Supermodularity in Two-Stage Distributionally Robust Optimization

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

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

Aiqi Zhang
Department of Systems Engineering and Engineering Management, Chinese University of Hong Kong, Shatin, Hong Kong
aqzhang@se.cuhk.edu.hk,

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 of these two-stage problems. We also show that the optimal scenario-wise segregated affine decision rule returns the same optimal value in our setting. Further, we provide a necessary and sufficient condition for checking whether any given two-stage optimization problem has the supermodularity property. We apply this framework to several classic problems, including the multi-item newsvendor problem, the facility location design 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 using 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 the 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 operational examples include multi-item newsvendor, facility location, unit commitment problems, etc.

One classic solution approach to two-stage optimization problems is stochastic programming (Shapiro et al. 2009, Birge and Louveaux 2011), in which uncertainties are assumed to follow some given probability distributions. To incorporate ambiguity, robust optimization, which was introduced by Soyster (1973) and promoted by Ben-Tal and Nemirovski (1998), El Ghaoui et al. (1998), Bertsimas and Sim (2004), 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 and Zhang 2009) 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 the 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 $\mathcal{F}$ 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. These sets are often based on moment information (Delage and Ye 2010, Zymler et al. 2013a,b, Mehrotra and Zhang 2014) or statistical measures, including the $\phi$-divergence (Ben-Tal et al. 2013) and Wasserstein distance (Gao and Kleywegt 2016, Esfahani and Kuhn 2018, Hanasusanto and Kuhn 2018). 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 (see for instance, Goh and Sim 2010, Bertsimas et al. 2019), there is no theoretical result for the performance of these approximations. Instead, examples are given such that the linear decision rule approximation can be infeasible even for problems with complete recourse (Bertsimas et al. 2019).

Further, very 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 convex piecewise linear 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 additionally 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. Consequently, scholars are also interested in exploring supermodularity in their 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 the 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. We also discuss the optimality of the segregated affine decision rules when problems have the property of supermodularity. Further, we provide a necessary and sufficient condition to check whether any given two-stage optimization problem has this property. We then identify 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 using our approach.

Our key contributions are summarized as follows.
1. We show that for a specific distributional uncertainty set with moment information, the second-stage problem has an explicit common worst-case distribution whenever it has the property of supermodularity. By inserting this worst-case distribution, the original two-stage problem can be reduced to a deterministic optimization problem of polynomial size.
2. Leveraging the benefits of the polynomial size support of the worst-case distribution, we show that when the property of supermodularity holds, the scenario-wise segregated affine decision rules can return the same optimal value as the original problem.
3. When the second-stage problem has a linear programming formulation, we provide a necessary and sufficient condition to check its supermodularity. A simple algorithm is proposed to determine whether the condition is satisfied.
4. We provide several extensions to generalize the results and further apply them to several important operational 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 provide 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 of the objective function in the second stage. We then provide several extensions in Section 4 and discuss applications in Section 5. We finally conclude the paper in Section 6. For the sake of readability, all proofs are relegated to the appendix.

Notation and convention: For any integer $K \geq 1$, we define $[K] = \{1, \ldots, K\}$, which is the set of positive running indices to $K$. We represent column vectors and matrices by lower- and upper-case boldface characters, respectively. An $n$-dimensional column vector $x$ 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. $[A \ B]$ or $[A \ B]^\top$. Given any matrix $A = (a_{ij})_{i \in [m], j \in [n]} \in \mathbb{R}^{m \times n}$, we let $a_i^\top$ and $A_j$ be its $i$-th row vector and $j$-th column vector, respectively. Further, we use $A_{\mathcal{I}}$ to represent its submatrix $(a_{ij})_{i \in \mathcal{I}, j \in [n]} \in \mathbb{R}^{|\mathcal{I}| \times n}$ for any $\mathcal{I} \subseteq [m]$, and we use $|\cdot|$ to represent the cardinality of a set. We denote $\text{span}(A)$ to be the column space of $A$. For any two vectors $x', x'' \in \mathbb{R}^n$, we denote by $x' \leq x''$ if $x'_i \leq x''_i$ for all $i \in [n]$; moreover, we say $x', x''$ are ordered if either $x' \leq x''$ or $x'' \leq x'$, and they are unordered otherwise. We also define two operations join ("$\vee$") and meet ("$\wedge$") such that $x' \vee x'' = (\max\{x'_i, x''_i\})_{i=1,\ldots,n}$ and $x' \wedge x'' = (\min\{x'_i, x''_i\})_{i=1,\ldots,n}$ for any vectors $x', x'' \in \mathbb{R}^n$.

We let $e_i$ be the vector with only the $i$-th entry being 1 and all others being 0, and 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 $x \in \mathbb{R}^l$ before the uncertainty $\tilde{z}$, 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 $y \in \mathbb{R}^m$. For a given first-stage decision $x$ and an uncertainty realization $z$, we denote the second-stage cost by $g(x, z)$. It can be evaluated by the following linear program,

$$
g(x, z) = \min b^\top y$$
$$\text{s.t.} \ Wx + Uy \geq Vz + v^0, \quad (1)$$
where $b \in \mathbb{R}^m$, $W \in \mathbb{R}^{r \times l}$, $U \in \mathbb{R}^{r \times m}$, $V \in \mathbb{R}^{r \times n}$ and $v^0 \in \mathbb{R}^r$ are given constants. 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, Bertsimas and Goyal 2012, Zeng and Zhao 2013, Gupta et al. 2014, Bertsimas and Bidkhori 2015, Xu and Burer 2018, Bertsimas and Shtern 2018, El Housni and Goyal 2021) and is intractable in general (Feige et al. 2007, Bertsimas and Goyal 2012). Though it only has uncertainties on the right-hand side only, this model can cover a broad range of practical two-stage problems, which we will introduce in Section 5. Further, in Section 4, we will generalize our results to include left-hand-side uncertainty as well. We let $g(x, z) = \infty$ if Problem (1) is infeasible.

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

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

By choosing the first-stage decision $x$, 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

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

where $a \in \mathbb{R}^l$ is a given constant vector, $\mathcal{X} \subseteq \mathbb{R}^l$ is the set of all feasible first-stage decisions. We assume that Problem (2) has finite optimal value.

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

$$\mathcal{F} = \left\{ \mathbb{P} \left| \begin{array}{l} \mathbb{E}_P [\tilde{z} | \tilde{k} = k] = \mu^k, \quad \forall k \in [K] \\ \mathbb{P} [\tilde{z} - \mu^k | \tilde{k} = k] \leq \delta^k, \quad \forall k \in [K], \forall i \in [n] \\ \mathbb{P} (\tilde{z}^k \leq \tilde{z} \leq \tilde{z}^k | \tilde{k} = k) = 1, \forall k \in [K] \\ \mathbb{P} (\tilde{k} = k) = q_k, \quad \forall k \in [K] \\ q \in Q \end{array} \right. \right\}.$$

Here a random scenario $\tilde{k}$ is introduced and its realization affects the distributional information of $\tilde{z}$. In particular, if the random scenario is realized as $k \in [K]$, we have corresponding distributional information for $\tilde{z}$: mean being $\mu^k$, MAD of $\tilde{z}$ being bounded by $\delta^k$ for all $i \in [n]$, and support being $[\tilde{z}^k, \tilde{z}^k]$. The probability that $\tilde{k}$ is realized as $k$ is denoted by $q_k$. We also allow ambiguity in $q = (q_k)_{k \in [K]}$ and only know that $q$ is in a given polyhedron $Q = \{ q \mid Rq \leq \nu, q \geq 0 \}$. Since $q$ represents the probability mass, we assume $Q \subseteq \{ q \in \mathbb{R}^K_+ \mid 1^\top q = 1 \}$. Without loss of generality (WLOG), we make the following assumptions about $\mathcal{F}$ to avoid trivial cases. If there are $i, k$ such that $\delta^k_i = 0$, by the constraint on MAD, $\tilde{z}_i$ realizes at $\mu^k_i$ almost surely when the random scenario
\( \tilde{k} \) takes value at \( k \), and hence we can let \( \tilde{z}_i^k = \mu_i^k = \bar{z}_i^k \) for notational simplification. Similarly, for any \( i, k \) with \( \mu_i^k \in \{ \tilde{z}_i^k, z_i^k \} \), by the constraint on mean and MAD, we can also let \( \tilde{z}_i^k = \mu_i^k = z_i^k \) and \( \delta_i^k = 0 \) for notational simplification. Moreover, \( Q \) is such that \( \forall k \in [K] \), there exists \( q \in Q \) with \( q_k > 0 \), otherwise the scenario \( k \) almost surely does not happen and we can ignore it.

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. When the uncertain variable is one-dimensional, the worst-case expectation has a decision independent expression if the objective function is convex (Ben-Tal and Hochman [1972]). Postek et al. [2019] and Den Hertog and van Leeuwaarden [2019] use MAD information to focus on a special case where all random variables are independent. In practice, the MAD information is also easy to estimate (Postek et al. [2018]). Other examples of applying the MAD information can be also seen in Qi [2017] and Conejo et al. [2021]. 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 \( Q \) 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.

To explore the solvability of Problem (2), we will first investigate the worst-case distribution of \( \tilde{z} \) 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,

\[
\mathcal{F}^k = \left\{ P^k \left| \begin{array}{l}
E_{P^k}[\tilde{z}] = \mu^k, \\
E_{P^k}[|\tilde{z}_i - \mu_i^k|] \leq \delta_i^k, \\
P^k(\tilde{z}_i \leq \bar{z} \leq z_i^k) = 1
\end{array} \right., \quad \forall i \in [n] \right\}.
\]

(4)

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

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

\[
P^{k^*}(\tilde{z}_i = w) = \begin{cases} 
\frac{\delta_i^k}{2(\mu_i^k - z_i^k)} & \text{if } w = \tilde{z}_i^k \\
1 - \frac{\delta_i^k(\mu_i^k - \tilde{z}_i^k)}{2(\mu_i^k - \mu_i^k)(\mu_i^k - z_i^k)} & \text{if } w = \mu_i^k \\
\frac{\delta_i^k}{2(\mu_i^k - \mu_i^k)} & \text{if } w = z_i^k \\
0 & \text{otherwise,}
\end{cases}
\]

(5)
where $\delta_k^i = \min\left\{ \delta_k^i, \frac{2(z_i^k - \mu_i^k)(\alpha_i^k - \beta_i^k)}{\lambda_i^k - \gamma_i^k} \right\}$ for all $i \in [n]$ with $z_i^k > \alpha_i^k$.

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 $z_i^k = \alpha_i^k$, obviously $\mathbb{P}_k(\tilde{z}_i = \alpha_i^k) = \mathbb{P}_k(\tilde{z}_i = \mu_i^k) = \mathbb{P}_k(\tilde{z}_i = \beta_i^k) = 1$). Therefore, to evaluate $\sup_{\tilde{z}\in\mathbb{R}^n} \mathbb{E}_{\tilde{z}}[g(x, \tilde{z})]$, it suffices to focus on the distributions with support $\{ z \mid z_i \in \{ \alpha_i^k, \mu_i^k, \beta_i^k \}, i \in [n] \}$. Unfortunately, the number of points in this set is exponentially large in $n$, which essentially renders the two-stage problem computationally challenging to solve. We next show that if the function $g(x, z)$ is supermodular in $z$, the computational burden can be eased. We first define supermodularity as follows.

**Definition 1** A function $f : \mathbb{R}^n \rightarrow \mathbb{R}$ is supermodular if $f(w') + f(w'') \leq f(w' \wedge w'') + f(w' \vee w'')$ for all $w', w'' \in \mathbb{R}^n$.

In transportation theory 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.

**Lemma 1** (Rachev and Rüschendorf 1998) Consider any supermodular function $f : \mathbb{R}^2 \rightarrow \mathbb{R}$, and any two-dimensional random vector $\tilde{w}$ with the marginal cumulative distribution function for $\tilde{w}_1, \tilde{w}_2$ being $F_1, F_2$, respectively. Let $\mathcal{P} = \{ \mathbb{P} \mid \mathbb{P}(\tilde{w}_i \leq u) = F_i(u) \ \forall x \in \mathbb{R}, i = 1, 2 \}$ be the set of all possible distributions for $\tilde{w}$. Then

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

Clearly, the upper bound in Lemma 1 is achieved when $(\tilde{w}_1, \tilde{w}_2) \overset{d}{=} (F_1^{-1}(\tilde{u}), F_2^{-1}(\tilde{u}))$ with $\tilde{u}$ being uniformly distributed on $[0, 1]$ . In this worst-case distribution, considering any two realizations $w', w''$, we then have $u', u'' \in [0, 1]$ such that $w' = (F_1^{-1}(u'), F_2^{-1}(u'))$ and $w'' = (F_1^{-1}(u''), F_2^{-1}(u''))$. This implies $w', w''$, 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 ($\vee$) and meet ($\wedge$), such that the marginal distribution is unchanged and the expectation of $f(\tilde{w})$ 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.

**Proposition 2** Consider any function $f : \mathbb{R}^n \rightarrow \mathbb{R}$. The following statements are equivalent.

1. It is brought to our attention that a concurrent work-in-progress (Chen et al. 2018) 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.
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 $\mathcal{P} = \{P \mid P(\tilde{w}_i = x_{ij}) = p_{ij}, j \in [m_i], i \in [n]\}$. Then there exists $P^* \in \arg\sup_{P \in \mathcal{P}} E_P[f(\tilde{w})]$ such that the set $W_{P^*} = \{w \in \mathbb{R}^n \mid P^*(\tilde{w} = w) > 0\}$ forms a chain of at most $(\sum_{i \in [n]} (m_i - 1) + 1)$ points.

Here, 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.

![Figure 1](image)

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

Intuitively, when moving the same amount of probability mass from any two points $w', w''$ to $w' \land w'', w' \lor w''$, the marginal distribution does not change but the expectation of $f(\tilde{w})$ 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 $g(x, z)$ is supermodular in $z$, then the worst-case distribution for $\sup_{\mathbb{P}_k \in \mathcal{F}_k} \mathbb{E}_\mathbb{P}_k[g(x, \tilde{z})]$ 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 worst-case distribution for $\sup_{\mathbb{P}_k \in \mathcal{F}_k} \mathbb{E}_\mathbb{P}_k[g(x, \tilde{z})]$ has an explicit three-point distribution in each
Algorithm 1: Algorithm for worst-case distribution

1: **Input:** $\mathcal{F}^k$ in Equation (4) with given $\mu^k, \delta^k, z^k, \bar{z}^k$

2: **Initialization:**
   - denote $P^{k*}$ as the worst-case distribution in Proposition 1 and calculate $P^{k*}(\tilde{z}^k = w)$ for $w \in \{z^k, \mu^k, \bar{z}^k\}$, $i \in [n]$ using Equation (5)
   - $z^1 = z^k$, $q^1 = (P^{k*}(\tilde{z}^1 = z^k), P^{k*}(\tilde{z}^2 = z^k), \ldots, P^{k*}(\tilde{z}^n = z^k))$, $p_1 = \min\{q_1^1, \ldots, q_n^1\}$ and $j = 1$

3: while $j \leq 2n$ do
   4: choose $r_j$ as the minimal index in $[n]$ such that $q^j_{r_j} = p_j$
   5: $z^{j+1} = z^j$, $q^{j+1} = q^j - p_j 1$
   6: update $q_{r_j}^{j+1} = \mu_{r_j}^k$, if its existing value is $z_{r_j}^j$, and $z_{r_j}^{j+1} = \bar{z}_{r_j}^k$, if its existing value is $\mu_{r_j}^k$
   7: update $q_{r_j}^{j+1} = P^{k*}(\tilde{z}_{r_j} = z_{r_j}^{j+1})$
   8: $p_{j+1} = \min\{q_1^{j+1}, q_2^{j+1}, \ldots, q_n^{j+1}\}$
   9: update $j = j + 1$

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

By moving from $z^k$ to $\bar{z}^k$, 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 behind the results 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.

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 $x$, 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 uncertain scenario has only one possible realization. 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

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

We denote by \( z^{k,1}, \ldots, z^{k,2n+1}, p^k \) 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 \( g(x, z) \) is supermodular in \( z \) for any \( x \), Problem (2) is equivalent to the following linear program,

\[
\begin{align*}
\min & \quad a^\top x + \nu^\top l \\
\text{s.t.} & \quad R_k^\top l \geq \sum_{i \in [2n+1]} p_k^i b_i^\top y^{k,i}, \quad k \in [K] \\
& \quad W x + U y^{k,i} \geq V z^{k,i} + v^0, \quad k \in [K], i \in [2n+1] \\
& \quad l \geq 0, \quad x \in \mathcal{X}.
\end{align*}
\]

Intuitively, the reformulation in Theorem 1 incorporates all possible scenarios in the worst-case distribution, and assigns a corresponding second-stage decision to each of those scenarios. Therefore, the two-stage problem can be formulated as a static linear programming problem. Nevertheless, the classic approach using this idea has to handle an exponential number of scenarios, leading
to computational intractability. Here, by exploring the potential property of supermodularity in the uncertainties, we reduce the number of scenarios 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 (7) is an equivalent reformulation of the original problem, and hence Problem (7) maintains the same feasibility for any given first-stage decision $x$. 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 $g(x, z)$ already. In particular, if Problem (7) has a finite optimal value, then at the optimal $x$, the second-stage problem is feasible when $\tilde{z}$ takes any pre-determined realizations ($z_{k,i}$ in Problem (7)). By supermodularity, these pre-determined realizations constitute the worst-case distribution. It implies that when $\tilde{z}$ 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 $x_{opt}$ is optimal to Problem (7), then for all $z \in \bigcup_{k \in [K]} [z_k^+, z_k^-]$, $g(x_{opt}, z)$ is finite, i.e., the second-stage problem is feasible when $x = x_{opt}$.

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. We next illustrate this with the following example. Consider $\tilde{z} = (\tilde{z}_1, \tilde{z}_2)$ with three different distributional uncertainty sets as follows,

\[
F_{sup} = \{ P \mid \mathbb{E}_P[\tilde{z}] = 0.5, \mathbb{E}_P[|\tilde{z} - 0.5|] \leq (0.3, 0.45), P(0 \leq \tilde{z} \leq 1) = 1 \},
\]

\[
F_{ind} = \{ P \mid \mathbb{E}_P[\tilde{z}] = 0.5, \mathbb{E}_P[|\tilde{z} - 0.5|] \leq (0.3, 0.45), P(0 \leq \tilde{z} \leq 1) = 1, \tilde{z}_1 \perp \perp \tilde{z}_2 \},
\]

\[
F_{sce} = \left\{ P \mid \begin{align*}
\mathbb{E}_P[\tilde{z} | \tilde{k} = 1] &= (0.2, 0.95), \mathbb{E}_P[\tilde{z} | \tilde{k} = 2] = (0.8, 0.05), \\
\Pr \left(0, 0.9 \leq \tilde{z} \leq (0.4, 1) | \tilde{k} = 1 \right) &= 1, \Pr \left((0.6, 0) \leq \tilde{z} \leq (1, 0.1) | \tilde{k} = 2 \right) = 1,
\end{align*} \right\},
\]

where $\tilde{z}_1 \perp \perp \tilde{z}_2$ represents the stochastic independence and we assume $0 \leq \delta < (0.2, 0.05)$ to guarantee that the three-point marginal distribution exists when the MAD equals $\delta$ (see Proposition 1). Compared with $F_{sup}$, $F_{ind}$ further imposes the assumption of independence between $\tilde{z}_1$ and $\tilde{z}_2$. The scenario-based uncertainty set $F_{sce}$ specifies two possible scenarios with corresponding means, supports and upper bounds of MADs. We can easily show that $F_{ind} \subseteq F_{sup}$ and $F_{sce} \subseteq F_{sup}$.

For both $F_{sup}$ and $F_{ind}$, we calculate their marginal distributions following Proposition 1 as follows,

\[
\Pr(\tilde{z}_1 = w) = \begin{cases} 
0.3, & \text{if } w = 0 \\
0.4, & \text{if } w = 0.5 \\
0.3, & \text{if } w = 1
\end{cases}, \quad \Pr(\tilde{z}_2 = w) = \begin{cases} 
0.45, & \text{if } w = 0 \\
0.1, & \text{if } w = 0.5 \\
0.45, & \text{if } w = 1
\end{cases},
\]
For $\mathcal{F}_{\text{sc}}$, we let $\mathbb{P}^1, \mathbb{P}^2$ be the conditional distributions on $\tilde{k} = 1$ and $\tilde{k} = 2$, respectively. Their marginal distributions can be represented as

$$
\mathbb{P}^1(\tilde{z}_1 = w) = \begin{cases} 
2.5\delta_1, & \text{if } w = 0 \\
1 - 5\delta_1, & \text{if } w = 0.2 \\
2.5\delta_1, & \text{if } w = 0.4 
\end{cases}, \\
\mathbb{P}^1(\tilde{z}_2 = w) = \begin{cases} 
10\delta_2, & \text{if } w = 0.9 \\
1 - 20\delta_2, & \text{if } w = 0.95 \\
10\delta_2, & \text{if } w = 1 
\end{cases}; \\
\mathbb{P}^2(\tilde{z}_1 = w) = \begin{cases} 
2.5\delta_1, & \text{if } w = 0 \\
1 - 5\delta_1, & \text{if } w = 0.8 \\
2.5\delta_1, & \text{if } w = 1 
\end{cases}, \\
\mathbb{P}^2(\tilde{z}_2 = w) = \begin{cases} 
10\delta_2, & \text{if } w = 0 \\
1 - 20\delta_2, & \text{if } w = 0.05 \\
10\delta_2, & \text{if } w = 0.1 
\end{cases}.
$$

If the objective function of the second-stage problem is supermodular in the uncertain parameters and $\delta_1 < 4\delta_2$, we derive their worst-case distributions according to Proposition 1 and Algorithm 1 and show them in Figure 3.

![Figure 3](image_url)

**Figure 3** The worst-case distribution of $\sup_{\mathcal{F} \in \mathcal{F}} \mathbb{E}_\mathcal{F} [g(x, \tilde{z})]$ when $\mathcal{F} = \mathcal{F}_{\text{ind}}$ (Figure (a)), $\mathcal{F}_{\text{sup}}$ (Figure (b)), and $\mathcal{F}_{\text{sc}}$ (Figure (c)). The support for each case is marked by nodes with the exact probability masses.

For the worst-case distributions derived from sets $\mathcal{F}_{\text{sup}}$ and $\mathcal{F}_{\text{sc}}$, we can calculate their corresponding correlation coefficients, which are $\rho_{\text{sup}} = 0.816$ and $\rho_{\text{sc}} = \frac{-0.135 + 0.05\delta_1}{\sqrt{(0.09 + 0.28)(0.2025 + 0.05\delta_2)}}$, respectively. Since $0 \leq \delta < (0.2, 0.05)$, we get $\rho_{\text{sc}} \in (-1, -0.77)$. Hence, with the scenario-based uncertainty set, we can have negative correlation between random variables. In addition, we remark that for the worst-case distribution derived from $\mathcal{F}_{\text{ind}}$, we can also construct a scenario-based uncertainty set to recover this distribution.
2.4. Optimality of 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).

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

$$\min_{x, y(k, z)} a^\top x + \sup_{P \in \mathcal{F}} E_P [b^\top y(\tilde{k}, \tilde{z})]$$

s.t. $Wx + Uy(k, z) \geq Vz + v^0 \quad \forall z \in [z^k, \bar{z}^k], k \in [K], x \in \mathcal{X}. \quad (8)$

In general, the above formulation involves a functional decision $y(\tilde{k}, \tilde{z})$ and hence induces computational complexity. We now prove that in our setting, it suffices 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 $z^{k,1}, \ldots, z^{k,2n+1} \in \mathbb{R}^n$. We first lift the support to $\mathbb{R}^{2n}$ by defining

$$\zeta^{k,i} = \begin{bmatrix} \omega^{k,i} \\ \nu^{k,i} \end{bmatrix} \quad (9)$$

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

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

Using the above property, we are now ready to show the optimality of a segregated affine decision rule. Specifically, we restrain the recourse decision $y(k, z)$ to be an affine function in $[\mu^k - z]^+, (z - \mu^k)^+$, and obtain the following problem based on Problem (8),

$$\min_{x, \Theta^k, \phi^k, k \in [K]} a^\top x + \sup_{P \in \mathcal{F}} E_P \left[ b^\top \left( \Theta^k (\mu^k - z)^+ + \phi^k \right) \right]$$

subject to $Wx + Uy(k, z) \geq Vz + v^0 \quad \forall z \in \{z^{k,1}, \ldots, z^{k,2n+1}\}, k \in [K], x \in \mathcal{X}. \quad (10)$
Denote the optimal solution for $x, y^{k,i}$ to Problem (7) by $x^{\text{opt}}, y^{k,i \text{opt}}$, $k \in [K], i \in [2n + 1]$, which can be considered as given constants. Further, for any $k \in [K]$, we define a matrix $D^k \in \mathbb{R}^{2n \times 2n}$, a matrix $\Theta^k_{\text{opt}} \in \mathbb{R}^{m \times 2n}$ and a vector $\phi^k_{\text{opt}} \in \mathbb{R}^m$ as follows,

$$D^k = [\zeta^{k,1} - \zeta^{k,2n+1} \ldots \zeta^{k,2n} - \zeta^{k,2n+1}],$$

$$\Theta^k_{\text{opt}} = [y^{k,1}_{\text{opt}} - y^{k,2n+1}_{\text{opt}} \ldots y^{k,2n}_{\text{opt}} - y^{k,2n+1}_{\text{opt}}] (D^k)^{-1},$$

$$\phi^k_{\text{opt}} = y^{k,2n+1}_{\text{opt}} - \Theta^k_{\text{opt}} \zeta^{k,2n+1},$$

where $\zeta^{k,i}$ is the lifted uncertainty realization defined in Equation (9). Note that $D^k$ is invertible since by Lemma 2, $\zeta^{k,1}, \ldots, \zeta^{k,2n+1}$ are affinely independent, and hence $\Theta^k_{\text{opt}}$ is well defined. We then have the following result.

**Proposition 4** If $g(x, z)$ is supermodular in $z$ for any $x$, then Problem (10) and Problem (8) have the same optimal value. Specifically, $x = x^{\text{opt}}, \Theta^k = \Theta^k_{\text{opt}}$ and $\phi^k = \phi^k_{\text{opt}}, k \in [K]$ is an optimal solution for Problem (10).

By Proposition 4 with the supermodularity of $g(x, z)$ in $z$, 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 $[(\mu^k - z)^+ + (z - \mu^k)^+]$. It is worth mentioning that when we change the original optimization problem (8) to the affine decision rule formulation (10), we do not enforce the constraint to be feasible for all possible $z$. Instead, we only enforce the constraint for the $(2n + 1)$ realizations of $\bar{z}$ at each scenario. This is for two reasons. First, if we enforce the constraint for all possible $z$, the affine decision rule formulation, though having affine structure, is still computationally intractable. Interested readers can refer to Goh and Sim (2010), who propose a conservative approximation approach for such problems. Without computational tractability, it is meaningless to investigate the corresponding affine decision rule formulation. The second reason is that as shown in Proposition 4, it does not change the optimal value as well as the optimal first-stage decision, which is the essential focus in two-stage problems.

