A Column Generation Scheme for Distributionally Robust Multi-Item Newsvendor Problems

Shanshan Wang, Erick Delage
GERAD & Department of Decision Sciences, HEC Montréal, Montréal, QC Canada H3T 2A7
shanshan.wang@hec.ca,erick.delage@hec.ca

In this paper, we study a distributionally robust multi-item newsvendor problem, where the demand distribution is unknown but specified with a general event-wise ambiguity set. Using the event-wise affine decision rules, we can obtain a conservative approximation formulation of the problem, which can typically be further reformulated as a linear program. In order to efficiently solve the resulting large-scale linear program, we develop a column generation based decomposition scheme and speed up the computational efficiency by exploiting a special column selection strategy and stopping early based on a Karush–Kuhn–Tucker condition test. Focusing on the Wasserstein ambiguity set and the event-wise mean absolute deviation set, a computational study demonstrates the computational efficiency of the proposed algorithm over a set of 540 randomly generated instances, significantly outperforming the commercial solver and a Benders decomposition method.

Key words: Distributionally robust optimization, column generation, multi-item newsvendor problem, event-wise ambiguity set

History: November 8, 2021

1. Introduction

The multi-item newsvendor problem consists in deciding the order quantities for products that should be stocked before observing the uncertain demand so as to maximize the expected profit. The newsvendor problem is a fundamental operations management problem with various applications (see, e.g., a recent review by Qin et al. 2011), such as the reservation of operating room time in hospital (e.g., Rahimian et al. 2019) and production of fashion goods (e.g., Donohue 2000). In the latter problem, a decision maker faces opportunity costs of lost sales if the order quantity is lower than its demand, otherwise, the unsold inventory cost will incur. Thus, it is crucial to find a solution to this problem that optimally trades off between understocking and overstocking.

In practice, the exact distribution of each item's demand may not be accessible, and only historical data is available, which can be used to estimate the distribution. This has typically been addressed by using the stochastic programming paradigm (Shapiro et al.

2014). However, the solutions of stochastic programming problems can be sensitive to the misestimation of the probability distribution thus producing solutions that may result in a poor out-of-sample performance. In distributionally robust optimization (DRO), one instead seeks a solution that performs best under the worst-case distribution within an ambiguity set of distributions (e.g., Delage and Ye 2010, Wiesemann et al. 2014, Mohajerin Esfahani and Kuhn 2018). With proper design of the ambiguity set, the solution of a DRO can be robust to estimation error while converging to the true optimum as more distribution information is obtained. This motivates us to develop a distributionally robust optimization model over the event-wise ambiguity set proposed by Chen et al. (2020). This framework integrates stochastic programming and robust optimization, where the uncertainty consists of discrete random scenarios and continuous random variables. The event-wise ambiguity set can capture several traditional statistic-based ambiguity sets such as Wasserstein ambiguity set (e.g., Mohajerin Esfahani and Kuhn 2018, Gao and Kleywegt 2016) and ϕ-divergence measure-based set (e.g., Ben-Tal et al. 2013, Bayraksan and Love 2015).

To obtain a tractable conservative approximation of hard two-stage DRO problems, a common way is to use approximate policies using linear decision rules (LDR), which leads to formulations that can be solved with state-of-art commercial solvers for small/mid-sized problems. Unfortunately, larger sized problems, which arise in realistic decision-making settings, still constitutes a great numerical challenge. Most of the recent literature has focused on increasing the quality of LDR approximation by making the decision rules more flexible (e.g., Kuhn et al. 2011, Postek and den Hertog 2016, Gauvin et al. 2017, Bertsimas et al. 2019). Unfortunately, such improvement necessarily comes at the price of additional computational burden, even when most of this flexibility remains unused at optimum. For this reason, this paper proposes a column generation (CG) based decomposition method to accelerate the resolution of large-scale LDR-based DRO models when an optimal sparse LDR solution exists. More specifically, our contributions are summarized as follows:

- We propose for the first time a CG solution scheme for the conservative LDR approximation of a two-stage DRO model. Specifically, we address a classical distributionally robust (DR) multi-item newsvendor problem where the ambiguity takes the form of a general event-wise ambiguity set. In doing so, we identify which are the columns needed in the restricted master problem (RMP) to ensure its feasibility, and design a special column selection strategy to accelerate convergence.

- We propose a novel early stopping criterion based on the Karush–Kuhn–Tucker (KKT) condition test that can be used in CG when the RMP is highly degenerate. Our numerical experiments involving a Wasserstein-based DRO model showed that the early stopping criterion can significantly improve solution time.
- We perform a comprehensive numerical study involving up to 20 items and 8000 scenarios, and using two popular ambiguity sets, namely a set based on the Wasserstein metric and a set based on a mixture of distributions for which the mean and support are known while the mean absolute deviation is upper bounded. In the latter case, CG decreases the solution time by an average of about 67% when compared to CPLEX, while this reduction is of about 32% in the former case. The acceleration strategies are also shown to be responsible for most of this improvements.

The remainder of this paper is organized as follows. Section 2 reviews the literature on algorithms for solving DRO problems and DR multi-item newsvendor problems. Section 3 formulates the DR multi-item newsvendor problem with event-wise ambiguity set as a linear programming model by using the event-wise affine decision rules technique. Subsequently, we provide reformulations under two specific ambiguity sets. We describe our column generation decomposition scheme to solve the problem in Section 4, and further propose several strategies to accelerate the algorithm. In Section 5, we give some computational results on the newsvendor problems and demonstrate the efficiency of CG scheme developed in this paper. Section 6 concludes this paper with a summary of the important findings.

Notation: We use boldface lowercase (e.g. $\boldsymbol{x}$) characters to denote vectors. We use $\mathscr{P}_0(\mathbb{R}^I)$ to represent the set of all probability distributions on $\mathbb{R}^I$. Given a function $f : \mathbb{R}^m \mapsto \mathbb{R}$, the conjugate of f is defined as $f^*(\boldsymbol{z}) := \sup_{\boldsymbol{x}\in\mathbb{R}^m} \boldsymbol{x}^\top \boldsymbol{z} - f(\boldsymbol{x})$. The support function of a set $\Xi \subseteq \mathbb{R}^m$ is denoted by $\delta^*(\boldsymbol{z} \mid \Xi)$ and defined as $\delta^*(\boldsymbol{z} \mid \Xi) := \sup_{\boldsymbol{\xi}\in\Xi} \boldsymbol{\xi}^\top \boldsymbol{z}$.

2. Literature Review

The concept of DRO was first introduced by Scarf (1957) for a single-item newsvendor problem where the uncertainty set includes all probability distributions with a given mean and standard deviation. Since then, a number of variations of this framework have been widely investigated in the literature (e.g., Dupacová 2008, Delage and Ye 2010, Goh and Sim 2010, Wiesemann et al. 2014, Mohajerin Esfahani and Kuhn 2018, Bertsimas et al. 2019, Rahimian and Mehrotra 2019, Delage and Saif 2021). Most of the popular DRO problem formulations can be reformulated as finite dimensional convex programs by using duality theory (see, e.g., Goh and Sim 2010, Ben-Tal et al. 2013, Postek et al. 2018) and are solved by an off-the-shelf commercial solver (e.g., CPLEX, GUROBI). Alternatively, many other DRO formulations are known to be intractable, thus requiring the use of more elaborated schemes if exact solutions are needed. Mehrotra and Papp (2014) considered a DRO problem whose uncertainty set is defined using bounds on the moments of arbitrary order. They solved it using a central cutting surface algorithm combined with a randomized column generation method. A similar approach was used in Luo and Mehrotra (2019), where the authors studied intractable DRO problems that employ the type-1 Wasserstein ambiguity set. Intractable DRO problems also arise in finite outcome spaces. For instance, Wang et al. (2021) investigated a distributionally robust chance-constrained assignment problem and proposed a decomposition scheme that identifies a new probability distribution at each iteration to add to the finite relaxation of a mixed-integer semi-infinite reformulation. Namkoong and Duchi (2016) developed a two-player bandit mirror descent algorithm for a DRO empirical risk minimization problem with f-divergences. Finally, Chen et al. (2021) employed a progressive hedging method combined with a primal-dual hybrid gradient method to solve a two-stage DRO problems with discrete scenario support.

It is well known that two-stage DRO problems are typically intractable because of the difficulty of evaluating the worst-case expectation, which motivates the development of both exact and approximate solution method. Bansal et al. (2018) studied a two-stage mixed binary distributionally robust optimization with finite support. They developed an L-shaped algorithm which adopts a distribution separation procedure and parametric cuts and identifies conditions under which the algorithm converges finitely to an exact solution. Saif and Delage (2021) reformulated a two-stage data-driven DR capacitated facility location problem as a mixed-integer linear program with an exponential number of variables and constraints, and employed a column-and-constraint generation algorithm to solve it. Gamboa et al. (2021) investigated a two-stage Wasserstein-based DRO problem with right-hand-sided uncertainty and rectangular support. They identified a finite yet intractable reformulation and developed different decomposition schemes including a single-cut and multicut Benders decomposition (with and without regularization). Chen et al. (2021) proposed a decomposition algorithm based on Lagrange decomposition and primal-dual hybrid gradient method for solving a two-stage DRO problem after discretizing the outcome space. Unlike all the above work that has suggested exact solution schemes for solving two-stage DRO problems, it is much more common in the literature to circumvent the intractability of such models by solving a tractable conservative approximation based on LDRs (see, e.g., Kuhn et al. 2011, Bertsimas et al. 2019, Chen et al. 2020, Saif and Delage 2021). Under the right conditions, this approach produces a finite dimensional convex optimization model that can once again be solved using off-the-shelf software. In this paper, we propose for the first time a column generation algorithm to accelerate the solution time of conservative approximation of two-stage DRO problem obtained by exploiting such LDRs.

