatrix}\boldsymbol{\beta}\in\operatorname{span}(\boldsymbol{U}_{\hat{\mathcal{I}}})$ and $\begin{bmatrix}\boldsymbol{0}\\ \beta_i d_{ia}\end{bmatrix}\notin\operatorname{span}(\boldsymbol{U}_{\hat{\mathcal{I}}})$. According to $\begin{bmatrix}\boldsymbol{0}\\ \boldsymbol{d}_i^\top\end{bmatrix}\boldsymbol{\beta}\in\operatorname{span}(\boldsymbol{U}_{\hat{\mathcal{I}}})$, there exists $\boldsymbol{\alpha}\in\Re^{r_0}$ with $\boldsymbol{U}_{\mathcal{I}}\boldsymbol{\alpha}=\boldsymbol{0},\boldsymbol{u}_i^\top\boldsymbol{\alpha}=\boldsymbol{d}_i^\top\boldsymbol{\beta}$. Since $\boldsymbol{U}_{\mathcal{I}}$ is invertible, $\boldsymbol{\alpha}=\boldsymbol{0}$ and hence $\boldsymbol{d}_i^\top\boldsymbol{\beta}=\boldsymbol{u}_i^\top\boldsymbol{\alpha}=0$. According to $\begin{bmatrix}\boldsymbol{0}\\ \beta_i d_{ia}\end{bmatrix}\notin\operatorname{span}(\boldsymbol{U}_{\hat{\mathcal{I}}})$, we obtain $\beta_a d_{ia}\neq 0$. As $\boldsymbol{\beta}\geq\boldsymbol{0}$, $\beta_a d_{ia}\neq 0$ and $\boldsymbol{d}_i^\top\boldsymbol{\beta}=0$, we must have an index $b\in[r_0]$ such that d_{ib},d_{ia} are of different signs. Hence the algorithm returns $s=0$.

□

C.20. Proof of Theorem 6

We first reformulate the second-stage problem as

$$\begin{aligned} g^W(\boldsymbol{x},\boldsymbol{z})=\min\ & \boldsymbol{b}^\top\boldsymbol{y}\\ \text{s.t. } & \boldsymbol{U}\boldsymbol{y}-\left(\boldsymbol{V}-\left[\boldsymbol{W}^1\boldsymbol{x}\cdots\boldsymbol{W}^n\boldsymbol{x}\right]\right)\boldsymbol{z}\geq-\boldsymbol{W}^0\boldsymbol{x}+\boldsymbol{v}^0,\end{aligned}$$

where $[\boldsymbol{W}^1\boldsymbol{x}\cdots\boldsymbol{W}^n\boldsymbol{x}]$ stands for an $r\times n$ matrix with its i-th column being $\boldsymbol{W}^i\boldsymbol{x}$. We denote $\bar{\boldsymbol{V}}^{\boldsymbol{x}}=\boldsymbol{V}-[\boldsymbol{W}^1\boldsymbol{x}\cdots\boldsymbol{W}^n\boldsymbol{x}]$ for convenience. Following Theorem 2, it suffices to show that the proposed conditions hold if and only if $\boldsymbol{U},\bar{\boldsymbol{V}}^{\boldsymbol{x}}$ satisfy the conditions in Theorem 2 for any $\boldsymbol{x}$. The case of $\operatorname{rank}(\boldsymbol{U})=r$ is straightforward. Hence, in the rest of the prove, we only focus on the case of $\operatorname{rank}(\boldsymbol{U})<r$, i.e., the second condition in this theorem and that in Theorem 2, which are called Condition 2) and Condition $\tilde{2}$) throughout this proof. In particular, Condition $\tilde{2}$) can be stated as

$\tilde{2}$) for all $\mathcal{I}\subseteq[r]$, $\boldsymbol{\beta}\in\Re^n_+$, $\boldsymbol{x}\in\Re^l$ with $|\mathcal{I}|=\operatorname{rank}(\boldsymbol{U})+1$, $\operatorname{rank}(\boldsymbol{U}_{\mathcal{I}})=\operatorname{rank}(\boldsymbol{U})$ and $\bar{\boldsymbol{V}}^{\boldsymbol{x}}_{\mathcal{I}}\boldsymbol{\beta}\in\operatorname{span}(\boldsymbol{U}_{\mathcal{I}})$, we must have $\beta_i(\bar{\boldsymbol{V}}^{\boldsymbol{x}}_{\mathcal{I}})_i\in\operatorname{span}(\boldsymbol{U}_{\mathcal{I}})$ holds for every $i\in[n]$.

We now prove that Condition 2) is equivalent to Condition $\tilde{2}$).

First, we make an equivalent interpretation for Condition 2) and Condition $\tilde{2}$). Notice that both conditions are for the same set of index sets. We consider any such index set $\mathcal{I}$. Since $|\mathcal{I}|=\operatorname{rank}(\boldsymbol{U}_{\mathcal{I}})+1$, $\operatorname{span}(\boldsymbol{U}_{\mathcal{I}})$ is a hyperplane in $\Re^{|\mathcal{I}|}$. Therefore, there exists a unit vector $\boldsymbol{\eta}\in\Re^{|\mathcal{I}|}$ such that it is orthogonal to all vectors in span($\boldsymbol{U}_{\mathcal{I}}$), and all elements in $\Re^{|\mathcal{I}|}$ can be represented as linear combinations of $\boldsymbol{\eta}$ and a vector in span($\boldsymbol{U}_{\mathcal{I}}$). Therefore, for any $i,j \in [n]$,

$$\begin{array}{ll} (\boldsymbol{V}_{\mathcal{I}})_i = \boldsymbol{\xi}_i + \lambda_i \boldsymbol{\eta}, & \boldsymbol{W}_{\mathcal{I}}^i \boldsymbol{x} = \boldsymbol{\zeta}_i^{\boldsymbol{x}} + \mu_i^{\boldsymbol{x}} \boldsymbol{\eta}, \\ (\boldsymbol{V}_{\mathcal{I}})_j = \boldsymbol{\xi}_j + \lambda_j \boldsymbol{\eta}, & \boldsymbol{W}_{\mathcal{I}}^j \boldsymbol{x} = \boldsymbol{\zeta}_j^{\boldsymbol{x}} + \mu_j^{\boldsymbol{x}} \boldsymbol{\eta}. \end{array} \tag{50}$$

for some $\lambda_i, \lambda_j, \mu_i^{\boldsymbol{x}}, \mu_j^{\boldsymbol{x}} \in \Re$ and $\boldsymbol{\xi}_i, \boldsymbol{\xi}_j, \boldsymbol{\zeta}_i^{\boldsymbol{x}}, \boldsymbol{\zeta}_j^{\boldsymbol{x}} \in \text{span}(\boldsymbol{U}_{\mathcal{I}})$. Since $\boldsymbol{\eta}$ is a unit vector, we have

$$0 = \boldsymbol{\eta}^\top \left((\boldsymbol{V}_{\mathcal{I}})_i - \boldsymbol{\eta}\boldsymbol{\eta}^\top (\boldsymbol{V}_{\mathcal{I}})_i\right) = \boldsymbol{\eta}^\top (\boldsymbol{V}_{\mathcal{I}})_i - \boldsymbol{\eta}^\top \boldsymbol{\eta}\boldsymbol{\eta}^\top (\boldsymbol{\xi}_i + \lambda_i \boldsymbol{\eta}) = \boldsymbol{\eta}^\top (\boldsymbol{V}_{\mathcal{I}})_i - \lambda_i,$$

and hence $\lambda_i = \boldsymbol{\eta}^\top (\boldsymbol{V}_{\mathcal{I}})_i$. The same logic applies to $(\boldsymbol{V}_{\mathcal{I}})_j$ and $\boldsymbol{W}_{\mathcal{I}}^i \boldsymbol{x}, \boldsymbol{W}_{\mathcal{I}}^j \boldsymbol{x}$.

In Condition 2), we notice that $(\boldsymbol{\eta}^\top (\boldsymbol{V}_{\mathcal{I}})_i) \cdot (\boldsymbol{\eta}^\top (\boldsymbol{V}_{\mathcal{I}})_j) \geq 0$ is equivalent to $\lambda_i \lambda_j \geq 0$; moreover, $(\boldsymbol{W}_{\mathcal{I}}^i)^\top \boldsymbol{\eta}\boldsymbol{\eta}^\top \boldsymbol{W}_{\mathcal{I}}^j$ is positive semidefinite if and only if $(\boldsymbol{\eta}^\top \boldsymbol{W}_{\mathcal{I}}^i \boldsymbol{x}) \cdot \left(\boldsymbol{\eta}^\top \boldsymbol{W}_{\mathcal{I}}^j \boldsymbol{x}\right) \geq 0$ for all $\boldsymbol{x}$. Hence, we conclude that Condition 2a) is equivalent to "$\lambda_i \lambda_j \geq 0, \mu_i^{\boldsymbol{x}} \mu_j^{\boldsymbol{x}} \geq 0$ for all $\boldsymbol{x}$ and $i,j \in [n]$". For Condition 2b), since the equality holds if and only if $(\boldsymbol{\eta}^\top (\boldsymbol{V}_{\mathcal{I}})_i) \cdot \left(\boldsymbol{\eta}^\top \boldsymbol{W}_{\mathcal{I}}^j \boldsymbol{x}\right) = (\boldsymbol{\eta}^\top (\boldsymbol{V}_{\mathcal{I}})_j) \cdot (\boldsymbol{\eta}^\top \boldsymbol{W}_{\mathcal{I}}^i \boldsymbol{x})$ for all $\boldsymbol{x}$, we conclude that it is equivalent to the condition "$\lambda_i \mu_j^{\boldsymbol{x}} = \lambda_j \mu_i^{\boldsymbol{x}}$ for all $\boldsymbol{x}$ and $i,j \in [n]$".

For Condition $\tilde{2}$), by the definition of $\bar{\boldsymbol{V}}_{\mathcal{I}}^{\boldsymbol{x}}$,

$$\begin{array}{l} (\bar{\boldsymbol{V}}_{\mathcal{I}}^{\boldsymbol{x}})_i = (\boldsymbol{V}_{\mathcal{I}})_i - \boldsymbol{W}_{\mathcal{I}}^i \boldsymbol{x} = (\boldsymbol{\xi}_i - \boldsymbol{\zeta}_i^{\boldsymbol{x}}) + (\lambda_i - \mu_i^{\boldsymbol{x}}) \boldsymbol{\eta}, \\ (\bar{\boldsymbol{V}}_{\mathcal{I}}^{\boldsymbol{x}})_j = (\boldsymbol{V}_{\mathcal{I}})_j - \boldsymbol{W}_{\mathcal{I}}^j \boldsymbol{x} = (\boldsymbol{\xi}_j - \boldsymbol{\zeta}_j^{\boldsymbol{x}}) + (\lambda_j - \mu_j^{\boldsymbol{x}}) \boldsymbol{\eta}. \end{array}$$

Observing that Condition $\tilde{2}$) is violated if and only if there exist $i,j \in [n]$ with $(\lambda_i - \mu_i^{\boldsymbol{x}})(\lambda_j - \mu_j^{\boldsymbol{x}}) < 0$, we obtain its equivalent condition as

$$(\lambda_i - \mu_i^{\boldsymbol{x}})(\lambda_j - \mu_j^{\boldsymbol{x}}) \geq 0 \qquad \forall \boldsymbol{x} \in \Re^l, i,j \in [n]. \tag{51}$$

We now prove the direction "Condition $\tilde{2}$) $\implies$ Condition 2)". Consider any two distinct indexes $i,j \in [n]$. We first choose $\boldsymbol{x} = \boldsymbol{0}$, hence $\bar{\boldsymbol{V}}^{\boldsymbol{x}} = \boldsymbol{V}$. We assume the contrary to the first argument of Condition 2a), i.e., $\lambda_i \lambda_j < 0$, WLOG, $\lambda_i > 0$, $\lambda_j < 0$. We then choose $\boldsymbol{\beta} = -\lambda_j \boldsymbol{e}_i + \lambda_i \boldsymbol{e}_j \in \Re_+^n$, and have $\boldsymbol{V}_{\mathcal{I}} \boldsymbol{\beta} = -\lambda_j \boldsymbol{\xi}_i + \lambda_i \boldsymbol{\xi}_j \in \text{span}(\boldsymbol{U}_{\mathcal{I}})$. However, $\beta_j (\boldsymbol{V}_{\mathcal{I}})_j = \lambda_i \boldsymbol{\xi}_j + \lambda_i \lambda_j \boldsymbol{\eta} \notin \text{span}(\boldsymbol{U}_{\mathcal{I}})$. We hence have contradiction with Condition $\tilde{2}$), and conclude $\lambda_i \lambda_j \geq 0$, the first argument of Condition 2a) is true.