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 4 we show that by lifting the uncertainty realization $z \in \mathbb{R}^n$ to $[(\mu^k - z)^+ + (z - \mu^k)^+] \in \mathbb{R}^{2n}$, 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 $g(x, z)$.

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.

3. Conditions for supermodularity of the second-stage problems

If the second-stage cost, \( g(x, z) \), is supermodular in \( z \), Section 2 has shown that a tractable formulation can be achieved. Unfortunately, supermodularity in \( z \) 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 \( g(x, z) \) as

\[
g(x, z) = \min_b b^\top y \\
\text{s.t. } Uy - Vz \geq -Wx + v^0. 
\]

(11)

It is the optimal value of a parametric optimization problem which is parametrized by \( z \), and we need to explore the supermodularity in this parameter. Note that since we only focus on the supermodularity in \( z \) but not in \( x \), we do not consider \( x \) as a parameter in this parametric optimization problem. Hence, we move \( x \) 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 (11), 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 \( g(x, z) = -\max \{-b^\top y \mid Uy - Vz \geq -Wx + v^0\} \), then the supermodularity condition for \( g \) is reduced to the submodularity condition for the inner maximization problem. It is then again different from the literature which is on supermodularity for maximization problem. Therefore, we cannot rely on the literature of maximization problems to resolve the challenge of the minimization problem.

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 \( S \subseteq \mathbb{R}^m \times \mathbb{R}^n \) is an inverse additive lattice if for any \( p, q \in \mathbb{R}^m \), \( z', z'' \in \mathbb{R}^n \) with \((p, z' \land z'')\), \((q, z' \lor z'')\) \( \in S \), there exist \( y', y'' \in \mathbb{R}^m \) such that \((y', z')\), \((y'', z'')\) \( \in S \) and \( y' + y'' = p + q \).
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 $x$, we denote the set of all feasible pairs of $(y, z)$ as $S(x)$, i.e., $S(x) = \{(y, z) \mid Uy - Vz \geq -Wx + v^0\}$.

**Theorem 2** Given any $x$, $g(x, z)$ is supermodular in $z$ for any $b$ if and only if $S(x)$ is an inverse additive lattice.

Theorem 2 presents a necessary and sufficient condition for the second-stage cost being supermodular in the uncertainty realization $z$ for a given first-stage decision $x$. Now it remains to characterize the structure of the second-stage problem such that the condition can always be satisfied for any $x$.

**Theorem 3** $g(x, z)$ is supermodular in $z$ for any $x, b$ and $v^0$ if and only if $U \in \mathbb{R}^{r \times m}$ and $V \in \mathbb{R}^{r \times n}$ satisfy one of the following conditions:

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

2) for all $I \subseteq [r]$, $\beta \in \mathbb{R}^n_+$ satisfying $|I| = \text{rank}(U) + 1$, $\text{rank}(U|_I) = \text{rank}(U)$ and $V|_I \beta \in \text{span}(U|_I)$,

   we must have $\beta_i (V|_I)_i \in \text{span}(U|_I)$ holds for every $i \in [n]$.

The above result can be considered to correspond to Theorem 11 of [Chen et al., 2021], which focuses on the supermodularity in a parametric maximization problem. Nevertheless, as we mentioned at the beginning of this section, due to the different nature of minimization and maximization problems, the results differ essentially and do not apply to each other.

For any given matrices $U, V$, we provide the following algorithm to check explicitly whether the condition in Theorem 3 is met.

**Algorithm 2** algorithm for checking supermodularity

1: Input: $U \in \mathbb{R}^{r \times m}$, $V \in \mathbb{R}^{r \times n}$

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

3: if $r_0 < r$ then

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

5: for all $I \subseteq [r]$ with $|I| = r_0$ and $U|_I$ invertible, do

6: for $i \in [r]\setminus I$ do

7: $d_i^T = v_i^T - u_i^T U|_I^{-1} V|_I$

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 4 The condition in Theorem 3 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 \( U \) and \( V \) can also be a necessary condition. If the condition is violated for any tested index set \( I \), then the function \( g(x, z) \) must not be supermodular for all \( x, b, v^0 \). On the other hand, if we aim at sufficient conditions only, some matrices with simple structures can be easily shown to satisfy the conditions. We provide the following examples for illustration.

- \( U = I_{r \times r} \) or \( U = [I_{r \times r}, U^0] \) for some \( U^0 \in \mathbb{R}^{r \times (m-r)} \). The first condition of Theorem 3 is satisfied, hence \( g(x, z) \) is supermodular in \( z \) for any arbitrary \( V \in \mathbb{R}^{r \times n} \).
- \( U = \begin{bmatrix} I_{m \times m} \\ U^0 \end{bmatrix} \) for some \( u_i \in \mathbb{R}^m \). In this case, \( \text{span} (U) = \{ \xi \in \mathbb{R}^r \mid \sum_{i \in [m]} u_i \xi_i = \xi_r \} \). Correspondingly, based on the second condition of Theorem 3, we can prove \( g(x, z) \) is supermodular in \( z \) if and only if \( (u_r, -1)^\top V_1, \ldots, (u_r, -1)^\top V_n \) have the same sign.
- Given any \( g(x, z) = \min \{ b^\top y \mid Uy - Vz \geq -Wx + v^0 \} \), we consider the problem with partial constraints, i.e., \( g^2(x, z) = \min \{ b^\top y \mid Uz - Vz \geq -Wz x + v^0 \} \) for some \( I \subseteq [r] \). If \( g(x, z) \) is supermodular in \( z \), so is \( g^2(x, z) \).
- \( U = \begin{bmatrix} I_{m \times m} \\ U^0 \end{bmatrix} \in \mathbb{R}^{r \times m}, V = \begin{bmatrix} V^1 \\ 0_{(r-m) \times n} \end{bmatrix} \in \mathbb{R}^{r \times n} \). This choice of \( U \) and \( V \) includes the ATO system, the details of which will be discussed later, as a special case. The corresponding result can be formalized as follows.

Proposition 5 Assume \( U = \begin{bmatrix} I_{m \times m} \\ U^0 \end{bmatrix} \in \mathbb{R}^{r \times m}, V = \begin{bmatrix} V^1 \\ 0_{(r-m) \times n} \end{bmatrix} \in \mathbb{R}^{r \times n} \) and for each row of \( U^0 \), all components have the same sign. The function \( g(x, z) \) is supermodular in \( z \) for any \( x, b, v^0 \) and \( V^1 \), if and only if every \( 2 \times 3 \) submatrix of \( U^0 \) 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 3 or Algorithm 2 to verify whether the second-stage cost function is supermodular in \( z \). If the answer is positive, we can use the result in Theorem 1 to obtain an equivalent formulation as Problem 7, and derive the optimal solution efficiently.

4. Extensions

In the previous sections, we have discussed distributionally robust two-stage problems where the second-stage cost can be evaluated by solving a linear program. In this section we consider three possible extensions and show that, when the property of supermodularity holds, the exact tractable reformulation can be applied to more general settings.
4.1. Left-hand-side uncertainties in the constraints

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

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

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

**Theorem 5** \( g^W(x, z) \) is supermodular in \( z \) for any \( x, b \) and \( v^0 \) if and only if \( U \in \mathbb{R}^{r \times m} \), \( V \in \mathbb{R}^{r \times n} \) and \( W_i \in \mathbb{R}^{r \times l}, i \in [n] \) satisfy one of the following conditions:

1) \( \text{rank}(U) = r \),

2) for all \( I \subseteq [r], \eta \in \mathbb{R}^{\left| I \right|} \) with \( \left| I \right| = \text{rank}(U) + 1 \), \( \text{rank}(U_I) = \text{rank}(U) \) and \( U_I^\top \eta = 0 \), we have

\[
2a) (\eta^\top (V_I)_i) \cdot (\eta^\top (V_I)_j) \geq 0, \quad (W^j_I)^\top \eta \eta^\top W^j_I \text{ is positive semidefinite, for all } i, j \in [n];
\]

\[
2b) (\eta^\top (V_I)_i) \cdot (\eta^\top W^j_I) = (\eta^\top (V_I)_j) \cdot (\eta^\top W^j_I), \quad \text{for all } i, j \in [n].
\]

For Condition 2) in Theorem 5, considering any concerned \( I \), i.e., \( \left| I \right| = \text{rank}(U) + 1 \) and \( \text{rank}(U_I) = \text{rank}(U) \), the null space of \( U_I \) is of dimension 1. That is, \( \exists \eta^o \) such that for all \( \eta \) with \( U_I^\top \eta = 0 \) we have \( \eta = k \eta^o \) for some \( k \in \mathbb{R} \). We can easily observe that both Conditions 2a) and 2b) hold for all such \( \eta \) if and only if they hold for \( \eta^o \). Therefore, to verify whether Conditions 2a) and 2b) hold, it suffices to check for \( \eta^o \) only. Hence, as in Theorem 4, we can similarly build a corresponding algorithm to check the supermodularity of \( g^W \).

4.2. 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 \( u: \mathbb{R} \to \mathbb{R} \) and any monotone supermodular function \( h: \mathbb{R}^n \to \mathbb{R} \), the function \( \phi: \mathbb{R}^n \to \mathbb{R} \) defined as \( \phi(z) = u(h(z)) \) 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

\[
\min_{x \in X} \sup_{P \in \mathcal{P}} \mathbb{E}_P \left[ u \left( a^\top x + g(x, \tilde{z}) \right) \right],
\]
where \( g(x, z) \) is the second-stage cost function defined by (1), and \( u : \mathbb{R} \to \mathbb{R} \) is a piecewise convex and non-decreasing disutility function defined as
\[
u(w) = \max_{j \in [J]} \{ c_j w + d_j \}, \quad \forall w \in \mathbb{R}, \tag{13}\]
for some constants \( c_j \geq 0 \) and \( d_j, j \in [J] \).

**Proposition 6** If \( g(x, z) \) is monotone and supermodular in \( z \) for all \( x \in \mathcal{X} \), then Problem (12) is equivalent to the following problem
\[
\min \nu^T l \\
\text{s.t.} \quad R_k^T l \geq \sum_{i \in [2n+1]} p_{k,i} f_{k,i}, \quad k \in [K] \\
W x + U y^{k,i} \geq V z^{k,i} + v^0, \quad k \in [K], i \in [2n+1] \\
f^{k,i} \geq c_j (a^T x + b^T y^{k,i}) + d_j, \quad k \in [K], i \in [2n+1], j \in [J] \\
l \geq 0, \quad x \in \mathcal{X},
\tag{14}\]
where \( p_{k,i}, z^{k,i}, i \in [2n+1] \) 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) \cite{Ben-TalTeboulle1986}. It is shown that the OCE models a broad range of risk measures \cite{Ben-TalTeboulle2007}, 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,
\[
\min \sup_{x \in \mathcal{X}} \text{OCE}_u(a^T x + g(x, z)) = \min \sup_{x \in \mathcal{X}} \inf_{\theta \in \mathbb{R}} \{ \theta + \mathbb{E}_v [u(a^T x + g(x, z) - \theta)] \}, \tag{15}\]
where \( u(\cdot) \) is a piecewise convex and non-decreasing disutility function taken the form of (13). We now show that our method is applicable to Problem (15).

**Corollary 2** If \( g(x, z) \) is monotone and supermodular in \( z \) for all \( x \in \mathcal{X} \), then the OCE minimization problem (15) is equivalent to the following linear program
\[
\min \theta + \nu^T l \\
\text{s.t.} \quad R_k^T l \geq \sum_{i \in [2n+1]} p_{k,i} f_{k,i}, \quad k \in [K] \\
W x + U y^{k,i} \geq V z^{k,i} + v^0, \quad k \in [K], i \in [2n+1] \\
f^{k,i} \geq c_j (a^T x + b^T y^{k,i} - \theta) + d_j, \quad k \in [K], i \in [2n+1], j \in [J] \\
l \geq 0, \quad x \in \mathcal{X},
\tag{16}\]
where \( p_{k,i}, z^{k,i}, i \in [2n+1] \) are obtained from Algorithm 1 given the ambiguity sets \( \mathcal{F}^k, k \in [K] \) defined by (4).
4.3. 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 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,

$$F^G = \left\{ \mathbb{P} \left| \begin{array}{cl}
\mathbb{P}(z \leq \tilde{z} \leq z) &= 1 \\
\mathbb{E}_{\mathbb{P}}[\tilde{z}] &= \mu \\
\mathbb{E}_{\mathbb{P}}[h^i_j(\tilde{z})] &\leq \delta^i_j, \quad i \in [n], j \in [J_i]
\end{array} \right. \right\},$$

(17)

where $J_i \geq 1$ is an integer, $h^i_j$ is a given convex piecewise linear 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 $F^G$ generalizes $F^K$, defined in Equation (4), as it replaces the MAD information in $F^K$ by the expectations of several convex piecewise linear functions. Obviously, $F^G$ includes $F^K$ as a special case by choosing $J^i = 1$ and $h^i_j(z) = |z - \mu_i|$ for all $z \in \mathbb{R}$.

Unfortunately, as we will show later in this subsection, not all ambiguity sets $F^G$ defined by Equation (17) can lead to a tractable reformulation using the procedures we discussed in Section 2. Here we aim to identify the conditions for $F^G$ 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 $z^1_i = \underline{z}_i$, $z^2_i = \overline{z}_i$ and denote $z^3_i, z^4_i, \ldots, z^{S^i_i}_i \in (\underline{z}_i, \overline{z}_i)$ 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 6** The following two statements are equivalent.

1. Given any $\delta^i_j$, $i \in [n]$, $j \in [J_i]$ satisfying $F^G \neq \emptyset$, there exists $p_i = (p_{i1}, \ldots, p_{iS^i_i}) \in \mathbb{R}^{S^i_i}_+$, $i \in [n]$ such that for all convex function $f : \mathbb{R}^n \to \mathbb{R}$, we have $p^* \in \arg\sup_{p \in F^G} \mathbb{E}_{\mathbb{P}}[f(\tilde{z})]$ and for any $i \in [n],$

$$p^*(\tilde{z}_i = w) = \begin{cases} 
p_{is} & \text{if } w = z^s_i, \ s \in [S^i_i], \\
0 & \text{otherwise.}
\end{cases}$$

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 $p^*$ provided in Theorem 6 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 6) or the first-stage decision $x$ (in Proposition 1). Therefore, if $F^G$ satisfies the condition in Theorem 6 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 $F^G$ violates the condition in Theorem 6 there are two-stage problems such that the worst-case distribution would depend on the first-stage decision $x$, and hence our method cannot work.

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, where we minimize the expected sum of waiting time and overtime. Finally, a general formulation of ATO systems is discussed in Section 5.5 where we identify a class of systems which are tractable using our method. 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 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. They also develop a semidefinite program to derive the 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 $s_i$, respectively. We assume $r_i > s_i$ to avoid trivial solutions. Before the random demand $\tilde{z}$ is resolved, we need to decide the
ordering quantity $x$, which is subject to a budget $\Gamma$. Our goal is to minimize the worst-case expected disutility of cost. This yields the following optimization problem

$$
\min_{x \in X_{\text{news}}} \sup_{P \in \mathcal{F}} \mathbb{E}_P \left[ u \left( -r^\top x + (r - s) (x - \tilde{z})^+ \right) \right],
$$

where $X_{\text{news}} = \{ x \in \mathbb{R}^n \mid t^\top x \leq \Gamma, x \geq 0 \}$ and $u$ is a convex and increasing piecewise linear disutility function as defined in (13).

**Corollary 3** Problem (18) is equivalent to the following problem,

$$
\min \nu^\top l \\
\text{s.t. } R_k^\top l \geq \sum_{i \in [2n+1]} \frac{1}{\rho} p_i^k f^{k,i}, \quad k \in [K] \\
y^{k,i} \geq x - z^{k,i}, \quad k \in [K], i \in [2n+1] \\
y^{k,i} \geq 0, \quad k \in [K], i \in [2n+1] \\
f^{k,i} \geq c_j \left( -r^\top x + (r - s)^\top y^{k,i} \right) + d_j, \quad k \in [K], i \in [2n+1], j \in [J] \\
l \geq 0, \quad x \in X_{\text{news}}.
$$

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

$$
\min_{x \in X_{\text{news}}} \sup_{P \in \mathcal{F}} \mathbb{E}_{\text{CVaR}_\rho} \left[ -r^\top x + (r - s) (x - \tilde{z})^+ \right].
$$

By the definition of CVaR, the above problem is equivalent to

$$
\min_{x \in X_{\text{news}}} \sup_{P \in \mathcal{F}} \inf_{\theta \in \mathbb{R}} \left\{ \theta + \mathbb{E}_P \left[ \frac{1}{\rho} \cdot \left( -r^\top x + (r - s)^\top (x - z)^+ - \theta \right)^+ \right] \right\}.
$$

**Corollary 4** Problem (19) is equivalent to the following problem,

$$
\min \theta + \nu^\top l \\
\text{s.t. } R_k^\top l \geq \sum_{i \in [2n+1]} \frac{1}{\rho} p_i^k f^{k,i}, \quad k \in [K] \\
y^{k,i} \geq x - z^{k,i}, \quad k \in [K], i \in [2n+1] \\
y^{k,i} \geq 0, \quad k \in [K], i \in [2n+1] \\
f^{k,i} \geq -r^\top x + (r - s)^\top y^{k,i} - \theta, \quad k \in [K], i \in [2n+1] \\
f^{k,i} \geq 0, \quad k \in [K], i \in [2n+1] \\
l \geq 0, \quad x \in X_{\text{news}}.
$$

With the objective of minimizing CVaR and the multi-modal demand assumption, we notice that 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, here we focus on partial marginal information and obtain an exact linear programming reformulation of the original problem. However, our model cannot account for the stockout costs as in Hanasusanto et al. (2015). When stockout costs are included, the total cost no longer decreases with the demand \( z \). As a result, despite the total cost itself still being supermodular in \( z \), the supermodularity is not preserved when a general convex disutility is considered. We illustrate using the following example. Consider a 2-item problem with the selling price \( r = (6, 3) \), salvage value \( s = (2, 2) \), stockout cost \( b = (2, 2) \) and disutility function \( u(w) = (w + 5)^+ \). Then the total cost is

\[
- r^\top \min \{x, z\} - s^\top (x - z)^+ + b^\top (z - x)^+
\]

\[
= - (r + b)^\top x + b^\top z + (r - s + b)^\top (x - z)^+
\]

\[
= - (8, 5)^\top x + (2, 2)^\top z + (6, 3)^\top (x - z)^+.
\]

Let \( g(x, z) = (2, 2)^\top z + (6, 3)^\top (x - z)^+ \) and consider \( x = (1, 1), z' = (2, 0), z'' = (0, 2) \),

\[
g(x, z') = g((1, 1), (0, 0)) = 0 + 9 = 9, \quad u(-8 - 5 + 9) = 1,
g(x, z') = g((1, 1), (2, 0)) = 4 + 3 = 7, \quad u(-8 - 5 + 7) = 0,
g(x, z'') = g((1, 1), (0, 2)) = 4 + 6 = 10, \quad u(-8 - 5 + 10) = 2,
g(x, z' \lor z'') = g((1, 1), (2, 2)) = 8 + 0 = 8, \quad u(-8 - 5 + 8) = 0.
\]

Clearly we have \( u(-(r + b)^\top x + g(x, z' \lor z'')) = 1 + 0 < 0 + 2 = u(-(r + b)^\top x + g(x, z')) + u(-(r + b)^\top x + g(x, z'')) \), which violates supermodularity.

### 5.2. Reliable facility location design

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 \( x = (x_1, \ldots, x_n) \) is made, where

\[ x_i = 1 \text{ if facility is opened at location } i, \quad x_i = 0 \text{ otherwise.} \]

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 \( \tilde{z} = (\tilde{z}_1, \ldots, \tilde{z}_n) \), which are realized after the first-stage decision \( x \) is made. \( \tilde{z}_i = 0 \) if location \( i \) is disrupted, and \( \tilde{z}_i = 1 \) otherwise. The disruption happens at location \( i \) with probability \( M_i \). The cost minimization problem is formulated as

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

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

\[
\begin{align*}
& g(x, z) = \min \sum_{i \in [n], j \in [m]} c_{ij} y_{ij} \\
& \text{s.t.} \quad \sum_{i \in [n]} y_{ij} = 1, \quad j \in [m], \\
& \quad 0 \leq y_{ij} \leq x_i z_i, \quad i \in [n], j \in [m].
\end{align*}
\]
We remark that the above second-stage problem always has an optimal solution \( y_{ij} \in \{0, 1\} \quad \forall i \in [n], j \in [m], \) and hence we do not include the binary constraints on \( y_{ij} \) in the problem explicitly. Lu et al. (2015) prove the supermodularity of \( g(x, z) \) by verifying the definition. We show that the same result can be obtained by a direct application of Theorem 3.

**Proposition 7** The function \( g(x, z) \) defined by (21) is supermodular in \( z \) for all \( x \in \mathcal{X}^{\text{fac}} \).

### 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, prior to the realization of the demands, we determine an allocation \( x_i \) for the \( i \)-th store. The feasible set \( \mathcal{X}^{\text{lot}} \) describes the capacity of the stores, i.e. \( 0 \leq x_i \leq K_i \) for some \((K_1, \ldots, K_n) \in \mathbb{R}_+^n \). 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 as follows,

\[
\min_{x \in \mathcal{X}^{\text{lot}}} \left\{ \alpha^T x + \sup_{P \in \mathcal{F}} \mathbb{E}_P \left[ \min \left\{ \sum_{s,j \in [n]} b_{sj} y_{sj} \right| \sum_{j \in [n]} y_{js} - \sum_{j \in [n]} y_{sj} \geq z_s - x_s, \quad s \in [n], j \in [n] \right\} \right\}.
\]  

(21)

While Bertsimas and de Ruiter (2016) derive an approximation for the robust version of Problem (21) when the uncertainty set is a polyhedron, we next show that the problem can be solved exactly in polynomial time within our setting. Let \( g(x, z) \) be the second-stage cost for a given allocation \( x \) and realized demand \( z \).

**Proposition 8** The function \( g(x, z) \) defined by the inner optimization problem in (21) is supermodular in \( z \) for all \( x \). Hence, Problem (21) is equivalent to the following linear optimization problem

\[
\min \alpha^T x + \nu^T l \\
\text{s.t.} \quad R_k^T l \geq \sum_{i \in [2n+1]} p^k_i \sum_{s,j \in [n]} b_{sj} y^{k,i}_{sj}, \quad k \in [K] \\
\sum_{j \in [n]} y^{k,i}_{js} - \sum_{j \in [n]} y^{k,i}_{sj} \geq z^{k,i}_s - x_s, \quad s \in [n], k \in [K], i \in [2n+1] \\
y^{k,i}_{sj} \geq 0, \quad s \in [n], j \in [n], k \in [K], i \in [2n+1] \\
l \geq 0, \quad x \in \mathcal{X}^{\text{lot}},
\]

where \( p^k_i, z^{k,i}, k \in [K], i \in [2n+1] \) are the output of Algorithm 1 given the ambiguity sets \( \mathcal{F}^k, k \in [K] \) defined by Equation (4).
When the transported amount \(y_{sj}\) is bounded by a capacity \(c_{sj}\), as in Bertsimas and Shtern (2018), the second-stage cost becomes

\[
\hat{g}(x, z) = \min \sum_{s,j \in [n]} b_{sj} y_{sj},
\]

subject to

\[
\sum_{j \in [n]} y_{js} - \sum_{j \in [n]} y_{sj} \geq z_s - x_s, \quad s \in [n]
\]

\[
0 \leq y_{sj} \leq c_{sj}, \quad s \in [n], j \in [n].
\]

By a similar analysis we can verify that \(\hat{g}(x, z)\) defined by (22) 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 (Ho and Lau 1992, Cayirli and Veral 2003, Gupta and Denton 2008). In this section, we focus on robust appointment scheduling problems where no-shows are possible. When random no-shows are considered, Jiang et al. (2017) assume 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 all previous studies.

We consider to schedule \(n\) appointments within a given time period \(\Gamma\). For all \(i \in [n]\), we assume customer \(i\) shows up with probability \(\theta_i\) and use \(\tilde{\xi}_i \in \{0, 1\}\) to characterize this event, i.e., \(P(\tilde{\xi}_i = 1) = \theta_i, P(\tilde{\xi}_i = 0) = 1 - \theta_i\). Let \(\tilde{z}_i\) be the actual duration of the \(i\)-th service. At the beginning of the planning horizon, 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

\[
\tilde{w}_{i+1} = \max \left\{ \tilde{w}_i + \tilde{\xi}_i \tilde{z}_i - x_i, 0 \right\}, \quad \forall i \in [n - 1].
\]

We follow Jiang et al. (2017) and formulate the problem using a two-stage optimization structure as

\[
\min_{x \in X_{opt}} \sup_{P \in \mathcal{G}} \mathbb{E}_P \left[ g(x, \tilde{\xi}, \tilde{z}) \right], \quad (23)
\]
Here the feasible set is defined as $X^{app} = \{ x \in \mathbb{R}_+^n \mid 1^T x \leq \Gamma \}$, and the second-stage problem can be written as

$$g(x, \xi, z) = \min \left\{ 1^T y \mid y_t \geq \sum_{s=j}^{t} (\xi_s z_s - x_s), \ j \in [t], t \in [n] \right\}, \quad (24)$$

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

$$F_{\xi} = \left\{ P_{\xi} \mid \mathbb{E}_{P_{\xi}} \left[ \tilde{\xi} \mid \tilde{k} = k \right] = \theta_k, \ P_{\xi} \left( \tilde{k} = k \right) = q_k, q \in \mathcal{Q} \right\}$$

with $\Xi^k = \{0,1\}^n$ and $F$ is defined by Equation (3). Here we do not incorporate the MAD information for random no-shows since the support of no-show contains only two points and hence the MAD information is redundant. Though the information set $G$ differs slightly from that in Equation (3), the key idea and process of our approach are still applicable. Using the condition in Theorem 3, we next demonstrate the supermodularity of the function $g(x, \xi, z)$.

**Proposition 9** Function $g(x, \xi, z)$ is supermodular in $(\xi, z)$ for all $x$. Hence, Problem (23) has a polynomial size linear programming reformulation.