We now wish to summarize the work that has focused on the distributionally robust multi-item newsvendor problem. Gallego and Moon (1993) extended Scarf's work to the multi-item setting, which can be solved using a simple line search. Ardestani-Jaafari and Delage (2016) provided a tractable linear reformulation for the distributionally robust multi-item newsvendor problems when the demand is supported on a budgeted uncertainty set, and with information about the mean vector and first order partial moments. Ben-Tal et al. (2013) presented a tractable convex reformulation of the problem where the unknown probabilities of the different scenarios lie within a ϕ-divergences uncertainty set. Hanasusanto et al. (2015) considered demand to follow a mixture model where only the mean and covariance matrix are known for each distribution in the mixture. They conservatively approximated the problem using quadratic decision rules, which reduces it to a semidefinite program. Bertsimas et al. (2018) proposed to use hypothesis testing tools for constructing the ambiguity sets based on data and to approximate the resulting DRO model using conic programming. Kamburowski (2015) identified closed-form expressions for the worst-case and best-case order quantities in a single-item model with symmetric or symmetric and unimodal. Again for the single-item problem, Natarajan et al. (2018) obtained closed-form expression for the worst-case expected profit when distribution information includes mean, variance, semivariance, and second-order partitioned statistics to capture asymmetry. Finally, Rahimian et al. (2019) studied the multi-item model using total variation distance to a reference discrete distribution with a special attention to the regions of the demand that are critical to the optimal worst-case cost. Broadly speaking, there is no doubt that the newsvendor problem is the foundation of many operations management problems, which explains the high interest in the development of new formulations and solution schemes for this class of problems. In this regard, our work considers the event-wise ambiguity set proposed by Chen et al. (2020), which encompasses most of the ambiguity sets used in the literature. We appear to be the first to study the efficiency of decomposition schemes for very general model. This suggests that our findings should benefit a vast part of the DRO literature.

3. Two-Stage DR Multi-Item Newsvendor Problem

In a classical version of multi-item newsvendor problem, there are a set of different items $\mathcal{I} := \{1, \cdots, I\}$ with positive unit selling price $\boldsymbol{b} \in \mathbb{R}^I_+$ and ordering cost $\boldsymbol{c} \in \mathbb{R}^I_+$. For simplicity, we assume that the salvage value which is the value of unsold units at the end of the sales period is equal to zero. The decision maker places the orders with a total budget d before the random demand $\tilde{\boldsymbol{u}} \sim \mathbb{P}$, $\mathbb{P} \in \mathscr{P}_0(\mathbb{R}^I)$ is realized. We let a vector $\boldsymbol{x} \in \mathbb{R}^I_+$ denote the ordering quantities. After the demand is observed, the sold quantity is $\min\{x_i, u_i\}$ for each item $i \in \mathcal{I}$. The DRO model for the multi-item newsvendor problem can be expressed as follows:

$$\text{maximize} \inf_{\mathbb{P}\in\mathscr{F}} \mathbb{E}_{\mathbb{P}}\left[\sum_{i\in\mathcal{I}} b_i \min\{x_i, \tilde{u}_i\}\right] \tag{1a}$$

$$\text{subject to } \boldsymbol{c}^\top \boldsymbol{x} = d, \tag{1b}$$

$$x_i \geq 0, \qquad \forall i \in \mathcal{I}, \tag{1c}$$

where $\mathbb{P}$ is the joint probability distribution of $\tilde{\boldsymbol{u}}$, and $\mathscr{F}$ is an ambiguity set that consists of a family of probability distributions. Objective (1a) maximizes the worst-case expected operating revenue, constraint (1b) denotes that the total ordering cost is equal to the budget, and (1c) defines the domain of $\boldsymbol{x}$. Problem (1) can be further rewritten as the following two-stage minimization problem:

$$\text{minimize } -\boldsymbol{b}^\top \boldsymbol{x} + \sup_{\mathbb{P}\in\mathscr{F}} \mathbb{E}_{\mathbb{P}}\left[Q(\boldsymbol{x}, \boldsymbol{u})\right] \tag{2a}$$

$$\text{subject to (1b), (1c),}$$

where

$$Q(\boldsymbol{x}, \boldsymbol{u}) := \text{minimize} \sum_{i\in\mathcal{I}} b_i y_i \tag{3a}$$

$$\text{subject to } y_i \geq x_i - u_i, \qquad \forall i \in \mathcal{I}, \tag{3b}$$

$$y_i \geq 0, \qquad \forall i \in \mathcal{I} \tag{3c}$$

represents the revenue adjustments for unsold items. Problem (3) is always feasible for any $\boldsymbol{x} \in \mathbb{R}^I$ and $\boldsymbol{u} \in \mathbb{R}^I$.

3.1. Event-Wise Ambiguity Set

In this section, we use the event-wise ambiguity set to address the incomplete knowledge of demand distribution, which was proposed by Chen et al. (2020). More specifically, the event-wise ambiguity set can be expressed in the following form:

$$\mathscr{F} = \left\{ \mathbb{P} \in \mathscr{P}_0(\mathbb{R}^{I_z} \times \mathcal{S}) : \begin{array}{ll} (\tilde{\boldsymbol{z}}, \tilde{s}) \sim \mathbb{P},\ \boldsymbol{p} \in \mathcal{P} & \\ \mathbb{E}_{\mathbb{P}}[\tilde{\boldsymbol{z}} | \tilde{s} \in \mathcal{E}_k] \in \mathcal{Q}_k, & \forall k \in \mathcal{K} \\ \mathbb{P}[(\tilde{\boldsymbol{z}} \in \mathscr{W}_s | \tilde{s} = s)] = 1, & \forall s \in \mathcal{S} \\ \mathbb{P}(\tilde{s} = s) = p_s, & \forall s \in \mathcal{S} \end{array} \right\}$$

with set $\mathcal{K} = \{1, \cdots, K\}$ and $\mathcal{S} = \{1, \cdots, S\}$, the set $\mathcal{E}_k \subseteq \mathcal{S}$ for each $k \in \mathcal{K}$, is an event for which we have conditional moment information, and p_s is the probability of scenario s such that $\boldsymbol{p}$ belongs to an ambiguity set $\mathcal{P}$, satisfying $\mathcal{P} \subseteq \{\boldsymbol{p} \in \mathbb{R}^S | \boldsymbol{p} > 0,\ \sum_{s \in \mathcal{S}} p_s = 1\}$. As in Bertsimas et al. (2019) and Chen et al. (2020), the random variable $\tilde{\boldsymbol{z}} := (\tilde{\boldsymbol{u}}, \tilde{\boldsymbol{v}})$ includes the primary random variable $\tilde{\boldsymbol{u}} \in \mathbb{R}^I$ and the auxiliary random variable $\tilde{\boldsymbol{v}} \in \mathbb{R}^{I_v}$. Using the primary and auxiliary random variables has the ability of generating ambiguity sets with richer structure (Wiesemann et al. 2014). For simplicity of exposition, for each $s \in \mathcal{S}$, we focus on support sets of the form $\mathscr{W}_s = \left\{(\boldsymbol{u}, \boldsymbol{v}) \in \mathbb{R}^I \times \mathbb{R}^{I_v} \middle| \boldsymbol{u} \in [\underline{\boldsymbol{u}}, \bar{\boldsymbol{u}}],\ \boldsymbol{g}_s(\boldsymbol{u}) \leq \boldsymbol{v}\right\}$, where $\boldsymbol{g}_s : \mathbb{R}^I \mapsto \mathbb{R}^{I_v}$, although any bounded polyhedral set for $\boldsymbol{u}$ could be accommodate in what follows.

In this paper, we will focus on LP representable models, hence we make the following assumption.

Assumption 1. *The epigraph of $\boldsymbol{g}_s$ for all $s \in \mathcal{S}$, the set $\mathcal{Q}_k$ for all $k \in \mathcal{K}$ and $\mathcal{P}$ are polyhedral representable sets.*

We will also make use in due time of the following technical assumption about each $\boldsymbol{g}_s$.

Assumption 2. *For all $s \in \mathcal{S}$, we have a set $\bar{\mathcal{I}}^v \subseteq \{1, \ldots, I_v\}$ such that $g_{j,s} : \mathbb{R}^I \to \mathbb{R}_+$, for all $j \in \bar{\mathcal{I}}^v$, and $\sum_{j \in \bar{\mathcal{I}}^v} g_{j,s}(\boldsymbol{u}) \geq \tau_1 + \tau_2 \|\boldsymbol{u}\|_1$ for some $\tau_1 \in \mathbb{R}$ and $\tau_2 > 0$.*

Note that Assumption 2 can be made without loss of generality given that the fact that $\boldsymbol{u}$ is bounded implies that adding the constraint $\|\boldsymbol{u}\|_1 \leq v_{I_v+1}$ in $\mathscr{W}_s$ with $\mathbb{E}_{\mathbb{P}}[\tilde{v}_{I_v+1} | \tilde{s} \in \mathcal{E}_k] = \max_{\boldsymbol{u} \in [\underline{\boldsymbol{u}}, \bar{\boldsymbol{u}}]} \|\boldsymbol{u}\|_1$ in $\mathscr{F}$ effectively imposes that $\mathbb{E}_{\mathbb{P}}[\|\boldsymbol{u}\|_1 | \tilde{s} \in \mathcal{E}_k] \leq \max_{\boldsymbol{u} \in [\underline{\boldsymbol{u}}, \bar{\boldsymbol{u}}]} \|\boldsymbol{u}\|_1$ which is always redundant for $\mathscr{F}$. When these constraints are added, we can get that Assumption 2 is satisfied since $g_{I_v+1,s}(\boldsymbol{u}) = \|\boldsymbol{u}\|_1$.

The above ambiguity set $\mathcal{F}$ can capture several existing ambiguity sets such as the Wasserstein ambiguity set and the event-wise mean absolute deviation set.