Next we show the second argument of Condition 2a) by contradiction. Notice for any constant $\theta \in \Re$, $(\bar{\boldsymbol{V}}_{\mathcal{I}}^{\theta \boldsymbol{x}})_i = (\boldsymbol{\xi}_i - \theta \boldsymbol{\zeta}_i^{\boldsymbol{x}}) + (\lambda_i - \theta \mu_i^{\boldsymbol{x}}) \boldsymbol{\eta}, (\bar{\boldsymbol{V}}_{\mathcal{I}}^{\theta \boldsymbol{x}})_j = (\boldsymbol{\xi}_j - \theta \boldsymbol{\zeta}_j^{\boldsymbol{x}}) + (\lambda_j - \theta \mu_j^{\boldsymbol{x}}) \boldsymbol{\eta}$. If $\mu_i^{\boldsymbol{x}} \mu_j^{\boldsymbol{x}} < 0$, we can always find θ such that $(\lambda_i - \theta \mu_i^{\boldsymbol{x}})(\lambda_j - \theta \mu_j^{\boldsymbol{x}}) = \mu_i^{\boldsymbol{x}} \mu_j^{\boldsymbol{x}} \theta^2 - (\lambda_i \mu_j^{\boldsymbol{x}} + \mu_i^{\boldsymbol{x}} \lambda_j)\theta + \lambda_i \lambda_j < 0$. Therefore, the equivalent condition for Condition $\tilde{2}$), i.e., (51), is violated for $\theta \boldsymbol{x}$. Hence, we have contradiction, and conclude that $\mu_i^{\boldsymbol{x}} \mu_j^{\boldsymbol{x}} \geq 0$, the second argument of Condition 2a) is true.

We now prove Condition 2b). By Condition 2a), we already have $\lambda_i \lambda_j \geq 0, \mu_i^{\boldsymbol{x}} \mu_j^{\boldsymbol{x}} \geq 0$. WLOG, we assume $\lambda_i, \lambda_j, \mu_i^{\boldsymbol{x}}, \mu_j^{\boldsymbol{x}} \geq 0$. Assume the opposite to Condition 2b), i.e., $\lambda_i \mu_j^{\boldsymbol{x}} \neq \lambda_j \mu_i^{\boldsymbol{x}}$. WLOG, we let $0 \leq \lambda_i \mu_j^{\boldsymbol{x}} < \lambda_j \mu_i^{\boldsymbol{x}}$, which implies $\lambda_j, \mu_i^{\boldsymbol{x}} > 0$. By Condition $\tilde{2}$), i.e., (51), we have $(\lambda_i - \mu_i^{\boldsymbol{x}})(\lambda_j - \mu_j^{\boldsymbol{x}}) \geq 0$. Combining with $\lambda_j, \mu_i^{\boldsymbol{x}} > 0$, we know that at least one of $\lambda_i, \mu_j^{\boldsymbol{x}}$ is nonzero. Consider the case that $\mu_j^{\boldsymbol{x}} \neq 0$. Define $\theta_i = \lambda_i/\mu_i^{\boldsymbol{x}}, \theta_j = \lambda_j/\mu_j^{\boldsymbol{x}}$, then following the assumptions of $\lambda_i\mu_j^{\boldsymbol{x}} < \lambda_j\mu_i^{\boldsymbol{x}}$ we have $\theta_i < \theta_j$. Choosing any $\theta \in (\theta_i, \theta_j)$, we have $\lambda_i < \theta\mu_i^{\boldsymbol{x}}, \lambda_j > \theta\mu_j^{\boldsymbol{x}}$. Hence, the equivalent condition for Condition $\tilde{2}$), i.e., (51), is violated for $\theta\boldsymbol{x}$. The case of $\lambda_j \neq 0$ can be proved similarly. Hence, we always have contradiction, and conclude that Condition 2b) is true.

Now it remains to prove the direction "Condition 2) $\implies$ Condition $\tilde{2}$)". Given any $\boldsymbol{x} \in \Re^l$, we let $\lambda_i, \lambda_j, \mu_i^{\boldsymbol{x}}, \mu_j^{\boldsymbol{x}} \in \Re$ and $\boldsymbol{\xi}_i, \boldsymbol{\xi}_j, \boldsymbol{\zeta}_i^{\boldsymbol{x}}, \boldsymbol{\zeta}_j^{\boldsymbol{x}} \in \operatorname{span}(\boldsymbol{U}_{\mathcal{I}})$ be constants as defined in (50). By Condition 2), we know $\lambda_i\lambda_j \geq 0, \mu_i^{\boldsymbol{x}}\mu_j^{\boldsymbol{x}} \geq 0$ and $\lambda_i\mu_j^{\boldsymbol{x}} = \lambda_j\mu_i^{\boldsymbol{x}}$. WLOG, we let $\lambda_i, \lambda_j \geq 0$. Possible realizations of the parameters are as follows.

- $\lambda_i = \lambda_j = 0$. Then either $\mu_i^{\boldsymbol{x}}, \mu_j^{\boldsymbol{x}} \geq 0$ or $\mu_i^{\boldsymbol{x}}, \mu_j^{\boldsymbol{x}} \leq 0$, it always implies $(\lambda_i - \mu_i^{\boldsymbol{x}})(\lambda_j - \mu_j^{\boldsymbol{x}}) \geq 0$.
- $\lambda_i = 0, \lambda_j > 0$ (or $\lambda_i > 0, \lambda_j = 0$). Then $\lambda_i\mu_j^{\boldsymbol{x}} = \lambda_j\mu_i^{\boldsymbol{x}} = 0$, implying $\mu_i^{\boldsymbol{x}} = 0$ (or $\mu_j^{\boldsymbol{x}} = 0$). In either case we have $(\lambda_i - \mu_i^{\boldsymbol{x}})(\lambda_j - \mu_j^{\boldsymbol{x}}) \geq 0$.
- $\lambda_i > 0, \lambda_j > 0$. Denote $\theta = \lambda_i/\mu_i^{\boldsymbol{x}} = \lambda_j/\mu_j^{\boldsymbol{x}}$, then $(\lambda_i - \mu_i^{\boldsymbol{x}})(\lambda_j - \mu_j^{\boldsymbol{x}}) = \mu_i^{\boldsymbol{x}}\mu_j^{\boldsymbol{x}}(\theta - 1)^2 \geq 0$.

So we always have $(\lambda_i - \mu_i^{\boldsymbol{x}})(\lambda_j - \mu_j^{\boldsymbol{x}}) \geq 0$, Condition $\tilde{2}$) holds. □

C.21. Proof of Lemma 3

Consider any $\boldsymbol{z}', \boldsymbol{z}'' \in \Re^n$, we denote $a = h(\boldsymbol{z}' \wedge \boldsymbol{z}''), b = h(\boldsymbol{z}'), c = h(\boldsymbol{z}''), d = h(\boldsymbol{z}' \vee \boldsymbol{z}'')$ and $d_0 = b + c - a$. From the supermodularity of f we know $b + c \leq a + d$; hence, $d_0 \leq d$. We then have

$$\phi(\boldsymbol{z}') + \phi(\boldsymbol{z}'') = u(b) + u(c) \leq u(a) + u(d_0) \leq u(a) + u(d) = \phi(\boldsymbol{z}' \wedge \boldsymbol{z}'') + \phi(\boldsymbol{z}' \vee \boldsymbol{z}''), \tag{52}$$

where the second inequality holds since u is non-decreasing, and the first equality can be proved as follows. We notice that either $a \leq \min\{b, c\} \leq \max\{b, c\} \leq d_0$ (if h is increasing) or $a \geq \max\{b, c\} \geq \min\{b, c\} \geq d_0$ (if h is decreasing) holds; since $a + d_0 = b + c$ and u is convex, we then have the first inequality in Equation (52). That proves the supermodularity of ϕ. □

C.22. Proof of Proposition 12

Applying Lemma 3, we have that $u\left(\boldsymbol{a}^\top\boldsymbol{x} + g(\boldsymbol{x}, \boldsymbol{z})\right)$ is supermodular in $\boldsymbol{z}$ for all $\boldsymbol{x} \in \mathcal{X}$. Hence, following Theorem 1 by treating $u\left(\boldsymbol{a}^\top\boldsymbol{x} + g(\boldsymbol{x}, \tilde{\boldsymbol{z}})\right)$ in Problem (27) as the $g(\boldsymbol{x}, \tilde{\boldsymbol{z}})$ in Problem (2), Problem (27) can be solved equivalently by

$$\begin{aligned}
\min \quad & \boldsymbol{\nu}^\top \boldsymbol{l} \\
\text{s.t.} \quad & \boldsymbol{R}_k^\top \boldsymbol{l} \geq \sum_{i \in [2n+1]} p_i^k u\left(\boldsymbol{a}^\top\boldsymbol{x} + \boldsymbol{b}^\top\boldsymbol{y}^{k,i}\right), && k \in [K] \\
& \boldsymbol{W}\boldsymbol{x} + \boldsymbol{U}\boldsymbol{y}^{k,i} \geq \boldsymbol{V}\boldsymbol{z}^{k,i} + \boldsymbol{v}^0, && k \in [K], i \in [2n+1] \\
& \boldsymbol{l} \geq \boldsymbol{0}, \ \boldsymbol{x} \in \mathcal{X},
\end{aligned}$$

Introducing auxiliary variables $f^{k,i}$ with $f^{k,i} \geq u\left(\boldsymbol{a}^\top\boldsymbol{x} + \boldsymbol{b}^\top\boldsymbol{y}^{k,i}\right) = \max_{j \in [j]}\{c_j\left(\boldsymbol{a}^\top\boldsymbol{x} + \boldsymbol{b}^\top\boldsymbol{y}^{k,i}\right) + d_j\}$, we then obtain the equivalent reformulation as in (29). □

C.23. Proof of Corollary 3

By the minimax Theorem in Sion (1958), in Problem (30), we can interchange the maximization over $\mathbb{P}\in\mathcal{F}$ and the minimization over $\theta\in\Re$. Hence, Problem (30) is equivalent to

$$\begin{aligned} \min \quad & \theta + \sup_{\mathbb{P}\in\mathcal{F}} \mathbb{E}_{\mathbb{P}}\left[u\left(\boldsymbol{a}^\top\boldsymbol{x} + g(\boldsymbol{x},\tilde{\boldsymbol{z}}) - \theta\right)\right] \\ \text{s.t.} \quad & \boldsymbol{x}\in\mathcal{X}. \end{aligned}$$

Its equivalent reformulation (31) can be obtained as a direct application of Proposition 12. □

C.24. Proof of Theorem 7

We first prove the direction of "1" → "2", by contradiction. Assume "2" is false, i.e., there exists $i\in[n],\ j\in[J]$ such that h_i^j has at least three pieces on $[\underline{z}_i,\overline{z}_i]$, then it suffices to show there are f^1,f^2 that are associated with worst-case distributions with distinct marginals.