Therefore, given the information set $G$, we can reformulate Problem (23) 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 $\Xi^k$ in Equation (25). For example, if we do not want to incorporate the scenarios in which all patients are absent, we can let $K = n$ and consider $\Xi^k = \{\xi \in \{0,1\}^n \mid \xi_k = 1\} = \{0,1\} \times \cdots \times \{1\} \times \cdots \times \{0,1\}$ for all $k \in [n]$. The problem remains tractable, since the support of $\xi$ 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. In this case, we exclude the case that all customers do not show up (i.e. $\xi = 0$).

It is brought to our attention that Chen et al. (2018) prove when no-shows are not considered, the objective function is supermodular in the uncertain appointment durations $z$. 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 3.
5.5. Assemble-to-order systems

The ATO system is an important operational problem. We refer interested readers to Song and Zipkin (2003) and Atan et al. (2017) for a comprehensive review. 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 to the ATO problem 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 $i$, first we decide the order-up-to inventory level $x_i$. Then the uncertain demand $\tilde{z}_j$ for end product $j$, $j \in [n]$, is realized as $z_j$. After that, we make the second-stage decision $y_j$, which is the quantity of product $j$ to be assembled. With the goal of minimizing the worst-case expected cost, we have the following formulation,

$$
\min \ c^T (x - x_{int}) + \sup_{P \in \mathcal{F}} \mathbb{E}_P [g(x, \tilde{z})]
$$

s.t. $x \geq x_{int}$,

(26)

where

$$
g(x, z) = \min \ h^T (x - Ay) + p^T (z - y) - r^T y
$$

s.t. $Ay \leq x$, $y \leq z$, $y \geq 0$.

(27)

Here $c$ and $x_{int}$ are the per-unit ordering cost and initial inventory level of the components, respectively; $h$ is the per-unit inventory holding cost of the leftover components; $p$ and $r$ are the per-unit penalty cost of the shortage and per-unit selling price of the end products, respectively. The elements in matrix $A$, 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 $A \in \mathbb{R}^{l \times n}$. We next provide a condition on $A$ such that the function $g(x, z)$ is supermodular in $z$, hence the optimal order-up-to level for each component can be derived efficiently based on Theorem 1.

Theorem 7 The function $g(x, z)$ is supermodular in $z$ for any $x, h, p, r$ if and only if every $2 \times 3$ submatrix of the matrix $A$ contains at least one pair of column vectors which are linearly dependent.

We next test the condition of Theorem 7 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 $g(x, z)$, 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 $j \in S_i \cap S_{i'}$.

Applying Theorem 7, we conclude that Proportional Tree Family has the property of supermodularity.

Corollary 5 The function $g(x, z)$ is supermodular in $z$ for any $x, h, p, r$ if the system belongs to Proportional Tree Family.

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

- $A \in \mathbb{R}^{l \times 1}$ or $A \in \mathbb{R}^{l \times 2}$, 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 7.
- Binary Tree Family (Zipkin 2016): a system belonging to Tree Family and with all elements in $A$ being binary. We can show it is in the Proportional Tree Family.
- The generalized W System (Zipkin 2016, Chen et al. 2021): $A = [D c^\top] \in \mathbb{R}^{(n+1) \times n}$ with $D$ 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, Doğru et al. 2017): $A = [D c] \in \mathbb{R}^{n \times (n+1)}$ with $D$ 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 7 and hence $g(x, z)$ is not supermodular in $z$.

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. 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 is independent of the first-stage
decision and 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. We also show that the scenario-wise segregated affine decision rules can provide an optimal value of the original 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.

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 news-vendor, reliable facility location design, lot-sizing, appointment scheduling with random no-shows, and general ATO systems. While these problems are considered to be computationally challenging in general, with our approach, they can be solved exactly and efficiently.

References


Appendix

Proof of Proposition 1

For a one-dimensional random variable, when its MAD is known exactly, Ben-Tal and Hochman (1972) have derived the worst-case distribution. Here since in $F^k$, 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 $\tau_i > \bar{z}_i$. For the $i$-th marginal, given the support $[\bar{z}_i, \tau_i]$ and mean $\mu_i$, the maximum possible value of MAD is $\frac{2(\tau_i - \mu_i)(\mu_i - \bar{z}_i)}{\tau_i - \bar{z}_i}$ (Lemma 1 in Ben-Tal and Hochman (1972)). Hence, we let $\hat{\delta}_i = \min \left\{ \delta_i, \frac{2(\tau_i - \mu_i)(\mu_i - \bar{z}_i)}{\tau_i - \bar{z}_i} \right\}$. Then, the worst-case expectation of $g(x, \hat{z})$ under the $F^k$ defined in Equation (4) can be reformulated as

$$
\sup_{P \in F} E_P [g(x, \hat{z})] = \sup \left\{ E_P [g(x, \hat{z})] \middle| \begin{array}{l}
E_P [\hat{z}] = \mu, \\
E_P [\hat{z} - \mu] \leq \hat{\delta}, \\
P(\bar{z} \leq \hat{z} \leq \tau) = 1
\end{array} \right\}
= \sup \left\{ E_P [g(x, \hat{z})] \middle| \begin{array}{l}
E_P [\hat{z}] = \mu, \\
E_P [\hat{z} - \mu] \leq \hat{\delta}, \\
P(\bar{z} \leq \hat{z} \leq \tau) = 1
\end{array} \right\}
= \sup_{0 \leq d \leq \hat{\delta}} V(d),
$$

where

$$
V(d) = \sup \left\{ E_P [g(x, \hat{z})] \middle| \begin{array}{l}
E_P [\hat{z}] = \mu, \\
E_P [\hat{z} - \mu] = d, \\
P(\bar{z} \leq \hat{z} \leq \tau) = 1
\end{array} \right\}.
$$

We prove our proposition by two steps.

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

$$
P^* (\hat{z}_i = w) = \begin{cases} 
\frac{d_i}{2(\mu_i - \bar{z}_i)} & \text{if } w = \bar{z}_i \\
\frac{1 - d_i(\tau_i - \bar{z}_i)}{2(\tau_i - \mu_i)(\mu_i - \bar{z}_i)} & \text{if } w = \mu_i \\
\frac{d_i}{2(\tau_i - \mu_i)} & \text{if } w = \tau_i \\
0 & \text{otherwise.}
\end{cases} \quad (28)
$$

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

Consider the first case where $V(d) = \infty$. In this case, there exists $P'$ such that $E_{P'} [g(x', \hat{z})] = \infty, E_{P'} [\hat{z}] = \mu, E_{P'} [\hat{z} - \mu] = d$ and $P'(\bar{z} \leq \hat{z} \leq \tau) = 1$. We denote by $\text{supp}(P)$ the support of any probability distribution $P$. Observing that the feasible set of $V(d)$ is nonempty (any distribution with marginal distribution as in (28) is feasible), $E_{P'} [g(x', \hat{z})] = \infty$ implies that there must exist $z' \in \text{supp}(P') \subseteq [\bar{z}, \tau]$ such that $g(x, z') = \infty$. We let $P''$ be any probability distribution with marginal distribution as defined in (28), then $\text{supp}(P'') = \prod_{i \in [n]} S_i$ where $S_i = \{\mu_i\}$ if $d_i = 0$, $S_i = \{\bar{z}_i, \mu_i, \tau_i\}$ if $d_i \in (0, \hat{\delta}_i)$ and $S_i = \{\bar{z}_i, \tau_i\}$ if $d_i = \hat{\delta}_i$, $i \in [n]$. Now, consider any $i \in [n]$. If $d_i = 0$, we must have $P'(\hat{z}_i = \mu_i) = 1$; hence, $z'_i = \mu_i \in \text{conv}(S_i)$; if $d_i > 0$, $z'_i \in \text{conv}(S_i) = [\bar{z}_i, \tau_i]$ since $z_i \in [\bar{z}, \tau]$. Hence, in
any case, \( z_i' \in \text{conv}(S_i) \). Consequently, we have \( z' \in \text{conv}(\text{supp}(P'')) \). Since function \( g(x, z) \) is convex in \( z \) (see Theorem 2, Section 3.1 in Birge and Louveaux [2011]), there must exist \( z'' \in \text{supp}(P'') \) such that \( g(x', z'') = \infty \). Hence, \( P'' \) is also a worst-case distribution.

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

\[
V(d) = \min \left\{ s + \mu^\top t + d^\top r \mid s + z^\top t + (|z - \mu|)^\top r \geq g(x, z), \ \forall z \leq z \leq z \right\} .
\]  

(29)

For any given \( t, \mu, r \), since \( g(x, z) \) is convex in \( z \), the function \( g(x, z) - z^\top t - (|z - \mu|)^\top r \) is convex in \( z \) if \( z \in [a, b] \) 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 (29) is equivalent to

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

(30)

Substituting the constraints in Problem (29) by (30) and writing its dual form again, we obtain

\[
V(d) = \sup \left\{ \sum_{\tau=1}^{3^n} p_\tau g(x, z^\tau) \mid \sum_{\tau=1}^{3^n} p_\tau z^{\tau}_i = \mu_i, \ i \in [n] \right. \\
\sum_{\tau=1}^{3^n} p_\tau |z^{\tau}_i - \mu_i| = d_i, \ i \in [n] \\
\sum_{\tau=1}^{3^n} p_\tau = 1 \\
\left. p_\tau \geq 0, \ \tau \in [3^n] \right\},
\]

(31)

where \( z^1, \ldots, z^{3^n} \) represent all \( z \in \prod_{i \in [n]} \{\underline{z}_i, \mu_i, \overline{z}_i\} \), and \( p_1, \ldots, p_{3^n} \) are the associated decision variables. Therefore, we can find a distribution \( P^* \) which is optimal to \( V(d) \), with its support being \( z^1, \ldots, z^{3^n} \). This implies \( P^*(\tilde{z}_i = w) = 0 \) 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 \( \mu_i \) and MAD value \( d_i \in [0, \delta_i] \), we observe that a distribution which places \( \frac{d_i}{2(\mu_i - \underline{z}_i)} \) amount of mass at \( \underline{z}_i \), \( 1 - \frac{d_i(\overline{z}_i - \mu_i)}{2(\overline{z}_i - \mu_i)} \) at \( \mu_i \) and \( \frac{d_i}{2(\overline{z}_i - \mu_i)} \) at \( \overline{z}_i \), is uniquely determined. Hence, \( P^* \) must have a marginal distribution as in (28).

**Step 2.** We next show that function \( V(d) \) is non-decreasing in \( d \).

Consider any \( 0 \leq d' \leq d'' \leq \delta \) with \( d'' - d' = \theta e_i \) for some \( \theta > 0, i^* \in [n] \), and the probability distribution \( P' \) with \( P'(\tilde{z} = z^\tau) = p'_\tau, \tau \in [3^n] \) such that \( p'_1, \ldots, p'_{3^n} \) is the worst-case distribution in Problem (31) when \( d = d' \). WLOG, we let \( i^* = 1 \). We define another distribution \( P'' \) with \( P''(\tilde{z} = z^\tau) = p''_\tau, \tau \in [3^n] \) as

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

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

\[
\epsilon(z) = \epsilon(z + (\underline{z}_1 - \mu_1)e_1) = \epsilon(z + (\overline{z}_1 - \mu_1)e_1) = \frac{\theta/2}{1 - d_i^2 \left(\frac{1}{\mu_1 - \underline{z}_1} + \frac{1}{\overline{z}_1 - \mu_1}\right)} \epsilon'(\tilde{z} = z)
\]

(32)
for all $z \in \prod_{i \in [n]} \{\tilde{z}_i, \mu_i, \tau_i\}$ with $z_1 = \mu_1$. We next verify that $p''_1, \ldots, p''_n$ satisfy the constraints in (31) when replacing $d$ by $d''$.

From the definition of $\theta$, we observe that for all $z$ such that $z_1 = \mu_1$,

$$
\frac{\epsilon(z)}{\mu_1 - \tilde{z}_1} + \frac{\epsilon(z)}{\tau_1 - \mu_1} = \frac{d'_1 - d'_2}{2} \left( \frac{1}{\mu_1 - \tilde{z}_1} + \frac{1}{\tau_1 - \mu_1} \right) + \left( 1 - \frac{d'_2}{2} \left( \frac{1}{\mu_1 - \tilde{z}_1} + \frac{1}{\tau_1 - \mu_1} \right) \right) \mathbb{P}'(\tilde{z} = z).
$$

Because $0 < d'_1 < d''_1 < \hat{d}_1$ and the three-point distribution is uniquely determined, we have $1 - \frac{d'_2}{2} \left( \frac{1}{\mu_1 - \tilde{z}_1} + \frac{1}{\tau_1 - \mu_1} \right) < 0$. Hence $\mathbb{P}''(\tilde{z} = z) \in [0, \mathbb{P}'(\tilde{z} = z)]$ for all $z$ with $z_1 = \mu_1$. By the definition of $\mathbb{P}''$ we notice that $\sum_{z \in \prod_{i \in [n]} \{\tilde{z}_i, \mu_i, \tau_i\}} \mathbb{P}''(\tilde{z} = z) = \sum_{z \in \prod_{i \in [n]} \{\tilde{z}_i, \mu_i, \tau_i\}} \mathbb{P}'(\tilde{z} = z) = 1$, we have $\mathbb{P}''(\tilde{z} = z) \in [0, 1]$ for all $z \in \prod_{i \in [n]} \{\tilde{z}_i, \mu_i, \tau_i\}$.

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

$$
\sum_{\tau = 1}^{3^n} p''_{\tau} z_1^\tau = \sum_{z: z_1 = \mu_1} \mathbb{P}''(\tilde{z} = z) \mu_1 + \sum_{z: z_1 = \tilde{z}_1} \mathbb{P}''(\tilde{z} = z) \tilde{z}_1 + \sum_{z: z_1 = \tau_1} \mathbb{P}''(\tilde{z} = z) \tau_1
$$

$$
= \sum_{z: z_1 = \mu_1} \left( \mathbb{P}'(\tilde{z} = z) \mu_1 + \mathbb{P}'(\tilde{z} = z) \tilde{z}_1 + \mathbb{P}'(\tilde{z} = z) \tau_1 \right)
$$

$$
= \sum_{z: z_1 = \mu_1} \left( -\frac{\epsilon(z)}{\mu_1 - \tilde{z}_1} - \frac{\epsilon(z)}{\tau_1 - \mu_1} \right) \mu_1 + \frac{\epsilon(z)}{\mu_1 - \tilde{z}_1} \tilde{z}_1 + \frac{\epsilon(z)}{\tau_1 - \mu_1} \tau_1
$$

$$
= \mu_1 + \sum_{z: z_1 = \mu_1} \left( \frac{\epsilon(z)}{\mu_1 - \tilde{z}_1} (\tilde{z}_1 - \mu_1) + \frac{\epsilon(z)}{\tau_1 - \mu_1} (\tau_1 - \mu_1) \right)
$$

$$
= \mu_1,
$$

where the third equality is due to the property of $\epsilon$ stated in (32). 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, $\sum_{\tau = 1}^{3^n} p''_{\tau} z_i^\tau = \mu$.

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

$$
\sum_{z: z_1 = \mu_1} \mathbb{P}''(\tilde{z} = z) = \sum_{z: z_1 = \mu_1} \mathbb{P}'(\tilde{z} = z) - \left( \frac{1}{\mu_1 - \tilde{z}_1} + \frac{1}{\tau_1 - \mu_1} \right) \sum_{z: z_1 = \mu_1} \epsilon(z)
$$

$$
= \left( 1 - \frac{\theta/2}{\mu_1 - \tilde{z}_1 + \frac{1}{\tau_1 - \mu_1}} \right) \sum_{z: z_1 = \mu_1} \mathbb{P}'(\tilde{z} = z)
$$

$$
= \left( 1 - \frac{\theta/2}{\mu_1 - \tilde{z}_1 + \frac{1}{\tau_1 - \mu_1}} \right) \left( 1 - \frac{d'_1}{2} \left( \frac{1}{\mu_1 - \tilde{z}_1} + \frac{1}{\tau_1 - \mu_1} \right) \right)
$$

$$
= 1 - \frac{d''_1}{2} \left( \frac{1}{\mu_1 - \tilde{z}_1} + \frac{1}{\tau_1 - \mu_1} \right),
$$

where the second equality is due to (32), the third equality holds since the marginal distribution is uniquely determined in $\mathbb{P}'$. Similarly, we have $\sum_{z: z_1 = \tilde{z}_1} \mathbb{P}''(\tilde{z} = z) = \frac{d''_1}{2(\mu_1 - \tilde{z}_1)}$ and $\sum_{z: z_1 = \tau_1} \mathbb{P}''(\tilde{z} = z) = \frac{d''_1}{2(\tau_1 - \mu_1)}$.}

Long, Qi, and Zhang: Supermodularity in Two-Stage DRO

37
Consider any \( \mathbb{P} \in \mathcal{P} \) such that there exists an unordered pair \( \mathbf{w}', \mathbf{w}'' \) with \( p' = \mathbb{P}(\mathbf{w} = \mathbf{w}') > 0 \), \( p'' = \mathbb{P}(\mathbf{w} = \mathbf{w}'') > 0 \). WLOG, assume \( p' \leq p'' \). We construct a new probability distribution \( \mathbb{P}^o \), such that

\[
\mathbb{P}^o(\mathbf{w} = \mathbf{w}) = \begin{cases} 
0 & \text{if } \mathbf{w} = \mathbf{w}' \\
 p'' - p' & \text{if } \mathbf{w} = \mathbf{w}' \\
 p' & \text{if } \mathbf{w} = \mathbf{w}'' \\
 p'(\mathbf{w} = \mathbf{w}') + p'' & \text{if } \mathbf{w} = \mathbf{w}' \land \mathbf{w}'' \\
 p'(\mathbf{w} = \mathbf{w}') & \text{if } \mathbf{w} = \mathbf{w}' \\
 \mathbb{P}(\mathbf{w} = \mathbf{w}) & \text{otherwise.}
\end{cases}
\]

In particular, based on \( \mathbb{P} \), we move the probability mass \( p' \) from the realization of \( \mathbf{w}', \mathbf{w}'' \) to \( \mathbf{w}' \land \mathbf{w}'', \mathbf{w}' \lor \mathbf{w}'' \). 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

\[
\mathbb{E}_{\mathbb{P}^o}[f(\mathbf{w})] - \mathbb{E}_{\mathbb{P}}[f(\mathbf{w})] = p'(f(\mathbf{w} \land \mathbf{w}'') + f(\mathbf{w} \lor \mathbf{w}'') - f(\mathbf{w}') - f(\mathbf{w}'')) \geq 0,
\]

where the first inequality holds because \( p', \ldots, p'' \) is a feasible solution, the second equality is based on (32), and the second inequality follows from the convexity of \( g \). Hence, \( V(\mathbf{d}) \) is non-decreasing on \([0, \hat{\delta}]\). Therefore, \( \sup_{\mathbf{d} \in \mathcal{F}} \mathbb{E}_\mathbf{v}[g(\mathbf{x}, \mathbf{z})] = \sup_{0 \leq d \leq \hat{\delta}} V(\mathbf{d}) = V(\hat{\delta}) \). The worst-case is with the form of (28) when \( \mathbf{d} = \hat{\delta} \), which is as proposed in our Proposition.
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 \( f(\tilde{w}) \) either increases or remains unchanged. Since any \( P \in \mathcal{P} \) 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 \( P^* \in \arg \sup_{P \in \mathcal{P}} E_\mathcal{F} [f(\tilde{w})] \) such that the support of \( P^* \) 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) \( \implies \) 1). Assuming the contrary of 1), i.e., \( f \) is not supermodular, then there exists a pair of unordered \( w', w'' \in \mathbb{R}^n \) such that \( f(w') + f(w'') > f(w' \land w'') + f(w' \lor w'') \). We denote \( I' = \{ i \in [n] \mid w'_i < w''_i \} \), \( I'' = \{ i \in [n] \mid w'_i > w''_i \} \) and \( I_e = \{ i \in [n] \mid w'_i = w''_i \} \). As \( w', w'' \) are unordered, we know \( I', I'' \) are both nonempty. For all \( i \in I' \cup I'' \), we let \( m_i = 2, x_{i1} = w'_i \land w''_i, x_{i2} = w'_i \lor w''_i \); \( p_{11} = p_{12} = 0.5 \); for all \( i \in I_e \), we let \( m_i = 1, x_{i1} = w'_i = w''_i \) and \( p_{11} = 1 \). Consequently, \( \mathcal{P} = \{ P \mid P(\tilde{w}_i = x_{ij}) = p_{ij}, j \in [m_i], i \in [n] \} \), and any \( P \in \mathcal{P} \) must has its support in \( \mathcal{W} = \Pi_{i \in I' \cup I''} \{ x_{i1}, x_{i2} \} \times \Pi_{i \in I_e} \{ x_{i1} \} \). Consider any \( P^0 \in \mathcal{P} \) such that its support \( \mathcal{W}_{p^0} = \{ w \in \mathbb{R}^n \mid P^0(\tilde{w} = w) > 0 \} \) forms a chain. We now show that \( P^0 \not\in \arg \sup_{P \in \mathcal{P}} E_\mathcal{F} [f(\tilde{w})] \) and hence statement 2) in the proposition is false, then the proof can be completed. To this end, we notice since \( \mathcal{W}_{p^0} \) forms a chain, we can label the elements in \( \mathcal{W}_{p^0} \) in ascending order, i.e., \( w^1 \leq w^2 \leq \ldots \).

We first show \( w^1 = w' \land w'' \). Consider any \( i \in [n] \). If \( w^1_i < x_{i1} \), then \( w^1_i \not\in \{ x_{ij} \mid j \in [m_i] \} \), contradicts with \( P^0 \in \mathcal{P} \). If \( w^1_i > x_{i1} \), then \( w_i > x_{i1} \) for all \( w \in \mathcal{W}_{p^0}, P^0(\tilde{w}_i = x_{i1}) = 0 \), which also contradicts with \( P^0 \in \mathcal{P} \). Therefore, we must have \( w^1_i = x_{i1} \forall i \in [n] \), i.e., \( w^1 = (x_{11}, \ldots, x_{n1}) = w' \land w'' \).

We next show \( w^2 = w' \lor w'' \). Assume that \( \exists i \in I' \cup I'' \) with \( w^2_i = w^1_i \). Since \( w^2 \geq w^1 \) and \( w^2 \not\neq w^1 \), we know that \( \exists j \in I' \cup I'' \) with \( w^2_j > w^1_j \). By \( w^2_j = w^2_i = x_{i1}, P^0(\tilde{w} = w^1) + P^0(\tilde{w} = w^2) < P^0(\tilde{w}_i = x_{i1}) = p_{11} = 0.5 \), we know \( P^0(\tilde{w} = w^1) < 0.5 \); by \( w^2_j > w^1_j = x_{j1} \), we know \( P^0(\tilde{w} = w^1) = p_{11} = 0.5 \). Hence, we have contradiction. That implies that \( \forall i \in I' \cup I'' \), \( w^2_i > w^1_i \), i.e., \( w^2 = w' \lor w'' \). Moreover, we have \( |\mathcal{W}| = 2 \) since \( w^2 \) is the maximum element of \( \mathcal{W} \).

Therefore, \( P^0 \) is such that \( P^0(\tilde{w} = w' \land w'') = P^0(\tilde{w} = w' \lor w'') = 0.5 \). Consider another distribution \( P^* \) such that \( P^*(\tilde{w} = w') = P^*(\tilde{w} = w'') = 0.5 \). We can easily have \( P^* \in \mathcal{P} \), and

\[
E_{P^0} [f(\tilde{w})] = 0.5 \times (f(w' \land w'') + f(w' \lor w'')) < 0.5 \times (f(w') + f(w'')) = E_{P^*} [f(\tilde{w})].
\]
Proof of Proposition 3

For notational simplicity, we drop the superscript \( k \) which represents the scenario \( k \); we also assume \( \bar{z}_i = -1, \mu_i = 0, \bar{z}_1 = 1 \forall i \in [n] \), 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

\[
\mathbf{mp}^{j,i} = \begin{cases} 
q^j_i & \text{if } z^j_i = 1 \ (\mathbf{mp}^{j,i} \in \mathbb{R} \text{ in this case}) \\
(q^j_i, P^*(\bar{z}_i = 1)) & \text{if } z^j_i = 0 \ (\mathbf{mp}^{j,i} \in \mathbb{R}^2 \text{ in this case}) \\
(q^j_i, P^*(\bar{z}_i = 0), P^*(\bar{z}_i = 1)) & \text{if } z^j_i = -1 \ (\mathbf{mp}^{j,i} \in \mathbb{R}^3 \text{ in this case})
\end{cases}
\]

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 \( \mathcal{I}^j = \{ z^j, \mathbf{mp}^{j,1}, \ldots, \mathbf{mp}^{j,n}, c_j \} \).

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

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 \( \mathbf{1}^\top \mathbf{mp}^{1,i} = P^*(\bar{z}_i = -1) + P^*(\bar{z}_i = 0) + P^*(\bar{z}_i = 1) = 1 \) 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, \( p_j = \min \{ \mathbf{mp}^{j,i}_1, \ldots, \mathbf{mp}^{j,n}_1 \} \), \( r_j = \min \{ i \in [n] \mid \mathbf{mp}^{j,i} = p_j \} \). After that, \( \mathbf{z}^{j+1} = \mathbf{z}^j + \mathbf{e}_{r_j} \).

We now prove that \( z^j_{r_j} \neq 1 \) by contradiction. Assume to the contrary, i.e., \( z^j_{r_j} = 1 \), then \( \mathbf{mp}^{j,r_j} \in \mathbb{R} \), we have \( c_j = \mathbf{1}^\top \mathbf{mp}^{j,r_j} = \mathbf{mp}^{j,r_j}_1 = p_j \). For any \( i \in [n] \setminus \{ r_j \} \), we observe i) \( \mathbf{mp}^{j,i}_1 \geq p_j = c_j \) (the inequality is because of our choice of \( p_j \)); ii) \( \mathbf{mp}^{j,end}_i > 0 \); and iii) \( \mathbf{1}^\top \mathbf{mp}^{j,i} = c_j + \mathbf{mp}^{j,i} \geq 0 \).

The last two observations are because \( \mathcal{I}^j \) is valid and hence satisfies conditions 2), 3) and 4).

Hence, we have \( \mathbf{mp}^{j,i} \in \mathbb{R} \) and then \( z^j_i = 1 \). That implies \( \mathbf{z}^j = 1 \). We notice that for any \( t \in [j-1] \), \( \mathbf{z}^{t+1} = \mathbf{z}^t + \mathbf{e}_i \) for some \( i \in [n] \). So moving from \( \mathbf{z}^1 = -1 \) to \( \mathbf{z}^j = 1 \) requires \( 2n \) steps, i.e., \( j = 2n+1 \), which contradicts \( j \in [2n] \). Hence, \( z^j_{r_j} = 1 \) is false, and we must have \( z^j_{r_j} \in \{-1,0\} \). We can conclude that \( \mathbf{z}^{j+1} = \mathbf{z}^j + \mathbf{e}_{r_j} \in \{-1,0,1\}^n \), 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, \( \mathbf{mp}^{j+1,r_j}_i = \mathbf{mp}^{j,r_j}_i > 0 \), the condition 3) is satisfied when \( i = r_j \). Moreover, \( \mathbf{1}^\top \mathbf{mp}^{j+1,r_j} = \mathbf{1}^\top \mathbf{mp}^{j,r_j} - \mathbf{mp}^{j,r_j}_1 = c_j - p_j \). We also observe \( c_j - p_j = \mathbf{1}^\top \mathbf{mp}^{j+1,r_j} \geq \mathbf{mp}^{j,end,r_j}_i > 0 \) and hence \( c_j > p_j \).