EXAMPLE 1. (Wasserstein ambiguity set) The type-1 Wasserstein ambiguity set (see, e.g., Mohajerin Esfahani and Kuhn (2018)) is defined as

$$\mathscr{F}_W := \left\{ \mathbb{P} \in \mathscr{P}_0(\mathbb{R}^I) : d_W(\mathbb{P}, \hat{\mathbb{P}}) \le \theta, \mathbb{P}(\tilde{\boldsymbol{u}} \in [\underline{\boldsymbol{u}}, \bar{\boldsymbol{u}}]) = 1 \right\},$$

where the empirical distribution $\hat{\mathbb{P}} = \frac{1}{S} \sum_{s \in \mathcal{S}} \delta_{\hat{\boldsymbol{u}}_s}$, and Wasserstein distance

$$d_W(\mathbb{P}, \hat{\mathbb{P}}) := \inf_{\mathbb{Q} \in \mathcal{M}(\mathbb{P}, \hat{\mathbb{P}})} \left\{ \mathbb{E}_{\mathbb{Q}} \left[\|\tilde{\boldsymbol{u}} - \tilde{\boldsymbol{u}}^\dagger\|_p \right] : \begin{array}{c} \mathbb{Q} \text{ is the joint distribution of } \tilde{\boldsymbol{u}} \text{ and } \tilde{\boldsymbol{u}}^\dagger \\ \text{with marginals } \mathbb{P} \text{ and } \hat{\mathbb{P}}, \text{ respectively} \end{array} \right\},$$

which computes the minimum transportation cost of transforming the distribution $\mathbb{P}$ into the empirical one $\hat{\mathbb{P}}$. The parameter θ represents the radius of the Wasserstein ball centered at $\hat{\mathbb{P}}$. The set $\mathscr{F}_W$ can also be formulated as the following the event-wise ambiguity set:

$$\mathscr{F}_W = \left\{ \mathbb{P} \in \mathscr{P}_0(\mathbb{R}^{I+1} \times \mathcal{S}) : \begin{array}{ll} ((\tilde{\boldsymbol{u}}, \tilde{v}), \tilde{s}) \sim \mathbb{P}, & \\ \mathbb{E}_{\mathbb{P}}[\tilde{v} | \tilde{s} \in \mathcal{S}] = \theta, & \\ \mathbb{P}[(\tilde{\boldsymbol{u}}, \tilde{v}) \in \mathscr{W}_s] = 1, & \forall s \in \mathcal{S} \\ \mathbb{P}(\tilde{s} = s) = \dfrac{1}{S}, & \forall s \in \mathcal{S} \end{array} \right\}$$

where $\mathscr{W}_s = \{(\boldsymbol{u}, v) | \boldsymbol{u} \in [\underline{\boldsymbol{u}}, \bar{\boldsymbol{u}}], ||\boldsymbol{u} - \hat{\boldsymbol{u}}_s||_p \le v\}$, for $s \in \mathcal{S}$. In recent years, the Wasserstein ambiguity set has attracted a lot of attention in the field of optimization and machine learning, which exhibits a number of attractive properties (Mohajerin Esfahani and Kuhn 2018). For example, one can adjust the Wasserstein radius to control the robustness. Moreover, as the number of observed empirical data points increases, the radius can be set to ensure that the DRO problems under $\mathscr{F}_W$ converges to the true stochastic programming.

EXAMPLE 2. (Event-wise mean absolute deviation set) An event-wise mean absolution deviation set (see, e.g., Hanasusanto et al. (2015)) can be specified by $\mathscr{F}_M$ as follows:

$$\mathscr{F}_M := \sum_{s \in \mathcal{S}} p_s \mathscr{F}_M^s$$

where

$$\mathscr{F}_M^s := \left\{ \mathbb{P} \in \mathscr{P}_0(\mathbb{R}^I) : \mathbb{E}_{\mathbb{P}}[\tilde{\boldsymbol{u}}] = \boldsymbol{\mu}_s, \mathbb{E}_{\mathbb{P}}[|\tilde{\boldsymbol{u}} - \boldsymbol{\mu}_s|] \le \boldsymbol{\sigma}_s, \mathbb{P}(\tilde{\boldsymbol{u}} \in [\underline{\boldsymbol{u}}, \bar{\boldsymbol{u}}]) = 1 \right\}.$$

Hence, any distribution $\mathbb{P} \in \mathscr{F}_M$ is a mixture of S distinct distribution $\mathbb{P}_1, \cdots, \mathbb{P}_S$ with probabilities $p_1, \cdots, p_S \in [0,1], \sum_{s\in\mathcal{S}} p_s = 1$. Each distribution $\mathbb{P}_s$ belongs to the ambiguity set $\mathscr{F}_M^s$ that imposes an upper bound on the mean absolute deviation conditioned on knowing which $\mathbb{P}_s$ generates the realization. The set $\mathscr{F}_M$ can also be rewritten in the format of an event-wise ambiguity set:

$$\mathscr{F}_M = \left\{ \mathbb{P} \in \mathscr{P}_0(\mathbb{R}^{I+I_v} \times \mathcal{S}) : \begin{array}{ll} ((\tilde{\boldsymbol{u}}, \tilde{\boldsymbol{v}}), \tilde{s}) \sim \mathbb{P}, & \\ \mathbb{E}_{\mathbb{P}}[\tilde{\boldsymbol{u}} | \tilde{s} = s] = \boldsymbol{\mu}_s, & \forall s \in \mathcal{S} \\ \mathbb{E}_{\mathbb{P}}[\tilde{\boldsymbol{v}} | \tilde{s} = s] = \boldsymbol{\sigma}_s, & \forall s \in \mathcal{S} \\ \mathbb{P}[(\tilde{\boldsymbol{u}}, \tilde{\boldsymbol{v}}) \in \mathscr{W}_s | \tilde{s} = s] = 1, & \forall s \in \mathcal{S} \\ \mathbb{P}(\tilde{s} = s) = p_s, & \forall s \in \mathcal{S} \end{array} \right\}$$

where $\mathscr{W}_s = \{(\boldsymbol{u}, \boldsymbol{v}) | \boldsymbol{u} \in [\underline{\boldsymbol{u}}, \bar{\boldsymbol{u}}], |u_i - \mu_{is}| \le v_i, \forall i \in \mathcal{I}\}$. This set contains the multimodal distributions, which can be observed in many situations. For example, the decision maker has aggregate demand observations coming from many small customers with independent demands, then the distribution of demand is multimodal (Hanasusanto et al. 2015).

Note that problem (2) is known to be generally intractable due to the difficulty of evaluating the worst-case expectation. A common way of deriving tractable reformulations is to use approximate policies such as linear decision rules. Hence, in the following section, we will present a tractable approximation scheme based on the event-wise affine decision rule for solving problem (2).

3.2. Event-Wise Affine Decision Rules

Under a measurable decision function map, problem (2) can be rewritten as the following minimization problem:

$$\text{minimize} \; -\boldsymbol{b}^\top \boldsymbol{x} + \sup_{\mathbb{P} \in \mathscr{F}} \mathbb{E}_{\mathbb{P}} \left[\sum_{i \in \mathcal{I}} b_i y_i(\tilde{\boldsymbol{u}}, \tilde{\boldsymbol{v}}, \tilde{s}) \right] \tag{4a}$$

subject to (1b), (1c),

$$y_i(\boldsymbol{u}, \boldsymbol{v}, s) \ge x_i - u_i, \qquad \forall (\boldsymbol{u}, \boldsymbol{v}) \in \mathscr{W}_s, s \in \mathcal{S}, i \in \mathcal{I}, \tag{4b}$$

$$y_i(\boldsymbol{u}, \boldsymbol{v}, s) \ge 0, \qquad \forall (\boldsymbol{u}, \boldsymbol{v}) \in \mathscr{W}_s, s \in \mathcal{S}, i \in \mathcal{I}. \tag{4c}$$

The recourse decision $\boldsymbol{y}(\tilde{\boldsymbol{u}}, \tilde{\boldsymbol{v}}, \tilde{s})$ depends on the realized uncertainty $(\tilde{\boldsymbol{u}}, \tilde{\boldsymbol{v}})$ and the scenario. We now introduce the event-wise affine decision rule proposed by Chen et al. (2020) for the variable $y_i(\cdot)$ as follows:

$$\mathscr{L}_i(\mathcal{C}_i) := \left\{ y_i : \mathbb{R}^I \times \mathbb{R}^{I_v} \times \mathcal{S} \mapsto \mathbb{R} \;\middle|\; \begin{array}{l} y_i(\boldsymbol{u}, \boldsymbol{v}, s) = y_i^0(s) + \sum_{k \in \mathcal{I}} y_{ik}^u(s) u_k + \sum_{j \in \mathcal{I}_v} y_{ij}^v(s) v_j \\ \text{for some } (y_i^0, y_{ik}^u, y_{ij}^v) \in \mathscr{A}(\mathcal{C}_i),\ k \in \mathcal{I}, j \in \mathcal{I}_v \end{array} \right\},$$

where $\mathcal{C}_i$ is a collection of mutually exclusive and collectively exhaustive events, and

$$\mathscr{A}(\mathcal{C}_i) = \left\{ x : \mathcal{S} \mapsto \mathbb{R} \middle| \, \forall \mathcal{E} \in \mathcal{C}_i, \exists x^{\mathcal{E}} \in \mathbb{R} \text{ such as } x(s) = x^{\mathcal{E}} \right\}.$$

In this paper, the recourse decision $\boldsymbol{y}(\cdot,\cdot,\cdot)$ is a wait-and-see decision, i.e. $\mathcal{C}_i = \left\{ \{s\} \middle| s \in \mathcal{S} \right\}$, for $i \in \mathcal{I}$.

Using the event-wise affine decision rule, we can obtain the upper bound for problem (4) by solving the following problem:

$$\text{minimize } -\boldsymbol{b}^\top \boldsymbol{x} + \sup_{\mathbb{P} \in \mathscr{F}} \mathbb{E}_{\mathbb{P}} \left[\sum_{i \in \mathcal{I}} b_i y_i(\tilde{\boldsymbol{u}}, \tilde{\boldsymbol{v}}, \tilde{s}) \right] \tag{5a}$$

$$\text{subject to (1b), (1c), (4b), (4c),}$$

$$y_i \in \mathscr{L}_i(\mathcal{C}), \qquad \forall i \in \mathcal{I}. \tag{5b}$$

The following lemma is obtained by applying Theorem 1 in Chen et al. (2020), which shows that problem (5) can be reformulated as a classical robust optimization problem.

Lemma 1. *Assuming the Slater's condition holds on the worst-case expectation:*

$$\sup_{\mathbb{P} \in \mathscr{F}} \mathbb{E}_{\mathbb{P}} \left[\sum_{i \in \mathcal{I}} b_i y_i(\tilde{\boldsymbol{u}}, \tilde{\boldsymbol{v}}, \tilde{s}) \right],$$

then problem (5) is equivalent to the following robust optimization problem:

$$\text{minimize } -\boldsymbol{b}^\top \boldsymbol{x} + \gamma \tag{6a}$$

$$\text{subject to (1b), (1c), (4b), (4c), (5b),}$$

$$\gamma \geq \sum_{s \in \mathcal{S}} p_s \alpha_s + \sum_{k \in \mathcal{K}} \boldsymbol{\mu}_k^{u\top} \boldsymbol{\beta}_k^u + \sum_{k \in \mathcal{K}} \boldsymbol{\mu}_k^{v\top} \boldsymbol{\beta}_k^v, \qquad \forall \left\{ \frac{\boldsymbol{\mu}_k}{\sum_{s \in \mathcal{E}_k} p_s} \right\}_{k \in \mathcal{K}} \in \prod_{k \in \mathcal{K}} \mathcal{Q}_k, \boldsymbol{p} \in \mathcal{P}, \tag{6b}$$

$$\alpha_s + \sum_{k\in\mathcal{K}_s} \boldsymbol{u}^\top \boldsymbol{\beta}_k^u + \sum_{k\in\mathcal{K}_s} \boldsymbol{v}^\top \boldsymbol{\beta}_k^v \geq \boldsymbol{b}^\top \boldsymbol{y}(\boldsymbol{u},\boldsymbol{v},s), \quad \forall(\boldsymbol{u},\boldsymbol{v})\in\mathscr{W}_s,\ s\in\mathcal{S}, \tag{6c}$$

$$\gamma\in\mathbb{R}, \boldsymbol{\alpha}\in\mathbb{R}^S, \boldsymbol{\beta}_k^u\in\mathbb{R}^I, \boldsymbol{\beta}_k^v\in\mathbb{R}^{I_v}, \quad \forall k\in\mathcal{K}, \tag{6d}$$

where $\boldsymbol{\mu}_k = (\boldsymbol{\mu}_k^u, \boldsymbol{\mu}_k^v)$ *and* $\mathcal{K}_s = \{k\in\mathcal{K}\mid s\in\mathcal{E}_k\}$ *for all* $s\in\mathcal{S}$.