WLOG, let h_1^1 be the function which at least three pieces on $[\underline{z}_1,\overline{z}_1]$. We choose functions $f^1,f^2:\Re^n\to\Re$ such that $f^1(\boldsymbol{z})=g^1(z_1), f^2(\boldsymbol{z})=g^2(z_1), \boldsymbol{z}\in\Re^n$ for some $g^1,g^2:\Re\to\Re$. Moreover, for all $j\in\{2,\ldots,J\}$, we choose δ_1^j to be sufficiently large such that $\mathbb{E}_{\mathbb{P}}\left[h_1^j(\tilde{z}_1)\right]\le\delta_1^j$ holds for all $\mathbb{P}\in\{\mathbb{P}\mid\mathbb{P}(\underline{z}_1\le\tilde{z}_1\le\overline{z}_1)=1,\mathbb{E}_{\mathbb{P}}[\tilde{z}_1]=\mu_1\}$. We then have for $i=1,2$,

$$\sup_{\mathbb{P}\in\mathcal{F}^G}\mathbb{E}_{\mathbb{P}}\left[f^i(\tilde{\boldsymbol{z}})\right]=\sup_{\mathbb{P}\in\mathcal{G}}\mathbb{E}_{\mathbb{P}}\left[g^i(\tilde{z}_1)\right]$$

where

$$\mathcal{G}=\left\{\mathbb{P}\ \middle|\ \begin{array}{l}\mathbb{P}\big(\underline{z}_1\le\tilde{z}_1\le\overline{z}_1\big)=1\\ \mathbb{E}_{\mathbb{P}}[\tilde{z}_1]=\mu_1\\ \mathbb{E}_{\mathbb{P}}\big[h_1^1(\tilde{z}_1)\big]\le\delta_1^1.\end{array}\right\},$$

For notational simplicity, we omit the superscript and subscript of h and δ, as well as the subscript of z and μ. That is, we consider $\sup_{\mathbb{P}\in\mathcal{G}}\mathbb{E}_{\mathbb{P}}\left[g^i(\tilde{z})\right]$ with $\mathcal{G}=\left\{\mathbb{P}\ \middle|\ \begin{array}{l}\mathbb{P}\big(\underline{z}\le\tilde{z}\le\overline{z}\big)=1\\ \mathbb{E}_{\mathbb{P}}[\tilde{z}]=\mu\\ \mathbb{E}_{\mathbb{P}}\big[h(\tilde{z})\big]\le\delta\end{array}\right\}$, where h has at least three pieces on $[\underline{z},\overline{z}]$. Now it suffices to find $g^1,g^2:\Re\to\Re$ such that there does not exist a common worst-case distribution.

Let $J+1$ be the number of pieces of h on $[\underline{z},\overline{z}]$ for some $J\ge2$, and denote the corresponding breakpoints by $\underline{z}=z^0<\cdots<z^{J+1}=\overline{z}$. We define two functions $l^1,l^2:\Re_+\to\Re$ such that

$$\begin{aligned} l^1(p_1)&\in\left\{p_0h(z^0)+p_1h(z^1)+p_{J+1}h(z^{J+1})\mid p_0+p_1+p_{J+1}=1,p_0z^0+p_1z^1+p_{J+1}z^{J+1}=\mu\right\}\\ l^2(p_2)&\in\left\{p_0h(z^0)+p_2h(z^2)+p_{J+1}h(z^{J+1})\mid p_0+p_2+p_{J+1}=1,p_0z^0+p_2z^2+p_{J+1}z^{J+1}=\mu\right\}\end{aligned}\tag{53}$$

Notice that the sets in Equation (53) are singleton since for any given p_1 or p_2, we have unique p_0 and p_{J+1}. Therefore, the functions l^1,l^2 are indeed uniquely determined by Equation (53). We have two observations on l^1,l^2. First, $l^1(0)=l^2(0)$, and when $p_1=p_2=0$, their corresponding p_0 and p_{J+1} (in the set defined in Equation (53)) are strictly positive. Second, both l^1,l^2 are continuous function, and they are also increasing function due to the convexity of h. By the two observations, we can find $\epsilon_1, \epsilon_2 > 0$ which are sufficiently small and such that $l^1(\epsilon_1) = l^2(\epsilon_2)$, and when $p_1 = \epsilon_1$ and $p_2 = \epsilon_2$, their corresponding p_0 and p_{J+1} are strictly positive. Define

$$\boldsymbol{H}^1 = \begin{bmatrix} 1 & 1 & 1 \\ z^0 & z^1 & z^{J+1} \\ h(z^0) & h(z^1) & h(z^{J+1}) \end{bmatrix}, \qquad \boldsymbol{H}^2 = \begin{bmatrix} 1 & 1 & 1 \\ z^0 & z^2 & z^{J+1} \\ h(z^0) & h(z^2) & h(z^{J+1}) \end{bmatrix}.$$

We hence can find $\boldsymbol{p}^1, \boldsymbol{p}^2 \in \Re^3_{++}$ and choose $\delta \in \Re$ such that $\boldsymbol{H}^1\boldsymbol{p}^1 = \boldsymbol{H}^2\boldsymbol{p}^2 = (1, \mu, \delta)$. Let the discrete probability $\mathbb{P}^1, \mathbb{P}^2$ be with

$$\mathbb{P}^1(\tilde{z} = w) = \begin{cases} p_1^1 \text{ if } w = z^0 \\ p_2^1 \text{ if } w = z^1 \\ p_3^1 \text{ if } w = z^{J+1} \\ 0 \ \ \text{otherwise} \end{cases}, \qquad \mathbb{P}^2(\tilde{z} = w) = \begin{cases} p_1^2 \text{ if } w = z^0 \\ p_2^2 \text{ if } w = z^2 \\ p_3^2 \text{ if } w = z^{J+1} \\ 0 \ \ \text{otherwise} \end{cases}.$$

Then $\mathbb{P}^1, \mathbb{P}^2$ have the support $\mathcal{Z}^1 = \{z^0, z^1, z^{J+1}\}, \mathcal{Z}^2 = \{z^0, z^2, z^{J+1}\}$, respectively.

Consider any $i \in \{1, 2\}$. Since h is piecewise linear convex, we can choose a convex function g^i such that $g^i(z) = h(z)$ for $z \in \mathcal{Z}^i$ and $g^i(z) < h(z)$ for all $z \in [\underline{z}, \overline{z}] \setminus \mathcal{Z}^i$. Therefore, we have

$$\mathbb{E}_{\mathbb{P}^i}\left[g^i(z)\right] = \sum_{z \in \mathcal{Z}^i} \mathbb{P}(\tilde{z} = z) g^i(z) = \sum_{z \in \mathcal{Z}^i} \mathbb{P}(\tilde{z} = z) h(z) = \mathbb{E}_{\mathbb{P}^i}[h(z)] = \delta,$$

where the first and and third equalities are by the definition of $\mathbb{P}^i$, the second equality holds since $g^i(z) = h(z)$ when $z \in \mathcal{Z}^i$, and the last equality is due to the way we choose $\boldsymbol{p}^i$. Since $g^i(z) \le h(z)$, we have $\mathbb{E}_{\mathbb{P}}[g^i(z)] \le \mathbb{E}_{\mathbb{P}}[h(z)] \le \delta$ for any $\mathbb{P} \in \mathcal{G}$. Hence $\mathbb{P}^i$ is a worst-case distribution to $\sup_{\mathbb{P} \in \mathcal{G}} \mathbb{E}_{\mathbb{P}}[g^i(\tilde{z})]$. In what follows, we show that $\mathbb{P}^i$ is the only worst-case distribution.

We first consider any $\mathbb{P} \in \mathcal{G}$ with support $\mathcal{Z}$ such that $\mathcal{Z} \setminus \mathcal{Z}^i \neq \emptyset$, then there exists $[z', z''] \subseteq [\underline{z}, \overline{z}] \setminus \mathcal{Z}^i$ such that $\mathbb{P}(\tilde{z} \in [z', z'']) > 0$. Therefore,

$$\mathbb{E}_{\mathbb{P}}\left[g^i(z)\right] = \int_{[\underline{z}, \overline{z}]} g^i(z) d\mathbb{P} < \int_{[\underline{z}, \overline{z}]} h(z) d\mathbb{P} = \mathbb{E}_{\mathbb{P}}[h(z)] \le \delta,$$

where the first inequality follows from that $g^i(z) < h(z)$ for all $z \in [\underline{z}, \overline{z}] \setminus \mathcal{Z}^i$, the last inequality is due to $\mathbb{P} \in \mathcal{G}$. Hence, $\mathbb{P} \notin \arg\sup_{\mathbb{P} \in \mathcal{G}} \mathbb{E}_{\mathbb{P}}[g^i(\tilde{z})]$. It implies that for any $\mathbb{P}^* \in \arg\sup_{\mathbb{P} \in \mathcal{G}} \mathbb{E}_{\mathbb{P}}[g^i(\tilde{z})]$, the support of $\mathbb{P}^*$ must be a subset of $\mathcal{Z}^i$, and $\mathbb{P}^*$ can be fully characterized by a vector $\boldsymbol{p}^* \in \Re^3_+$ such that $\boldsymbol{H}^i\boldsymbol{p}^* = (1, \mu, \delta)$. Observing that $\boldsymbol{H}^i$ is invertible (due to that h is not linear), $\boldsymbol{p}^\star$ is unique and is exactly the aforementioned $\boldsymbol{p}^i$. Therefore, $\mathbb{P}^i$ is the unique worst-case distribution to $\sup_{\mathbb{P} \in \mathcal{G}} \mathbb{E}_{\mathbb{P}}[g^i(\tilde{z})]$. Hence, there does not exist a common worst-case distribution to $\sup_{\mathbb{P} \in \mathcal{G}} \mathbb{E}_{\mathbb{P}}[g^1(\tilde{z})]$ and $\sup_{\mathbb{P} \in \mathcal{G}} \mathbb{E}_{\mathbb{P}}[g^2(\tilde{z})]$. "1" is false.

We next prove the direction of "2"→"1".

By strong duality,

$$\sup_{\mathbb{P} \in \mathcal{F}^G} \mathbb{E}_{\mathbb{P}}[f(\tilde{\boldsymbol{z}})] = \inf \left\{ s + \boldsymbol{\mu}^\top \boldsymbol{t} + \sum_{i=1}^n \sum_{j=1}^{J_i} \delta_i^j r_i^j \;\middle|\; \begin{array}{l} s + \boldsymbol{z}^\top \boldsymbol{t} + \sum_{i=1}^n \sum_{j=1}^{J_i} h_i^j(z_i) r_i^j \ge f(\boldsymbol{z}), \ \forall \boldsymbol{z} \in [\underline{\boldsymbol{z}}, \overline{\boldsymbol{z}}] \\ r_i^j \ge 0, \ i \in [n], j \in [J_i] \end{array} \right\}.$$

Let $\mathcal{Z} = \{ \boldsymbol{z} \mid z_i \in \{z_i^1, \ldots, z_i^{S_i}\}, i \in [n] \}$ which contains all $\boldsymbol{z}$ such that each of its dimension is on the breakpoints. Then we observe that $[\underline{\boldsymbol{z}}, \overline{\boldsymbol{z}}]$ can be decomposed as $[\underline{\boldsymbol{z}}, \overline{\boldsymbol{z}}] = \cup_{i=1}^{S} \mathcal{Z}^i$ for some S and disjoint $\mathcal{Z}^1, \ldots, \mathcal{Z}^S$ such that all $\mathcal{Z}^i$ are boxes with extreme points in $\mathcal{Z}$ and $\sum_{i=1}^n \sum_{j=1}^{J_i} h_i^j(z_i)$ are linear within each $\mathcal{Z}^i$. Together with the convexity of f, the dual problem is equivalent to

$$\inf \left\{ s + \boldsymbol{\mu}^\top \boldsymbol{t} + \sum_{i=1}^{n} \sum_{j=1}^{J_i} \delta_i^j r_i^j \;\middle|\; \begin{array}{l} s + \boldsymbol{z}^\top \boldsymbol{t} + \sum_{i=1}^n \sum_{j=1}^{J_i} h_i^j(z_i) r_i^j \geq f(\boldsymbol{z}),\ \forall \boldsymbol{z} \in \mathcal{Z} \\ r_i^j \geq 0,\ i \in [n], j \in [J_i] \end{array} \right\}.$$

Writing its dual form again, we conclude that there exists a worst-case distribution with its support as $\mathcal{Z}$. Hence, for $\sup_{\mathbb{P} \in \mathcal{F}^G} \mathbb{E}_{\mathbb{P}}[f(\tilde{\boldsymbol{z}})]$, it suffices to consider the probability distributions with support $\mathcal{Z}$.