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

\[
\mathbf{mp}^{j+1,i} = \begin{cases} 
\mathbf{mp}^{j,i} - p_j & \text{if } s = 1 \\
\mathbf{mp}^{j,i} & \text{if } z^j_{i+1} \in \{-1,0\} \text{ and } s \neq 1
\end{cases}
\] (33)
We note that if \( z_i^{j+1} = z_i^j = 1 \), then \( mp_i^{j,i}, mp_i^{j+1,i} \in \mathbb{R}_+ \), and \( mp_i^{j,i} = 1^T mp_i^{j,i} = c_j > p_j \), \( mp_i^{j+1,i} = mp_i^{j+1,i} = mp_i^{j,i} - p_j > 0 \). If \( z_i^{j+1} = z_i^j \in \{-1, 0\} \), obviously \( mp_i^{j+1,i} = mp_i^{j,i} > 0 \). Therefore, condition 3) is satisfied for \( i \). Moreover, by Equation (33) we also know \( 1^T mp_i^{j+1,i} = 1^T mp_i^{j,i} - p_j = c_j - p_j \). Since we have previously obtained \( 1^T mp_i^{j+1,i} = 1^T mp_i^{j,i} - p_j = c_j - p_j \), 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 \( Q^j \) as the set of all mass functions with the marginal mass given by \( mp^{j,1}, \ldots, mp^{j,n} \) and the possible realizations forming a chain. More specifically, define \( w^{j,i} \in \{-1, 0, 1\}^{2-\delta_j^i} \) by

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

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

\[
    Q^j = \left\{ f^j : W^j \rightarrow [0, 1] \mid \sum_{z} f^j(z) = mp^{j,i}, i \in [n], s \in [2-\delta_j^i] \right\}.
\]

Noticing that \( W^{j+1} = \{ z \in W^j \mid z_{r_j} \neq z_j^0 \} \), we define another set \( \hat{Q}^j \) by

\[
    \hat{Q}^j = \left\{ f^j : W^j \rightarrow [0, 1] \mid f^j(z) = p_j, \forall z \in W^j \text{ such that } z_{r_j} = z_j^0, z \neq z^j \right\}.
\]

We next prove \( Q^j = \hat{Q}^j \).

First, consider any \( f^j \in Q^j \). Suppose \( \exists z^o \in W^j \) with \( z^o = z_{r_j}^j \) and \( z^o \neq z^j \) such that \( f^j(z^o) > 0 \). That implies the existence of \( s \in [n] \setminus \{ r_j \} \) such that \( z^o_s > z_s^j \). Hence,

\[
    \sum_{z_{r_j}^j = z_{r_j}^o} f^j(z) = \sum_{z_{r_j}^j = z_{r_j}^o} f^j(z) - \sum_{z_{r_j}^j = z_{r_j}^o, z_s^j = z_s^o} f^j(z) = p_j - \sum_{z_{r_j}^j = z_{r_j}^o, z_s^j = z_s^o} f^j(z) < p_j,
\]

\[
    \sum_{z_{r_j}^j > z_{r_j}^o} f^j(z) = \sum_{z_{r_j}^j > z_{r_j}^o} f^j(z) - \sum_{z_{r_j}^j > z_{r_j}^o, z_s^j = z_s^o} f^j(z) \leq p_j - \sum_{z_{r_j}^j > z_{r_j}^o, z_s^j = z_s^o} f^j(z) < p_j.
\]

Therefore, we have \( z^* \in W^j \) such that \( z_{r_j}^* > z_{r_j}^o, z_s^* = z_s^o \neq z_s^j \) and \( f^j(z^*) > 0 \), contradicting that \( \{ z \mid f^j(z) > 0 \} \) forms a chain. Therefore, \( f^j(z) = 0 \forall z \in W^j \) with \( z_{r_j} = z_{r_j}^j \) and \( z \neq z^j \), and \( f^j(z^j) = mp_{r_j}^{j,r_j} - \sum_{z_{r_j} = z_{r_j}^j, z_s \neq z_s^j} f^j(z) = p_j - 0 = p_j \). Therefore, \( f^j \) satisfies the first two conditions in \( \hat{Q}^j \). The corresponding \( f_j^{j+1} \) in \( Q_j^{j+1} \) can be easily verified by showing the chain structure and checking the equality constraints on the marginal mass. Hence, we have \( f^j \in \hat{Q}^j \).

We now prove the reverse. Consider any \( f^j \in \hat{Q}^j \) and we check whether it satisfies the two conditions in \( Q^j \). 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 \( f^j \in \mathcal{Q}' \). We can conclude that \( \mathcal{Q}' = \hat{\mathcal{Q}}' \) for all \( j \in [2n+1] \).

Finally, by representing \( \mathcal{Q}' \) in the form of \( \hat{\mathcal{Q}}' \), with recursion we can easily get

\[
\mathcal{Q}' = \left\{ f : \mathcal{W}^j \rightarrow [0,1] \mid \begin{array}{l}
  f(z^i) = p_i, \quad i \in [2n] \\
  f(z) = 0, \quad \forall z \in \mathcal{W}^1 \setminus \{z^i, i \in [2n]\} \setminus \mathcal{W}_{2n+1}^1 \\
  f(z) = \hat{f}(z), \quad \forall z \in \mathcal{W}_{2n+1}^1 \\
  \hat{f} \in \mathcal{Q}_{2n+1}^i
\end{array} \right\} \tag{34}
\]

We note that since \( z^j \in \{-1,0,1\}^n, \ z^1 = -1 \), and any time the movement from \( z^j \) to \( z^{j+1} \) is to increase one dimension by 1 while maintaining other dimensions unchanged, and hence we have \( z_{2n+1} = 1 \). Therefore, \( \mathcal{W}_{2n+1} = \{z_{2n+1}\} \). Then by Equation (34), we have

\[
\mathcal{Q}' = \left\{ f : \mathcal{W}^j \rightarrow [0,1] \mid \begin{array}{l}
  f(z^i) = p_i, \quad i \in [2n+1] \\
  f(z) = 0, \quad \forall z \in \mathcal{W}^1 \setminus \{z^i, i \in [2n+1]\}
\end{array} \right\}
\]

Hence, the result is proved. \( \square \)

**Proof of Theorem 1**

We define the function \( f(x) \) as

\[
f(x) = \min_{\nu^T l} \quad \text{s.t.} \quad R_k^T l \geq \sum_{i \in [2n+1]} p_i b^k y^{k,i}, \quad k \in [K] \\
Wx + Uy^{k,i} \geq Vz^{k,i} + v^0, \quad k \in [K], i \in [2n+1] \\
l \geq 0,
\]

then Problem (7) is equivalent with \( \min_{x \in \mathcal{X}} f(x) \).

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

Consider any \( x \in \mathcal{X} \setminus \mathcal{X}_{fcea} \), we have

\[
\infty = \sup_{P \in \mathcal{P}} \mathbb{E}_P [g(x, \bar{z})] = \max \left\{ \sum_{k \in [K]} q_k \sup_{p^k \in \mathcal{P}^k} \mathbb{E}_{p^k} [g(x, \bar{z})] \mid q \in \mathcal{Q} \right\}.
\]

Since any feasible \( q \in \mathcal{Q} \subseteq \{q \in \mathbb{R}_{+}^K \mid \sum_{k \in [K]} q_k = 1\} \) is bounded, there must be \( k \in [K] \) such that

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

where the last equality follows from Proposition 3. Again, since \( p_i^k \in [0,1] \) \( \forall i \in [2n+1] \), there exists \( i \in [2n+1] \) such that \( g(x, z^{k,i}) = \infty \). It is equivalent to the infeasibility of the constraint \( Wx + Uy^{k,i} \geq Vz^{k,i} + v^0 \), which is involved in the problem defining \( f(x) \). Hence, \( f(x) = \infty \).
Therefore, Problem (7) is equivalent with \( \min_{x \in X_{\text{feas}}} f(x) \). We notice that Problem (2) is equivalent to \( \min_{x \in X_{\text{feas}}} \sup_{p \in F} \mathbb{E}_p [g(x, \hat{z})] \). Hence, for proving this theorem, now it suffices to show that \( \forall x \in X_{\text{feas}} \), we have \( \sup_{p \in F} \mathbb{E}_p [g(x, \hat{z})] = f(x) \). To this end, consider any \( x \in X_{\text{feas}} \), we then know \( \sup_{p \in F} \mathbb{E}_p [g(x, \hat{z})] \) is finite. Notice that 1) \( \sup_{p \in F} \mathbb{E}_p [g(x, \hat{z})] = \max \left\{ \sum_{k \in [K]} q_k \sup_{p^k \in F^k} \mathbb{E}_{p^k} [g(x, \hat{z})] \left| q \in Q \right. \right\} \) and 2) by the assumption on \( Q \), for any \( k \in [K] \) there exists \( q \in Q \) with \( q_k > 0 \). Hence, for all \( k \in [K] \), \( \sup_{p^k \in F^k} \mathbb{E}_{p^k} [g(x, \hat{z})] \) must be finite. It implies that \( g(x, z) \) is finite for all \( z \in [\underline{z}^k, \overline{z}^k] \). Moreover,

\[
\sup_{p \in F} \mathbb{E}_p [g(x, \hat{z})] = \max \left\{ \sum_{k \in [K]} q_k \sup_{p^k \in F^k} \mathbb{E}_{p^k} [g(x, \hat{z})] \left| Rq \leq \nu, \nu \geq 0 \right. \right\} = \min \left\{ \nu^\top l \left| R^k_l \geq \sup_{p^k \in F^k} \mathbb{E}_{p^k} [g(x, \hat{z})], k \in [K] \right. \right\} = f(x),
\]

where the second equality is due to strong duality.

Proof of Corollary 1

It has been proved in the proof for Theorem 1.

Proof of Lemma 2

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

\[
\{ \tilde{\omega}^i - \tilde{\omega}^{i+1}, \tilde{\nu}^{i+1} - \tilde{\nu}^i \} = \{ 0, e_{\kappa_i} \}
\]

for some \( \kappa_i \in [n], i \in [2n] \). This follows from that \( z^{i+1} - z^i \) has exactly one nonzero entry, the index of which is denoted as \( \kappa_i \). Specifically, for the \( \kappa_i \)-th entry where \( z^i \) moves to \( z^{i+1} \), 1) if the move is from the lower bound to the mean, then \( \tilde{\omega}^{i+1} = \tilde{\omega}^i - e_{\kappa_i}, \tilde{\nu}^{i+1} = \tilde{\nu}^i \) and hence \( \tilde{\zeta}^{i+1} = \tilde{\zeta}^i - \left[ \begin{array}{c} -e_{\kappa_i} \\ 0 \end{array} \right] \); 2) if the move is from the mean to the upper bound, then \( \tilde{\omega}^{i+1} = \tilde{\omega}^i, \tilde{\nu}^{i+1} = \tilde{\nu}^i + e_{\kappa_i} \) and hence \( \tilde{\zeta}^{i+1} = \tilde{\zeta}^i = \left[ \begin{array}{c} 0 \\ e_{\kappa_i} \end{array} \right] \). 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
This implies that the matrix $[\zeta^2 - \zeta^1 \ldots \zeta^{2n+1} - \zeta^{2n}]$ are also invertible. \hfill \Box

**Proof of Proposition 4**

We first let $V_{\text{adapt}}$ and $V_{\text{idr}}$ represent the optimal values for Problems (8) and (10), respectively. Our aim is to show that $V_{\text{adapt}} = V_{\text{idr}}$.

We first prove $V_{\text{adapt}} \leq V_{\text{idr}}$. To show this, we define a new problem by relaxing Problem (8) such that the constraints of second-stage problem apply only to the realizations $z^{k,i}, k \in [K], i \in [2n+1]$ and denote the optimal value as $V_{\text{relax}}$, i.e.,

$$V_{\text{relax}} = \min_{x} a^\top x + \sup_{\mathcal{p} \in \mathcal{F}} \mathbb{E}_{\mathcal{p}} \left[ b^\top y(\tilde{k}, \tilde{z}) \right]
\quad \text{s.t.} \quad W x + U y(k, z^{k,i}) \geq V z^{k,i} + v^0, \quad k \in [K], i \in [2n+1] \quad (35)
$$

Equivalently, we have $V_{\text{relax}} = \min_{x \in \mathcal{X}} a^\top x + \sup_{\mathcal{p} \in \mathcal{F}} \mathbb{E}_{\mathcal{p}} [g'(x, z)]$, where

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

Fixing any $x \in \mathcal{X}$, we recall that $p^{k,i}_k, z^{k,i}, k \in [K], i \in [2n+1]$ returned by Algorithm 1 gives a worst-case distribution to Problem (8), and, at the same time, is an admissible probability distribution to Problem (10) because the two problems share the same ambiguity set. It follows that

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

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

Further, we observe that Problem (10) can be directly obtained from Problem (35) by imposing a restriction of linearity structure on \( y(k, z) \). This implies any feasible \( \Theta^k, \phi^k \) to Problem (10) determines a function \( y(k, z) \) that is feasible to Problem (35). Hence, \( V_{\text{relax}} \leq V_{\text{idr}} \). We then conclude that \( V_{\text{adapt}} \leq V_{\text{relax}} \leq V_{\text{idr}} \).

We next show \( V_{\text{adapt}} \geq V_{\text{idr}} \). To this end, we construct a recourse decision rule that is feasible to Problem (10) and returns the optimal value of Problem (8).

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 \( \tilde{\omega} = (\mu - \tilde{z})^+, \tilde{\nu} = (\tilde{z} - \mu)^+, \tilde{\zeta} = (\tilde{\omega}, \tilde{\nu}) \), and let \( \omega, \nu, \zeta \) be the counterpart when \( \tilde{z} \) is realized as \( z \). Then \( \tilde{z} = \mu - \tilde{\omega} + \tilde{\nu} = \mu + [-I_{n \times n} \ I_{n \times n}] \tilde{\zeta}, \ |	ilde{z} - \mu| = \tilde{\omega} + \tilde{\nu} = [I_{n \times n} \ I_{n \times n}] \tilde{\zeta} \).

Define

\[
y_{\text{opt}}(z) = \Theta_{\text{opt}} \left[ (\mu - z)^+ \right] + \phi_{\text{opt}} = \Theta_{\text{opt}} \zeta + \phi_{\text{opt}}.
\]

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

\[
y_{\text{opt}}(z)^i = y_{\text{opt}}^{2n+1} + \Theta_{\text{opt}}(\zeta^i - \zeta^{2n+1}) = y_{\text{opt}}^{2n+1} + y_{\text{opt}}^1 \cdot y_{\text{opt}}^{2n+1} \ldots y_{\text{opt}}^2 \cdot y_{\text{opt}}^{2n+1} D^{-1}(\zeta^i - \zeta^{2n+1}) = y_{\text{opt}}^i,  
\]

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

\[
\sup_{\mathcal{F}} \mathbb{E}_{\nu} [b^\top y_{\text{opt}}(\tilde{z})] = \sum_{i \in [2n+1]} p_i b^\top y_{\text{opt}}(z)^i = \sum_{i \in [2n+1]} p_i b^\top y_{\text{opt}}^i = \sum_{i \in [2n+1]} p_i \min \{ b^\top y \mid W x_{\text{opt}} + U y \geq V z + v \} = \sup_{\mathcal{F}} \mathbb{E}_{\nu} [g(x_{\text{opt}}, \tilde{z})]. 
\]

The first and last equalities follow from the supermodularity of \( b^\top y_{\text{opt}}(z) \) and \( g(x_{\text{opt}}, z) \) defined by (1), respectively. The second equality holds since \( y_{\text{opt}}(z)^i = y_{\text{opt}}^i, i \in [2n+1] \) (as shown in (36)), while the third one follows from the definition of \( y_{\text{opt}}^i \). It follows that the worst-case expected cost returned by \( x_{\text{opt}} \), \( y_{\text{opt}}(z) \) is the same as the optimal value of Problem (8). Further, we can observe easily that the solution \( x_{\text{opt}} \), \( y_{\text{opt}}(z) \) is feasible for Problem (10).

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

\[
y_{\text{opt}}(k, z) = \Theta_{\text{opt}}^k \left[ (\mu^k - z)^+ \right] + \phi_{\text{opt}}^k,
\]
It is supermodular in \( z \) and for any realized scenario \( k \), \( \sup_{p_k \in F_k} E_{p_k} [b^\top y_{opt}(k, \tilde{z})] = \sup_{p_k \in F_k} E_{p_k} [\min \{b^\top y \mid W x_{opt} + U y \geq V z + v^0\}] \). Hence
\[
\begin{align*}
    a^\top x_{opt} + \sup_{p \in F} E_p \left[ b^\top y_{opt}(k, \tilde{z}) \right] \\
    = a^\top x_{opt} + \max_{q \in Q} \sum_{k \in [K]} q_k \sup_{p_k \in F_k} E_{p_k} [b^\top y_{opt}(k, \tilde{z})] \\
    = a^\top x_{opt} + \max_{q \in Q} \sum_{k \in [K]} q_k \sup_{p_k \in F_k} E_{p_k} [\min \{b^\top y \mid W x_{opt} + U y \geq V z + v^0\}] \\
    = a^\top x_{opt} + \sup_{p \in F} E_p [g(x_{opt}, \tilde{z})] \\
    = V_{adap}. 
\end{align*}
\]
Similar to Equation (36), we can check that \( y_{opt}(k, z^{k,i}) \in \min \{b^\top y \mid W x_{opt} + U y \geq V z + v^0\} \) for all \( k \in [K], i \in [2n + 1] \). It follows that \( x_{opt}, \Theta_{opt}^k, \phi_{opt}^k, k \in [K] \) is a feasible solution to Problem (10).

Therefore, we can conclude that \( V_{ldr} \leq V_{adap} \). Hence, we have \( V_{adap} = V_{ldr} \) and \( x_{opt}, \Theta_{opt}^k, \phi_{opt}^k, k \in [K] \) is an optimal solution to Problem (10).

**Proof of Theorem 2**

We first prove the “if” part. Suppose \( S(x) \) is an inverse additive lattice, then given any \( z', z'', p, q \) with \((p, z' \land z''), (q, z' \lor z'') \in S(x)\), there exist \( y', y'' \) such that \((y', z'), (y'', z'') \in S(x)\) and \( y' + y'' = p + q \). We then have
\[
g(x, z') + g(x, z'') \leq b^\top y' + b^\top y'' = b^\top p + b^\top q.
\]
Taking the minimum on the right-hand-side over all \( p, q \) with \((p, z' \land z''), (q, z' \lor z'') \in S(x)\), we obtain \( g(x, z') + g(x, z'') \leq g(x, z' \land z'') + g(x, z' \lor z'') \).

Next we prove the “only if” part by contradiction. Suppose \( S(x) \) is not an inverse additive lattice, then there exist \( z', z'', p, q \) with \((p, z' \land z''), (q, z' \lor z'') \in S(x)\) but \( p + q \notin W \), where the set \( W \) is defined as \( W = \{r + s \mid (r, z'), (s, z'') \in S(x)\} \). According to the definition of \( S(x) \), we can easily see that \( W \) is convex and closed. By the Hyperplane Separation Theorem, there exist a vector \( \eta \) and a real number \( \lambda \) such that,
\[
\eta^\top (p + q) < \lambda < \eta^\top w \quad \forall w \in W.
\]
Consider the second-stage cost function \( g(x, z) \) (defined in Equation (11)) with coefficient \( b = \eta \). We have
\[
g(x, z') + g(x, z'') = \min \{\eta^\top y \mid (y, z') \in S(x)\} + \min \{\eta^\top y \mid (y, z'') \in S(x)\} \\
= \min \{\eta^\top (r + s) \mid (r, z'), (s, z'') \in S(x)\} \\
= \min \{\eta^\top w \mid w \in W\} > \lambda,
\]
\[
g(x, z' \land z'') + g(x, z' \lor z'') = \min \{\eta^\top y \mid (y, z' \land z'') \in S(x)\} + \min \{\eta^\top y \mid (y, z' \lor z'') \in S(x)\} \\
\leq \eta^\top (p + q) < \lambda.
\]
Therefore, \( g(x, z') + g(x, z'') > g(x, z' \land z'') + g(x, z' \lor z'') \), which contradicts the supermodularity. The “only if” part is completed.

**Proof of Theorem 3**

Based on Theorem 2, the above theorem is equivalent to this statement: \( S(x) \) is an additive inverse lattice for all \( x \) and \( v^0 \) if and only if \( U \) and \( V \) 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 \( x \) and \( v^0 \) such that \( S(x) \) is not an additive inverse lattice, i.e., we have \( z', z'', p, q \) with \( z^\wedge = z' \land z'' \), \( z^\lor = z' \lor z'' \) and \((p, z^\wedge), (q, z^\lor) \in S(x)\), such that \( y' + y'' \neq p + q \) holds for all \( y', y'' \) with \((y', z'), (y'', z'') \in S(x)\).

We denote \( c = -Wx + v^0 \), \( t^1 = Up - Vz^\wedge \geq c \), \( t^2 = Uq - Vz^\lor \geq c \). Here the two inequalities are due to \((p, z^\wedge), (q, z^\lor) \in S(x)\) and the definition of \( S(x) \). We define a set \( W \) as

\[
W = \{ y \in \mathbb{R}^m \mid (t^1 \land t^2) + Vz' \leq Uy \leq (t^1 \lor t^2) + Vz' \}.
\]

Note that \( W \) should be an empty set, otherwise there exists a \( y^0 \in W \) and hence

\[
Uy^0 - Vz' \geq (t^1 \land t^2) \geq c,
\]

\[
U(p + q - y^0) - Vz'' = Up - Vz^\wedge + Uq - Vz^\lor - (Uy^0 - Vz') \geq t^1 + t^2 - (t^1 \lor t^2) = t^1 \land t^2 \geq c,
\]

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

We now show that the first part of the condition in our theorem is not true. If \( \text{rank}(U) = r \), we can solve \( y \) with \( Uy = (t^1 \land t^2) + Vz' \leq (t^1 \lor t^2) + Vz' \), which contradicts the emptiness of \( W \). Therefore, \( \text{rank}(U) < r \).

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

\[
\begin{align*}
\text{max} & \quad 0 \\
\text{s.t.} & \quad [U \quad -U] \begin{bmatrix} y \\ 0 \end{bmatrix} \leq \begin{bmatrix} (t^1 \lor t^2) + Vz' \\ -(t^1 \land t^2) - Vz' \end{bmatrix}.
\end{align*}
\]

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

\[
\begin{align*}
\text{max} & \quad 0 \\
\text{s.t.} & \quad [U^{\mathcal{I}} \quad -U^{\mathcal{I}}] \begin{bmatrix} y \\ 0 \end{bmatrix} \leq \begin{bmatrix} (t^1_{\mathcal{I}} \lor t^2_{\mathcal{I}}) + V_{\mathcal{I}}z' \\ -(t^1_{\mathcal{I}} \land t^2_{\mathcal{I}}) - V_{\mathcal{I}}z' \end{bmatrix}
\end{align*}
\]

is also infeasible. We write the dual of (38) as follows,

\[
\begin{align*}
\min & \quad r^\top ((t^1_{\mathcal{I}} \lor t^2_{\mathcal{I}}) + V_{\mathcal{I}}z') - s^\top ((t^1_{\mathcal{I}} \land t^2_{\mathcal{I}}) + V_{\mathcal{I}}z') \\
\text{s.t.} & \quad U^{\mathcal{I}}(r - s) = 0 \\
& \quad r, s \geq 0.
\end{align*}
\]
Observing that \( r = s = 0 \) gives a feasible solution of (39), the infeasibility of the primal problem implies the unboundedness of the above dual problem. Therefore, there exist \( r, s \geq 0 \) with \( U_T^T (r - s) = 0 \) such that the following inequalities holds,

\[
0 > r^T ((t^1_T \lor t^2_T) + V_T z') - s^T ((t^1_T \land t^2_T) + V_T z')
\]

\[
= r^T ((t^1_T \lor t^2_T) + V_T z' - U_T q) - s^T ((t^1_T \land t^2_T) + V_T z' - U_T q)
\]

\[
\geq r^T (t^1_T + V_T z' - U_T q) - s^T (t^2_T + V_T z' - U_T q)
\]

\[
= (r - s)^T V_T (z' - z^\lor),
\]

where the first inequality is obtained from the unboundedness of (39), the first equality is due to \( U_T^T (r - s) = 0 \), the second inequality follows from \( t^1_T \land t^2_T \leq t^1_T \lor t^2_T \), and the second equality comes from \( t^2_T = U_T q - V_T z^\lor \). We remark that in the above equation, if we use \( U_T p \) instead of \( U_T q \) in the first equality, and \( t^1_T \) instead of \( t^2_T \) in the second inequality, then \( 0 > (r - s)^T V_T (z' - z^\lor) \) can be obtained similarly.

We define \( \Delta_1 = (r - s)^T V_T (z' - z^\lor) \), \( \Delta_2 = (r - s)^T V_T (z' - z^\land) \), and

\[
\beta = \frac{z' - z^\lor}{\Delta_1} - \frac{z' - z^\land}{\Delta_2}.
\]

We have three observations on \( \beta \). First, \( \beta \geq 0 \) since \( \Delta_1, \Delta_2 < 0 \) and \( z^\land \leq z' \leq z^\lor \).

Second, \( V_T \beta \in \text{span}(U_T) \). 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 \( \text{null}(U_T^T) \subseteq \text{null}((V_T \beta)^T) \), where \( \text{null}(\cdot) \) is the null space of a given matrix. Since \( |I| = \text{rank}(U) + 1 = \text{rank}(U_T) + 1 = \text{rank}(U_T^T) + 1 \), \( \text{null}(U_T^T) \) is of dimension 1. That implies \( \forall w \in \text{null}(U_T^T) \), we have \( w = k(r - s) \) for some \( k \in \mathbb{R} \). Therefore,

\[
(V_T \beta)^T w = k(r - s)^T V_T \beta = k \left( \frac{(r - s)^T V_T (z' - z^\lor)}{\Delta_1} - \frac{(r - s)^T V_T (z' - z^\land)}{\Delta_2} \right) = k(1 - 1) = 0.
\]

That is, \( w \in \text{null}((V_T \beta)^T) \). Hence, \( \text{null}(U_T^T) \subseteq \text{null}((V_T \beta)^T) \) and then we have \( V_T \beta \in \text{span}(U_T) \).