We now give an exact LP reformulation of problem (6) as shown in the following theorem:

THEOREM 1. *If Assumption 1 holds, then the robust optimization problem (6) is equivalent to the following LP representable problem:*

$$\text{minimize } -\boldsymbol{b}^\top\boldsymbol{x}+\gamma \tag{7a}$$

subject to (1b), (1c), (6d),

$$\gamma \geq \delta^*\left(\boldsymbol{\alpha} + \sum_{k\in\mathcal{K}} \mathbf{1}_{\mathcal{E}_k}\nu_k \middle| \mathcal{P}\right), \tag{7b}$$

$$\nu_k \geq \delta^*(\boldsymbol{\beta}_k|\mathcal{Q}_k), \quad \forall k\in\mathcal{K}, \tag{7c}$$

$$\alpha_s - \sum_{i\in\mathcal{I}} b_i y_i^0(s) \geq \sum_{j\in\mathcal{I}_v} \lambda_{js}(g_{js})^*(\boldsymbol{w}_{js}/\lambda_{js}) + \bar{\boldsymbol{u}}^\top\boldsymbol{\eta}_s^1 - \underline{\boldsymbol{u}}^\top\boldsymbol{\eta}_s^2, \quad \forall s\in\mathcal{S}, \tag{7d}$$

$$\boldsymbol{\lambda}_s = \sum_{k\in\mathcal{K}_s} \boldsymbol{\beta}_k^v - \sum_{i\in\mathcal{I}} b_i \boldsymbol{y}_i^v(s), \quad \forall s\in\mathcal{S}, \tag{7e}$$

$$\sum_{j\in\mathcal{I}_v} \boldsymbol{w}_{js} = \sum_{i\in\mathcal{I}} b_i \boldsymbol{y}_i^u(s) - \sum_{k\in\mathcal{K}_s} \boldsymbol{\beta}_k^u - \boldsymbol{\eta}_s^1 + \boldsymbol{\eta}_s^2, \quad \forall s\in\mathcal{S}, \tag{7f}$$

$$y_i^0(s) \geq x_i + \sum_{j\in\mathcal{I}_v} \lambda'_{ijs}(g_{js})^*\left(\boldsymbol{w}'_{ijs}/\lambda'_{ijs}\right) + \bar{\boldsymbol{u}}^\top\boldsymbol{\eta}_{is}^3 - \underline{\boldsymbol{u}}^\top\boldsymbol{\eta}_{is}^4, \quad \forall i\in\mathcal{I}, s\in\mathcal{S}, \tag{7g}$$

$$\sum_{j\in\mathcal{I}_v} \boldsymbol{w}'_{ijs} = \boldsymbol{e}_i - \boldsymbol{y}_i^u(s) - \boldsymbol{\eta}_{is}^3 + \boldsymbol{\eta}_{is}^4, \quad \forall i\in\mathcal{I}, s\in\mathcal{S}, \tag{7h}$$

$$\lambda'_{ijs} = y_{ij}^v(s), \quad \forall i\in\mathcal{I}, j\in\mathcal{I}_v, s\in\mathcal{S}, \tag{7i}$$

$$y_i^0(s) \geq \sum_{j\in\mathcal{I}_v} \lambda''_{ijs}(g_{js})^*\left(\boldsymbol{w}''_{ijs}/\lambda''_{ijs}\right) + \bar{\boldsymbol{u}}^\top\boldsymbol{\eta}_{is}^5 - \underline{\boldsymbol{u}}^\top\boldsymbol{\eta}_{is}^6, \quad \forall i\in\mathcal{I}, s\in\mathcal{S}, \tag{7j}$$

$$\sum_{j\in\mathcal{I}_v} \boldsymbol{w}''_{ijs} = -\boldsymbol{y}_i^u(s) - \boldsymbol{\eta}_{is}^5 + \boldsymbol{\eta}_{is}^6, \quad \forall i\in\mathcal{I}, s\in\mathcal{S}, \tag{7k}$$

$$\lambda''_{ijs} = y_{ij}^v(s), \quad \forall i\in\mathcal{I}, j\in\mathcal{I}_v, s\in\mathcal{S}, \tag{7l}$$

$$\boldsymbol{\eta}_s^1, \boldsymbol{\eta}_s^2, \boldsymbol{\lambda}_s, \boldsymbol{\eta}_{is}^3, \boldsymbol{\eta}_{is}^4, \lambda'_{ijs}, \boldsymbol{\eta}_{is}^5, \boldsymbol{\eta}_{is}^6, \lambda''_{ijs} \geq 0, \quad \forall i\in\mathcal{I}, j\in\mathcal{I}_v, s\in\mathcal{S}, \tag{7m}$$

where $\mathbf{1}_{\mathcal{E}_k} \in \mathbb{R}^S$ *is a vector with one for all* $s \in \mathcal{E}_k$ *and zero elsewhere, and* $\boldsymbol{e}_i$ *is a vector in* $\mathbb{R}^I$ *with a one in the ith component and zero elsewhere.*

Proof. For constraint (6b), we have

$$\begin{aligned}
&\sup_{\boldsymbol{p}\in\mathcal{P},\boldsymbol{\mu}_k/\left(\sum_{s\in\mathcal{E}_k} p_s\right)\in\mathcal{Q}_k,k\in\mathcal{K}} \boldsymbol{\alpha}^\top \boldsymbol{p} + \sum_{k\in\mathcal{K}} \boldsymbol{\beta}_k^\top \boldsymbol{\mu}_k \\
&= \sup_{\boldsymbol{p}\in\mathcal{P}} \boldsymbol{\alpha}^\top \boldsymbol{p} + \sum_{k\in\mathcal{K}} \left(\sum_{s\in\mathcal{E}_k} p_s\right) \sup_{\boldsymbol{\mu}_k/\left(\sum_{s\in\mathcal{E}_k} p_s\right)\in\mathcal{Q}_k} \boldsymbol{\beta}_k^\top \boldsymbol{\mu}_k / \left(\sum_{s\in\mathcal{E}_k} p_s\right) \\
&= \sup_{\boldsymbol{p}\in\mathcal{P}} \boldsymbol{\alpha}^\top \boldsymbol{p} + \sum_{k\in\mathcal{K}} \left(\sum_{s\in\mathcal{E}_k} p_s\right) \delta^*\left(\boldsymbol{\beta}_k | \mathcal{Q}_k\right) \\
&= \sup_{\boldsymbol{p}\in\mathcal{P}} \boldsymbol{\alpha}^\top \boldsymbol{p} + \sum_{k\in\mathcal{K}} \left(\sum_{s\in\mathcal{E}_k} p_s\right) \inf_{\nu_k:\nu_k\geq\delta^*(\boldsymbol{\beta}_k|\mathcal{Q}_k)} \nu_k \\
&= \inf_{\nu_k:\nu_k\geq\delta^*(\boldsymbol{\beta}_k|\mathcal{Q}_k),k\in\mathcal{K}} \sup_{\boldsymbol{p}\in\mathcal{P}} \boldsymbol{\alpha}^\top \boldsymbol{p} + \sum_{k\in\mathcal{K}} \left(\sum_{s\in\mathcal{E}_k} p_s\right) \nu_k \qquad (8) \\
&= \inf_{\nu_k:\nu_k\geq\delta^*(\boldsymbol{\beta}_k|\mathcal{Q}_k),k\in\mathcal{K}} \delta^*\left(\boldsymbol{\alpha} + \sum_{k\in\mathcal{K}} \mathbf{1}_{\mathcal{E}_k} \nu_k \middle| \mathcal{P}\right),
\end{aligned}$$

where we exploited in (8) Sion's minimax theorem to reverse the order of the $\sup_{\boldsymbol{p}}$ and $\inf_{\nu_k}$, which relied on the fact that $\mathcal{P}$ is bounded.

We then obtain the equivalent counterpart of constraint (6c). We can rewrite constraint (6c) as

$$\alpha_s + \sum_{k\in\mathcal{K}_s} \boldsymbol{u}^\top \boldsymbol{\beta}_k^u + \sum_{k\in\mathcal{K}_s} \boldsymbol{v}^\top \boldsymbol{\beta}_k^v \geq \sum_{i\in\mathcal{I}} b_i \left(y_i^0(s) + \boldsymbol{u}^\top \boldsymbol{y}_i^u(s) + \boldsymbol{v}^\top \boldsymbol{y}_i^v(s)\right), \quad \forall (\boldsymbol{u},\boldsymbol{v}) \in \mathscr{W}_s,\ s\in\mathcal{S},$$

which is equivalent to:

$$\alpha_s - \sum_{i\in\mathcal{I}} b_i y_i^0(s) \geq \boldsymbol{u}^\top \left(\sum_{i\in\mathcal{I}} b_i \boldsymbol{y}_i^u(s) - \sum_{k\in\mathcal{K}_s} \boldsymbol{\beta}_k^u\right) + \boldsymbol{v}^\top \left(\sum_{i\in\mathcal{I}} b_i \boldsymbol{y}_i^v(s) - \sum_{k\in\mathcal{K}_s} \boldsymbol{\beta}_k^v\right), \forall (\boldsymbol{u},\boldsymbol{v}) \in \mathscr{W}_s,\ s\in\mathcal{S}. \qquad (9)$$