Assuming “2” is true, i.e., $h_i^j, i \in [n], j \in [J_i]$ are piecewise linear convex functions with exactly two pieces on $[\underline{z}_i, \overline{z}_i]$, we will show “1” is true. In other words, we will show the existence of a $\mathbb{P}^* \in \arg\sup_{\mathbb{P} \in \mathcal{F}^G} \mathbb{E}_{\mathbb{P}}[f(\tilde{\boldsymbol{z}})]$ such that for any dimension i, $\mathbb{P}^*(\tilde{z}_i = w)$ has the structure as in “1”. WLOG, we let such i be n. Further, for notational simplicity, we drop the subscript n for $\tilde{z}_n, z_n, \underline{z}_n, \overline{z}_n, \mu_n, h_n^j, \delta_n^j, J_n$. Hence, we have $\tilde{\boldsymbol{z}} = (\tilde{z}_1, \ldots, \tilde{z}_{n-1}, \tilde{z})$, $\boldsymbol{z} = (z_1, \ldots, z_{n-1}, z)$ and so on, and we will prove that $\mathbb{P}^*(\tilde{z} = w)$ has the structure as in “1”.

The proof will be done by induction. Starting from the case of $J = 1$, with an approach almost the same as that in the proof for Proposition 1, we can show that $\mathbb{P}^*$ has the structure in “1”. More specifically, denoting the breakpoint of h^1 by $\hat{z} \in (\underline{z}, \overline{z})$, then we move the probability mass on $\boldsymbol{z}$ with $z = \hat{z}$ to $\boldsymbol{z} - (\hat{z} - \underline{z})\boldsymbol{e}_n$ and $\boldsymbol{z} + (\overline{z} - \hat{z})\boldsymbol{e}_n$ until we cannot move any further. Such move will terminate at a probability distribution which has marginals in the form given by “1”, and the expected value of $\mathbb{E}_{\mathbb{P}}[f(\tilde{\boldsymbol{z}})]$ is no less.

Suppose when $J = \hat{J} - 1$ for some $\hat{J} \geq 2$, we have “1” being true. We now consider the case of $J = \hat{J}$. We separately analyze the following two scenarios.

- **Scenario I**: There are distinct $i, j \in [J]$ such that h^i, h^j have the same breakpoint in $(\underline{z}, \overline{z})$. WLOG, we let h^1, h^2 be both with breakpoint $\hat{z} \in (\underline{z}, \overline{z})$. Define $\hat{\mathcal{G}} = \{\mathbb{P} \mid \mathbb{P}(\underline{z} \leq \tilde{z} \leq \overline{z}) = 1, \mathbb{E}_{\mathbb{P}}[\tilde{z}] = \mu\}$, $\mathcal{G}^i = \{\mathbb{P} \mid \mathbb{E}_{\mathbb{P}}[h^i(\tilde{z})] \leq \delta^i\}$, $i \in \{1, 2\}$.

If $\hat{\mathcal{G}} \cap \mathcal{G}^1 \cap \mathcal{G}^2 = \hat{\mathcal{G}} \cap \mathcal{G}^2$, then denote $\mathcal{G}'$ to be the ambiguity set obtained from $\mathcal{F}^G$ by removing the constraint on $\mathbb{E}_{\mathbb{P}}[h^1(\tilde{z})]$. We have $\mathcal{G}' = \mathcal{F}^G$, and hence $\sup_{\mathbb{P} \in \mathcal{F}^G} \mathbb{E}_{\mathbb{P}}[f(\tilde{\boldsymbol{z}})] = \sup_{\mathbb{P} \in \mathcal{G}'} \mathbb{E}_{\mathbb{P}}[f(\tilde{\boldsymbol{z}})]$. Therefore, we have a problem with $J = \hat{J} - 1$, in which case we know “1” is true by the induction assumption.

If $\hat{\mathcal{G}} \cap \mathcal{G}^1 \cap \mathcal{G}^2 \neq \hat{\mathcal{G}} \cap \mathcal{G}^2$, we next show $\hat{\mathcal{G}} \cap \mathcal{G}^1 \cap \mathcal{G}^2 = \hat{\mathcal{G}} \cap \mathcal{G}^1$.

Consider any $\mathbb{P} \in \hat{\mathcal{G}}$, we define a vector $(s_1^{\mathbb{P}}, s_2^{\mathbb{P}}, \underline{s}^{\mathbb{P}}, \hat{s}^{\mathbb{P}}, \overline{s}^{\mathbb{P}})$ which is uniquely determined by the following system of equations,

$$\begin{cases} \displaystyle\int_{z \le \hat{z}} z d\mathbb{P}(\boldsymbol{z}) = \underline{z}\underline{s}^{\mathbb{P}} + \hat{z} s_1^{\mathbb{P}} \\ \mathbb{P}(\tilde{z} \le \hat{z}) = \underline{s}^{\mathbb{P}} + s_1^{\mathbb{P}} \\ \displaystyle\int_{z > \hat{z}} z d\mathbb{P}(\boldsymbol{z}) = \hat{z} s_2^{\mathbb{P}} + \overline{z}\overline{s}^{\mathbb{P}}. \\ \mathbb{P}(\tilde{z} > \hat{z}) = s_2^{\mathbb{P}} + \overline{s}^{\mathbb{P}} \\ \hat{s}^{\mathbb{P}} = s_1^{\mathbb{P}} + s_2^{\mathbb{P}} \end{cases} \tag{54}$$

In this case, for any piecewise linear convex function with two pieces and with breakpoint at $\hat{z}$, which can be denoted by $h(z) = \begin{cases} \underline{a}z + \underline{b} \text{ if } z \le \hat{z} \\ \overline{a}z + \overline{b} \text{ if } z \ge \hat{z} \end{cases}$ where $\underline{a} < \overline{a}$, we have

$$\begin{aligned} \mathbb{E}_{\mathbb{P}}[h(\tilde{z})] &= \int_{z \le \hat{z}} (\underline{a}z + \underline{b})\, d\mathbb{P}(\boldsymbol{z}) + \int_{z > \hat{z}} (\overline{a}z + \overline{b})\, d\mathbb{P}(\boldsymbol{z}) \\ &= \underline{a} \int_{z \le \hat{z}} z d\mathbb{P}(\boldsymbol{z}) + \underline{b}\mathbb{P}(\tilde{z} \le \hat{z}) + \overline{a} \int_{z > \hat{z}} z d\mathbb{P}(\boldsymbol{z}) + \overline{b}\mathbb{P}(\tilde{z} > \hat{z}) \\ &= \underline{a}\left(\underline{z}\underline{s}^{\mathbb{P}} + \hat{z}s_1^{\mathbb{P}}\right) + \underline{b}\left(\underline{s}^{\mathbb{P}} + s_1^{\mathbb{P}}\right) + \overline{a}\left(\hat{z}s_2^{\mathbb{P}} + \overline{z}\overline{s}^{\mathbb{P}}\right) + \overline{b}\left(s_2^{\mathbb{P}} + \overline{s}^{\mathbb{P}}\right) \\ &= \underline{s}^{\mathbb{P}} h(\underline{z}) + \hat{s}^{\mathbb{P}} h(\hat{z}) + \overline{s}^{\mathbb{P}} h(\overline{z}), \end{aligned}$$

where the third and fourth inequalities are due to (54). Moreover, by (54) we can easily have $\underline{s}^{\mathbb{P}} + \hat{s}^{\mathbb{P}} + \overline{s}^{\mathbb{P}} = 1$ and $\underline{s}^{\mathbb{P}}\underline{z} + \hat{s}^{\mathbb{P}}\hat{z} + \overline{s}^{\mathbb{P}}\overline{z} = \mu$, which imply

$$\underline{s}^{\mathbb{P}} = \frac{\overline{z} - \mu - (\overline{z} - \hat{z})\hat{s}^{\mathbb{P}}}{\overline{z} - \underline{z}}, \ \overline{s}^{\mathbb{P}} = \frac{\mu - \underline{z} - (\hat{z} - \underline{z})\hat{s}^{\mathbb{P}}}{\overline{z} - \underline{z}}.$$

Therefore,

$$\mathbb{E}_{\mathbb{P}}[h(\tilde{z})] = \frac{\overline{z} - \mu - (\overline{z} - \hat{z})\hat{s}^{\mathbb{P}}}{\overline{z} - \underline{z}} h(\underline{z}) + \hat{s}^{\mathbb{P}} h(\hat{z}) + \frac{\mu - \underline{z} - (\hat{z} - \underline{z})\hat{s}^{\mathbb{P}}}{\overline{z} - \underline{z}} h(\overline{z}) = c^h + \Delta^h \hat{s}^{\mathbb{P}}, \tag{55}$$

where c^h, Δ^h are constants depending on h but independent from $\mathbb{P}$; moreover,

$$\Delta^h = h(\hat{z}) - \left(\frac{\overline{z} - \hat{z}}{\overline{z} - \underline{z}} h(\underline{z}) + \frac{\hat{z} - \underline{z}}{\overline{z} - \underline{z}} h(\overline{z})\right) < h(\hat{z}) - h\left(\frac{\overline{z} - \hat{z}}{\overline{z} - \underline{z}} \underline{z} + \frac{\hat{z} - \underline{z}}{\overline{z} - \underline{z}} \overline{z}\right) = h(\hat{z}) - h(\hat{z}) = 0, \tag{56}$$

where the inequality follows from the convexity of h.

Recall that $\hat{\mathcal{G}} \cap \mathcal{G}^1 \cap \mathcal{G}^2 \neq \hat{\mathcal{G}} \cap \mathcal{G}^2$, then there exists $\mathbb{P}^o \in \left(\hat{\mathcal{G}} \cap \mathcal{G}^2\right) \setminus \mathcal{G}^1$. Therefore, consider any $\hat{\mathbb{P}} \in \hat{\mathcal{G}} \cap \mathcal{G}^1$,

$$c^{h^1} + \Delta^{h^1} \hat{s}^{\mathbb{P}^o} = \mathbb{E}_{\mathbb{P}^o}\left[h^1(\tilde{z})\right] > \delta^1 \ge \mathbb{E}_{\hat{\mathbb{P}}}\left[h^1(\tilde{z})\right] = c^{h^1} + \Delta^{h^1} \hat{s}^{\hat{\mathbb{P}}},$$

where the two equalities follow from (55), the two inequalities are due to $\mathbb{P}^o \notin \mathcal{G}^1$ and $\hat{\mathbb{P}} \in \mathcal{G}^1$. Hence, we have $\hat{s}^{\mathbb{P}^o} < \hat{s}^{\hat{\mathbb{P}}}$ since (56) results in $\Delta^{h^1} < 0$. It then implies

$$\mathbb{E}_{\hat{\mathbb{P}}}\left[h^2(\tilde{z})\right] = c^{h^2} + \Delta^{h^2} \hat{s}^{\hat{\mathbb{P}}} < c^{h^2} + \Delta^{h^2} \hat{s}^{\mathbb{P}^o} = \mathbb{E}_{\mathbb{P}^o}\left[h^2(\tilde{z})\right] \le \delta^2,$$

where the last inequality holds since $\mathbb{P}^o \in \mathcal{G}^2$. Therefore, $\hat{\mathbb{P}} \in \mathcal{G}^2$, and we have $\hat{\mathcal{G}} \cap \mathcal{G}^1 \cap \mathcal{G}^2 = \hat{\mathcal{G}} \cap \mathcal{G}^1$. Similar to the case of $\hat{\mathcal{G}} \cap \mathcal{G}^1 \cap \mathcal{G}^2 = \hat{\mathcal{G}} \cap \mathcal{G}^2$, we now can reduce the problem $\sup_{\mathbb{P}\in\mathcal{F}^G} \mathbb{E}_{\mathbb{P}}[f(\tilde{\boldsymbol{z}})]$ to one with $J = \hat{J} - 1$, and hence "1" is true by induction.