The third observation is that there exists some \( i \in [n] \) such that \((V_T)_{i \beta} \notin \text{span}(U_T)\). To show this, we denote \( \mathcal{H} = \{i \in [n] \mid z_i' \leq z_i''\} \). We then have for every \( i \in \mathcal{H} \), \( z_i' = z_i' \), \( z_i'' = z_i'' \) and hence \( \beta_i = \frac{z_i' - z_i''}{\Delta_1} \). In addition, since for every \( i \in [n] \setminus \mathcal{H} \), \( z_i' > z_i'' \), \( \frac{z_i' - z_i''}{\Delta_1} = \frac{z_i' - z_i''}{\Delta_1} = 0 \), we have

\[
(r - s)^T \sum_{i \in \mathcal{H}} \beta_i (V_T)_{i} = (r - s)^T \left( \sum_{i \in \mathcal{H}} \beta_i (V_T)_{i} + \sum_{i \in [n] \setminus \mathcal{H}} 0 \cdot (V_T)_{i} \right) = (r - s)^T V_T z' - z'^\lor = 1.
\]

Hence, \( (r - s) \notin \text{null} \left( \sum_{i \in \mathcal{H}} \beta_i (V_T)_{i} \right) \), which implies that \( \text{null}(U_T^T) \) is not a subset of \( \text{null} \left( \sum_{i \in \mathcal{H}} \beta_i (V_T)_{i} \right) \). Consequently we have \( \sum_{i \in \mathcal{H}} \beta_i (V_T)_{i} \notin \text{span}(U_T) \), implying that there exists some \( i \in \mathcal{H} \) such that \((V_T)_{i \beta} \notin \text{span}(U_T)\).

With the three observations, we have a contradiction of the second condition in Theorem 3.

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

We arbitrarily choose $z^\top \in \mathbb{R}^n, p \in \mathbb{R}^m$ and let $z^\top = z^\top + \beta \geq z^\top, q = p + \alpha$, then $U_Ip - V_Iz^\top = U_I(q - \alpha) - V_I(z^\top - \beta) = U_Iq - V_Iz^\top$. We also arbitrarily choose $x$, and then choose $v^0$ such that $c = -Wx + v^0$ is with $c_I = U_Ip - V_Iz^\top$ and $c_j$ being sufficiently small for all $j \notin I$. Then we have $(p, z^\top), (q, z^\top) \in S(x)$. We further define $z' = z^\top + \beta, z'' = z^\top - \beta, e_i$ so that $z' \land z'' = z^\top, z' \lor z'' = z^\top$. Then we have

$$c_I + V_Iz' = c_I + V_I(z^\top + \beta, e_i) = U_Ip - V_Iz^\top + V_I(z^\top + \beta, e_i) = U_Ip + \beta_i(V_I), \notin \text{span}(U_I),$$

where the last relationship holds since $U_Ip \in \text{span}(U_I)$ but $\beta_i(V_I), \notin \text{span}(U_I)$.

Hence, $\{y \in \mathbb{R}^m \mid U_Iy = c_I + V_Iz'\} = \emptyset$, i.e. for any $y'$ satisfying $U_Iy' - V_Iz' \geq c$, there exists $j \in I$ such that $u_j^\top y' - v_j^\top z' > c_j$. If there exists some $y''$ with $U_Iy'' - V_Iz'' \geq c$ satisfies $y' + y'' = p + q$,

$$U_Iy'' - V_Iz'' = U_I(p + q - y'') - V_I(z^\top + z'' - z')$$

$$= U_Ip - V_Iz^\top + U_Iq - V_Iz^\top - (U_Iy' - V_Iz')$$

$$= 2c_I - (U_Iy' - V_Iz'),$$

then we should have $2c_j - (u_j^\top y' - v_j^\top z') < c_j$ for the above mentioned $j$, which contradicts the assumption $(y'', z'') \in S(x)$. Hence we prove the necessity of the conditions on $U, V$.

**Proof of Theorem 4**

The case for $\text{rank}(U) = r$ is straightforward, so we only consider the case where $\text{rank}(U) < r$. In that case, we only need to verify whether $U, V$ satisfy the second part of the condition in Theorem 3, which depends solely on the relationship between $V$ and $\text{span}(U)$. Thus, removing the dependent columns in $U$ 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 $U$ 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 $I \subseteq [r]$ and indices $i \in [r] \setminus I, a, b \in [r_0]$ with $|I| = r_0, U_I$ invertible and $d_{ia}d_{ib} < 0$. WLOG, we let $d_{ia} > 0, d_{ib} < 0$.

Denote $\beta = \begin{bmatrix} e_a \end{bmatrix}_{d_{ia}} - \begin{bmatrix} e_b \end{bmatrix}_{d_{ib}} \geq 0, \alpha = U_I^{-1}\left( \begin{bmatrix} V_I \end{bmatrix}_{d_{ia}} - \begin{bmatrix} V_I \end{bmatrix}_{d_{ib}} \right)$, then

$$\begin{bmatrix} V_I \\ v_i^\top \end{bmatrix} \beta = \begin{bmatrix} \begin{bmatrix} V_I \end{bmatrix}_{d_{ia}} - \begin{bmatrix} V_I \end{bmatrix}_{d_{ib}} \\ v_i^\top \end{bmatrix} \end{bmatrix}_{d_{ia}} - \begin{bmatrix} \begin{bmatrix} V_I \end{bmatrix}_{d_{ia}} - \begin{bmatrix} V_I \end{bmatrix}_{d_{ib}} \\ v_i^\top \end{bmatrix} \end{bmatrix}_{d_{ib}} = \begin{bmatrix} \begin{bmatrix} V_I \end{bmatrix}_{d_{ia}} + u_i^\top U_I^{-1}(V_I) \\ \begin{bmatrix} V_I \end{bmatrix}_{d_{ib}} + u_i^\top U_I^{-1}(V_I) \end{bmatrix} \end{bmatrix} = U_I^\top \begin{bmatrix} \begin{bmatrix} V_I \end{bmatrix}_{d_{ia}} - \begin{bmatrix} V_I \end{bmatrix}_{d_{ib}} \\ v_i^\top \end{bmatrix} \end{bmatrix},$$
We let $\tilde{I} = \mathcal{I} \cup \{i\}$. The above equality implies $V_{\tilde{I}} \beta = U_{\tilde{I}} \alpha \in \text{span}(U_{\tilde{I}})$. On the other hand, for $\beta_a(V_{\tilde{I}}) a$ we have

$$
\beta_a \left[ \begin{array}{c} (V_{\tilde{I}}) a \\ v_{ia} \end{array} \right] = \left[ \begin{array}{c} (V_{\tilde{I}}) a \\ d_{ia} + u_i U_{\tilde{I}}^{-1}(V_{\tilde{I}}) a \end{array} \right] = \left[ \begin{array}{c} U_{\tilde{I}} \\ u_i \end{array} \right] U_{\tilde{I}}^{-1}(V_{\tilde{I}}) a + \left[ \begin{array}{c} 0 \\ 1 \end{array} \right] = U_{\tilde{I}} U_{\tilde{I}}^{-1}(V_{\tilde{I}}) a + \left[ \begin{array}{c} 0 \\ 1 \end{array} \right].
$$

Since $U_{\tilde{I}}$ is invertible, there is no $\gamma \in \mathbb{R}^n$ such that $U_{\tilde{I}} \gamma = \left[ \begin{array}{c} U_{\tilde{I}} \\ u_i \end{array} \right] \gamma = \left[ \begin{array}{c} 0 \\ 1 \end{array} \right]$. Hence $\beta_a(V_{\tilde{I}}) a \notin \text{span}(U_{\tilde{I}})$ and the second part of the condition in Theorem 3 is violated.

We now investigate the case where the second part of the condition in Theorem 3 is violated. That means, there exist $\hat{I} \subseteq [r]$, $\beta \geq 0$ and $a \in [r_0]$ such that $|\hat{I}| = r_0 + 1$, rank($U_{\hat{I}}$) = $r_0$, $V_{\hat{I}} \beta \notin \text{span}(U_{\hat{I}})$ but $\beta_a(V_{\tilde{I}}) a \notin \text{span}(U_{\tilde{I}})$. We choose $I \subseteq \hat{I}$ such that $|I| = r_0$ and $U_{\tilde{I}}$ is invertible, and denote $i$ as the unique index in $\hat{I} \setminus I$. It follows that

$$
V_{\tilde{I}} \beta = \left[ \begin{array}{c} V_{\tilde{I}} \\ v_i \end{array} \right] \beta = \left[ \begin{array}{c} v_i^\top - u_i^\top U_{\tilde{I}}^{-1} V_{\tilde{I}} \\ 0 \end{array} \right] \beta + \left[ \begin{array}{c} U_{\tilde{I}} \\ u_i \end{array} \right] U_{\tilde{I}}^{-1} V_{\tilde{I}} \beta = \left[ \begin{array}{c} 0 \\ d_i^\top \end{array} \right] \beta + U_{\tilde{I}} U_{\tilde{I}}^{-1} V_{\tilde{I}} \beta,
$$

$$
\beta_a(V_{\tilde{I}}) = \beta_a \left[ \begin{array}{c} (V_{\tilde{I}}) a \\ v_{ia} \end{array} \right] = \beta_a \left[ \begin{array}{c} v_{ia} - u_i U_{\tilde{I}}^{-1}(V_{\tilde{I}}) a \\ d_{ia} \end{array} \right] + \left[ \begin{array}{c} U_{\tilde{I}} \\ u_i \end{array} \right] U_{\tilde{I}}^{-1}(V_{\tilde{I}}) a = \left[ \begin{array}{c} 0 \\ \beta_a d_{ia} \end{array} \right] + \beta_a U_{\tilde{I}} U_{\tilde{I}}^{-1}(V_{\tilde{I}}) a.
$$

Since $V_{\tilde{I}} \beta, U_{\tilde{I}} U_{\tilde{I}}^{-1} V_{\tilde{I}} \beta, \beta_a U_{\tilde{I}} U_{\tilde{I}}^{-1}(V_{\tilde{I}}) a \in \text{span}(U_{\tilde{I}})$ and $\beta_a(V_{\tilde{I}}) a \notin \text{span}(U_{\tilde{I}})$, the above equations imply $\left[ \begin{array}{c} 0 \\ d_i^\top \end{array} \right] \beta \in \text{span}(U_{\tilde{I}})$ and $\left[ \begin{array}{c} 0 \\ \beta_a d_{ia} \end{array} \right] \notin \text{span}(U_{\tilde{I}})$. According to $\left[ \begin{array}{c} 0 \\ d_i^\top \end{array} \right] \beta \in \text{span}(U_{\tilde{I}})$, there exists $\alpha \in \mathbb{R}^n$ with $U_{\tilde{I}} \alpha = 0, u_i^\top \alpha = d_i^\top \beta$. Since $U_{\tilde{I}}$ is invertible, $\alpha = 0$ and hence $d_i^\top \beta = u_i^\top \alpha = 0$.

According to $\left[ \begin{array}{c} 0 \\ \beta_a d_{ia} \end{array} \right] \notin \text{span}(U_{\tilde{I}})$, we obtain $\beta_a d_{ia} \neq 0$. As $\beta \geq 0$, $\beta_a d_{ia} \neq 0$ and $d_i^\top \beta = 0$, we must have an index $b \in [r_0]$ such that $d_{ib}, d_{ia}$ are of different signs. Hence the algorithm returns $s = 0$.

Proof of Proposition 5

“$\Leftarrow$” Assume there exists a $2 \times 3$ submatrix of $U^\circ$ such that any pair of columns in it are linearly independent. WLOG, let $U^\circ_{(1,2),(1,3)} \in \mathbb{R}^{2\times3}$ be such matrix and we denote it by $A = [A_1 A_2 A_3] \in \mathbb{R}^{2\times3}$. WLOG, assume $A_3 = t_1 A_1 + t_2 A_2$ with $t_1, t_2 > 0$. Choose $V^1 = I_{m \times m}, \mathcal{I} = [m + 2 \setminus \{3\}]$, $\beta = t_1 e_1 + t_2 e_2 \geq 0, \alpha = t_1 e_1 + t_2 e_2 - e_3$. We then have $V_{\mathcal{I}} \beta = t_1 e_1 + t_2 e_2$; at the same time, $U_{\mathcal{I}} \alpha = t_1 e_1 + t_2 e_2$ since $A_3 = t_1 A_1 + t_2 A_2$. Hence, $V_{\mathcal{I}} \beta = U_{\mathcal{I}} \alpha \in \text{span}(U_{\mathcal{I}})$. However, $(V_{\mathcal{I}})_{11} \beta_1 = \beta_1 e_1 \notin \text{span}(U_{\mathcal{I}})$. Therefore, the second condition in Theorem 3 is violated, there exists an instance of $g(x, z)$ which is not supermodular in $z$.

“$\Rightarrow$” Assume that every $2 \times 3$ submatrix of $U^\circ$ contains at least one pair of column vectors which are linearly dependent. We prove the result by showing the second condition in Theorem 3 is always satisfied. To see this, consider any $\mathcal{I} \subseteq [r]$ such that $|\mathcal{I}| = m + 1$, rank($U_{\mathcal{I}}$) = $m$. 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 $U_{\mathcal{I}_1}$ is extracted
from $I_{m \times m}$ and $U_0$ is from $U^\circ$. We further let $J_1, J_0$ be a partition of $[m]$ such that $U_{I_1, J_1}$ contains all nonzero columns in $U_{I_1}$ and hence $U_{I_1, J_0} = 0$. Noting that $U_{I_1}$ contains rows extracted from $I_{m \times m}$, we know $I_1 = J_1$. Hence, $|I_0| = m \pm 1 - |J_1| = m \pm 1 - (m - |J_0|) = |J_0| + 1$. We illustrate the partition of $U_I$ as follows,

$$U_I = \begin{bmatrix} U_{I_1} \\ U_{I_0} \end{bmatrix} = \begin{bmatrix} I_{|J_1| \times |J_1|} & 0_{|J_1| \times |J_0|} \\ U_{I_0, J_1} & U_{I_0, J_0} \end{bmatrix}.$$

We first prove that there exists a unit vector $p \in \mathbb{R}^{|J_0|}$, such that it is orthogonal to $\text{span}(U_{I_0, J_0})$ and $\text{span}(U_{I_0, J_0}, p) = \mathbb{R}^{|J_0|}$. Notice that $U_I$ is of full column rank and hence so does its submatrix $U_{I, J_0} = \begin{bmatrix} 0_{|J_1| \times |J_0|} \\ U_{I_0, J_0} \end{bmatrix}$, which implies $U_{I_0, J_0} \in \mathbb{R}^{|J_0| \times |J_0|}$ is also of full column rank. Therefore, $\text{span}(U_{I_0, J_0})$ is of dimension $|J_0| = |I_0| - 1$, the existence of $p$ can be proved.

We now show that the orthogonal unit vector $p$ can be chosen such that $\forall i \in J_1$, $(U_{I_0})_i = s_i + \gamma_i p$ for some $s_i \in \text{span}(U_{I_0, J_0})$ and $\gamma_i \geq 0$. For those $i \in J_1$ with $(U_{I_0})_i \in \text{span}(U_{I_0, J_0})$, we always have $\gamma_i = 0$ regardless of the choice of orthogonal vector $p$. Now we consider any given $j \in J_1$ with $(U_{I_0})_j \notin \text{span}(U_{I_0, J_0})$. Since $\text{span}(U_{I_0, J_0}, p) = \mathbb{R}^{|J_0|}$, we can surely represent $(U_{I_0})_j = s_j + \gamma_j p$ for some $s_j \in \text{span}(U_{I_0, J_0})$ and $\gamma_j \neq 0$. Moreover, the unit vector $p$ can be chosen (as $-p$, if necessary) to make $\gamma_j > 0$. Consider any $k \in J_1 \setminus \{j\}$ with $(U_{I_0})_k \notin \text{span}(U_{I_0, J_0})$. Denote $Q = \begin{bmatrix} (U_{I_0})_j \\ (U_{I_0})_k \\ U_{I_0, J_0} \end{bmatrix}$. Notice that every $2 \times 3$ submatrix of $U^\circ$, and hence that of $Q$, contains at least one pair of column vectors which are linearly dependent. By Lemma [5] there are at least one pair of columns in $Q$ which are linearly dependent. Since $U_{I_0, J_0}$ is of full column rank and $(U_{I_0, J_1}, (U_{I_0})_k \notin \text{span}(U_{I_0, J_0})$, the two linearly dependent columns can only be $(U_{I_0})_j, (U_{I_0})_k$, i.e., $(U_{I_0})_k = \zeta(U_{I_0})_j$, for some $\zeta \neq 0$ (recall that both $(U_{I_0})_k$ and $(U_{I_0})_j$ are nonzero vector since they are not in $\text{span}(U_{I_0, J_0})$). As all components in the same row of $U$ are with the same sign, we know $\zeta > 0$. Therefore, $(U_{I_0})_k = \zeta (s_j + \gamma_j p) = \zeta s_j + \zeta \gamma_j p$ where $\gamma_j > 0$.

We are now ready to prove the second condition in Theorem [3] holds. Consider any $\beta \geq 0$ and $\alpha$ such that $V_I \beta = U_I \alpha$. Observing the first block, characterized by $I_1$, we have $V_{I_1} \beta = U_{I_1} \alpha = (I_{m \times m})_{I_1} \alpha = \alpha_{I_1}$; since $V_I \beta$ are both nonnegative, we have $\alpha_{I_1} \geq 0$. Observing the second block, characterized by $I_0$, by $V_{I_0} = 0$, we have

$$0 = V_{I_0} \beta = U_{I_0} \alpha = [U_{I_0, J_1} U_{I_0, J_0}] \begin{bmatrix} \alpha_{J_1} \\ \alpha_{J_0} \end{bmatrix} = U_{I_0, J_1} \alpha_{I_1} + U_{I_0, J_0} \alpha_{I_0} = s + p \sum_{i \in N} \alpha_i \gamma_i \quad (40)$$

for some $s \in \text{span}(U_{I_0, J_0})$. Here we denote the index set $N = \{i \in J_1 \mid (U_{I_0})_i \notin \text{span}(U_{I_0, J_0})\}$ and hence the last equality above holds due to the argument proved in the last paragraph. Moreover, since $s$ and $p$ are orthogonal, by (40) we have $\sum_{i \in N} \alpha_i \gamma_i = 0$, which implies $\alpha_i = 0 \forall i \in N$, as we have already known $\gamma_i > 0, \alpha_i \geq 0 \forall i \in N$ (recall that $N \subseteq I_1 = I_1$, and $\alpha_{I_1} \geq 0$). Therefore, $\forall i \in N, 0 = \alpha_i = u_i^T \alpha = v_i^T \beta$, where the last equality is due to $V_I \beta = U_I \alpha$. As $V_I \beta \geq 0, \forall i \in N, v_i^T \beta = 0$
implies $v_k \beta_k = 0 \ \forall k \in [m]$. We now consider any $j \in [m]$ and it remains to show $(V_I)_j \beta_j = U_I \eta$ for some $\eta \in \mathbb{R}^m$. To this end, we choose $\eta \in \mathbb{R}^m$ with $\eta_i = v_j \beta_j \ \forall i \in J_1 = I_1$ and we determine $\eta_{J_0}$ later. Then $\forall i \in I_1$, we have $u_i^\top \eta = \eta_i = v_j \beta_j$. We additionally observe that $\forall i \in \mathcal{N}$, $\eta_i = 0$ since $v_j \beta_j = 0$. We now move on to $I_0$, and have

$$U_{I_0} \eta = [U_{I_0, J_1} \ U_{I_0, J_0}] \begin{bmatrix} \eta_{J_1} \\ \eta_{J_0} \end{bmatrix} = U_{I_0, J_1} \eta_{J_1} + U_{I_0, J_0} \eta_{J_0} = \sum_{i \in J_1} s_i \eta_i + U_{I_0, J_0} \eta_{J_0},$$

where the last equality is due to that when $i \in \mathcal{N}$, $\eta_i = 0$ and when $j \in J_1 \ \forall \eta = (U_{I_0})_j = s_j + \gamma_j \ u$ with $\gamma_j = 0$. Since $s_i \in \text{span}(U_{I_0, J_0})$, we can choose $\eta_{J_0}$ such that $\sum_{i \in J_1 \ \forall \eta} s_i \eta_i + U_{I_0, J_0} \eta_{J_0} = 0$. In this case, $U_{I_0} \eta = 0 = (V_{I_0})_j \beta_j$. Hence, we conclude $(V_{I})_j \beta_j = U_{I} \eta \in \text{span}(U_{I}).$ \hfill $\Box$

**Proof of Theorem 5**

We first reformulate the second-stage problem as

$$g^W(x, z) = \min \ b^\top y$$

$$\text{s.t. } Uy - (V - [W^1 x \cdots W^nx]) z \geq -W^0 x + v^0,$$

where $[W^1 x \cdots W^nx]$ stands for an $r \times n$ matrix with its $i$-th column being $W^i x$. We denote $\bar{V}^x = V - [W^1 x \cdots W^nx]$ for convenience. Following Theorem 3, it suffices to show that the proposed conditions hold if and only if $U, \bar{V}^x$ satisfy the conditions in Theorem 3 for any $x$. The case of $\text{rank}(U) = r$ is straightforward. Hence, in the rest of the prove, we only focus on the case of $\text{rank}(U) < r$, i.e., the second condition in this theorem and that in Theorem 3 which are called Condition 2) and Condition 2) throughout this proof. In particular, Condition 2) can be stated as

2) for all $I \subseteq [r]$, $\beta \in \mathbb{R}^r$, $x \in \mathbb{R}^d$ with $|I| = \text{rank}(U) + 1$, $\text{rank}(U_I) = \text{rank}(U)$ and $\bar{V}^x_I \beta \in \text{span}(U_I)$, we must have $\beta_i (\bar{V}^x_I), \in \text{span}(U_I)$ holds for every $i \in [n]$.

We now prove that Condition 2) is equivalent to Condition 2).

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

$$(V_I)_i = \xi_i + \lambda_i \eta, \ \ W^i x = \xi_i^x + \mu_i^x \eta,$$

$$(V_I)_j = \xi_j + \lambda_j \eta, \ \ W^j x = \xi_j^x + \mu_j^x \eta.$$  \hfill (41)

for some $\lambda_i, \lambda_j, \mu_i^x, \mu_j^x \in \mathbb{R}$ and $\xi_i, \xi_j, \xi_i^x, \xi_j^x \in \text{span}(U_I)$. Since $\eta$ is a unit vector, we have

$$0 = \eta^\top ((V_I)_i - \eta \bar{V}^x_I (V_I)_i) = \eta^\top (V_I)_i - \eta^\top \eta \eta^\top (\xi_i + \lambda_i \eta) = \eta^\top (V_I)_i - \lambda_i,$$

and hence $\lambda_i = \eta^\top (V_I)_i$. The same logic applies to $(V_I)_j$ and $W^i x, W^j x.$
In Condition 2), we notice that \((\eta^\top (V_Z)_i) \cdot (\eta^\top (V_Z)_j) \geq 0\) is equivalent to \(\lambda_i \lambda_j \geq 0\); moreover, 
\((W_z^i)^\top \eta \eta^\top W_z^j\) is positive semidefinite if and only if 
\((\eta^\top W_z^i x) \cdot (\eta^\top W_z^j x) \geq 0\) for all \(x\). Hence, we conclude that Condition 2a) is equivalent to “\(\lambda_i \lambda_j \geq 0, \mu^i_j \mu^j_i \geq 0\) for all \(x\) and \(i,j \in [n]\)”.

For Condition 2b), since the equality holds if and only if 
\((\eta^\top (V_Z)_i) \cdot (\eta^\top W_z^j x) = (\eta^\top (V_Z)_j) \cdot (\eta^\top W_z^i x)\) for all \(x\), we conclude that it is equivalent to the condition “\(\lambda_i \mu^x_j = \lambda_j \mu^x_i\) for all \(x\) and \(i,j \in [n]\)”.

For Condition 2), by the definition of \(V^x_z\),
\[
(V^x_z)_i = (V_Z)_i - W_z^i x = (\xi_i - \xi^x_i) + (\lambda_i - \mu^x_i) \eta,
\]
\[
(V^x_z)_j = (V_Z)_j - W_z^j x = (\xi_j - \xi^x_j) + (\lambda_j - \mu^x_j) \eta.
\]

Observing that Condition 2) is violated if and only if \(\exists i,j \in [n]\) with \((\lambda_i - \mu^x_i)(\lambda_j - \mu^x_j) < 0\), we obtain its equivalent condition as
\[
(\lambda_i - \mu^x_i)(\lambda_j - \mu^x_j) \geq 0 \quad \forall x \in \mathbb{R}^I, i,j \in [n].
\]

We now prove the direction “Condition 2) \implies Condition 2)”.

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 choose \(\beta = -\lambda_j e_1 + \lambda_i e_j \in \mathbb{R}^n_+\), and have \(V_z \beta = -\lambda_j \xi_i + \lambda_i \xi_j \in \text{span}(U_Z)\). However, if \(\beta_j (V_Z)_j = \lambda_i \xi_j + \lambda_j \eta \notin \text{span}(U_Z)\), we have contradiction with Condition 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 \(\theta \in \mathbb{R}\), 
\[
(V^\theta z)_i = (\xi_i - \theta \xi^x_i) + (\lambda_i - \theta \mu^x_i) \eta, (V^\theta z)_j = (\xi_j - \theta \xi^x_j) + (\lambda_j - \theta \mu^x_j) \eta.
\]
If \(\mu^x_i \mu^x_j < 0\), we can always find \(\theta\) such that 
\((\lambda_i - \theta \mu^x_i)(\lambda_j - \theta \mu^x_j) = \mu_i^x \mu_j^x \theta^2 - (\lambda_i \mu_j^x + \mu^x_i \lambda_j) \theta + \lambda_i \lambda_j < 0\). Therefore, the equivalent condition for Condition 2), i.e., (42), is violated for \(\theta x\). Hence, we have contradiction, and conclude that \(\mu^x_i \mu^x_j \geq 0\), the second argument of Condition 2a) is true.