Note that, for any fixed $s \in \mathcal{S}$, the right-hand side of constraint (9) can be further reformulated as

$$\sup_{\boldsymbol{u},\boldsymbol{v},\{\boldsymbol{\rho}_j\}_{j\in\mathcal{I}_v}:\underline{\boldsymbol{u}}\leq\boldsymbol{u}\leq\bar{\boldsymbol{u}},\, g_{sj}(\boldsymbol{\rho}_j)\leq v_j,\, \boldsymbol{\rho}_j=\boldsymbol{u},\forall j} \boldsymbol{u}^\top \left(\sum_{i\in\mathcal{I}} b_i \boldsymbol{y}_i^u(s) - \sum_{k\in\mathcal{K}_s} \boldsymbol{\beta}_k^u\right) + \boldsymbol{v}^\top \left(\sum_{i\in\mathcal{I}} b_i \boldsymbol{y}_i^v(s) - \sum_{k\in\mathcal{K}_s} \boldsymbol{\beta}_k^v\right),$$

which is a convex optimization problem. Hence, since Slater's condition is satisfied by $\boldsymbol{u} := (\bar{\boldsymbol{u}} + \underline{\boldsymbol{u}})/2$, $\boldsymbol{\rho}_j = \boldsymbol{u}$, and $\boldsymbol{v} := \boldsymbol{g}_s(\boldsymbol{u}) + 1$, we can conclude that strong duality holds and replace the supremum with the dual infimum. The latter is derived as:

$$
\begin{aligned}
&\sup_{\boldsymbol{u},\boldsymbol{v},\{\boldsymbol{\rho}_j\}_{j\in\mathcal{I}_v}:\underline{\boldsymbol{u}}\le\boldsymbol{u}\le\bar{\boldsymbol{u}},\, g_{sj}(\boldsymbol{\rho}_j)\le v_j,\, \boldsymbol{\rho}_j=\boldsymbol{u},\,\forall j\in\mathcal{I}_v} \boldsymbol{c}_1^\top \boldsymbol{u} + \boldsymbol{c}_2^\top \boldsymbol{v} \\
&= \sup_{\boldsymbol{u},\boldsymbol{v},\{\boldsymbol{\rho}_j\}_{j\in\mathcal{I}_v}} \inf_{\boldsymbol{\eta}_s^1,\boldsymbol{\eta}_s^2,\boldsymbol{\lambda}_s\ge 0,\{\boldsymbol{w}_{js}\}_{j\in\mathcal{I}_v}} \boldsymbol{c}_1^\top \boldsymbol{u} + \boldsymbol{c}_2^\top \boldsymbol{v} + (\bar{\boldsymbol{u}} - \boldsymbol{u})^\top \boldsymbol{\eta}_s^1 - (\underline{\boldsymbol{u}} - \boldsymbol{u})^\top \boldsymbol{\eta}_s^2 \\
&+ \sum_{j\in\mathcal{I}_v} \lambda_{js}(v_j - g_{js}(\boldsymbol{\rho}_j)) + \boldsymbol{w}_{js}^\top(\boldsymbol{\rho}_j - \boldsymbol{u}) \\
&= \inf_{\boldsymbol{\eta}_s^1,\boldsymbol{\eta}_s^2,\boldsymbol{\lambda}_s\ge 0,\{\boldsymbol{w}_{js}\}_{j\in\mathcal{I}_v}} \sup_{\boldsymbol{u},\boldsymbol{v},\{\boldsymbol{\rho}_j\}_{j\in\mathcal{I}_v}} \boldsymbol{c}_1^\top \boldsymbol{u} + \boldsymbol{c}_2^\top \boldsymbol{v} + (\bar{\boldsymbol{u}} - \boldsymbol{u})^\top \boldsymbol{\eta}_s^1 - (\underline{\boldsymbol{u}} - \boldsymbol{u})^\top \boldsymbol{\eta}_s^2 \\
&+ \sum_{j\in\mathcal{I}_v} \lambda_{js}(v_j - g_{sj}(\boldsymbol{\rho}_j)) + \boldsymbol{w}_{js}^\top(\boldsymbol{\rho}_j - \boldsymbol{u}) \\
&= \inf_{\boldsymbol{\eta}_s^1,\boldsymbol{\eta}_s^2,\boldsymbol{\lambda}_s\ge 0,\{\boldsymbol{w}_{js}\}_{j\in\mathcal{I}_v},\boldsymbol{c}_2+\boldsymbol{\lambda}_s=0,\,\sum_{j\in\mathcal{I}_v}\boldsymbol{w}_{js}=\boldsymbol{c}_1-\boldsymbol{\eta}_s^1+\boldsymbol{\eta}_s^2} \bar{\boldsymbol{u}}^\top \boldsymbol{\eta}_s^1 - \underline{\boldsymbol{u}}^\top \boldsymbol{\eta}_s^2 + \sum_{j\in\mathcal{I}_v} \sup_{\boldsymbol{\rho}_j} \boldsymbol{w}_{js}^\top \boldsymbol{\rho}_j - \lambda_{js} g_{js}(\boldsymbol{\rho}_j) \\
&= \inf_{\boldsymbol{\eta}_s^1,\boldsymbol{\eta}_s^2,\boldsymbol{\lambda}_s\ge 0,\{\boldsymbol{w}_{js}\}_{j\in\mathcal{I}_v},\boldsymbol{c}_2+\boldsymbol{\lambda}_s=0,\,\sum_{j\in\mathcal{I}_v}\boldsymbol{w}_{js}=\boldsymbol{c}_1-\boldsymbol{\eta}_s^1+\boldsymbol{\eta}_s^2} \bar{\boldsymbol{u}}^\top \boldsymbol{\eta}_s^1 - \underline{\boldsymbol{u}}^\top \boldsymbol{\eta}_s^2 + \sum_{j\in\mathcal{I}_v} \lambda_{js} g_{js}^*(\boldsymbol{w}_{js}/\lambda_{js}),
\end{aligned}
$$

where $\boldsymbol{c}_1 := (\sum_{i\in\mathcal{I}} b_i \boldsymbol{y}_i^u(s) - \sum_{k\in\mathcal{K}_s} \boldsymbol{\beta}_k^u)$, $\boldsymbol{c}_2 := (\sum_{i\in\mathcal{I}} b_i \boldsymbol{y}_i^v(s) - \sum_{k\in\mathcal{K}_s} \boldsymbol{\beta}_k^v)$, and $\lambda_{js} g_{js}^*(\boldsymbol{w}_{js}/\lambda_{js})$ is the perspective function of the conjugate function of g_{sj} applied on $(\boldsymbol{w}_{sj}, \lambda_j)$.

Similarly, we can obtain the reformulation of constraint (4b) and (4c). Therefore, the robust optimization problem (6) is equivalent to problem (7). $\square$

3.3. Reformulations under Specific Ambiguity Sets

In this section, we will provide several linear programming reformulations of problem (5) based on the specific ambiguity sets discussed in Section 3.1.

In order to identify a linear programming reformulation, we make the following assumption.

Assumption 3. *The Wasserstein ambiguity set uses the Wasserstein metric with ℓ_1–norm, i.e.* $\mathscr{W}_s = \{(\boldsymbol{u}, v) | \boldsymbol{u} \in [\bar{\boldsymbol{u}}, \underline{\boldsymbol{u}}], ||\boldsymbol{u} - \hat{\boldsymbol{u}}_s||_1 \le v\}$.

In the rest of the paper, we will focus on the Wasserstein ambiguity set with $p = 1$, yet it is easy to extend the results to the one with $p = \infty$. The following corollary gives a reformulation of problem (5) when using the Wasserstein ambiguity set.

Corollary 1. *Under a Wasserstein ambiguity set, and Assumption 3, problem (5) is equivalent to the following LP problem*

$$\text{minimize } -\boldsymbol{b}^\top \boldsymbol{x} + \gamma \tag{10a}$$

$$\text{subject to } (1b), (1c), (6d),$$

$$\gamma \geq \frac{1}{S}\sum_{s\in\mathcal{S}} \alpha_s + \theta\beta, \tag{10b}$$

$$\alpha_s \geq \sum_{i\in\mathcal{I}} b_i y_i^0(s) + \left(\sum_{i\in\mathcal{I}} b_i \boldsymbol{y}_i^u(s) - \boldsymbol{\eta}_s^1 + \boldsymbol{\eta}_s^2\right)^\top \hat{\boldsymbol{u}}_s + \bar{\boldsymbol{u}}^\top \boldsymbol{\eta}_s^1 - \underline{\boldsymbol{u}}^\top \boldsymbol{\eta}_s^2, \quad \forall s\in\mathcal{S}, \tag{10c}$$

$$w_s^1 + \sum_{i\in\mathcal{I}} b_i y_i^v(s) - \beta \leq 0, \quad \forall s\in\mathcal{S}, \tag{10d}$$

$$\sum_{i\in\mathcal{I}} b_i y_{ij}^u(s) - \eta_{js}^1 + \eta_{js}^2 \leq w_s^1, \quad \forall j\in\mathcal{I}, s\in\mathcal{S}, \tag{10e}$$

$$-\sum_{i\in\mathcal{I}} b_i y_{ij}^u(s) + \eta_{js}^1 - \eta_{js}^2 \leq w_s^1, \quad \forall j\in\mathcal{I}, s\in\mathcal{S}, \tag{10f}$$

$$y_i^0(s) \geq x_i - \hat{u}_{is} + \left(-\boldsymbol{y}_i^u(s) - \boldsymbol{\eta}_{is}^3 + \boldsymbol{\eta}_{is}^4\right)^\top \hat{\boldsymbol{u}}_s + \bar{\boldsymbol{u}}^\top \boldsymbol{\eta}_{is}^3 - \underline{\boldsymbol{u}}^\top \boldsymbol{\eta}_{is}^4, \quad \forall i\in\mathcal{I}, s\in\mathcal{S}, \tag{10g}$$

$$w_{is}^2 - y_i^v(s) \leq 0, \quad \forall i\in\mathcal{I}, s\in\mathcal{S}, \tag{10h}$$

$$-y_{ij}^u(s) - \eta_{ijs}^3 + \eta_{ijs}^4 - e_{ij} \leq w_{is}^2, \quad \forall i,j\in\mathcal{I}, s\in\mathcal{S}, \tag{10i}$$

$$y_{ij}^u(s) + \eta_{ijs}^3 - \eta_{ijs}^4 + e_{ij} \leq w_{is}^2, \quad \forall i,j\in\mathcal{I}, s\in\mathcal{S}, \tag{10j}$$

$$y_i^0(s) \geq \left(-\boldsymbol{y}_i^u(s) - \boldsymbol{\eta}_{is}^5 + \boldsymbol{\eta}_{is}^6\right)^\top \hat{\boldsymbol{u}}_s + \bar{\boldsymbol{u}}^\top \boldsymbol{\eta}_{is}^5 - \underline{\boldsymbol{u}}^\top \boldsymbol{\eta}_{is}^6, \quad \forall i\in\mathcal{I}, s\in\mathcal{S}, \tag{10k}$$

$$w_{is}^3 - y_i^v(s) \leq 0, \quad \forall i\in\mathcal{I}, s\in\mathcal{S}, \tag{10l}$$

$$-y_{ij}^u(s) - \eta_{ijs}^5 + \eta_{ijs}^6 \leq w_{is}^3, \quad \forall i,j\in\mathcal{I}, s\in\mathcal{S}, \tag{10m}$$

$$y_{ij}^u(s) + \eta_{ijs}^5 - \eta_{ijs}^6 \leq w_{is}^3, \quad \forall i,j\in\mathcal{I}, s\in\mathcal{S}, \tag{10n}$$

$$\boldsymbol{\eta}_s^1, \boldsymbol{\eta}_s^2, \boldsymbol{\eta}_{is}^3, \boldsymbol{\eta}_{is}^4, \boldsymbol{\eta}_{is}^5, \boldsymbol{\eta}_{is}^6 \geq 0, \quad \forall i\in\mathcal{I}, s\in\mathcal{S}. \tag{10o}$$