- **Scenario II**: All h^j, $j \in [\hat{J}]$, have distinct breakpoints in $(\underline{z}, \overline{z})$. In this case, denote the breakpoint of h^j by z^j, $j \in [\hat{J}]$. WLOG, assume $\underline{z} = z^0 < z^1 < \cdots < z^{\hat{J}} < z^{\hat{J}+1} = \overline{z}$. Consider any $\mathbb{P} \in \mathcal{F}^G$. Denote by $p_j = \mathbb{P}(\tilde{z} = z^j)$ the marginal probability mass at $z = z^j, j = 0, \ldots, \hat{J}+1$. Recalling that we just focus on the distribution with support at the breakpoints, then the constraint $\mathbb{P} \in \mathcal{F}^G$ is equivalent to the following system,

$$\sum_{j=0}^{\hat{J}+1} p_j = 1, \tag{57a}$$

$$\sum_{j=0}^{\hat{J}+1} z^j p_j = \mu, \tag{57b}$$

$$\sum_{j=0}^{\hat{J}+1} h^i(z^j) p_j \le \delta^i, \qquad i \in [\hat{J}], \tag{57c}$$

$$p_j \ge 0, \qquad j \in \{0, \ldots, \hat{J}+1\}. \tag{57d}$$

By (57a) and (57b) we have

$$p_0 = \frac{\overline{z} - \mu - \sum_{j=1}^{\hat{J}} (\overline{z} - z^j) p_j}{\overline{z} - \underline{z}}, \; p_{\hat{J}+1} = \frac{\mu - \underline{z} - \sum_{j=1}^{\hat{J}} (z^j - \underline{z}) p_j}{\overline{z} - \underline{z}}, \tag{58}$$

which implies that $p_0, p_{\hat{J}+1}$ can be determined by $\boldsymbol{p} = (p_1, \ldots, p_{\hat{J}})$. In what follows, we simplify the constraints (57a)-(57d).

We first investigate the constraint (57c) for any given $i \in [\hat{J}]$. Since h^i is convex and has breakpoints $\{\underline{z}, z^i, \overline{z}\}$, we can denote $h^i(z) = \begin{cases} h^i(z^i) - \gamma_i(z^i - z) & \text{if } z \in [\underline{z}, z^i] \\ h^i(z^i) + \xi_i(z - z^i) & \text{if } z \in [z^i, \overline{z}] \end{cases}$ for some $\gamma_i < \xi_i$. It follows that

$$\begin{aligned}
&\sum_{j=0}^{\hat{J}+1} h^i(z^j) p_j \\
&= h^i(z^i) - \gamma_i(z^i - \underline{z}) p_0 - \gamma_i \sum_{j=1}^{i} (z^i - z^j) p_j + \xi_i \sum_{j=i}^{\hat{J}} (z^j - z^i) p_j + \xi_i(\overline{z} - z^i) p_{\hat{J}+1} \\
&= h^i(z^i) - \frac{\gamma_i}{\overline{z} - \underline{z}} \left((z^i - \underline{z}) \left(\overline{z} - \mu - \sum_{j=1}^{\hat{J}} (\overline{z} - z^j) p_j \right) + (\overline{z} - \underline{z}) \sum_{j=1}^{i} (z^i - z^j) p_j \right) \\
&\quad + \frac{\xi_i}{\overline{z} - \underline{z}} \left((\overline{z} - z^i) \left(\mu - \underline{z} - \sum_{j=1}^{\hat{J}} (z^j - \underline{z}) p_j \right) + (\overline{z} - \underline{z}) \sum_{j=i}^{\hat{J}} (z^j - z^i) p_j \right)
\end{aligned}$$

$$
\begin{aligned}
&= h^i(z^i) + \frac{\xi_i(\overline{z} - z^i)(\mu - \underline{z})}{\overline{z} - \underline{z}} - \frac{\gamma_i(\overline{z} - \mu)(z^i - \underline{z})}{\overline{z} - \underline{z}} \\
&\quad - \frac{\gamma_i}{\overline{z} - \underline{z}} \left(\sum_{j=1}^{i} \left((\overline{z} - \underline{z})(z^i - z^j) - (\overline{z} - z^j)(z^i - \underline{z}) \right) p_j - \sum_{j=i+1}^{\hat{J}} (\overline{z} - z^j)(z^i - \underline{z}) p_j \right) \\
&\quad + \frac{\xi_i}{\overline{z} - \underline{z}} \left(- \sum_{j=1}^{i-1} (\overline{z} - z^i)(z^j - \underline{z}) p_j + \sum_{j=i}^{\hat{J}} \left((z^j - z^i)(\overline{z} - \underline{z}) - (\overline{z} - z^i)(z^j - \underline{z}) \right) p_j \right) \\
&= h^i(z^i) + \frac{\xi_i(\overline{z} - z^i)(\mu - \underline{z})}{\overline{z} - \underline{z}} - \frac{\gamma_i(\overline{z} - \mu)(z^i - \underline{z})}{\overline{z} - \underline{z}} \\
&\quad + \frac{\gamma_i}{\overline{z} - \underline{z}} \left(- \sum_{j=1}^{i-1} (z^i - \overline{z})(z^j - \underline{z}) p_j + (\overline{z} - z^i)(z^i - \underline{z}) p_i + \sum_{j=i+1}^{\hat{J}} (\overline{z} - z^j)(z^i - \underline{z}) p_j \right) \\
&\quad - \frac{\xi_i}{\overline{z} - \underline{z}} \left(\sum_{j=1}^{i-1} (\overline{z} - z^i)(z^j - \underline{z}) p_j + (\overline{z} - z^i)(z^i - \underline{z}) p_i - \sum_{j=i+1}^{\hat{J}} (z^j - \overline{z})(z^i - \underline{z}) p_j \right) \\
&= h^i(z^i) + \frac{\xi_i(\overline{z} - z^i)(\mu - \underline{z})}{\overline{z} - \underline{z}} - \frac{\gamma_i(\overline{z} - \mu)(z^i - \underline{z})}{\overline{z} - \underline{z}} - \frac{\xi_i - \gamma_i}{\overline{z} - \underline{z}} \sum_{j=1}^{\hat{J}} \left(\overline{z} - z^{\max\{i,j\}} \right) \left(z^{\min\{i,j\}} - \underline{z} \right) p_j.
\end{aligned}
$$

Hence the i-th constraint of (57c) is equivalent to

$$
\sum_{j=1}^{\hat{J}} \left(\overline{z} - z^{\max\{i,j\}} \right) \left(z^{\min\{i,j\}} - \underline{z} \right) p_j \geq d_i,
$$

where $d_i = \frac{\overline{z} - \underline{z}}{\xi_i - \gamma_i} \left(h^i(z^i) + \frac{\xi_i(\overline{z} - z^i)(\mu - \underline{z})}{\overline{z} - \underline{z}} - \frac{\gamma_i(\overline{z} - \mu)(z^i - \underline{z})}{\overline{z} - \underline{z}} - \delta^i \right)$. Denote $\lambda_j = \overline{z} - z^j, \pi_j = z^j - \underline{z}$ for all $j \in [\hat{J}]$, and let

$$
\boldsymbol{A} = \left(\lambda_{\max\{i,j\}} \pi_{\min\{i,j\}} \right)_{i,j \in [\hat{J}]} = \begin{bmatrix} \lambda_1\pi_1 & \lambda_2\pi_1 & \lambda_3\pi_1 & \cdots & \lambda_{\hat{J}}\pi_1 \\ \lambda_2\pi_1 & \lambda_2\pi_2 & \lambda_3\pi_2 & \cdots & \lambda_{\hat{J}}\pi_2 \\ \lambda_3\pi_1 & \lambda_3\pi_2 & \lambda_3\pi_3 & \cdots & \lambda_{\hat{J}}\pi_3 \\ \vdots & \vdots & \vdots & & \vdots \\ \lambda_{\hat{J}}\pi_1 & \lambda_{\hat{J}}\pi_2 & \lambda_{\hat{J}}\pi_3 & \cdots & \lambda_{\hat{J}}\pi_{\hat{J}} \end{bmatrix}.
$$

Then (57c) is equivalent to $\boldsymbol{A}\boldsymbol{p} \geq \boldsymbol{d}$, where both $\boldsymbol{A}$ and $\boldsymbol{d}$ are constants determined by $\mathcal{F}^G$.

For (57d), by (58) we have that $p_0 \geq 0$ is equivalent to $\sum_{j=1}^{\hat{J}} (\overline{z} - z^j) p_j \leq \overline{z} - \mu$, and $p_{\hat{J}+1} \geq 0$ is equivalent to $\sum_{j=1}^{\hat{J}} (z^j - \underline{z}) p_j \leq \mu - \underline{z}$. Recalling the definition of $\boldsymbol{A}$, the constraints $p_0 \geq 0, p_{\hat{J}+1} \geq 0$ can be further reformulated as

$$
\boldsymbol{a}_1^\top \boldsymbol{p} \leq b_l, \boldsymbol{a}_{\hat{J}}^\top \boldsymbol{p} \leq b_u,
$$

respectively, where $b_l = (\overline{z} - \mu)\pi_1,\ b_u = (\mu - \underline{z})\lambda_{\hat{J}}$.

Therefore, $p_0, \ldots, p_{\hat{J}+1}$ satisfy (57a)-(57d) if and only if $(p_0, \boldsymbol{p}, p_{\hat{J}+1}) \in \mathcal{P}$ where

$$
\mathcal{P} = \left\{ (p_0, \boldsymbol{p}, p_{\hat{J}+1}) \in \Re^{\hat{J}+2} \;\middle|\; \begin{array}{l} \boldsymbol{A}\boldsymbol{p} \geq \boldsymbol{d},\ \boldsymbol{a}_1^\top \boldsymbol{p} \leq b_l,\ \boldsymbol{a}_{\hat{J}}^\top \boldsymbol{p} \leq b_u,\ \boldsymbol{p} \geq \boldsymbol{0} \\ p_0 = \frac{\overline{z} - \mu}{\overline{z} - \underline{z}} - \frac{\boldsymbol{a}_1^\top \boldsymbol{p}}{(\overline{z} - \underline{z})\pi_1}, \quad p_{\hat{J}+1} = \frac{\mu - \underline{z}}{\overline{z} - \underline{z}} - \frac{\boldsymbol{a}_{\hat{J}}^\top \boldsymbol{p}}{(\overline{z} - \underline{z})\lambda_{\hat{J}}} \end{array} \right\}, \tag{59}
$$

where the equalities on p_0 and $p_{\hat{J}+1}$ are from the equalities in (58). Note that $\mathcal{P}\neq\emptyset$ as we assume $\mathcal{F}^G\neq\emptyset$. Denote by $(\mathrm{C}_i), i\in[\hat{J}]$ the i-th constraint of $\boldsymbol{A}\boldsymbol{p}\geq\boldsymbol{d}$, i.e., $\boldsymbol{a}_i^\top\boldsymbol{p}\geq d_i$. We say a constraint (C_i) is *redundant* if the strict inequality $\boldsymbol{a}_i^\top\boldsymbol{p}>d_i$ holds for any $\boldsymbol{p}\in\hat{\mathcal{P}}=\{\boldsymbol{p}\in\mathcal{R}_+^{\hat{J}}\mid\boldsymbol{A}\boldsymbol{p}\geq\boldsymbol{d}\}$.