We now prove Condition 2b). By Condition 2a), we already have \(\lambda_i \lambda_j \geq 0, \mu^x_i \mu^x_j \geq 0\). WLOG, we assume \(\lambda_i, \lambda_j, \mu^x_i, \mu^x_j \geq 0\). Assume the opposite to Condition 2b), i.e., \(\lambda_i \mu^x_j \neq \lambda_j \mu^x_i\). WLOG, we let 
\(0 \leq \lambda_i \mu^x_j < \lambda_j \mu^x_i\), which implies \(\lambda_j, \mu^x_j > 0\). By Condition 2), i.e., (42), we have 
\((\lambda_i - \mu^x_i)(\lambda_j - \mu^x_j) \geq 0\). Combining with \(\lambda_j, \mu^x_j > 0\), we know that at least one of \(\lambda_i, \mu^x_i\) is nonzero. Consider the case that \(\mu^x_i \neq 0\). Define \(\theta_i = \lambda_i / \mu^x_i, \theta_j = \lambda_j / \mu^x_j\), then following the assumptions of \(\lambda_i \mu^x_j < \lambda_j \mu^x_i\) we have \(\theta_i < \theta_j\). Choosing any \(\theta \in (\theta_i, \theta_j)\), we have \(\lambda_i < \theta \mu^x_i, \lambda_j > \theta \mu^x_j\). Hence, the equivalent condition for Condition 2), i.e., (42), is violated for \(\theta x\). 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 2)”.

Given any \(x \in \mathbb{R}^I\), we let \(\lambda_i, \lambda_j, \mu^x_i, \mu^x_j \in \mathbb{R}\) and \(\xi_i, \xi_j, \xi^x_i, \xi^x_j \in \text{span}(U_Z)\) be constants as defined in (41). By Condition 2), we know \(\lambda_i \lambda_j \geq 0, \mu^x_i \mu^x_j \geq 0\) and \(\lambda_i \mu^x_j = \lambda_j \mu^x_i\). WLOG, we let \(\lambda_i, \lambda_j \geq 0\). Possible realizations of the parameters are as follows.
• $\lambda_i = \lambda_j = 0$. Then either $\mu_i^\pi, \mu_j^\pi \geq 0$ or $\mu_i^\pi, \mu_j^\pi \leq 0$, it always implies $(\lambda_i - \mu_i^\pi)(\lambda_j - \mu_j^\pi) \geq 0$.
• $\lambda_i = 0, \lambda_j > 0$ (or $\lambda_i > 0, \lambda_j = 0$). Then $\lambda_i \mu_i^\pi = \lambda_j \mu_j^\pi = 0$, implying $\mu_i^\pi = 0$ (or $\mu_j^\pi = 0$). In either case we have $(\lambda_i - \mu_i^\pi)(\lambda_j - \mu_j^\pi) \geq 0$.
• $\lambda_i > 0, \lambda_j > 0$. Denote $\theta = \lambda_i / \mu_i^\pi = \lambda_j / \mu_j^\pi$, then $(\lambda_i - \mu_i^\pi)(\lambda_j - \mu_j^\pi) = \mu_i^\pi \mu_j^\pi (\theta - 1)^2 \geq 0$.

So we always have $(\lambda_i - \mu_i^\pi)(\lambda_j - \mu_j^\pi) \geq 0$, Condition 2) holds. □

**Proof of Lemma 3**

Consider any $z', z'' \in \mathbb{R}^n$, we denote $a = h(z' \land z''), b = h(z'), c = h(z'')$, $d = h(z' \lor z'')$ 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

$$\phi(z') + \phi(z'') = u(b) + u(c) \leq u(a) + u(d_0) \leq u(a) + u(d) = \phi(z' \land z'') + \phi(z' \lor z''),$$

(43)

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 (43). That proves the supermodularity of $\phi$. □

**Proof of Proposition 6**

Applying Lemma 3, we have that $u(a^\top x + g(x, z))$ is supermodular in $z$ for all $x \in \mathcal{X}$. Hence, following Theorem 1 by treating $u(a^\top x + g(x, z))$ in Problem (12) as the $g(x, z)$ in Problem (2), Problem (12) can be solved equivalently by

$$\min \nu^\top l$$

s.t. $R_k^\top l \geq \sum_{i=2^{n+1}} p_i u(a^\top x + b^\top y^{k,i}), \ k \in [K]$

$Wx + Uy^{k,i} \geq Vz^{k,i} + v^0, \ k \in [K], i \in [2n + 1]$

$l \geq 0, \ x \in \mathcal{X},$

Introducing auxiliary variables $f^{k,i}$ with $f^{k,i} \geq u(a^\top x + b^\top y^{k,i}) = \max_{j \in [J]} \{c_j (a^\top x + b^\top y^{k,i}) + d_j\}$, we then obtain the equivalent reformulation as in (14).

□

**Proof of Corollary 2**

By the minimax Theorem in Sion et al. (1958), in Problem (15), we can interchange the maximization over $\mathbb{P} \in \mathcal{F}$ and the minimization over $\theta \in \mathbb{R}$. Hence, Problem (15) is equivalent to

$$\min \theta + \sup_{\mathbb{P} \in \mathcal{F}} \mathbb{E}_{\mathbb{P}} [u(a^\top x + g(x, z) - \theta)]$$

s.t. $x \in \mathcal{X}$. \hspace{1cm} (44)

Its equivalent reformulation (16) can be obtained as a direct application of Proposition 6. □
Proof of Theorem 6

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

WLOG, let $h_i^1$ be the function which at least three pieces on $[z_1, z_1]$. We choose functions $f^1, f^2 : \mathbb{R}^n \to \mathbb{R}$ such that $f^1(z) = g^1(z_1), f^2(z) = g^2(z_1) \forall z \in \mathbb{R}^n$ for some $g^1, g^2 : \mathbb{R} \to \mathbb{R}$. Moreover, for all $j \in \{2, \ldots, J\}$, we choose $\delta_i^j$ to be sufficiently large such that $\mathbb{E}_\rho \left[h_i^j(z_1)\right] \leq \delta_i^j$ holds for all $\rho \in \{\rho \mid \rho(z_1 \leq z_1 \leq z_1) = 1, \mathbb{E}_\rho [z_1] = \mu_1\}$. We then have for $i = 1, 2$,

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

where

$$\mathcal{G} = \left\{ \rho \mid \rho(z_1 \leq z_1 \leq z_1) = 1, \mathbb{E}_\rho [z_1] = \mu_1, \mathbb{E}_\rho [h_i^j(z_1)] \leq \delta_i^j \right\}.$$

For notational simplicity, we omit the superscript and subscript of $h$ and $\delta$, as well as the subscript of $z$ and $\mu$. That is, we consider $\sup_{\rho \in \mathcal{G}} \mathbb{E}_\rho [g^i(\tilde{z})]$ with $\mathcal{G} = \left\{ \rho \mid \rho(\tilde{z} \leq \tilde{z} \leq \tilde{z}) = 1, \mathbb{E}_\rho [\tilde{z}] = \mu, \mathbb{E}_\rho [h(\tilde{z})] \leq \delta \right\}$, where $h$ has at least three pieces on $[z, z]$. Now it suffices to find $g^1, g^2 : \mathbb{R} \to \mathbb{R}$ such that there does not exist a common worst-case distribution.

Let $J + 1$ be the number of pieces of $h$ on $[z, z]$ for some $J \geq 2$, and denote the corresponding breakpoints by $z = z_0 < \cdots < z_{J + 1} = z$. We define two functions $l^1, l^2 : \mathbb{R}_+ \to \mathbb{R}$ such that

$$l^1(p_1) \in \left\{ p_0 h(z_0^0) + p_1 h(z_1^0) + \cdots + p_{J + 1} h(z_{J + 1}^0) \mid p_0 + p_1 + \cdots = 1, p_0 z_0 + p_1 z_1 + \cdots + p_{J + 1} z_{J + 1} = \mu \right\}$$

$$l^2(p_2) \in \left\{ p_0 h(z_0^0) + p_1 h(z_1^0) + \cdots + p_{J + 1} h(z_{J + 1}^0) \mid p_0 + p_1 + \cdots = 1, p_0 z_0^2 + p_1 z_1^2 + \cdots + p_{J + 1} z_{J + 1}^2 = \mu \right\} \quad (45)$$

Notice that the sets in Equation (45) 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 (45). 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 (45)) 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

$$H^1 = \begin{bmatrix} 1 & 1 & 1 \\ z_0^0 & z_1^0 & z_{J + 1}^0 \\ h(z_0^0) & h(z_1^0) & h(z_{J + 1}^0) \end{bmatrix}, \quad H^2 = \begin{bmatrix} 1 & 1 & 1 \\ z_0^2 & z_1^2 & z_{J + 1}^2 \\ h(z_0^2) & h(z_1^2) & h(z_{J + 1}^2) \end{bmatrix},$$

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

$$\mathbb{P}^1(\tilde{z} = w) = \begin{cases} p_1^1 & \text{if } w = z_0^0 \\ p_2^1 & \text{if } w = z_1^0 \\ p_3^1 & \text{if } w = z_{J + 1}^0 \\ 0 & \text{otherwise} \end{cases}, \quad \mathbb{P}^2(\tilde{z} = w) = \begin{cases} p_1^2 & \text{if } w = z_0^0 \\ p_2^2 & \text{if } w = z_2^0 \\ p_3^2 & \text{if } w = z_{J + 1}^0 \\ 0 & \text{otherwise} \end{cases}$$
Then \( P^1, P^2 \) have the support \( Z^1 = \{ z^0, z^1, z^{j+1} \}, Z^2 = \{ z^0, z^2, z^{j+1} \} \), respectively.

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

\[
E_{P^i} [g^i(z)] = \sum_{z \in Z^i} P(z = z)g^i(z) = \sum_{z \in Z^i} P(z = z)h(z) = E_{P^i} [h(z)] = \delta,
\]

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

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

\[
E_P [g^i(z)] = \int_{[\underline{z}, \overline{z}]} g^i(z) dP < \int_{[\underline{z}, \overline{z}]} h(z) dP = E_P [h(z)] \leq \delta,
\]

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

We next prove the direction of “\( 2 \)” “\( 1 \)”.

By strong duality,

\[
\sup_{P \in G} E_P [f(\tilde{z})] = \inf \left\{ s + \mu^\top t + \sum_{i=1}^n \sum_{j=1}^{J_i} \delta_i^j r_i^j \left| \begin{array}{c} s + \mu^\top z + \sum_{i=1}^n \sum_{j=1}^{J_i} h_i^j(z_i) r_i^j \geq f(z), \ \forall z \in [\underline{z}, \overline{z}] \end{array} \right. \right\}.
\]

Let \( Z = \{ z \mid z_i \in \{ z_1^i, \ldots, z_{S_i}^i \}, i \in [n] \} \) which contains all \( z \) such that each of its dimension is on the breakpoints. Then we observe that \( [\underline{z}, \overline{z}] \) can be decomposed as \( [\underline{z}, \overline{z}] = \cup_{i=1}^S Z_i \) for some \( S \) and disjoint \( Z_1, \ldots, Z_S \) such that all \( Z_i \) are boxes with extreme points in \( Z \) and \( \sum_{i=1}^n \sum_{j=1}^{J_i} h_i^j(z_i) \) are linear within each \( Z_i \). Together with the convexity of \( f \), the dual problem is equivalent to

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

Writing its dual form again, we conclude that there exists a worst-case distribution with its support as \( Z \). Hence, for \( \sup_{P \in G} E_P [f(\tilde{z})] \), it suffices to consider only the probability distributions with support as \( Z \).
Assuming “2” is true, i.e., \( h_i^j, i \in [n], j \in [J] \) are convex piecewise linear functions with exactly two pieces on \([z_i, \overline{z}_i] \), we will show “1” is true. In other words, we will show the existence of a \( \mathbb{P}^* \in \text{arg sup}_{\mathbb{P} \in \mathcal{F}} \mathbb{E}_\mathbb{P}[f(\tilde{z})] \) 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 \( \tilde{z}_n, z_n, \overline{z}_n, \mu_n, h_n^i, \delta_n^i, J_n \). Hence, we have \( \tilde{z} = (\tilde{z}_1, \ldots, \tilde{z}_{n-1}, \tilde{z}) \), \( z = (z_1, \ldots, z_{n-1}, z) \) 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 \( \hat{\tilde{z}} \in (\overline{z}, \overline{z}) \), then we move the probability mass on \( z \) with \( z = \hat{\tilde{z}} \) to \( z - (\hat{\tilde{z}} - \tilde{z})e_n \) and \( z + (\overline{z} - \hat{\tilde{z}})e_n \) 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 \( \mathbb{E}_\mathbb{P}[f(\tilde{z})] \) 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 \((\overline{z}, \overline{z})\). WLOG, we let \( h^1, h^2 \) be both with breakpoint \( \hat{\tilde{z}} \in (\overline{z}, \overline{z}) \). Define \( \hat{\mathcal{G}} = \{ \mathbb{P} \mid \mathbb{P}(\tilde{z} \leq \hat{\tilde{z}} \leq \overline{z}) = 1, \mathbb{E}_\mathbb{P}[\tilde{z}] = \mu \} \), \( \mathcal{G}^i = \{ \mathbb{P} \mid \mathbb{E}_\mathbb{P}[h^i(\tilde{z})] \leq \delta^i \}, i \in \{1, 2\} \).

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 \( \mathbb{E}_\mathbb{P}[h^i(\tilde{z})] \). We have \( \mathcal{G}' = \mathcal{F}^G \), and hence \( \sup_{\mathbb{P} \in \mathcal{F}^G} \mathbb{E}_\mathbb{P}[f(\tilde{z})] = \sup_{\mathbb{P} \in \mathcal{G}'} \mathbb{E}_\mathbb{P}[f(\tilde{z})] \). 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 \((s_1^p, s_2^p, \overline{s}_p, s_p, \overline{s}_p)\) which is uniquely determined by the following system of equations,

\[
\begin{align*}
\int_{z \leq \hat{\tilde{z}}} zd\mathbb{P}(z) &= \overline{s}_p^p + \hat{\tilde{z}}s_1^p \\
\mathbb{P}(\hat{\tilde{z}} \leq \tilde{z}) &= \overline{s}_p^p + s_1^p \\
\int_{z > \hat{\tilde{z}}} zd\mathbb{P}(z) &= \overline{s}_p^p + \hat{\tilde{z}} \overline{s}_p \\
\mathbb{P}(\tilde{z} > \hat{\tilde{z}}) &= s_2^p + \overline{s}_p \\
s_p^p &= s_1^p + s_2^p
\end{align*}
\]
In this case, for any convex piecewise linear function with two pieces and with breakpoint at \( \hat{z} \), which can be denoted by \( h(z) = \begin{cases} \frac{a z + b}{z} & \text{if } z \leq \hat{z} \\ \alpha z + \beta & \text{if } z \geq \hat{z} \end{cases} \) where \( \alpha < \alpha \), we have

\[
\mathbb{E}_\mathbb{P}[h(\hat{z})] = \int_{z \leq \hat{z}} (a z + b) \, d\mathbb{P}(z) + \int_{z > \hat{z}} (\alpha z + \beta) \, d\mathbb{P}(z) \\
= a \int_{z \leq \hat{z}} z \, d\mathbb{P}(z) + b \mathbb{P}(\hat{z} \leq z) + \alpha \int_{z > \hat{z}} z \, d\mathbb{P}(z) + \beta \mathbb{P}(z > \hat{z}) \\
= a (\hat{z} s^p + \hat{z} s^p) + b (s^p + s^p) + \alpha (\hat{z} s^p + \hat{z} s^p) + \beta (s^p + s^p) \\
= s^p h(\hat{z}) + \bar{s}^p h(\bar{z}),
\]

where the third and fourth inequalities are due to (46). Moreover, by (46) we can easily have \( s^p + \bar{s}^p + \bar{s}^p = 1 \) and \( s^p \hat{z} + \bar{s}^p \hat{z} + \bar{s}^p \bar{z} = \mu \), which imply

\[
s^p = \frac{\bar{s} - \mu - (\bar{s} - \hat{z}) \bar{s}^p}{\bar{s} - \hat{z}}, \quad \bar{s}^p = \frac{\mu - \bar{s} - (\hat{z} - \bar{z}) \bar{s}^p}{\bar{s} - \hat{z}}.
\]

Therefore,

\[
\mathbb{E}_\mathbb{P}[h(\hat{z})] = \frac{\bar{s} - \mu - (\bar{s} - \hat{z}) \bar{s}^p}{\bar{s} - \hat{z}} h(\hat{z}) + \bar{s}^p h(\hat{z}) + \frac{\mu - \bar{s} - (\hat{z} - \bar{z}) \bar{s}^p}{\bar{s} - \hat{z}} h(\bar{z}) = c^h + \Delta^h \bar{s}^p,
\]

where \( c^h, \Delta^h \) are constants depending on \( h \) but independent from \( \mathbb{P} \); moreover,

\[
\Delta^h = h(\hat{z}) - \left( \frac{\bar{s} - \mu - (\bar{s} - \hat{z}) \bar{s}^p}{\bar{s} - \hat{z}} h(\hat{z}) + \frac{\mu - \bar{s} - (\hat{z} - \bar{z}) \bar{s}^p}{\bar{s} - \hat{z}} h(\bar{z}) \right) < h(\hat{z}) - h(\bar{z}) = 0,
\]

where the inequality follows from the convexity of \( h \).

Recall that \( \hat{G} \cap G^1 \cap G^2 = \hat{G} \cap G^2 \), then \( \exists P^o \in (\hat{G} \cap G^2) \setminus G^1 \). Therefore, consider any \( \hat{P} \in \hat{G} \cap G^1 \),

\[
c^h^1 + \Delta^h^1 \bar{s}^p = \mathbb{E}_{P^o} [h^1(\hat{z})] > \delta^1 \geq \mathbb{E}_{\hat{P}} [h^1(\hat{z})] = c^h^1 + \Delta^h^1 \bar{s}^p,
\]

where the two equalities follow from (47), the two inequalities are due to \( P^o \not\in G^1 \) and \( \hat{P} \in G^1 \). Hence, we have \( \bar{s}^p < \bar{s}^p \) since (48) results in \( \Delta^h^1 < 0 \). It then implies

\[
\mathbb{E}_{\hat{P}} [h^2(\hat{z})] = c^h^2 + \Delta^h^2 \bar{s}^p < c^h^2 + \Delta^h^2 \bar{s}^p = \mathbb{E}_{P^o} [h^2(\hat{z})] \leq \delta^2,
\]

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

- **Scenario II:** All \( h^j, j \in [\hat{J}] \), have distinct breakpoints in \( (\hat{z}, \bar{z}) \). In this case, denote the breakpoint of \( h^j \) by \( z^j, j \in [\hat{J}] \). WLOG, assume \( \hat{z} = z^0 < z^1 < \cdots < z^{\hat{J}} < z^{\hat{J}+1} = \bar{z} \). Consider any \( \mathbb{P} \in \mathcal{F}^G \). Denote by \( p_j = \mathbb{P}(\hat{z} = z^j) \) 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,

\[
\begin{aligned}
\sum_{j=0}^{j+1} p_j &= 1, \\
\sum_{j=0}^{j+1} z^j p_j &= \mu, \\
\sum_{j=0}^{j+1} h^i(z^j)p_j &\leq \delta^i, \quad i \in [\hat{J}], \\
p_j &\geq 0, \quad j \in \{0, \ldots, \hat{J} + 1\}.
\end{aligned}
\]

By (49a) and (49b) we have

\[
p_0 = \frac{\overline{z} - \mu - \sum_{j=1}^{\hat{J}} (\overline{z} - z^j)p_j}{\overline{z} - \underline{z}}, \quad p_{j+1} = \frac{\mu - \overline{z} - \sum_{j=1}^{\hat{J}} (z^j - \overline{z})p_j}{\overline{z} - \underline{z}},
\]

which implies that $p_0, p_{j+1}$ can be determined by $p = (p_1, \ldots, p_j)$. In what follows, we simplify the constraints (49a)-(49d).

We first investigate the constraint (49c) for any given $i \in [\hat{J}]$. Since $h^i$ is convex and has breakpoints $\{\underline{z}, z^i, \overline{z}\}$, we can denote $h^i(z) = \begin{cases} h^i(z^i) - \gamma_i(z^i - \underline{z}) & \text{if } z \in [\underline{z}, z^i] \\ h^i(z^i) + \xi_i(z^i - \overline{z}) & \text{if } z \in [z^i, \overline{z}] \end{cases}$ for some $\gamma_i < \xi_i$. It follows that

\[
\sum_{j=0}^{j+1} h^i(z^j)p_j = h^i(z^i) - \gamma_i(z^i - \underline{z})p_0 - \gamma_i \sum_{j=1}^{i} (z^i - z^j)p_j + \xi_i \sum_{j=1}^{i} (z^j - z^i)p_j + \xi_i(\overline{z} - z^i)p_{j+1}
\]

\[
= h^i(z^i) - \frac{\gamma_i}{\overline{z} - \underline{z}} \left((z^i - \underline{z})\left(\overline{z} - \mu - \sum_{j=1}^{i} (\overline{z} - z^j)p_j\right) + (\overline{z} - \underline{z})\sum_{j=1}^{i} (z^j - z^i)p_j\right)
\]

\[
+ \frac{\xi_i}{\overline{z} - \underline{z}} \left((\overline{z} - z^i)\left(\mu - \overline{z} - \sum_{j=1}^{i} (z^j - \overline{z})p_j\right) + (\overline{z} - \underline{z})\sum_{j=1}^{i} (z^j - z^i)p_j\right)
\]

\[
= h^i(z^i) + \frac{\xi_i(\overline{z} - z^i)(\mu - \overline{z}) - \gamma_i(\overline{z} - \mu)(z^i - \underline{z})}{\overline{z} - \underline{z}} - \frac{\gamma_i}{\overline{z} - \underline{z}} \sum_{j=1}^{i} ((\overline{z} - \underline{z})(z^i - z^j) - (\overline{z} - z^j)(z^i - \underline{z}))p_j - \xi_i \sum_{j=i+1}^{\hat{J}} (\overline{z} - z^j)(z^i - \underline{z})p_j
\]

\[
+ \frac{\xi_i}{\overline{z} - \underline{z}} \left(-\sum_{j=1}^{i-1} (\overline{z} - z^j)(z^i - \underline{z})p_j + \sum_{j=1}^{i} ((z^j - z^i)(\overline{z} - \underline{z}) - (\overline{z} - z^j)(z^i - \underline{z}))p_j\right)
\]
\[ h^i(z^i) + \frac{\xi_i(z^i - \bar{z})(\mu - \bar{z})}{\bar{z} - \bar{z}} - \frac{\gamma_i(z^i - \bar{z})(z^i - \bar{z})}{\bar{z} - \bar{z}} \]
\[ + \frac{\gamma_i}{\bar{z} - \bar{z}} \left( - \sum_{j=1}^{i-1} (z^i - \bar{z})(z^j - \bar{z}) p_j + (\bar{z} - z^i)(z^i - \bar{z}) p_i + \sum_{j=i+1}^{j} (\bar{z} - z^j)(z^i - \bar{z}) p_j \right) \]
\[ - \frac{\xi_i}{\bar{z} - \bar{z}} \left( \sum_{j=1}^{i-1} (z^i - \bar{z})(z^j - \bar{z}) p_j + (\bar{z} - z^i)(z^i - \bar{z}) p_i - \sum_{j=i+1}^{j} (z^j - \bar{z})(z^i - \bar{z}) p_j \right) \]
\[ = h^i(z^i) + \frac{\xi_i(z^i - \bar{z})(\mu - \bar{z})}{\bar{z} - \bar{z}} - \frac{\gamma_i(z^i - \bar{z})(z^i - \bar{z})}{\bar{z} - \bar{z}} \]
\[ - \frac{\xi_i - \gamma_i}{\bar{z} - \bar{z}} \sum_{j=1}^{j} (z^i - z^{\max(i,j)}) (z^{\min(i,j)} - \bar{z}) p_j. \]

Hence the \( i \)-th constraint of (49c) is equivalent to
\[ \sum_{j=1}^{j} (z^i - z^{\max(i,j)}) (z^{\min(i,j)} - \bar{z}) p_j \geq d_i, \]
where
\[ d_i = \frac{\bar{z} - \bar{z}}{\xi_i - \gamma_i} \left( h^i(z^i) + \frac{\xi_i(z^i - \bar{z})(\mu - \bar{z})}{\bar{z} - \bar{z}} - \frac{\gamma_i(z^i - \bar{z})(z^i - \bar{z})}{\bar{z} - \bar{z}} - \delta^i \right). \]
Denote \( \lambda_j = \bar{z} - z^j, \pi_j = z^j - \bar{z} \) for all \( j \in [\hat{J}] \), and let
\[ A = \left( \lambda_{\max(i,j)} \pi_{\min(i,j)} \right)_{i,j \in [\hat{J}]} = \begin{bmatrix} \lambda_1 \pi_1 & \lambda_2 \pi_1 & \cdots & \lambda_j \pi_1 \\ \lambda_2 \pi_1 & \lambda_2 \pi_2 & \cdots & \lambda_j \pi_2 \\ \vdots & \vdots & \ddots & \vdots \\ \lambda_j \pi_1 & \lambda_j \pi_2 & \cdots & \lambda_j \pi_j \end{bmatrix}. \]

Then (49c) is equivalent to \( Ap \geq d \), where both \( A \) and \( d \) are constants determined by \( F^G \).

For (49d), by (50) we have that \( p_0 \geq 0 \) is equivalent to \( \sum_{j=1}^{j} (z^i - \bar{z}) p_j \leq \bar{z} - \mu \), and \( p_{j+1} \geq 0 \) is equivalent to \( \sum_{j=1}^{j} (z^i - \bar{z}) p_j \leq \mu - \bar{z} \). Recalling the definition of \( A \), the constraints \( p_0 \geq 0, p_{j+1} \geq 0 \) can be further reformulated as
\[ a_1^T p \leq b_1, \quad a_j^T p \leq b_u, \]
respectively, where \( b_1 = (\bar{z} - \mu) \pi_1, \quad b_u = (\mu - \bar{z}) \lambda_j \).

Therefore, \( p_0, \ldots, p_{j+1} \) satisfy (49a)-(49d) if and only if \( (p_0, p, p_{j+1}) \in \mathcal{P} \) where
\[ \mathcal{P} = \left\{ (p_0, p, p_{j+1}) \in \mathbb{R}^{j+2} \left| \begin{array}{c} Ap \geq d, \ a_1^T p \leq b_1, \ a_j^T p \leq b_u, \ p_0 \geq 0 \\ p_0 = \frac{\bar{z} - \mu}{\bar{z} - \bar{z}} - \frac{a_1^T p}{\pi_1}, \ p_{j+1} = \frac{\mu - \bar{z}}{\pi_j} - \frac{a_j^T p}{\pi_j} \lambda_j \end{array} \right. \right\}, \]

where the equalities on \( p_0 \) and \( p_{j+1} \) are from the equalities in (50). Note that \( \mathcal{P} \neq \emptyset \) as we assume \( F^G \neq \emptyset \). Denote by \( (C_i), i \in [\hat{J}] \) the \( i \)-th constraint of \( Ap \geq d \), i.e., \( a_i^T p \geq d_i \). We say a constraint \( (C_i) \) is redundant if the strict inequality \( a_i^T p > d_i \) holds for any \( p \in \hat{P} = \{ p \in \mathbb{R}_{++}^j \mid Ap \geq d \} \).