Proof. Note that for the Wasserstein ambiguity set, $K = 1$, $\mathcal{E} = \mathcal{S}$, $\mathcal{Q} = \left\{(\boldsymbol{q}_u, q_v) \in \mathbb{R}^I \times \mathbb{R} \,\middle|\, q_v = \theta\right\}$, and $\mathcal{P} = \left\{\boldsymbol{p} \in \mathbb{R}_{++}^S \,\middle|\, p_s = \frac{1}{S}, \forall s\in\mathcal{S}\right\}$. The support functions are:

$$\delta^*\left((\boldsymbol{z}_u, z_v) \middle| \mathcal{Q}\right) = \begin{cases} \theta z_v, & \text{if } \boldsymbol{z}_u = \boldsymbol{0}, \\ +\infty, & \text{otherwise,} \end{cases}$$

for set $\mathcal{Q}$, and for set $\mathcal{P}$:

$$\delta^*(\boldsymbol{z}|\mathcal{P}) = \frac{1}{S}\sum_{s\in\mathcal{S}} z_s.$$

We have that constraints (7b) and (7c) can be rewritten as $\gamma \geq \frac{1}{S}\sum_{s\in\mathcal{S}} \alpha_s + \nu$ and $\nu \geq \theta\beta$ respectively. Therefore, constraints (7b) and (7c) can be rewritten as (10b). Since $I_v = 1$, $g_s(\boldsymbol{u}) = \|\boldsymbol{u} - \hat{\boldsymbol{u}}_s\|_1$, and

$$\lambda_s(g_s)^*(\frac{\boldsymbol{w}_s}{\lambda_s}) = \begin{cases} \boldsymbol{w}_s^\top \hat{\boldsymbol{u}}_s, & \text{if } \|\boldsymbol{w}_s\|_\infty \leq \lambda_s, \\ \infty, & \text{otherwise,} \end{cases}$$

we can rewrite constraints (7d) – (7f) as constraints (10c) – (10f). Similarly, we can obtain the other constraints. Therefore, we conclude that problem (5) under the Wasserstein ambiguity set is equivalent to problem (10). □

For the event-wise mean absolute deviation set, we have $g_{js}(\boldsymbol{u}) = |u_j - \mu_{js}| = g_{js}(u_j)$ for $j \in \mathcal{I}_v$ and $s \in \mathcal{S}$, hence, $g(\boldsymbol{u})$ is separable. A reformulation of model (5) under this ambiguity set is given in Corollary 2.

Corollary 2. *Under the event-wise mean absolute deviation set, problem (5) is equivalent to the following LP problem:*

$$\text{minimize } -\boldsymbol{b}^\top \boldsymbol{x} + \gamma \tag{11a}$$

subject to (1b), (1c), (6d),

$$\gamma \geq \sum_{s\in\mathcal{S}} p_s(\alpha_s + \boldsymbol{\mu}_s^\top \boldsymbol{\beta}_s^u + \boldsymbol{\sigma}_s^\top \boldsymbol{\beta}_s^v), \tag{11b}$$

$$\alpha_s \geq \sum_{i\in\mathcal{I}} b_i y_i^0(s) + \left(\sum_{i\in\mathcal{I}} b_i \boldsymbol{y}_i^u(s) - \boldsymbol{\eta}_s^1 + \boldsymbol{\eta}_s^2 - \boldsymbol{\beta}_s^u\right)^\top \boldsymbol{\mu}_s + \bar{\boldsymbol{u}}^\top \boldsymbol{\eta}_s^1 - \underline{\boldsymbol{u}}^\top \boldsymbol{\eta}_s^2, \quad \forall s \in \mathcal{S}, \tag{11c}$$

$$\boldsymbol{w}_s^1 + \sum_{i\in\mathcal{I}} b_i \boldsymbol{y}_i^v(s) - \boldsymbol{\beta}_s^v \leq 0, \quad \forall s \in \mathcal{S}, \tag{11d}$$

$$\sum_{i\in\mathcal{I}} b_i \boldsymbol{y}_i^u(s) - \boldsymbol{\eta}_s^1 + \boldsymbol{\eta}_s^2 - \boldsymbol{\beta}_s^u \leq \boldsymbol{w}_s^1, \quad \forall s \in \mathcal{S}, \tag{11e}$$

$$-\sum_{i\in\mathcal{I}} b_i \boldsymbol{y}_i^u(s) + \boldsymbol{\eta}_s^1 - \boldsymbol{\eta}_s^2 + \boldsymbol{\beta}_s^u \leq \boldsymbol{w}_s^1, \quad \forall s \in \mathcal{S}, \tag{11f}$$

$$y_i^0(s) \geq x_i + \left(-\boldsymbol{y}_i^u(s) - \boldsymbol{\eta}_{is}^3 + \boldsymbol{\eta}_{is}^4\right)^\top \boldsymbol{\mu}_s + \bar{\boldsymbol{u}}^\top \boldsymbol{\eta}_{is}^3 - \underline{\boldsymbol{u}}^\top \boldsymbol{\eta}_{is}^4 - \mu_{is}, \quad \forall i \in \mathcal{I}, s \in \mathcal{S}, \tag{11g}$$

$$\boldsymbol{w}_{is}^2 - \boldsymbol{y}_i^v(s) \leq 0, \qquad \forall i \in \mathcal{I}, s \in \mathcal{S}, \tag{11h}$$

$$-\boldsymbol{y}_i^u(s) - \boldsymbol{\eta}_{is}^3 + \boldsymbol{\eta}_{is}^4 - \boldsymbol{e}_i \leq \boldsymbol{w}_{is}^2, \qquad \forall i \in \mathcal{I}, s \in \mathcal{S}, \tag{11i}$$

$$\boldsymbol{y}_i^u(s) + \boldsymbol{\eta}_{is}^3 - \boldsymbol{\eta}_{is}^4 + \boldsymbol{e}_i \leq \boldsymbol{w}_{is}^2, \qquad \forall i \in \mathcal{I}, s \in \mathcal{S}, \tag{11j}$$

$$y_i^0(s) \geq \left(-\boldsymbol{y}_i^u(s) - \boldsymbol{\eta}_{is}^5 + \boldsymbol{\eta}_{is}^6\right)^\top \boldsymbol{\mu}_s + \bar{\boldsymbol{u}}^\top \boldsymbol{\eta}_{is}^5 - \underline{\boldsymbol{u}}^\top \boldsymbol{\eta}_{is}^6, \qquad \forall i \in \mathcal{I}, s \in \mathcal{S}, \tag{11k}$$

$$\boldsymbol{w}_{is}^3 - \boldsymbol{y}_i^v(s) \leq 0, \qquad \forall i \in \mathcal{I}, s \in \mathcal{S}, \tag{11l}$$

$$-\boldsymbol{y}_i^u(s) - \boldsymbol{\eta}_{is}^5 + \boldsymbol{\eta}_{is}^6 \leq \boldsymbol{w}_{is}^3, \qquad \forall i \in \mathcal{I}, s \in \mathcal{S}, \tag{11m}$$

$$\boldsymbol{y}_i^u(s) + \boldsymbol{\eta}_{is}^5 - \boldsymbol{\eta}_{is}^6 \leq \boldsymbol{w}_{is}^3, \qquad \forall i \in \mathcal{I}, s \in \mathcal{S}, \tag{11n}$$

$$\boldsymbol{\eta}_s^1, \boldsymbol{\eta}_s^2, \boldsymbol{\eta}_{is}^3, \boldsymbol{\eta}_{is}^4, \boldsymbol{\eta}_{is}^5, \boldsymbol{\eta}_{is}^6 \geq 0, \qquad \forall i \in \mathcal{I}, s \in \mathcal{S}. \tag{11o}$$

Proof. Note that for the event-wise mean absolute deviation ambiguity set, $K = S$, $I_v = I$, $\mathcal{E}_s = \{s\}$, $\mathcal{Q}_s = \left\{(\boldsymbol{q}_u, \boldsymbol{q}_v) \in \mathbb{R}^I \times \mathbb{R}^I \,\middle|\, \boldsymbol{q}_u = \boldsymbol{\mu}_s, \boldsymbol{q}_v = \boldsymbol{\sigma}_s\right\}$, for $s \in \mathcal{S}$ and $\mathcal{P} = \left\{\boldsymbol{p} \in \mathbb{R}_{++}^S \,\middle|\, p_s = \frac{1}{S}, \forall s \in \mathcal{S}\right\}$. The support function of set $\mathcal{Q}_s$ for $s \in \mathcal{S}$ is $\delta^*\left((\boldsymbol{z}_u, \boldsymbol{z}_v) \middle| \mathcal{Q}_s\right) = \boldsymbol{\mu}_s^\top \boldsymbol{z}_u + \boldsymbol{\sigma}_s^\top \boldsymbol{z}_v$. Therefore, constraints (7b) and (7c) can be rewritten as (11b).