Consider the case that there exists $i\in[\hat{J}]$ such that (C_i) is redundant. WLOG, we let the redundant constraint be $(\mathrm{C}_{\hat{J}})$. In this case, we define $\mathcal{P}^o=\{\boldsymbol{p}\in\mathcal{R}_+^I\mid\boldsymbol{a}_i^\top\boldsymbol{p}\geq d_i,\ i\in[\hat{J}-1]\}$ and will show that $\hat{\mathcal{P}}=\mathcal{P}^o$. Obviously, $\hat{\mathcal{P}}\subseteq\mathcal{P}^o$ since all constraints in defining $\mathcal{P}^o$ are also used in defining $\hat{\mathcal{P}}$. We now show $\mathcal{P}^o\subseteq\hat{\mathcal{P}}$ by contradiction. Assume that there exists $\boldsymbol{p}^o\in\mathcal{P}^o\setminus\hat{\mathcal{P}}$, we have $\boldsymbol{a}_{\hat{J}}^\top\boldsymbol{p}^o<d_{\hat{J}}$. Choosing any $\boldsymbol{p}\in\hat{\mathcal{P}}$, as $(\mathrm{C}_{\hat{J}})$ is redundant, $\boldsymbol{a}_{\hat{J}}^\top\boldsymbol{p}>d_{\hat{J}}$. Therefore, we can find $\lambda\in(0,1)$ such that $\boldsymbol{p}^\lambda=\lambda\boldsymbol{p}+(1-\lambda)\boldsymbol{p}^o$ such that $\boldsymbol{a}_{\hat{J}}^\top\boldsymbol{p}^\lambda=d_{\hat{J}}$. Moreover, by $\boldsymbol{p}^o\in\mathcal{P}^o$ and $\boldsymbol{p}\in\hat{\mathcal{P}}$, we have $\boldsymbol{p}^\lambda\geq\boldsymbol{0}$ and $\boldsymbol{a}_i^\top\boldsymbol{p}^\lambda\geq d_i,\ i\in[\hat{J}-1]$. Therefore, we conclude $\boldsymbol{p}^\lambda\in\hat{\mathcal{P}}$, which is a contradiction since we assume $(\mathrm{C}_{\hat{J}})$ is redundant. Hence, $\mathcal{P}^o\subseteq\hat{\mathcal{P}}$, and it implies $\mathcal{P}^o=\hat{\mathcal{P}}$. Consequently, removing the constraint $\boldsymbol{a}_{\hat{J}}^\top\boldsymbol{p}\geq d_{\hat{J}}$ from the constraints in (59) does not change the set $\mathcal{P}$. Investigating its reformulation back to the form as constraints (57a)-(57d), we can see that now the problem of $\sup_{\mathbb{P}\in\mathcal{F}^G}\mathbb{E}_{\mathbb{P}}[f(\tilde{\boldsymbol{z}})]$ is equivalent to $\sup_{\mathbb{P}\in\mathcal{G}'}\mathbb{E}_{\mathbb{P}}[f(\tilde{\boldsymbol{z}})]$ where $\mathcal{G}'$ is the ambiguity set obtained from $\mathcal{F}^G$ by removing the constraint on $h^{\hat{J}}$. Therefore, we have a problem with $J=\hat{J}-1$, in which case we already have "1" being true by induction.

Now it suffices to consider the case that there is no redundant constraint among $(\mathrm{C}_1),\ldots,(\mathrm{C}_{\hat{J}})$. We will prove that there exists a unique $(p_0^\star,\boldsymbol{p}^\star,p_{\hat{J}+1}^\star)\in\mathcal{P}$ with $\boldsymbol{A}\boldsymbol{p}^\star=\boldsymbol{d}$. Recall that the system $\boldsymbol{A}\boldsymbol{p}=\boldsymbol{d}$ is

$$\begin{cases}\lambda_1\pi_1p_1+\lambda_2\pi_1p_2+\lambda_3\pi_1p_3+\cdots+\lambda_{\hat{J}}\pi_1p_{\hat{J}}=d_1 & (\mathrm{B}_1)\\ \lambda_2\pi_1p_1+\lambda_2\pi_2p_2+\lambda_3\pi_2p_3+\cdots+\lambda_{\hat{J}}\pi_2p_{\hat{J}}=d_2 & (\mathrm{B}_2)\\ \lambda_3\pi_1p_1+\lambda_3\pi_2p_2+\lambda_3\pi_3p_3+\cdots+\lambda_{\hat{J}}\pi_3p_{\hat{J}}=d_3 & (\mathrm{B}_3)\\ \qquad\vdots \\ \lambda_{\hat{J}}\pi_1p_1+\lambda_{\hat{J}}\pi_2p_2+\lambda_{\hat{J}}\pi_3p_3+\cdots+\lambda_{\hat{J}}\pi_{\hat{J}}p_{\hat{J}}=d_{\hat{J}} & (\mathrm{B}_{\hat{J}})\end{cases}$$

Combining (B_1) and (B_2) we have $\pi_1p_1=\frac{d_1\pi_2-d_2\pi_1}{\lambda_1\pi_2-\lambda_2\pi_1}$. Combining (B_2) and (B_3) we obtain $\pi_2p_2=\frac{d_2\pi_3-d_3\pi_2}{\lambda_2\pi_3-\lambda_3\pi_2}-\pi_1p_1$. Continuing the same procedure, we have

$$\begin{aligned}p_1^\star&=\frac{1}{\pi_1}\frac{d_1\pi_2-d_2\pi_1}{\lambda_1\pi_2-\lambda_2\pi_1}\\ p_2^\star&=\frac{1}{\pi_2}\left(\frac{d_2\pi_3-d_3\pi_2}{\lambda_2\pi_3-\lambda_3\pi_2}-\pi_1p_1^\star\right)\\ p_3^\star&=\frac{1}{\pi_3}\left(\frac{d_3\pi_4-d_4\pi_3}{\lambda_3\pi_4-\lambda_4\pi_3}-\pi_1p_1^\star-\pi_2p_2^\star\right)\\ &\vdots\\ p_{\hat{J}-1}^\star&=\frac{1}{\pi_{\hat{J}-1}}\left(\frac{d_{\hat{J}-1}\pi_{\hat{J}}-d_{\hat{J}}\pi_{\hat{J}-1}}{\lambda_{\hat{J}-1}\pi_{\hat{J}}-\lambda_{\hat{J}}\pi_{\hat{J}-1}}-\pi_1p_1^\star-\cdots-\pi_{\hat{J}-2}p_{\hat{J}-2}^\star\right)\\ p_{\hat{J}}^\star&=\frac{1}{\lambda_{\hat{J}}}\frac{\lambda_{\hat{J}-1}d_{\hat{J}}-\lambda_{\hat{J}}d_{\hat{J}-1}}{\lambda_{\hat{J}-1}\pi_{\hat{J}}-\lambda_{\hat{J}}\pi_{\hat{J}-1}}\end{aligned}$$

is the unique solution to $\boldsymbol{A}\boldsymbol{p}=\boldsymbol{d}$. $p_0^\star$ and $p_{\hat{J}+1}^\star$ can be uniquely determined by the equalities in (59). Moreover, since $\mathcal{P}\neq\emptyset$, we must have $d_1\le b_l$, $d_{\hat{J}}\le b_u$, which implies $\boldsymbol{a}_1^\top\boldsymbol{p}^\star=d_1\le b_l$, $\boldsymbol{a}_{\hat{J}}^\top\boldsymbol{p}^\star=d_{\hat{J}}\le b_u$. To see $(p_0^\star,\boldsymbol{p}^\star,p_{\hat{J}+1}^\star)\in\mathcal{P}$, it remains to prove for any $j\in[\hat{J}]$, $p_j^\star\ge 0$. We show that this must be the case, otherwise the constraint (C$_j$) is redundant. Recall that by the definition of λ_j and π_j, $j\in[\hat{J}]$, we have $\lambda_1>\cdots>\lambda_{\hat{J}}>0$ and $0<\pi_1<\cdots<\pi_{\hat{J}}$.

We first show that $p_1^\star\ge 0$, i.e., $d_1\pi_2-d_2\pi_1\ge 0$. Assume to the contrary that $d_1\pi_2<d_2\pi_1$, then

$$\begin{aligned}\boldsymbol{a}_1^\top\boldsymbol{p}-d_1&=\lambda_1\pi_1p_1+\pi_1(\lambda_2p_2+\cdots+\lambda_{\hat{J}}p_{\hat{J}})-d_1\\&\ge\lambda_1\pi_1p_1+\frac{\pi_1}{\pi_2}(d_2-\lambda_2\pi_1p_1)-d_1\\&=\frac{\pi_1}{\pi_2}d_2+\left(\lambda_1-\frac{\pi_1}{\pi_2}\lambda_2\right)\pi_1p_1-d_1\\&\ge\frac{\pi_1}{\pi_2}d_2-d_1>0\end{aligned}$$

for all $\boldsymbol{p}\in\mathcal{P}$. Here the first inequality follows from $\boldsymbol{a}_2^\top\boldsymbol{p}\ge d_2$; the second inequality holds because $\lambda_1>\lambda_2,\pi_1<\pi_2$, and the last inequality follows from the assumption $d_1\pi_2<d_2\pi_1$. Hence (C$_1$) is redundant.

Next, for $p_{\hat{J}}^\star$, we show $\lambda_{\hat{J}-1}\pi_{\hat{J}}-\lambda_{\hat{J}}\pi_{\hat{J}-1}\ge 0$ by contradiction. Assume $\lambda_{\hat{J}-1}\pi_{\hat{J}}<\lambda_{\hat{J}}\pi_{\hat{J}-1}$, then similar as above we have

$$\begin{aligned}\boldsymbol{a}_{\hat{J}}^\top\boldsymbol{p}-d_{\hat{J}}&=\lambda_{\hat{J}}(\pi_1p_1+\cdots+\pi_{\hat{J}-1}p_{\hat{J}-1})+\lambda_{\hat{J}}\pi_{\hat{J}}p_{\hat{J}}-d_{\hat{J}}\\&\ge\frac{\lambda_{\hat{J}}}{\lambda_{\hat{J}-1}}(d_{\hat{J}-1}-\lambda_{\hat{J}}\pi_{\hat{J}-1}p_{\hat{J}})+\lambda_{\hat{J}}\pi_{\hat{J}}p_{\hat{J}}-d_{\hat{J}}\\&=\frac{\lambda_{\hat{J}}}{\lambda_{\hat{J}-1}}d_{\hat{J}-1}+\left(\pi_{\hat{J}}-\frac{\lambda_{\hat{J}}}{\lambda_{\hat{J}-1}}\pi_{\hat{J}-1}\right)\lambda_{\hat{J}}p_{\hat{J}}-d_{\hat{J}}\\&\ge\frac{\lambda_{\hat{J}}}{\lambda_{\hat{J}-1}}d_{\hat{J}-1}-d_{\hat{J}}>0\end{aligned}$$

for all $\boldsymbol{p}\in\mathcal{P}$. Here the first inequality follows from $\boldsymbol{a}_{\hat{J}-1}^\top\boldsymbol{p}\ge d_{\hat{J}-1}$; the second inequality holds because $\lambda_{\hat{J}}<\lambda_{\hat{J}-1},\pi_{\hat{J}}>\pi_{\hat{J}-1}$, and the last inequality follows from the assumption $\lambda_{\hat{J}-1}\pi_{\hat{J}}<\lambda_{\hat{J}}\pi_{\hat{J}-1}$. Hence (C$_{\hat{J}}$) is redundant.