Consider the case that there exists \( i \in [\hat{J}] \) such that \( (C_i) \) is redundant. WLOG, we let the redundant constraint be \( (C_j) \). In this case, we define \( \mathcal{P}^* = \{ p \in \mathbb{R}_{++}^j \mid a_i^T p \geq d_i, \ i \in [\hat{J} - 1] \} \) and will
show that \( \hat{P} = P^o \). Obviously, \( \hat{P} \subseteq P^o \) since all constraints in defining \( P^o \) are also used in defining \( \hat{P} \). We now show \( P^o \subseteq \hat{P} \) by contradiction. Assume that there exists \( p^o \in P^o \setminus \hat{P} \), we have \( a^T_j p^o < d_j \). Choosing any \( p \in \hat{P} \), as \((C_j)\) is redundant, \( a^T_j p > d_j \). Therefore, we can find \( \lambda \in (0, 1) \) such that \( p^\lambda = \lambda p + (1 - \lambda)p^o \) such that \( a^T_j p^\lambda = d_j \). Moreover, by \( p^o \in P^o \) and \( p \in \hat{P} \), we have \( p^\lambda \geq 0 \) and \( a^T_j p^\lambda \geq d_i, \ i \in [J - 1] \). Therefore, we conclude \( p^\lambda \in \hat{P} \), which is a contradiction since we assume \((C_j)\) is redundant. Hence, \( P^o \subseteq \hat{P} \), and it implies \( P^o = \hat{P} \). Consequently, removing the constraint \( a^T_j p \geq d_j \) from the constraints in \((52)\) does not change the set \( P \). Investigating its reformulation back to the form as constraints \((49a)-(49d)\), we can see that now the problem of \( \sup_{p \in F^G} E_{p} [f(\tilde{z})] \) is equivalent to \( \sup_{p \in F'} E_{p} [f(\tilde{z})] \) where \( F' \) is the ambiguity set obtained from \( F^G \) by removing the constraint on \( h^j \). Therefore, we have a problem with \( J = J - 1 \), 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 \((C_1), \ldots, (C_j)\). We will prove that there exists a unique \((p^*_0, p^*, p^*_j, p^*_j + i) \in P \) with \( Ap^* = d \). Recall that the system \( Ap = d \) is

\[
\begin{align*}
\lambda_1 \pi_1 p_1 + \lambda_2 \pi_1 p_2 + \lambda_3 \pi_3 p_3 + \cdots + \lambda_j \pi_1 p_j &= d_1 \quad \text{(B1)} \\
\lambda_2 \pi_1 p_1 + \lambda_2 \pi_2 p_2 + \lambda_3 \pi_2 p_3 + \cdots + \lambda_j \pi_2 p_j &= d_2 \quad \text{(B2)} \\
\lambda_3 \pi_1 p_1 + \lambda_3 \pi_2 p_2 + \lambda_3 \pi_3 p_3 + \cdots + \lambda_j \pi_3 p_j &= d_3 \quad \text{(B3)} \\
\vdots \\
\lambda_j \pi_1 p_1 + \lambda_j \pi_2 p_2 + \lambda_j \pi_3 p_3 + \cdots + \lambda_j \pi_j p_j &= d_j \quad \text{(B_j)}
\end{align*}
\]

Combining \((B_1)\) and \((B_2)\) we have \( \pi_1 p_1 = \frac{d_1 \pi_2 - d_2 \pi_1}{\lambda_1 \pi_2 - \lambda_2 \pi_1} \). Combining \((B_2)\) and \((B_3)\) we obtain \( \pi_2 p_2 = \frac{d_2 \pi_3 - d_3 \pi_2}{\lambda_2 \pi_3 - \lambda_3 \pi_2} - \pi_1 p_1 \). Continuing the same procedure, we have

\[
\begin{align*}
p_1^* &= \frac{1}{\pi_1} \frac{d_1 \pi_2 - d_2 \pi_1}{\lambda_1 \pi_2 - \lambda_2 \pi_1} \\
p_2^* &= \frac{1}{\pi_2} \frac{d_2 \pi_3 - d_3 \pi_2}{\lambda_2 \pi_3 - \lambda_3 \pi_2} - \pi_1 p_1^* \\
p_3^* &= \frac{1}{\pi_3} \frac{d_3 \pi_4 - d_4 \pi_3}{\lambda_3 \pi_4 - \lambda_4 \pi_3} - \pi_1 p_1^* - \pi_2 p_2^* \\
&\vdots \\
p_{j-1}^* &= \frac{1}{\pi_{j-1}} \frac{d_{j-1} \pi_j - d_j \pi_{j-1}}{\lambda_{j-1} \pi_j - \lambda_j \pi_{j-1}} - \pi_1 p_1^* - \cdots - \pi_{j-2} p_{j-2}^* \\
p_j^* &= \frac{1}{\lambda_j} \frac{\lambda_j \pi_j - \lambda_j d_j}{\lambda_j \pi_j - \lambda_j \pi_{j-1}}
\end{align*}
\]

is the unique solution to \( Ap = d \). \( p_0^* \) and \( p_{j+1}^* \) can be uniquely determined by the equalities in \((52)\). Moreover, since \( P \neq \emptyset \), we must have \( d_1 \leq b_1, d_j \leq b_u \), which implies \( a^T_1 p^* = d_1 \leq b_1, a^T_j p^* = d_j \leq b_u \). To see \((p^*_0, p^*, p^*_j, p^*_j + i) \in P \), it remains to prove for any \( j \in [J], p_j^* \geq 0 \). We show that this must be the case, otherwise the constraint \((C_j)\) is redundant. Recall that by the definition of \( \lambda_j \) and \( \pi_j, j \in [J] \), we have \( \lambda_1 > \cdots > \lambda_j > 0 \) and \( 0 < \pi_1 < \cdots < \pi_j \).
We first show that $p_1^* \geq 0$, i.e., $d_1 \pi_2 - d_2 \pi_1 \geq 0$. Assume to the contrary that $d_1 \pi_2 < d_2 \pi_1$, then
\[
a_j^\top p - d_j = \lambda_j p_1 + \cdots + \lambda_j p_j - d_j
\]
\[
\geq \lambda_j (p_1 + \cdots + p_j) - d_j
\]
\[
\geq \lambda_j (p_1 + \cdots + p_j) - d_j
\]
\[
= \frac{\pi_1}{\pi_2} d_2 + \left( \lambda_1 - \frac{\pi_1}{\pi_2} \lambda_2 \right) \pi_1 p_1 - d_1
\]
\[
\geq \frac{\pi_1}{\pi_2} d_2 - d_1 > 0
\]
for all $p \in \mathcal{P}$. Here the first inequality follows from $a_j^\top p \geq d_2$; 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 (C1) is redundant.

Next, for $p_j^*$, we show $\lambda_{j-1} \pi_j - \lambda_j \pi_{j-1} \geq 0$ by contradiction. Assume $\lambda_{j-1} \pi_j < \lambda_j \pi_{j-1}$, then similar as above we have
\[
a_j^\top p - d_j = \lambda_j (p_1 + \cdots + p_j) - d_j
\]
\[
\geq \lambda_j (p_1 + \cdots + p_j) - d_j
\]
\[
= \frac{\lambda_j}{\lambda_{j-1}} d_{j-1} + \left( \pi_j - \frac{\lambda_j}{\lambda_{j-1}} \pi_{j-1} \right) \lambda_j p_j - d_j
\]
\[
\geq \frac{\lambda_j}{\lambda_{j-1}} d_{j-1} - d_j > 0
\]
for all $p \in \mathcal{P}$. Here the first inequality follows from $a_j^\top p \geq d_{j-1}$; the second inequality holds because $\lambda_j < \lambda_{j-1}, \pi_j > \pi_{j-1}$, and the last inequality follows from the assumption $\lambda_{j-1} \pi_j < \lambda_j \pi_{j-1}$. Hence (C1) is redundant.

Finally, for all $j \in \{2, \ldots, \hat{J} - 1\}$, we show that $\pi_j p_j^* = \frac{d_j \pi_{j+1} - d_{j+1} \pi_j}{\lambda_j \pi_{j+1} - \lambda_{j+1} \pi_j} - \sum_{k=1}^{j-1} \pi_k p_k^* \geq 0$. Suppose not, i.e., $\frac{d_j \pi_{j+1} - d_{j+1} \pi_j}{\lambda_j \pi_{j+1} - \lambda_{j+1} \pi_j} < \sum_{k=1}^{j-1} \pi_k p_k^*$ Consider any $p \in \mathcal{P}$. We then have
\[
a_j^\top p - d_j = \lambda_j (p_1 + \cdots + p_j) - d_j
\]
\[
\geq \lambda_j (p_1 + \cdots + p_j) - d_j
\]
\[
= \frac{1}{\pi_{j+1}} (d_j \pi_j - d_j \pi_{j+1} - (\lambda_{j+1} \pi_j - \lambda_j \pi_j) (p_1 + \cdots + p_j))
\]
where the inequality follows from $a_{j+1}^\top p \geq d_{j+1}$. Further, we also have
\[
a_j^\top p - d_j = \lambda_j (p_1 + \cdots + p_j) - d_j
\]
\[
\geq \lambda_j (p_1 + \cdots + p_j) - d_j
\]
\[
= \frac{1}{\pi_{j-1}} (d_j \pi_{j-1} - d_j \pi_j - (\lambda_{j-1} \pi_j - \lambda_j \pi_{j-1}) (p_1 + \cdots + p_{j-1} - 1))
\]
where the inequality follows from $a_{j-1}^\top p \geq d_{j-1}$. Define two $\mathbb{R} \to \mathbb{R}$ functions $\phi'(t) = \frac{1}{\pi_{j+1}} (d_j \pi_j - d_j \pi_{j+1} - (\lambda_{j+1} \pi_j - \lambda_j \pi_{j+1}) t)$, $\phi''(t) = \frac{1}{\pi_{j+1}} (d_j \pi_j - d_j \pi_{j+1} - (\lambda_{j+1} \pi_j - \lambda_j \pi_{j+1}) t)$, then $a_j^\top p - d_j \geq$
max \{ \phi'(\pi_1 p_1 + \cdots + \pi_j p_j), \phi''(\pi_1 p_1 + \cdots + \pi_j p_j) \}. By definition, \( \lambda_{j+1} \pi_j - \lambda_j \pi_{j+1} < 0 \), hence \( \phi' \) is increasing, which implies \( \phi'(\pi_1 p_1 + \cdots + \pi_j p_j) \geq \phi'(\pi_1 p_1 + \cdots + \pi_j p_j) \). Thus

\[
\alpha_j p - d_j \geq \max \{ \phi'(\pi_1 p_1 + \cdots + \pi_j p_j), \phi''(\pi_1 p_1 + \cdots + \pi_j p_j) \}.
\]

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

\[
\max \{ \phi'(t), \phi''(t) \} \geq \phi'(t) > \phi' \left( \frac{d_j \pi_{j+1} - d_j \pi_j}{\lambda_j \pi_{j+1} - \lambda_j \pi_j} \right) = 0 \quad \text{if} \quad t \geq \tau,
\]

\[
\max \{ \phi'(t), \phi''(t) \} \geq \phi''(t) > \phi' \left( \frac{d_j \pi_{j-1} - d_j \pi_{j-1}}{\lambda_j \pi_{j-1} - \lambda_j \pi_{j-1}} \right) = 0 \quad \text{if} \quad t \leq \tau,
\]

which implies

\[
\alpha_j p - d_j \geq \max \{ \phi'(\pi_1 p_1 + \cdots + \pi_j p_j), \phi''(\pi_1 p_1 + \cdots + \pi_j p_j) \} > 0.
\]

Hence \( (C_j) \) is redundant. We then conclude \( p^* \geq 0 \).

In summary, we have a unique \((p_0^*, \mathbf{p}^*, \mathbf{p}_{j+1}^*) \in \mathcal{P} \) with \( \mathbf{A} \mathbf{p}^* = \mathbf{d} \).

Related to \((p_0^*, \mathbf{p}^*, \mathbf{p}_{j+1}^*) \), we next prove the following observation.

**Observation:** Considering any \((p_0, \mathbf{p}, \mathbf{p}_{j+1}) \in \mathcal{P} \) with \((p_0, \mathbf{p}, \mathbf{p}_{j+1}) \neq (p_0^*, \mathbf{p}^*, \mathbf{p}_{j+1}^*) \), there exists \( i \in \{0, 1, \ldots, j \} \), \( i + 1 \leq k \leq j \), 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. \( \alpha_{k}^\top \mathbf{p} > d_k \).

Specifically, parts 1) and 2) mean that \( i \) is the index of the first distinct component when comparing \((p_0, \mathbf{p}, \mathbf{p}_{j+1}) \) and \((p_0^*, \mathbf{p}^*, \mathbf{p}_{j+1}^*) \); parts 3) and 4) mean that \( k \) is the index of the first nonzero component in \((p_0, \mathbf{p}, \mathbf{p}_{j+1}) \) after \( p_i \).

To prove parts 1) and 2), we consider any \( i \in \{0, \ldots, j \} \), and have

\[
\frac{\mathbf{z} - \mu}{\mathbf{z} - \mathbf{z}} = \frac{\alpha_{i+1} \mathbf{p}}{\mathbf{z} - \mathbf{z}} = \frac{\alpha_{i+1} \mathbf{p}}{(\mathbf{z} - \mathbf{z}) \pi_{i+1}} \mathbf{p}
\]

\[
= \frac{\mathbf{z} - \mu}{\mathbf{z} - \mathbf{z}} = \frac{1}{\mathbf{z} - \mathbf{z}} \left( \sum_{j=1}^{i} \lambda_j p_j + \frac{1}{\pi_{i+1}} (\lambda_{i+1} \pi_{1} p_1 + \cdots + \lambda_{i+1} \pi_{i} p_i) \right)
\]

\[
= \frac{\mathbf{z} - \mu}{\mathbf{z} - \mathbf{z}} = \frac{1}{\mathbf{z} - \mathbf{z}} \left( \sum_{j=1}^{i} \lambda_j p_j + \frac{1}{\pi_{i+1}} (\lambda_{i+1} \pi_{1} p_1 + \cdots + \lambda_{i+1} \pi_{i} p_i - \lambda_1 \pi_{i+1} p_1 - \cdots - \lambda_i \pi_{i+1} p_i) \right)
\]

\[
= \mathbf{p}_0 + \frac{1}{(\mathbf{z} - \mathbf{z}) \pi_{i+1}} \sum_{j=1}^{i} \alpha_{i+1} j p_j,
\]
where we define \( \alpha_{i+1,j} = \lambda_j \pi_{i+1} - \lambda_{i+1} \pi_j > 0 \) for all \( j \leq i \) since in this case \( \lambda_j > \lambda_{i+1} \) and \( \pi_j < \pi_{i+1} \). Hence,

\[
a_{i+1}^T p = (\bar{z} - \mu) \pi_{i+1} - (\bar{z} - \bar{z}) \pi_{i+1} p_0 - \sum_{j=1}^{i} \alpha_{i+1,j} p_j. \tag{54}
\]

Consider \( i = 0 \), by (54) we have

\[
p_0 = \frac{\bar{z} - \mu}{\bar{z} - \bar{z}} - \frac{a_{i+1}^T p}{(\bar{z} - \bar{z}) \pi_1} \leq \frac{\bar{z} - \mu}{\bar{z} - \bar{z}} - \frac{d_i}{(\bar{z} - \bar{z}) \pi_1} = \frac{\bar{z} - \mu}{\bar{z} - \bar{z}} - \frac{a_{i+1}^T p^*}{(\bar{z} - \bar{z}) \pi_1} = p_0^*,
\]

where the first inequality is due to \( A p \geq d \), the second equality follows from Equation (54) also applies to \( (p^*_0, p, p^*_j, p^*_j) \). Hence, if \( p_0 \neq p_0^* \), we must have \( p_0 < p_0^* \). Now, consider the case where \( p_0 = p_0^* \), we then denote \( i \geq 1 \) as the index of the first distinct component, i.e., \( p_j = p^*_j \) \( \forall j \in \{0, \ldots, i-1\} \), and \( p_i \neq p_i^* \). Note that \( i \leq \hat{j} - 1 \), otherwise the only distinct components are the last two dimension, i.e., the marginal masses at \( z^j \) and \( z^{j+1} \), which is impossible since \( (p_0, p, p, p^*_j, p^*_j) \) correspond to the same mean. As \( i \leq \hat{j} - 1 \), by (54),

\[
a_{i+1}^T p = (\bar{z} - \mu) \pi_{i+1} - (\bar{z} - \bar{z}) \pi_{i+1} p_0^* - \sum_{j=1}^{i} \alpha_{i+1,j} p_j^* + \alpha_{i+1,1} (p^*_j - p_i)
\]

\[
= a_{i+1}^T p^* + \alpha_{i+1,1} (p^*_j - p_i)
\]

\[
= d_{i+1} + \alpha_{i+1,1} (p^*_j - p_i)
\]

\[
\leq a_{i+1}^T p + \alpha_{i+1,1} (p^*_j - p_i),
\]

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

Parts 3) and 4) in Observation are straightforward. Specifically,

\[
\sum_{j=1}^{J+1} p_j = 1 - \sum_{j=0}^{i} p_j = 1 - \sum_{j=0}^{i} p_j^* + (p_i^* - p_i) \geq p_i^* - p_i > 0.
\]

Hence, there must be a nonzero component in \( p_{i+1}, \ldots, p_{j+1} \). 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 (54), which leads to

\[
a_k^T p = (\bar{z} - \mu) \pi_k - (\bar{z} - \bar{z}) \pi_k p_0 - \sum_{j=1}^{k-1} \alpha_{k,j} p_j > (\bar{z} - \mu) \pi_k - (\bar{z} - \bar{z}) \pi_k p_0 - \sum_{j=1}^{k-1} \alpha_{k,j} p_j^* = d_k.
\]

Here the inequality is due to parts 1) and 2), and \( 0 = p_j \leq p_j^* \) for all \( j \in \{i+1, \ldots, k-1\} \).

Now, base on Observation, we prove “1” is true by proposing a process to construct new distribution. Given any \( P \in \mathcal{F}^G \), let the associated marginals on \( \hat{z} \) at \( z^0, \ldots, z^{j+1} \) be \( (p_0, p, p_{j+1}) \). Consider the case where \( (p_0, p, p_{j+1}) \neq (p^*_0, p^*, p^*_{j+1}) \). We now construct a new probability distribution \( P' \) with support only at the breakpoints and defined as

\[
P' (\hat{z} = z) = \begin{cases} 
\mathbb{P} (\hat{z} = z) & \text{if } z \notin \{z^{k-1}, z^k, z^{k+1}\} \\
(1 - \theta) \mathbb{P} (\hat{z} = z) & \text{if } z = z^k \\
\mathbb{P} (\hat{z} = z^k) + \frac{z^{k+1} - z^k}{z^{k+1} - z^k} \theta \mathbb{P} (\hat{z} = z + (z^k - z^{k-1}) e_n) & \text{if } z = z^{k-1} \\
\mathbb{P} (\hat{z} = z) + \frac{z^{k+1} - z^k}{z^{k+1} - z^k} \theta \mathbb{P} (\hat{z} = z - (z^{k+1} - z^k) e_n) & \text{if } z = z^{k+1}
\end{cases}
\tag{55}
\]
for some \( \theta \in (0, 1) \). Intuitively, for all \( z_1, \ldots, z_{n-1} \), we move \( \theta \) portion of the probability mass at \((z_1, \ldots, z_{n-1}, z^k)\) to \((z_1, \ldots, z_{n-1}, z^{k-1})\) and \((z_1, \ldots, z_{n-1}, z^{k+1})\), keeping the mean unchanged. Hence \( \mathbb{P}' \) has the same marginal for \((\tilde{z}_1, \ldots, \tilde{z}_{n-1})\) as \( \mathbb{P} \). Denote the marginal of \( \mathbb{P}' \) on \( \tilde{z} \) by \( p_j', p_{j+1}' \) such that \( \mathbb{P}'(\tilde{z} = z^j) = p_j' \) for all \( j = 0, \ldots, \hat{J} + 1 \). By \ref{eq:55},

\[
\begin{cases}
  p_j' = p_j, & \forall j \notin \{k-1, k, k+1\}, \\
  p_k' = p_k - \theta p_k, \\
  p_{k-1}' = p_{k-1} + \frac{z_{k+1} - z_k}{z_{k+1} - z_{k-1}} \theta p_k, \\
  p_{k+1}' = p_{k+1} + \frac{z_{k+1} - z_{k-1}}{z_{k+1} - z_{k-1}} \theta p_k.
\end{cases}
\]

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

(P1) \( \mathbb{E}_{\mathbb{P}'}[f(\tilde{z})] \geq \mathbb{E}_{\mathbb{P}}[f(\tilde{z})] \). This is because

\[
\mathbb{E}_{\mathbb{P}'}[f(\tilde{z})] - \mathbb{E}_{\mathbb{P}}[f(\tilde{z})] = \sum_{z_i \in \{z_{1, \ldots, n}\}, i \in [n-1]} \theta \mathbb{P}(\tilde{z} = (z_1, \ldots, z_{n-1}, z^k)) \left( \frac{z_{k+1} - z_k}{z_{k+1} - z_{k-1}} f(z_1, \ldots, z_{n-1}, z^{k-1}) + \frac{z_k - z_{k-1}}{z_{k+1} - z_{k-1}} f(z_1, \ldots, z_{n-1}, z^{k+1}) - f(z_1, \ldots, z_{n-1}, z^k) \right) \geq 0,
\]

where the inequality is due to the convexity of \( f \).

(P2) \( a_j \mathbf{p}' = a_j \mathbf{p} \) for all \( j \neq k \) and \( a_k \mathbf{p}' < a_k \mathbf{p} \). To see this, for any \( j \in [\hat{J}] \), we observe

\[
\mathbb{E}_{\mathbb{P}'}[h^j(\tilde{z})] - \mathbb{E}_{\mathbb{P}}[h^j(\tilde{z})] = \theta p_k \left( \frac{z_{k+1} - z_k}{z_{k+1} - z_{k-1}} h^j(z^{k-1}) + \frac{z_k - z_{k-1}}{z_{k+1} - z_{k-1}} h^j(z^{k+1}) - h^j(z^k) \right) \geq 0, \tag{56}
\]

where the inequality is due to the convexity of \( h \). Moreover, the “\( \geq \)” takes “\( = \)” if \( j \neq k \) since \( h^j \) is linear on \([z^{k-1}, z^{k+1}]\) for such \( j \); by contrast, “\( \geq \)” becomes “\( \succ \)” for \( j = k \) since \( h^k \) has a breakpoint at \( z^k \). Therefore, by the definition of \( A \), this property is proved.

(P3) \( a_k \mathbf{p}' \) is continuously decreasing in \( \theta \), which is implied by \ref{eq:56} and the definition of \( A \).

Based on the Observation and (P1)-(P3), given any \( \mathbb{P} \in \mathcal{F}^G \) whose marginal on \( \tilde{z} \) is different from \( (p_0^*, \mathbf{p}^*, p_{J+1}^*) \), we can use the procedure as in \ref{eq:55} 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 \( a_k \mathbf{p}' \) drops to the value of \( d_k \) (note that when \( \theta = 0 \), \( a_k \mathbf{p}' = a_k \mathbf{p} > d_k \), 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 \( f(\tilde{z}) \) 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 \( z^j \) is moved towards \( z' \), the smallest breakpoint where the marginal mass of \( \mathbb{P} \) differs from \( (p_0^*, \mathbf{p}^*, p_{J+1}^*) \). Repeating such process, the margin converges to \( (p_0^*, \mathbf{p}^*, p_{J+1}^*) \). We hence conclude that there must be a worst-case distribution whose \( n \)-th marginal is \( (p_0^*, \mathbf{p}^*, p_{J+1}^*) \).
and indeed, there exists a unique

\[ \text{ward.} \]

We now consider only the case that \( \beta \)

\[ \text{We now show that for} \quad U \quad \text{and} \quad V \quad \text{such that} \quad g \]

\[ \text{supermodularity of} \quad T, S \]

\[ \text{ity of} \quad g \]

\[ \text{Then it can be verified that} \quad g \]

\[ \text{supermodularity of} \quad T, S \]

\[ \text{Proof of Proposition 7} \]

We prove the supermodularity by showing that Problem (20) satisfies the conditions in Theorem 6. We first reformulate Problem (20) as the sum of \( m \) sub-problems. Denote

\[ g_j(x, z) = \left\{ \sum_{i \in [n]} c_{ij} y_{ij} \left| \sum_{i \in [n]} y_{ij} = 1, 0 \leq y_{ij} \leq x_i z_i, i \in [n] \right. \right\}. \]

Then it can be verified that \( g(x, z) = \sum_{j \in [m]} g_j(x, z) \). Hence, it suffices to prove the supermodularity of \( g_j(x, z) \) for all \( j \in [m] \). Observing that \( x \in \{0, 1\}^n \), we denote \( S = \{i \in [n] \mid x_i = 1\} \) and \( T = [n] \setminus S \). It follows that the constraints in defining \( g_j(x, z) \) can be reformulated as \( U(y_{1j}, \ldots, y_{nj}) - V z \geq v^0, \)

\[ U = \begin{bmatrix} 1^T \\ -1^T \\ I_{n \times n} \\ -I_T \\ -I_S \end{bmatrix} \in \mathbb{R}^{(2+2n) \times n}, \quad V = \begin{bmatrix} 0_{(2+2n+|T|) \times n} \\ -I_S \end{bmatrix} \in \mathbb{R}^{(2+2n) \times n}, \quad v^0 = \begin{bmatrix} 1 \\ -1 \\ 0_{2n \times 1} \end{bmatrix} \in \mathbb{R}^{2+2n}. \]

Here \( I_T, I_S \) are the submatrices of \( I_{n \times n} \) consisting of rows which are indexed by elements in \( T, S \), separately. Note that \( \text{rank}(U) = n < 2n + 2 \). We hence can apply Theorem 6 to prove the supermodularity of \( g_j \). To this end, we consider any index set \( \mathcal{I} \) such that \( |\mathcal{I}| = n + 1 \) and \( \text{rank}(U_{\mathcal{I}}) = n \), any \( \beta \geq 0 \in \mathbb{R}^n, \alpha \in \mathbb{R}^n \) such that

\[ V_{\mathcal{I}} \beta = U_{\mathcal{I}} \alpha. \tag{57} \]

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

By \( (V_{\mathcal{I}})_j \neq 0 \), we have \( V_j \neq 0 \). Based on the structure of \( V \), \( V_j \) has only one nonzero element and indeed, there exists a unique \( i \) such that \( V_j = -e_i \in \mathbb{R}^{2n+2} \). Moreover, \( i \in \mathcal{I}, i > 2 + n + |\mathcal{I}| \), \( V_{ij} = -1 \) is the only nonzero element in the \( i \)th row, i.e., \( v_i = -e_j \in \mathbb{R}^n \). Therefore, by (57), we have \( u_i^T \alpha = v_i^T \beta = -\beta_j \), implying \( \alpha_j = \beta_j \) since \( U_{ij} = -1 \) is also the only nonzero element in \( u_i \).