Note that $g_{is}(\cdot)$ is separable, based on Ben-Tal et al. (2015), constraints (7d) – (7f) are equivalent to

$$\alpha_s - \sum_{i \in \mathcal{I}} b_i y_i^0(s) \geq \sum_{j \in \mathcal{I}} \lambda_{js} (g_{js})^* (w_{js}/\lambda_{js}) + \bar{\boldsymbol{u}}^\top \boldsymbol{\eta}_s^1 - \underline{\boldsymbol{u}}^\top \boldsymbol{\eta}_s^2, \qquad \forall s \in \mathcal{S},$$

$$\boldsymbol{\lambda}_s = \sum_{k \in \mathcal{K}_s} \boldsymbol{\beta}_k^v - \sum_{i \in \mathcal{I}} b_i \boldsymbol{y}_i^v(s), \qquad \forall s \in \mathcal{S},$$

$$\boldsymbol{w}_s = \sum_{i \in \mathcal{I}} b_i \boldsymbol{y}_i^u(s) - \sum_{k \in \mathcal{K}_s} \boldsymbol{\beta}_k^u - \boldsymbol{\eta}_s^1 + \boldsymbol{\eta}_s^2, \qquad \forall s \in \mathcal{S},$$

where

$$\lambda_{js}(g_{js})^*(w_{js}/\lambda_{js}) = \begin{cases} w_{js}\mu_{js}, & \text{if } |w_{js}| \leq \lambda_{js}, \\ \infty, & \text{otherwise.} \end{cases}$$

Thus, we can obtain constraints (11c) – (11f). Similarly, we can obtain other constraints of problem (11). This completes the proof. $\square$

4. A Column Generation Solution Scheme

In this section, we will develop a novel CG decomposition scheme to solve the large-scale LPs efficiently. In general, the CG method iteratively solves a restricted master problem (RMP). RMP will initially contain a subset of columns of the full master problem (FMP). At each iteration, we look in a step called the pricing step for new columns that have the potential to improve the current solution and add them to the RMP, until a RMP solution is proved optimal.

As indicated by the numerical study in Section 5, the optimal solutions of problem (7) are highly sparse (usually more than 70%), which means that most of the terms of the optimal solutions have a value of zero. This motivates us to propose a CG scheme. We remove a set of most likely ineffective columns (i.e. the optimal value of these columns is expected to be zero) from FMP and progressively identify the columns that are required to return to RMP. This method exploits the fact that the majority of these columns might not be part of an optimal solution.

In order to transform problem (7) to a standard form, first, we add slack variables to all inequality constraints, and change the constraints to equality constraints. For the unrestricted variables, we use the difference of two restricted variables to replace the variables, for example, we can write $\boldsymbol{y}^u = \boldsymbol{y}^{u+} - \boldsymbol{y}^{u-}$, and $\boldsymbol{y}^{u+}, \boldsymbol{y}^{u-} \geq 0$. Problem (7) can be reformulated in the following standard form:

$$\text{(FMP)} \quad \text{minimize} \sum_{j \in \mathcal{J}} c_j \lambda_j \tag{12a}$$

$$\text{subject to} \sum_{j \in \mathcal{J}} \boldsymbol{a}_j \lambda_j = \boldsymbol{q}, \tag{12b}$$

$$\lambda_j \geq 0, \qquad \forall j \in \mathcal{J}, \tag{12c}$$

where $\boldsymbol{\lambda}$ includes all the decision variables of problem (7), $\boldsymbol{a}_j, \boldsymbol{q} \in \mathbb{R}^N$ are the coefficient of λ_j and right-hand side vector, respectively, and N is the number of constraints in problem (7). To help with the discussion, we also present the dual LP associated to FMP:

$$\text{(Dual-FMP)} \quad \text{maximize}\, \boldsymbol{q}^\top \boldsymbol{\tau} \tag{13a}$$

$$\text{subject to}\, \boldsymbol{a}_j^\top \boldsymbol{\tau} \leq c_j, \qquad \forall j \in \mathcal{J}, \tag{13b}$$

$$\boldsymbol{\tau} \in \mathbb{R}^N, \tag{13c}$$

where $\boldsymbol{\tau}$ is the dual vector associated to constraint (12b).

In column generation, one can work with a subset $\mathcal{J}'$ of columns instead of the set of columns $\mathcal{J}$ in RMP. Thus RMP can be described as follows:

$$\text{(RMP)} \quad \text{minimize} \sum_{j \in \mathcal{J}'} c_j \lambda_j \tag{14a}$$

$$\text{subject to} \sum_{j\in\mathcal{J}'} \boldsymbol{a}_j\lambda_j = \boldsymbol{q}, \tag{14b}$$

$$\lambda_j \geq 0, \qquad \forall j \in \mathcal{J}'. \tag{14c}$$

Note that the feasibility of the RMP depends on the choice of the set $\mathcal{J}'$. On the other hand, the boundedness of RMP can be assumed since otherwise it implies that FMP is also unbounded. Assuming that RMP is both feasible and bounded, let $\boldsymbol{\lambda}$ and $\bar{\boldsymbol{\tau}}$ be an optimal primal dual solution pair for RMP. We are interested in verifying whether the solution pair $(\bar{\boldsymbol{\lambda}}, \bar{\boldsymbol{\tau}})$, where

$$\bar{\lambda}_j := \begin{cases} \lambda_j \text{ if } j \in \mathcal{J}', \\ 0 \;\; \text{otherwise} \end{cases}$$

is a solution pair to the FMP in (12), and if not to identify new columns that can potentially improve the current solution. This can be done by verifying whether constraint (13b) is satisfied by $\bar{\boldsymbol{\tau}}$, a process known as searching for strictly negative "reduced costs" defined as:

$$\bar{c}_j := c_j - \bar{\boldsymbol{\tau}}^\top \boldsymbol{a}_j \,,\, \forall\, j \in \mathcal{J}/\mathcal{J}'.$$

If no reduced cost is negative, then the optimality of the solution is confirmed. Otherwise, we can add a set of columns to RMP corresponding to a subset of the negative reduced costs, a process also referred as multiple pricing. The multiple pricing is a well-known strategy to reduce the total number of iterations and accelerate column generation (e.g., Desrochers and Soumis 1989, Desaulniers et al. 2002). We repeat the process until the positivity of all reduced costs are non-negative thus confirming the optimality of $(\bar{\boldsymbol{\lambda}}, \bar{\boldsymbol{\tau}})$ in FMP.

In problem (7), we can partition the FMP decision variables into the following five different classes:

- the "primary set" $\mathcal{J}_u$ includes $\boldsymbol{y}^u$ which is associated with the primary random variable $\boldsymbol{u}$;
- the "support set" $\mathcal{J}_\eta$ consists of $\boldsymbol{\eta}^1, \cdots, \boldsymbol{\eta}^6$, since these variables are introduced to reformulate the support set of $\boldsymbol{u}$;
- the "auxiliary set" $\mathcal{J}_v$ includes $\boldsymbol{y}^0$ and $\boldsymbol{y}^v$ which is defined under the auxiliary random variable $\boldsymbol{v}$;
- the "objective set" $\mathcal{J}_{obj}$ includes the variables that could directly affect the objective function, such as variables $\boldsymbol{x}$, γ, $\boldsymbol{\alpha}$, $\boldsymbol{\nu}$, $\boldsymbol{\beta}$;

- the "extra set" $\mathcal{J}_{extra}$ includes all remaining variables (i.e., $\boldsymbol{\lambda}, \boldsymbol{\lambda}', \boldsymbol{\lambda}'', \boldsymbol{w}$) together with the variables introduced to rewrite LP representable functions and sets to obtain a LP.

We note that the variables in set $\mathcal{J}_{obj}$ generally play a key role in the optimal objective value since their value directly decides the objective value, while the variables in $\mathcal{J}_v$ and $\mathcal{J}_{extra}$ play (as seen in the proof of Proposition 1) a key role in making the RMP feasible. We therefore further let $\mathcal{J}_0 := \mathcal{J}_v \cup \mathcal{J}_{obj} \cup \mathcal{J}_{extra}$ and show that RMP is always feasible with $\mathcal{J}_0$.

Proposition 1. *Given that Assumption 2 is satisfied and that* $\mathcal{Q}_k^v = \{\boldsymbol{q}^v \in \mathbb{R}^{I_v} | \exists \boldsymbol{q}^u \in \mathbb{R}^I, (\boldsymbol{q}^u, \boldsymbol{q}^v) \in \mathcal{Q}_k\}$ *is bounded for all* $k \in \mathcal{K}$*, if* $\mathcal{J}_0 = \mathcal{J}_v \cup \mathcal{J}_{obj} \cup \mathcal{J}_{extra}$*, then RMP with* $\mathcal{J}' = \mathcal{J}_0$ *is feasible.*

Proof. Note that RMP with $\mathcal{J}_0 = \mathcal{J}_v \cup \mathcal{J}_{obj} \cup \mathcal{J}_{extra}$ is equivalent to the following problem:

$$
\begin{aligned}
&\text{minimize } -\boldsymbol{b}^\top \boldsymbol{x} + \gamma \\
&\text{subject to } \boldsymbol{c}^\top \boldsymbol{x} = d, \\
&\qquad x_i \geq 0, && \forall i \in \mathcal{I} \\
&\qquad y_i^0(s) + \sum_{j \in \mathcal{I}_v} y_{ij}^v(s) v_j \geq 0, && \forall (\boldsymbol{u}, \boldsymbol{v}) \in \mathscr{W}_s', s \in \mathcal{S}, i \in \mathcal{I} \\
&\qquad y_i^0(s) + \sum_{j \in \mathcal{I}_v} y_{ij}^v(s) v_j \geq x_i - u_i, && \forall (\boldsymbol{u}, \boldsymbol{v}) \in \mathscr{W}_s', s \in \mathcal{S}, i \in \mathcal{I}, \\
&\qquad \gamma \geq \sum_{s \in \mathcal{S}} p_s \alpha_s + \sum_{k \in \mathcal{K}} \boldsymbol{\mu}_k^{v\top} \boldsymbol{\beta}_k^v + \sum_{k \in \mathcal{K}} \boldsymbol{\mu}_k^{u\top} \boldsymbol{\beta}_k^u, && \forall \left\{ \frac{\boldsymbol{\mu}_k}{\sum_{s \in \mathcal{E}_k} p_s} \right\}_{k \in \mathcal{K}} \in \prod_{k \in \mathcal{K}} \mathcal{Q}_k, \boldsymbol{p} \in \mathcal{P}, \\
&\qquad \alpha_s + \sum_{k \in \mathcal{K}_s} \boldsymbol{v}^\top \boldsymbol{\beta}_k^v + \sum_{k \in \mathcal{K}_s} \boldsymbol{u}^\top \boldsymbol{\beta}_k^u \geq \sum_{i \in \mathcal{I}} b_i (y_i^0(s) \\
&\qquad + \sum_{j \in \mathcal{I}_v} y_{ij}^v(s) v_j), && \forall (\boldsymbol{u}, \boldsymbol{v}) \in \mathscr{W}_s', \; s \in \mathcal{S}, \\
&\qquad \gamma \in \mathbb{R}, \alpha \in \mathbb{R}^S, \boldsymbol{\beta}_k^u \in \mathbb{R}^I, \boldsymbol{\beta}_k^v \in \mathbb{R}^{I_v}, && \forall k \in \mathcal{K},
\end{aligned}
$$

where $\mathscr{W}_s' := \left\{ (\boldsymbol{u}, \boldsymbol{v}) \in \mathbb{R}^I \times \mathbb{R}^{I_v} \middle| \boldsymbol{g}_s(\boldsymbol{u}) \leq \boldsymbol{v} \right\}$.