Finally, for all $j\in\{2,\ldots,\hat{J}-1\}$, we show that $\pi_jp_j^\star=\frac{d_j\pi_{j+1}-d_{j+1}\pi_j}{\lambda_j\pi_{j+1}-\lambda_{j+1}\pi_j}-\sum_{k=1}^{j-1}\pi_kp_k^\star\ge 0$. Suppose not, i.e., $\frac{d_j\pi_{j+1}-d_{j+1}\pi_j}{\lambda_j\pi_{j+1}-\lambda_{j+1}\pi_j}<\sum_{k=1}^{j-1}\pi_kp_k^\star=\frac{d_{j-1}\pi_j-d_j\pi_{j-1}}{\lambda_{j-1}\pi_j-\lambda_j\pi_{j-1}}$. Consider any $\boldsymbol{p}\in\mathcal{P}$. We then have

$$\begin{aligned}\boldsymbol{a}_j^\top\boldsymbol{p}-d_j&=\lambda_j(\pi_1p_1+\cdots+\pi_jp_j)+\pi_j(\lambda_{j+1}p_{j+1}+\cdots+\lambda_{\hat{J}}p_{\hat{J}})-d_j\\&\ge\lambda_j(\pi_1p_1+\cdots+\pi_jp_j)+\frac{\pi_j}{\pi_{j+1}}(d_{j+1}-\lambda_{j+1}(\pi_1p_1+\cdots+\pi_jp_j)-d_j\\&=\frac{1}{\pi_{j+1}}\left(d_{j+1}\pi_j-d_j\pi_{j+1}-(\lambda_{j+1}\pi_j-\lambda_j\pi_{j+1})(\pi_1p_1+\cdots+\pi_jp_j)\right),\end{aligned}$$

where the inequality follows from $\boldsymbol{a}_{j+1}^\top \boldsymbol{p} \geq d_{j+1}$. Further, we also have

$$\begin{aligned}\boldsymbol{a}_j^\top \boldsymbol{p} - d_j &= \lambda_j(\pi_1 p_1 + \cdots + \pi_{j-1}p_{j-1}) + \pi_j(\lambda_j p_j + \cdots + \lambda_{\hat{J}} p_{\hat{J}}) - d_j \\ &\geq \lambda_j(\pi_1 p_1 + \cdots + \pi_{j-1}p_{j-1}) + \frac{\pi_j}{\pi_{j-1}}(d_{j-1} - \lambda_{j-1}(\pi_1 p_1 + \cdots + \pi_{j-1}p_{j-1}) - d_j \\ &= \frac{1}{\pi_{j-1}}\left(d_{j-1}\pi_j - d_j\pi_{j-1} - (\lambda_{j-1}\pi_j - \lambda_j\pi_{j-1})(\pi_1 p_1 + \cdots + \pi_{j-1}p_{j-1})\right),\end{aligned}$$

where the inequality follows from $\boldsymbol{a}_{j-1}^\top \boldsymbol{p} \geq d_{j-1}$. Define two $\Re \to \Re$ functions $\phi'(t) = \frac{1}{\pi_{j+1}}\big(d_{j+1}\pi_j - d_j\pi_{j+1} - (\lambda_{j+1}\pi_j - \lambda_j\pi_{j+1})t\big), \phi''(t) = \frac{1}{\pi_{j-1}}\big(d_{j-1}\pi_j - d_j\pi_{j-1} - (\lambda_{j-1}\pi_j - \lambda_j\pi_{j-1})t\big)$, then $\boldsymbol{a}_j^\top \boldsymbol{p} - d_j \geq \max\big\{\phi'(\pi_1 p_1 + \cdots + \pi_j p_j), \phi''(\pi_1 p_1 + \cdots + \pi_{j-1}p_{j-1})\big\}$. By defintion, $\lambda_{j+1}\pi_j - \lambda_j\pi_{j+1} < 0$, hence ϕ' is increasing, which implies $\phi'(\pi_1 p_1 + \cdots + \pi_j p_j) \geq \phi'(\pi_1 p_1 + \cdots + \pi_{j-1}p_{j-1})$. Thus

$$\boldsymbol{a}_j^\top \boldsymbol{p} - d_j \geq \max\left\{\phi'(\pi_1 p_1 + \cdots + \pi_{j-1}p_{j-1}), \phi''(\pi_1 p_1 + \cdots + \pi_{j-1}p_{j-1})\right\}.$$

Notice that $\phi'(t) = 0$ if and only if $t = \frac{d_j\pi_{j+1} - d_{j+1}\pi_j}{\lambda_j\pi_{j+1} - \lambda_{j+1}\pi_j}$. Together with that $\phi'(t)$ is increasing, we have that $\phi'(t) > 0$ if $t > \frac{d_j\pi_{j+1} - d_{j+1}\pi_j}{\lambda_j\pi_{j+1} - \lambda_{j+1}\pi_j}$. Similarly, since $\lambda_{j-1}\pi_j - \lambda_j\pi_{j-1} > 0$, ϕ'' is decreasing, and we obtain $\phi''(t) > 0$ if $t < \frac{d_{j-1}\pi_j - d_j\pi_{j-1}}{\lambda_{j-1}\pi_j - \lambda_j\pi_{j-1}}$. By assumption we have $\frac{d_j\pi_{j+1} - d_{j+1}\pi_j}{\lambda_j\pi_{j+1} - \lambda_{j+1}\pi_j} < \frac{d_{j-1}\pi_j - d_j\pi_{j-1}}{\lambda_{j-1}\pi_j - \lambda_j\pi_{j-1}}$, therefore we can find some $\tau \in \left(\frac{d_j\pi_{j+1} - d_{j+1}\pi_j}{\lambda_j\pi_{j+1} - \lambda_{j+1}\pi_j}, \frac{d_{j-1}\pi_j - d_j\pi_{j-1}}{\lambda_{j-1}\pi_j - \lambda_j\pi_{j-1}}\right)$ such that $\phi'(\tau) = \phi''(\tau) > 0$. Now, for all $t \in \Re$,

$$\begin{aligned}\max\{\phi'(t), \phi''(t)\} &\geq \phi'(t) > \phi'\left(\frac{d_j\pi_{j+1} - d_{j+1}\pi_j}{\lambda_j\pi_{j+1} - \lambda_{j+1}\pi_j}\right) = 0 \quad \text{if } t \geq \tau, \\ \max\{\phi'(t), \phi''(t)\} &\geq \phi''(t) > \phi'\left(\frac{d_{j-1}\pi_j - d_j\pi_{j-1}}{\lambda_{j-1}\pi_j - \lambda_j\pi_{j-1}}\right) = 0 \quad \text{if } t \leq \tau,\end{aligned}$$

which implies

$$\boldsymbol{a}_j^\top \boldsymbol{p} - d_j \geq \max\left\{\phi'(\pi_1 p_1 + \cdots + \pi_{j-1}p_{j-1}), \phi''(\pi_1 p_1 + \cdots + \pi_{j-1}p_{j-1})\right\} > 0.$$

Hence (C_j) is redundant. We then conclude $\boldsymbol{p}^\star \geq \boldsymbol{0}$.

In summary, we have a unique $(p_0^\star, \boldsymbol{p}^\star, p_{\hat{J}+1}^\star) \in \mathcal{P}$ with $\boldsymbol{A}\boldsymbol{p}^\star = \boldsymbol{d}$.

Related to $(p_0^\star, \boldsymbol{p}^\star, p_{\hat{J}+1}^\star)$, we next prove the following observation.

Observation: Considering any $(p_0, \boldsymbol{p}, p_{\hat{J}+1}) \in \mathcal{P}$ with $(p_0, \boldsymbol{p}, p_{\hat{J}+1}) \neq (p_0^\star, \boldsymbol{p}^\star, p_{\hat{J}+1}^\star)$, there exists $i \in \{0, 1, \ldots, \hat{J}-1\}$, $i+1 \leq k \leq \hat{J}$, such that

1) $p_j = p_j^* \ \forall j \in \{0, \ldots, i-1\}$ if $i \geq 1$;
2) $p_i < p_i^*$;
3) $p_j = 0 \ \forall j \in \{i+1, \ldots, k-1\}$ if $k \geq i+2$;
4) $p_k > 0$;
5) $\boldsymbol{a}_k^\top \boldsymbol{p} > d_k$

Specifically, parts 1) and 2) mean that i is the index of the first distinct component when comparing $(p_0, \boldsymbol{p}, p_{\hat{J}+1})$ and $(p_0^\star, \boldsymbol{p}^\star, p_{\hat{J}+1}^\star)$; parts 3) and 4) mean that k is the index of the first nonzero component in $(p_0, \boldsymbol{p}, p_{\hat{J}+1})$ after p_i.

To prove parts 1) and 2), we consider any $i \in \{0, \ldots, \hat{J}-1\}$, and have

$$\begin{aligned}
&\frac{\overline{z}-\mu}{\overline{z}-\underline{z}} - \frac{\boldsymbol{a}_{i+1}^\top \boldsymbol{p}}{(\overline{z}-\underline{z})\pi_{i+1}} \\
&= \frac{\overline{z}-\mu}{\overline{z}-\underline{z}} - \frac{1}{\overline{z}-\underline{z}} \left(\sum_{j=i+1}^{\hat{J}} \lambda_j p_j + \frac{1}{\pi_{i+1}} (\lambda_{i+1}\pi_1 p_1 + \cdots + \lambda_{i+1}\pi_i p_i) \right) \\
&= \frac{\overline{z}-\mu}{\overline{z}-\underline{z}} - \frac{1}{\overline{z}-\underline{z}} \left(\sum_{j=1}^{\hat{J}} \lambda_j p_j + \frac{1}{\pi_{i+1}} (\lambda_{i+1}\pi_1 p_1 + \cdots + \lambda_{i+1}\pi_i p_i - \lambda_1 \pi_{i+1} p_1 - \cdots - \lambda_i \pi_{i+1} p_i) \right) \\
&= p_0 + \frac{1}{(\overline{z}-\underline{z})\pi_{i+1}} \sum_{j=1}^{i} \alpha_{i+1,j} p_j,
\end{aligned}$$

where we define $\alpha_{i+1,j} = \lambda_j \pi_{i+1} - \lambda_{i+1}\pi_j > 0$ for all $j \le i$ since in this case $\lambda_j > \lambda_{i+1}$ and $\pi_j < \pi_{i+1}$. Hence,

$$\boldsymbol{a}_{i+1}^\top \boldsymbol{p} = (\overline{z}-\mu)\pi_{i+1} - (\overline{z}-\underline{z})\pi_{i+1} p_0 - \sum_{j=1}^{i} \alpha_{i+1,j} p_j. \tag{60}$$