We now show that for \( U_j \), only zero element from blocks \( I_{n \times n} \) and \(-I_T \) are included in \( (U_{\mathcal{I}})_j \).
Assume to the contrary, i.e., \( \exists k \in \{3, \ldots, 2 + n + |T|\} \cap I \) with \( U_{kj} \neq 0 \). Note that \( U_{kj} \) is the only nonzero element in \( u_k \). Hence, \( u_k^T \alpha = u_{kj} \alpha_j \neq 0 \), \( v_k^T \beta = 0^T \beta = 0 \), contradicts with (57) and \( k \in I \). Therefore, \( U_{kj} = 0 \) for all \( k \in \{3, \ldots, 2 + n + |T|\} \cap I \). Now we consider two scenarios.

In the first scenario, \( \{1, 2\} \cap I = \emptyset \), then \( (U_2)_j \) has the only one nonzero element which is from \( -I_S, (U_2)_j = (V_2)_j \), and hence \( \beta_j(V_2)_j \in \text{span}(U_2) \).

In the second scenario, \( \{1, 2\} \cap I \neq \emptyset \). WLOG, let \( 1 \in I \). We then have \( \sum_{k \in [n]} \alpha_k = 1^T \alpha = u_1^T \alpha = v_1^T \beta = 0^T \beta = 0 \), where the third equality is due to (57). From the above analysis, we already have \( \alpha_j = \beta_j > 0 \), which implies that \( \exists k \neq j \) such that \( \alpha_k < 0 \). We now prove that for \( U_k \), only zero elements from blocks \( I_{n \times n} \) and \( -I_T \) are included in \( (U_T)_k \). This can be done with the same logic as that in above when we show only zero elements from blocks \( I_{n \times n} \) and \( -I_T \) are included in \( (U_T)_j \). We next show that for \( U_k \), only zero elements from blocks \( -I_S \) is included in \( (U_T)_k \).

Assume to the contrary, i.e., \( \exists l \in \{2 + n + |T| + 1, \ldots, 2 + 2n\} \cap I \) such that \( u_{lk} \neq 0 \). Notice that \( u_{lk} = -1 \) and \( v_{lk} = -1 \) are the only nonzero elements in \( u_l \) and \( v_l \), respectively. We have \( u_l^T \alpha = -\alpha_k \), \( v_l^T \alpha = -\beta_k \), and hence \( u_l^T \alpha \neq v_l^T \alpha \) since \( \alpha_k < 0 \) and \( \beta \geq 0 \). It contradicts with (57), and we have that only zero elements from blocks \( -I_S \) are included in \( (U_T)_k \). Therefore, from the two observations above we can conclude that \( (U_T)_k \) has all elements as zero from the blocks \( I_{n \times n}, -I_T \) and \( -I_S \). In other words, \( (U_T)_k \) and \( (U_T)_j \) only differ at \( u_{lk} = 0, u_{lj} = -1 \). We can then easily have \( \beta_j(V_2)_j = \beta_j(U_2) - \beta_j(U_T)_k \in \text{span}(U_T) \). □

**Proof of Proposition 8**

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

\[
g(x, z) = \min \sum_{s, j \in [n]} b_{sj} y_{sj}
\]

s.t. \( \begin{bmatrix} U^1 & \cdots & U^n \\ I_{n^2 \times n^2} \end{bmatrix} y - \begin{bmatrix} I_{n \times n} \\ 0_{n^2 \times n^2} \end{bmatrix} z \geq \begin{bmatrix} -x \\ 0_{n^2 \times 1} \end{bmatrix} \).

For any \( s \in [n] \), the matrix \( U^s \in \mathbb{R}^{n \times n} \) has \( e_s - e_j \) as its \( j \)-th column, \( j \in [n] \). Denote \( U^0 = [U^1 \cdots U^n], V^0 = I_{n \times n} \) and \( U = \begin{bmatrix} U^0 \\ I_{n^2 \times n^2} \end{bmatrix} \in \mathbb{R}^{(n + n^2) \times n^2}, V = \begin{bmatrix} V^0 \\ 0_{n^2 \times n^2} \end{bmatrix} \). Obviously, \( \text{rank}(U) = n^2 \) which is less than the number of rows in \( U \). Therefore, to complete the proof, we now show that \( U, V \) meet the second condition in Theorem 3.

Consider any index set \( I \) such that \( |I| = \text{rank}(U) + 1 = n^2 + 1, \text{rank}(U_T) = n^2 \). Denote \( I_0 = I \cap [n], I_1 = I \setminus I_0 \), then the rows of \( U_{I_0} \) (or \( V_{I_0} \)) are extracted from \( U^0 \) (or \( V^0 \)); the rows of \( U_{I_1} \) (or \( V_{I_1} \)) are extracted from \( I_{n^2 \times n^2} \) (or \( 0_{n^2 \times n^2} \)).

We first let the column index set \( J_0 \) be such that the submatrix \( U_{I_0,J_0} = 0 \), and let \( J_1 = [n^2] \setminus J_0 \). Hence, \( U_T \) can be decomposed into four submatrices \( U_{I_0,J_0}, U_{I_0,J_1}, U_{I_1,J_0}, U_{I_1,J_1} \). Recalling that \( U_{I_1} \) is a submatrix of \( I \), there is exactly one entry being one in its each row, and at most one
entry being 1 in its each column. Hence, in \( U_{I_1} \), the number of columns being 0 is \( n^2 - |I_1| = n^2 - (|I| - |I_0|) = |I_0| - 1 \). Noticing \( U_I \) is full column rank and \( U_{I_0,J_0} = 0 \), all of the \( |I_0| - 1 \) zero columns in \( U_{I_1} \) must be in \( U_{I_1,J_1} \). Denote the index set \( K_1 \) as the set of column index for those zero columns in \( U_{I_1} \), and \( K_2 = J_1 \setminus K_1 \). Then \( U_{I_1,J_1} \) can be further decomposed into two submatrices \( U_{I_1,K_1}, U_{I_1,K_2} \) where \( U_{I_1,K_1} = 0_{|I_1| \times (|I_0| - 1)} \).

Since \( U_{I_0,K_1} \in \mathbb{R}^{|I_0| \times (|I_0| - 1)} \) and it is of full column rank (otherwise it contradicts with \( U_I \) being full column rank and \( U_{I_2,K_1} = 0 \)), we have that \( \text{null}(U_{I_0,K_1}^T) \) is of dimension 1. Recalling that \( U_{I_0,K_1} \) is a submatrix of \( U^0 \), each column can only be either \( e_s \) or \( e_{s_1} - e_{s_2} \) for some \( s, s_1, s_2 \in [|I_0|] \). Let \( \mathcal{N}_j \subseteq [|I_0|] \) be the index set \( \{ s \mid \text{the s-th entry of } (U_{I_0})_j \text{ is non-zero} \} \) for any \( j \in K_1 \). We observe that

\[
\text{null}(U_{I_0,K_1}^T) = \left\{ \gamma \mid \forall j \in K_1 \text{ with } |N_j| = 1 : \gamma_s = 0 \quad \text{for } s \in \mathcal{N}_j \right\}.
\] (58)

Consider any nonzero \( \gamma \in \text{null}(U_{I_0,K_1}^T) \), we now prove that \( \exists s_1, s_2 \) 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., \( \exists s_1, s_2 \) such that \( \gamma_{s_1} \gamma_{s_2} \neq 0 \) and \( \gamma_{s_1} \neq \gamma_{s_2} \). We construct a vector \( \gamma \) such that \( \gamma_i = 0 \) for all \( i \) such that \( \gamma_i = 0 \), and \( \gamma_i = 1 \) for all \( i \) such that \( \gamma_i \neq 0 \). As \( \gamma \) satisfies the condition in (58), so does \( \hat{\gamma} \), and hence \( \hat{\gamma} \in \text{null}(U_{I_0,K_1}^T) \). Nevertheless, \( \gamma \) and \( \hat{\gamma} \) are obviously linearly independent, and hence we have contradiction to that \( \text{null}(U_{I_0,K_1}^T) \) is of dimension 1. Therefore, we can conclude that all nonzero elements in \( \gamma \) have the same value.

Consider any \( \eta^0 \in \mathbb{R}^{|I_0|}, \eta^1 \in \mathbb{R}^{|I_1|}, \eta = (\eta^0, \eta^1) \) such that \( U_I^T \eta = 0 \). It implies \( 0 = U_{I,K_1}^T \eta = U_{I_0,K_1}^T \eta^0 + U_{I_1,K_1}^T \eta^1 = U_{I_0,K_1}^T \eta^0 \), where the last equality is due to \( U_{I_1,K_1} = 0 \). Hence, \( \eta^0 \in \text{null}(U_{I_0,K_1}^T) \), whose dimension has been shown as 1. Therefore, \( \eta^0 = k \gamma \) for some \( k \in \mathbb{R} \). As we have shown above, all nonzero elements in \( \eta \) are equal, WLOG, we can have \( \eta^0 \geq 0 \). We are now ready to verify the second condition in Theorem 3.

Given any \( \beta \in \mathbb{R}^n_+ \) with \( V_I \beta \in \text{span}(U_I) \), as \( \eta \in \text{null}(U_I^T) \), we have \( 0 = \eta^T V_I \beta = (\eta^0)^T V_{I_0} \beta + (\eta^1)^T V_{I_1} \beta = (\eta^0)^T V_{I_0} \beta = \sum_{i \in [n]} \beta_i (\eta^0)^T (V_{I_0})_i \) where the third equality is due to \( V_{I_1} = 0 \). Since \( \eta^0 \geq 0, V_{I_0} \geq 0, \beta \geq 0 \), we have that \( \eta^0 \beta_i (V_{I_0})_i = \beta_i (\eta^0)^T (V_{I_0})_i = 0 \) for all \( i \in [n] \). Recall that \( \text{null}(U_I^T) \) is of dimension 1, we then have \( \beta_i (V_{I_0})_i \in \text{span}(U_I) \), and the second condition in Theorem 3 is satisfied. Thus \( g(x, z) \) is supermodular in \( z \) for all \( x \), and the reformulation is a simple corollary of Theorem 1.

**Proof of Proposition 9**

Let \( \hat{z} \in \mathbb{R}^n \) be such that \( \hat{z}_i = \xi_i z_i \) for all \( i \in [n] \), and we define \( \hat{g}(x, \hat{z}) = \min \left\{ 1^T y \mid y \geq \sum_{s,t} (\hat{z}_s - x_s), j \in [t], t \in [n] \right\} \). Notice that \( \hat{g}(x, \hat{z}) = \hat{g}(x, (\xi_1 z_1, \ldots, \xi_n z_n)) = g(x, \xi, z) \) defined by Equation (24). To prove the supermodularity of \( g \), we first show \( \hat{g}(x, \hat{z}) \) is supermodular in \( \hat{z} \), and then prove that \( g(x, \xi, z) = \hat{g}(x, (\xi_1 z_1, \ldots, \xi_n z_n)) \) is supermodular in \( (\xi_1, \ldots, \xi_n, z_1, \ldots, z_n) \).
To show the supermodularity of $\tilde{g}$, we first rewrite the problem defining $\tilde{g}$ in its matrix form, i.e.,

$$\tilde{g}(x, \tilde{z}) = \min \left\{ 1^T y \mid Uy - V\tilde{z} \geq -Wx \right\},$$

where $U = \begin{bmatrix} U^1 \\ \vdots \\ U^n \\ U_{n+1} \end{bmatrix} \in \mathbb{R}^{n^2+3n \times n}$, $V = W = \begin{bmatrix} V^1 \\ \vdots \\ V^n \end{bmatrix} \in \mathbb{R}^{n^2+3n \times n}$ are such that

$$\tilde{U}^t \in \mathbb{R}^{t \times n}$$

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

$$\tilde{V}^t \in \mathbb{R}^{t \times n}$$

are with elements of $\tilde{v}_{js}^t = \begin{cases} 1 & \text{if } j \leq s \leq t, \text{ for } j \in [t], s, t \in [n], \\ 0 & \text{otherwise} \end{cases}$

$$\tilde{U}^{n+1} = I_{n \times n}, \tilde{V}^{n+1} = 0_{n \times n}.$$

We prove $\tilde{g}(x, \tilde{z})$ is supermodular in $\tilde{z}$ by verify that $U, V$ satisfy the condition in Theorem 3.

To this end, consider any $I \subseteq [(n^2 + 3n)/2], \beta \in \mathbb{R}_+^n$ with $|I| = n + 1$, rank($U_I$) = $n$, and $V_I\beta \in \text{span}(U_I)$. Note that rank($U$) = $n$, and each row of $U_I \in \mathbb{R}^{(n+1) \times n}$ has only one nonzero element which takes the value of 1. Therefore, there exists $\omega \in [n]$ such that $U$ has two row vectors being $e_\omega$, and exactly one row vector being $e_i$ for each $i \in [n] \setminus \{\omega\}$. WLOG, we let $RI_1, \ldots, RI_{n+1}$ be the distinct row indices such that $I = \{RI_1, \ldots, RI_{n+1}\}$, $u_{RI_i} = e_i \ \forall i \in [n]$, $u_{RI_{n+1}} = e_\omega$, and $RI_\omega < RI_{n+1}$. Moreover, for notational brevity, we arrange the rows in $U_I, V_I$ with the order of $RI_1, \ldots, RI_{n+1}$, which would not change the satisfaction/violation of the condition in Theorem 3. Therefore, $U_I = \begin{bmatrix} I \\ e_\omega \end{bmatrix}$. In this case, for any $\alpha \in \mathbb{R}^n$, $U_I \alpha = \begin{bmatrix} \alpha \\ \alpha_\omega \end{bmatrix}$. This implies that, given any $\gamma \in \mathbb{R}^{n+1}$, we have $\gamma \in \text{span}(U)$ if and only if $\gamma_\omega = \gamma_{n+1}$. Therefore, consider any $\beta$ with $V_I\beta \in \text{span}(U_I)$, we know $\beta^T_{RI_\omega} \beta = \beta^T_{RI_{n+1}} \beta$. Our objective is to show $\beta_i(V_I)_i \in \text{span}(U_I)$, which is equivalent to $\beta_i v_{RI_\omega,i} = \beta_i v_{RI_{n+1},i}, \ \forall i \in [n]$. To see this, we consider two cases.

- Case 1: both $u_{RI_{n+1}}$ and $u_{RI_{n+1}}$ The $RI_\omega$-th and $RI_{n+1}$-th rows in $\tilde{U}^\omega$, respectively. By the structure of $\tilde{V}^\omega$, we know for all $s \in [n]$, $v_{RI_\omega,s} = \tilde{v}_{j_\omega,s}^\omega = \begin{cases} 1 & \text{if } s = j_\omega, \ldots, \omega \\ 0 & \text{otherwise} \end{cases}$

$$v_{RI_{n+1},s} = \tilde{v}_{j_{n+1},s}^\omega = \begin{cases} 1 & \text{if } s = j_{n+1}, \ldots, \omega \\ 0 & \text{otherwise} \end{cases}$$

In this case, $\beta^T_{RI_\omega} \beta = \beta^T_{RI_{n+1}} \beta$ implies $\sum_{j = j_\omega}^{\omega} \beta_j = \sum_{j = j_{n+1}}^{\omega} \beta_j$; and hence $\beta_j = 0, \ \forall j \in \{j_\omega, \ldots, j_{n+1} - 1\}$ since $\beta \geq 0$. 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: $\mathbf{u}_{RL}$ is extracted from $\bar{U}$ while $\mathbf{u}_{RI_{n+1}}$ is extracted from $\bar{U}^n$. The submatrix $V_{n+1} = 0_{n \times n}$ implies in this case $v_{RI_{n+1}} = 0$. Hence, $v_{RL}^T \beta = v_{RI_{n+1}}^T \beta$ implies $0 = v_{RL}^T \beta = \sum_{i \in \mathbb{N}} \beta_i v_{RL,i}$. Since $v_{RL} \geq 0$ and $\beta \geq 0$, we then have $\beta_i v_{RL,i} = 0 = \beta_i v_{RI_{n+1},i}$ for all $i \in [n]$.

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

We first consider argument as the pair $(\xi_i, z_i)$ for some $i \in [n]$ and fix all $\xi_s, z_s$ with $s \in [n] \setminus \{i\}$. As all the remaining elements are fixed, we define $g^\hat{} : \mathbb{R}^2 \to \mathbb{R}$ and $h : \mathbb{R} \to \mathbb{R}$ to be such that $g^\hat{}(\xi_i, z_i) = h(\hat{z}_i) = h(\xi_i z_i) = \hat{g}(\mathbf{x}, (\xi_1 z_1, \ldots, \xi_n z_n)) = g(\mathbf{x}, \xi_i, z_i)$. Hence, it is equivalent to show that $g^\hat{}$, as a function of $\xi_i, z_i$, is supermodular in its arguments. To this end, we first observe that $\xi_i z_i$ is increasing and supermodular in $(\xi_i, z_i)$ (recall that $\xi_i, z_i \geq 0$). Further, as a second-stage cost function, $\hat{g}(\mathbf{x}, \hat{z})$ has been shown as convex in $\hat{z}$ 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 [24]. Therefore, the supermodularity of $g^\hat{}$ follows as a corollary of Lemma $\mathbb{3}$ and $g(\mathbf{x}, \xi_i, z)$ is supermodular in $(\xi_i, z_i)$ for all $i \in [n]$. 

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

$$g^{ij}(\xi_i' \land \xi_i'', z_i' \lor z_i'') + g^{ij}(\xi_i' \lor \xi_i'', z_i' \lor z_i'') - g^{ij}(\xi_i', z_i') - g^{ij}(\xi_i'', z_i'')$$

$$= \hat{g}(\mathbf{x}, \hat{z}_{\text{min}}) + \hat{g}(\mathbf{x}, \hat{z}_{\text{max}}) - \hat{g}(\mathbf{x}, \hat{z}') - \hat{g}(\mathbf{x}, \hat{z}'')$$

$$= \hat{g}(\mathbf{x}, \hat{z} + \xi_j z_j' e_j) + \hat{g}(\mathbf{x}, \hat{z} + z_i e_i + \xi_j z_j' e_j) - \hat{g}(\mathbf{x}, \hat{z} + \xi_j z_j' e_j) - \hat{g}(\mathbf{x}, \hat{z} + z_i e_i + \xi_j z_j' e_j)$$

$$\geq 0,$$

where the inequality holds because $g(\mathbf{x}, \hat{z})$ is supermodular in $\hat{z}$. Hence, $g^{ij}$ is supermodular and therefore $g(\mathbf{x}, \xi_i, z_j)$ is supermodular in $(\xi_i, z_j)$.

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

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

$$\min \ \nu^T l$$

s.t. $R_k^T l \geq \sum_{i \in [3n+1]} p_i^k 1^T y^{k,i}, \ k \in [K]$ 

$$y^{k,i} \geq \sum_{s=j}^t (\xi^{k,i} z_s - x_s), \ j \in [t], t \in [n], k \in [K], i \in [3n+1]$$

$$y^{k,i} \geq 0, \ k \in [K], i \in [3n+1]$$

$$l \geq 0, x \in X_{\text{app}},$$

where $p_i^k, \xi^{k,i}, z^{k,i}, k \in [K], i \in [3n+1]$ are the output of Algorithm 1 given the ambiguity sets $G^k, k \in [K]$ defined as $G^k = \{P^k \mid \Pi_{\xi} P^k \in F_\xi^k, \Pi_z P^k \in F_z^k\}$, where $\Pi_{\xi} P^k, \Pi_z P^k$ denotes the marginal distribution of $\xi$ and $z$, respectively under $P^k$. $F_\xi^k$ is the conditional ambiguity set of $F_\xi$ when $k$ is realized as $k$, and $F_z^k$ is defined by 4.

**Proof of Theorem 7**

The constraint of the second-stage problem 27 can be represented as $U y - V z \geq -W x + v^0$, where $U = \begin{bmatrix} -I \\ I \\ -A \end{bmatrix}, V = \begin{bmatrix} -I \\ 0 \\ 0 \end{bmatrix}$ and $W = \begin{bmatrix} 0 \\ 0 \\ I \end{bmatrix}$.

We first prove the “if” direction. Suppose the condition for $A$ in this theorem is satisfied. By Theorem 3, whether $U, V$ lead to supermodularity of $g$ is equivalent to whether $-U, -V$ do so. Therefore, here we verify the supermodularity based on $-U, -V$. Observing that $-U$ and $-V$ have the structure as in Proposition 5 with $U^c = \begin{bmatrix} -I \\ A \end{bmatrix}$, we now show that every $2 \times 3$ submatrix of $U^c$ contains at least one pair of columns which are linearly dependent. If both rows of the $2 \times 3$ submatrix are extracted from $A$, then this submatrix must have two linearly dependent columns by the assumption on $A$. If at least one of the rows are from $-I$, since the rows from $-I$ 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 $A \in \mathbb{R}^{2 \times 3}_+$. Assume the contrary, i.e., every two columns in $A$ are in different directions. Given that $A \geq 0$, there must be one column in $A$ being a conical combination of the other two columns. WLOG, let $A_3$ be a conical combination of $A_1, A_2$. We remark that multiplying any strictly positive constant by a row/column in $A$, or switching rows, or switching columns does not affect whether the corresponding function $g$ is supermodular. Therefore, we can make the following simplification on $A$. Since $A_1, A_2$ are linearly independent, WLOG, we can let $A = \begin{bmatrix} 1 & a & c \\ b & 1 & d \end{bmatrix}$ with $ab < 1$. Since $A_3$ is a conical combination of $A_1, A_2$, we have $cd > 0$; WLOG, we can let $d = 1$, i.e., $A = \begin{bmatrix} 1 & a & c \\ b & 1 & 1 \end{bmatrix}$. Multiplying the first row by $1/c$, and then multiplying the first column by $c$, we have
\[ A = \begin{bmatrix} 1 & a/c & 1 \\ bc & 1 & 1 \end{bmatrix}. \] Let \( a/c, bc \) be the new \( a, b \), we have \( A = \begin{bmatrix} 1 & a & 1 \\ b & 1 & 1 \end{bmatrix} \) with \( ab < 1 \). Again, since \( A_3 \) is a conical combination of \( A_1, A_2 \), we have either \( a, b < 1 \) or \( a, b > 1 \). Together with \( ab < 1 \), we know \( a, b < 1 \). In summary, WLOG, we let \( A = \begin{bmatrix} 1 & a & 1 \\ b & 1 & 1 \end{bmatrix} \) with \( ab \in (0, 1) \).

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

\[
\bar{g}(x, z) = \min \left\{ -p^\top y \mid Ay \leq x, 0 \leq y \leq z \right\} = \min \left\{ -y_1 - y_2 - \epsilon y_3 \mid \begin{array}{c} y_1 + ay_2 + y_3 \leq 1 - ab \\ by_1 + y_2 + y_3 \leq 1 - ab \\ (0, 0, 0) \leq (y_1, y_2, y_3) \leq (z_1, z_2, z_3) \end{array} \right\}.
\]

Hence,

\[
\bar{g}(x, z') = \min \left\{ -y_1 - \epsilon y_3 \mid \begin{array}{c} y_1 + y_3 \leq 1 - ab, \\ 0 \leq y_1 \leq 1 - a, y_2 = 0, y_3 \geq 0 \end{array} \right\},
\]

\[
\bar{g}(x, z'') = \min \left\{ -y_2 - \epsilon y_3 \mid \begin{array}{c} y_2 + y_3 \leq 1 - ab, \\ y_1 = 0, 0 \leq y_2 \leq 1 - b, y_3 \geq 0 \end{array} \right\},
\]

\[
\bar{g}(x, z^\wedge) = \min \left\{ -\epsilon y_3 \mid \begin{array}{c} y_3 \leq 1 - ab, \\ y_1 = y_2 = 0, y_3 \geq 0 \end{array} \right\},
\]

\[
\bar{g}(x, z^\vee) = \min \left\{ -y_1 - y_2 - \epsilon y_3 \mid \begin{array}{c} y_1 + ay_2 + y_3 \leq 1 - ab, \\ by_1 + y_2 + y_3 \leq 1 - ab, \\ 0 \leq y_1 \leq 1 - a, 0 \leq y_2 \leq 1 - b, y_3 \geq 0 \end{array} \right\}.
\]

Since \( 0 < \epsilon < 1 \), in the optimization problem for \( \bar{g}(x, z') \), 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 \( \bar{g}(x, z'') \), the optimal \( y = (0, 1 - b, b(1 - a)) \); in that for \( \bar{g}(x, z^\wedge) \), the optimal \( y = (0, 0, 1 - ab) \); in that for \( \bar{g}(x, z^\vee) \), the optimal \( y = (1 - a, 1 - b, 0) \). We then have

\[
\bar{g}(x, z') + \bar{g}(x, z'') - \bar{g}(x, z^\wedge) - \bar{g}(x, z^\vee) = -((1 - a + \epsilon a(1 - b)) + (1 - b + \epsilon b(1 - a)) - \epsilon(1 - ab) - (1 - a + 1 - b)) = \epsilon(1 - a)(1 - b) > 0,
\]

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

For the general case of \( A \in \mathbb{R}_+^{n \times n} \), we can prove the result by the same contradiction. WLOG, we assume the \( 2 \times 3 \) submatrix of \( A \), 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 \( y \). We then have \( \bar{g}(x, z) \) with exactly the same expression in Equation (59). Therefore, we still have \( \bar{g}(x, z^\wedge) + \bar{g}(x, z^\vee) < \bar{g}(x, z') + \bar{g}(x, z'') \).
Proof of Corollary 5
We first prove the “if” direction using Theorem 7. Consider any $2 \times 3$ submatrix of $A$, which, WLOG, is $C = A_{\{1,2\},\{1,2,3\}}$. 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 $C$ has two zero elements, and hence $C$ 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 $C$ has two zero elements and hence $C$ 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 $C$ has at least one pair of columns which are linearly dependent. In summary, $C$ always have at least one pair of columns which are linearly dependent. □

Lemma 4 (Chen et al. (2021)) Consider any matrix $U \in \mathbb{R}^{r \times m}$ with $\text{rank}(U) < r$. Suppose that system
$$\begin{align*}
Ux & \leq \underline{c} \\
-Ux & \leq \overline{c}
\end{align*}$$
is infeasible. Then there exists $I \subseteq [r]$ with $|I| = \text{rank}(U) + 1$ and $\text{rank}(U_I) = \text{rank}(U)$ such that system
$$\begin{align*}
U_I x & \leq \underline{c}_I \\
-U_I x & \leq \overline{c}_I
\end{align*}$$
is also infeasible.

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