Consider $i = 0$, by (60) we have

$$p_0 = \frac{\overline{z}-\mu}{\overline{z}-\underline{z}} - \frac{\boldsymbol{a}_1^\top \boldsymbol{p}}{(\overline{z}-\underline{z})\pi_1} \le \frac{\overline{z}-\mu}{\overline{z}-\underline{z}} - \frac{d_1}{(\overline{z}-\underline{z})\pi_1} = \frac{\overline{z}-\mu}{\overline{z}-\underline{z}} - \frac{\boldsymbol{a}_1^\top \boldsymbol{p}^\star}{(\overline{z}-\underline{z})\pi_1} = p_0^\star,$$

where the first inequality is due to $\boldsymbol{A}\boldsymbol{p} \ge \boldsymbol{d}$, the second equality follows from $\boldsymbol{A}\boldsymbol{p}^\star = \boldsymbol{d}$ and the last equality holds since Equation (60) also applies to $(p_0^\star, \boldsymbol{p}^\star, p_{\hat{J}+1}^\star)$. Hence, if $p_0 \ne p_0^\star$, we must have $p_0 < p_0^*$. Now, consider the case where $p_0 = p_0^\star$, we then denote $i \ge 1$ as the index of the first distinct component, i.e., $p_j = p_j^\star$ for all $j \in \{0, \ldots, i-1\}$, and $p_i \ne p_i^\star$. Note that $i \le \hat{J}-1$, otherwise the only distinct components are the last two dimension, i.e., the marginal masses at $z^{\hat{J}}$ and $z^{\hat{J}+1}$, which is impossible since $(p_0, \boldsymbol{p}, p_{\hat{J}+1})$ and $(p_0^\star, \boldsymbol{p}^\star, p_{\hat{J}+1}^\star)$ correspond to the same mean. As $i \le \hat{J}-1$, by (60),

$$\begin{aligned}
\boldsymbol{a}_{i+1}^\top \boldsymbol{p} &= (\overline{z}-\mu)\pi_{i+1} - (\overline{z}-\underline{z})\pi_{i+1} p_0^\star - \textstyle\sum_{j=1}^{i} \alpha_{i+1,j} p_j^\star + \alpha_{i+1,i}(p_i^\star - p_i) \\
&= \boldsymbol{a}_{i+1}^\top \boldsymbol{p}^\star + \alpha_{i+1,i}(p_i^\star - p_i) \\
&= d_{i+1} + \alpha_{i+1,i}(p_i^\star - p_i) \\
&\le \boldsymbol{a}_{i+1}^\top \boldsymbol{p} + \alpha_{i+1,i}(p_i^\star - p_i),
\end{aligned}$$

which implies $p_i < p_i^\star$ since $p_i \ne p_i^\star$. Therefore, parts 1) and 2) in **Observation** are proved.

Parts 3) and 4) in **Observation** are straightfoward. Specifically,

$$\sum_{j=i+1}^{\hat{J}+1} p_j = 1 - \sum_{j=0}^{i} p_j = 1 - \sum_{j=0}^{i} p_j^\star + (p_i^\star - p_i) \ge p_i^\star - p_i > 0.$$

Hence, there must be a nonzero component in $p_{i+1}, \ldots, p_{\hat{J}+1}$. We then just let k be the index of the first nonzero component, parts 3) and 4) in **Observation** are proved.

Part 5) can be proved by the adoption of (60), which leads to

$$\boldsymbol{a}_k^\top \boldsymbol{p} = (\overline{z}-\mu)\pi_k - (\overline{z}-\underline{z})\pi_k p_0 - \sum_{j=1}^{k-1}\alpha_{k,j}p_j > (\overline{z}-\mu)\pi_k - (\overline{z}-\underline{z})\pi_k p_0^\star - \sum_{j=1}^{k-1}\alpha_{k,j}p_j^\star = d_k.$$

Here the inequality is due to parts 1) and 2), and $0 = p_j \le p_j^\star$ for all $j \in \{i+1,\dots,k-1\}$.

Now, base on **Observation**, we prove "1" is true by proposing a process to construct new distribution. Given any $\mathbb{P} \in \mathcal{F}^G$, let the associated marginals on $\tilde{z}$ at $z^0,\dots,z^{\hat{J}+1}$ be $(p_0, \boldsymbol{p}, p_{\hat{J}+1})$. Consider the case where $(p_0,\boldsymbol{p},p_{\hat{J}+1}) \neq (p_0^\star,\boldsymbol{p}^\star,p_{\hat{J}+1}^\star)$. We now construct a new probability distribution $\mathbb{P}'$ with support only at the breakpoints and defined as

$$\mathbb{P}'(\tilde{\boldsymbol{z}} = \boldsymbol{z}) = \begin{cases} \mathbb{P}(\tilde{\boldsymbol{z}} = \boldsymbol{z}) & \text{if } z \notin \{z^{k-1}, z^k, z^{k+1}\} \\ (1-\theta)\mathbb{P}(\tilde{\boldsymbol{z}} = \boldsymbol{z}) & \text{if } z = z^k \\ \mathbb{P}(\tilde{\boldsymbol{z}} = \boldsymbol{z}) + \frac{z^{k+1}-z^k}{z^{k+1}-z^{k-1}}\theta\mathbb{P}\left(\tilde{\boldsymbol{z}} = \boldsymbol{z} + (z^k - z^{k-1})\boldsymbol{e}_n\right) & \text{if } z = z^{k-1} \\ \mathbb{P}(\tilde{\boldsymbol{z}} = \boldsymbol{z}) + \frac{z^k-z^{k-1}}{z^{k+1}-z^{k-1}}\theta\mathbb{P}\left(\tilde{\boldsymbol{z}} = \boldsymbol{z} - (z^{k+1} - z^k)\boldsymbol{e}_n\right) & \text{if } z = z^{k+1} \end{cases} \tag{61}$$

for some $\theta \in (0,1)$. Intuitively, for all $z_1,\dots,z_{n-1}$, we move θ portion of the probability mass at $(z_1,\dots,z_{n-1},z^k)$ to $(z_1,\dots,z_{n-1},z^{k-1})$ and $(z_1,\dots,z_{n-1},z^{k+1})$, keeping the mean unchanged. Hence $\mathbb{P}'$ has the same marginal for $(\tilde{z}_1,\dots,\tilde{z}_{n-1})$ as $\mathbb{P}$. Denote the marginal of $\mathbb{P}'$ on $\tilde{z}$ by $p_0', \boldsymbol{p}', p_{\hat{J}+1}'$ such that $\mathbb{P}'(\tilde{z} = z^j) = p_j'$ for all $j = 0,\dots,\hat{J}+1$. By (61),

$$\begin{cases} p_j' = p_j, & \forall j \notin \{k-1,k,k+1\}, \\ p_k' = p_k - \theta p_k, & \\ p_{k-1}' = p_{k-1} + \frac{z^{k+1}-z^k}{z^{k+1}-z^{k-1}}\theta p_k, & \\ p_{k+1}' = p_{k+1} + \frac{z^k-z^{k-1}}{z^{k+1}-z^{k-1}}\theta p_k. & \end{cases}$$

There are three properties of $\mathbb{P}'$.

(P1) $\mathbb{E}_{\mathbb{P}'}[f(\tilde{\boldsymbol{z}})] \ge \mathbb{E}_{\mathbb{P}}[f(\tilde{\boldsymbol{z}})]$. This is because $\mathbb{E}_{\mathbb{P}'}[f(\tilde{\boldsymbol{z}})] - \mathbb{E}_{\mathbb{P}}[f(\tilde{\boldsymbol{z}})]$ equals

$$\sum_{z_i \in \left\{z_i^1,\dots,z_i^{S_i}\right\}, i\in[n-1]} \theta\mathbb{P}\left(\tilde{\boldsymbol{z}} = (z_1,\dots,z_{n-1},z^k)\right)\left(\frac{z^{k+1}-z^k}{z^{k+1}-z^{k-1}}f(z_1,\dots,z_{n-1},z^{k-1}) + \frac{z^k-z^{k-1}}{z^{k+1}-z^{k-1}}f(z_1,\dots,z_{n-1},z^{k+1}) - f(z_1,\dots,z_{n-1},z^k)\right),$$

which is nonnegative since f is convex.

(P2) $\boldsymbol{a}_j^\top \boldsymbol{p}' = \boldsymbol{a}_j^\top \boldsymbol{p}$ for all $j \ne k$ and $\boldsymbol{a}_k^\top \boldsymbol{p}' < \boldsymbol{a}_k^\top \boldsymbol{p}$. To see this, for any $j \in [\hat{J}]$, we observe

$$\mathbb{E}_{\mathbb{P}'}\left[h^j(\tilde{z})\right] - \mathbb{E}_{\mathbb{P}}\left[h^j(\tilde{z})\right] = \theta p_k\left(\frac{z^{k+1}-z^k}{z^{k+1}-z^{k-1}}h^j(z^{k-1}) + \frac{z^k-z^{k-1}}{z^{k+1}-z^{k-1}}h^j(z^{k+1}) - h^j(z^k)\right) \ge 0, \tag{62}$$

where the inequality is due to the convexity of h. Moreover, the "$\ge$" takes "$=$" if $j \ne k$ since h^j is linear on $[z^{k-1}, z^{k+1}]$ for such j; by contrast, "$\ge$" becomes "$>$" for $j = k$ since h^k has a breakpoint at z^k. Therefore, by the definition of $\boldsymbol{A}$, this property is proved.

(P3) $\boldsymbol{a}_k^\top \boldsymbol{p}'$ is continuously decreasing in θ, which is implied by (62) and the definition of $\boldsymbol{A}$.

Based on the **Observation** and **(P1)-(P3)**, given any $\mathbb{P} \in \mathcal{F}^G$ whose marginal on $\tilde{z}$ is different from $(p_0^\star, \boldsymbol{p}^\star, p_{\hat{J}+1}^\star)$, we can use the procedure as in (61) to construct a new probability distribution $\mathbb{P}'$. In this construction, we either choose $\theta = 1$ or the maximal value less than 1 such that $\boldsymbol{a}_k^\top \boldsymbol{p}'$ drops to the value of d_k (note that when $\theta = 0$, $\boldsymbol{a}_k^\top \boldsymbol{p}' = \boldsymbol{a}_k^\top \boldsymbol{p} > d_k$, where the inequality is due to the part 5) in **Observation**). Hence, $\mathbb{P}' \in \mathcal{F}^G$. Moreover, by **(P1)**, with $\mathbb{P}'$, the expectation of $f(\tilde{\boldsymbol{z}})$ is no less. Therefore, for any $\mathbb{P} \in \mathcal{F}^G$, by this procedure we construct a new probability distribution $\mathbb{P}' \in \mathcal{F}^G$ such that the objective is improved and the marginal masses after z^i is moved towards z^i, the smallest breakpoint where the marginal mass of $\mathbb{P}$ differs from $(p_0^\star, \boldsymbol{p}^\star, p_{\hat{J}+1}^\star)$. Repeating such process, the margin converges to $(p_0^\star, \boldsymbol{p}^\star, p_{\hat{J}+1}^\star)$. We hence conclude that there must be a worst-case distribution whose n-th marginal is $(p_0^\star, \boldsymbol{p}^\star, p_{\hat{J}+1}^\star)$. □

References

Ben-Tal, Aharon, Eithan Hochman. 1972. More bounds on the expectation of a convex function of a random variable. *Journal of Applied Probability* **9**(4) 803-812.

Ben-Tal, Aharon, Marc Teboulle. 1986. Expected utility, penalty functions, and duality in stochastic nonlinear programming. *Management Science* **32**(11) 1445-1466.

Ben-Tal, Aharon, Marc Teboulle. 2007. An old-new concept of convex risk measures: The optimized certainty equivalent. *Mathematical Finance* **17**(3) 449-476.

Bertsimas, Dimitris, Vineet Goyal. 2012. On the power and limitations of affine policies in two-stage adaptive optimization. *Mathematical Programming* **134**(2) 491-531.

Birge, John R, Francois Louveaux. 2011. *Introduction to Stochastic Programming*. Springer Science & Business Media.

Chen, Xin, Daniel Zhuoyu Long, Jin Qi. 2021. Preservation of supermodularity in parametric optimization: Necessary and sufficient conditions on constraint structures. *Operations Research* **69**(1) 1-12.

Shapiro, Alexander. 2001. On duality theory of conic linear problems. *Semi-infinite programming*. Springer, 135-165.

Sion, Maurice. 1958. On general minimax theorems. *Pacific Journal of Mathematics* **8**(1) 171-176.

von Neumann, John. 1928. Zur theorie der gesellschaftsspiele. *Mathematische Annalen* **100**(1) 295-320.