To construct a feasible assignment, we first identify a feasible order, e.g. $\boldsymbol{x} := (d/c_1)\boldsymbol{e}_1$. Thus, the first two constraints are satisfied. We then let for all $s \in \mathcal{S}$ and all $i \in \mathcal{I}$, $y_i^0(s) := x_i - \tau_1/\tau_2$ and for all $j \in \mathcal{I}_v$, $y_{ij}^v(s) := \mathbf{1}\{j \in \bar{\mathcal{I}}^v\}/\tau_2$. This guarantees that the next two constraints are satisfied given that:

$$
y_i^0(s) + \sum_{j \in \mathcal{I}_v} y_{ij}^v(s) v_j = x_i - \tau_1/\tau_2 + \sum_{j \in \bar{\mathcal{I}}^v} v_j/\tau_2 \geq x_i - \tau_1/\tau_2 + \sum_{j \in \bar{\mathcal{I}}^v} g_{j,s}(\boldsymbol{u})/\tau_2 \geq x_i + \|\boldsymbol{u}\|_1
$$

$$\geq \max(0,\, x_i - u_i)\,.$$

We can then simply set $\alpha_s := \sum_{i\in\mathcal{I}} b_i y_i^0(s)$, $\beta_{k,i}^u := 0$ and $\beta_{k,j}^v := (\sum_{i\in\mathcal{I}} b_i)\mathbf{1}\{j\in\bar{\mathcal{I}}^v\}/\tau_2$ for $s\in\mathcal{S}$, $k\in\mathcal{K}$, $i\in\mathcal{I}$ and $j\in\mathcal{I}_v$, so that:

$$\begin{aligned}
\alpha_s + \sum_{k\in\mathcal{K}_s} \boldsymbol{v}^\top \boldsymbol{\beta}_k^v + \sum_{k\in\mathcal{K}_s} \boldsymbol{u}^\top \boldsymbol{\beta}_k^u &= \sum_{i\in\mathcal{I}} b_i y_i^0(s) + (\sum_{i\in\mathcal{I}} b_i)|\mathcal{K}_s| \sum_{j\in\bar{\mathcal{I}}^v} v_j/\tau_2 \\
&\geq \sum_{i\in\mathcal{I}} b_i y_i^0(s) + \sum_{j\in\bar{\mathcal{I}}^v} (\sum_{i\in\mathcal{I}} b_i) v_j/\tau_2 \\
&= \sum_{i\in\mathcal{I}} b_i (y_i^0(s) + \sum_{j\in\mathcal{I}_v} y_{ij}^v(s) v_j)
\end{aligned}$$

Finally, let $\gamma := \max_{s\in\mathcal{S}} \alpha_s + \sum_{k\in\mathcal{K}} \sup_{(\boldsymbol{\mu}_k^u,\boldsymbol{\mu}_k^v)\in\mathcal{Q}_k} \boldsymbol{\mu}_k^{v\top} \boldsymbol{\beta}_k^v$, which exists since for all k's, the projection on $\boldsymbol{\mu}_k^v$ of $\mathcal{Q}_k$'s is assumed to be bounded. □

The following corollary shows that RMP with $\mathcal{J}_0$ is feasible under the Wasserstein ambiguity set and the event-wise mean absolute deviation set.

Corollary 3. *Under the Wasserstein ambiguity set and the and event-wise mean absolute deviation set, if* $\mathcal{J}_0 = \mathcal{J}_v \cup \mathcal{J}_{obj} \cup \mathcal{J}_{extra}$*, then RMP with* $\mathcal{J}' = \mathcal{J}_0$ *is feasible.*

Proof. Recall that for the Wasserstein ambiguity set, $I_v = 1$ and, with $\bar{I}^v = \{1\}$, we have that $\sum_{j\in\bar{\mathcal{I}}^v} g_{j,s}(\boldsymbol{u}) = \|\boldsymbol{u} - \hat{\boldsymbol{u}}_s\|_1 \geq \|\boldsymbol{u}\|_1 - \|\hat{\boldsymbol{u}}_s\|_1$, following from the triangle inequality. We therefore have that Assumption 2 is satisfied. Moreover, $K = 1$ and $\mathcal{Q}^v = \{\theta\}$, which is bounded. Thus, based on Proposition 1, RMP with $\mathcal{J}_0$ is feasible under the Wasserstein set.

For the even-wise mean absolute deviation set, one can let $\bar{\mathcal{I}}^v := \{1,\ldots,I_v\}$ and confirm that

$$\sum_{j\in\bar{\mathcal{I}}_v} g_{js}(\boldsymbol{u}) = \sum_{j=1}^{I_v} |u_j - \mu_{js}| = \|\boldsymbol{u} - \boldsymbol{\mu}_s\|_1 \geq \|\boldsymbol{u}\|_1 - \|\boldsymbol{\mu}_s\|_1,$$

hence Assumption 2 is satisfied. Besides, $\mathcal{Q}_k^v = \{\boldsymbol{\sigma}_k\}$ for all $k\in\mathcal{S}$ is again bounded. Therefore, RMP with $\mathcal{J}_0$ is feasible under the event-wise mean absolute deviation set. □

Algorithm 1 provides an outline of the CG algorithm.

4.1. Acceleration Strategy

CG algorithms are well known to suffer from convergence issues due to the potential high degeneracy of the RMP and dual variables' instability from one iteration to the next. This can lead to failure in achieving the expected computational efficiency (e.g., Vanderbeck 2000, Lübbecke and Desrosiers 2005). More specifically, a degenerate RMP often has multiple

Algorithm 1: Column Generation Algorithm

Initialize The number of iterations $\ell = 0$, and let L represents the maximum number of iteration.

Initialize $\mathcal{J}' = \mathcal{J}_0$ and COLUMN_FOUND=TRUE.

while $(\ell < L$ & $COLUMN\_FOUND=TRUE)$ **do**

  Set $\ell := \ell + 1$, COLUMN_FOUND=FALSE.

  Solve the following RMP:

$$\text{minimize} \sum_{j \in \mathcal{J}'} c_j \lambda_j \tag{15a}$$

$$\text{subject to} \sum_{j \in \mathcal{J}'} \boldsymbol{a}_j \lambda_j = \boldsymbol{q}, \tag{15b}$$

$$\lambda_j \geq 0. \qquad \forall j \in \mathcal{J}'. \tag{15c}$$

  Record the optimal solution $\boldsymbol{\lambda}^\ell$, and dual variables $\boldsymbol{\tau}^\ell$.

  for $j \in \mathcal{J} \backslash \mathcal{J}'$ **do**

    Compute the reduced cost $\bar{c}_j$ of λ_j: $\bar{c}_j = c_j - \boldsymbol{\tau}^{\ell\top} \boldsymbol{a}_j$.

    if $\bar{c}_j < 0$ **then**

      Add j to $\mathcal{J}'$.

      COLUMN_FOUND=TRUE

    end

  end

end

return The optimal solution.

optimal dual solutions, therefore, calculating the reduce cost on a "low quality" optimal dual solution might delay confirming the optimality of the primal solution. Moreover, a number of iterations may be performed without any or almost no improvement in the RMP objective function value (a.k.a. tailing-off effect).

Dual variable stabilization strategies have been proposed to circumvent these issues. Du Merle et al. (1999) proposed to set the dual variable in a relatively small region around the current dual values and to penalize if the dual variable goes outside the region. They adjusted the region and penalty parameters dynamically in the algorithm. Such methods can improve the computational effectiveness when the box imposed on the dual variables is set around a good choice of initial dual solution (Costa et al. 2019). However, we found that this approach does not work well for our problem. Instead, Bixby et al. (1992) and Rousseau et al. (2007) recommended using an interior point solver to obtain a dual solution that is in the relative interior of the optimal space of dual solutions to the RMP rather than using one of its extreme points, which should stabilizes the CG algorithm and decreases the number of iterations.

In addition, at each CG iteration, instead of adding the columns with negative reduced cost in the CG method, one can try to strategically select a subset of columns and add to RMP, which is called strategical column selection. The main idea behind this technique is that by exploiting the structure of problem, one could select more promising columns to RMP in each iteration without overburdening it. We did experiment with this idea in our numerical study. Specifically, for the set $\mathcal{J}_u$, if the reduced cost associated to $y^u_{ij}(s)$ for some i, j and s is negative, we add the columns $\{y^u_{ij}(1),\cdots,y^u_{ij}(S)\}$ to the RMP. For the set $\mathcal{J}_\eta$, our strategy for $i\in\mathcal{I}$ and $k=3,\cdots,6$ was as follows. If the reduced cost associated to $\eta^k_{ij^*s^*}$ is negative, where $(j^*,s^*)=\arg\min_{j,s} \bar{c}^{\eta^k}_{ijs}$ and $\bar{c}^{\eta^k}_{ijs}$ is the reduced cost associated to η^k_{ijs}, we add the columns $\{\eta^k_{ij^*1},\cdots,\eta^k_{ij^*S}\}$ to the RMP. The strategies in the case of $\boldsymbol{\eta}^1$ and $\boldsymbol{\eta}^2$ are similar to those for $\boldsymbol{\eta}^3,\cdots,\boldsymbol{\eta}^6$ except that the subscript i is dropped. As indicated by the numerical study, this strategy effectively reduces the total number of iterations and the overall solution time.

Finally, we propose an early stopping criterion based on a KKT condition test. The objective is to address the issue that CG might fail to prove the optimality of the primal solutions by using the sign of the reduced cost due to having identified a "poor quality" dual optimal solution. Instead, an efficient way to confirm optimality is to use the KKT condition test. We give the KKT condition in the following lemma.

LEMMA 2. *(KKT optimality conditions) Let $\boldsymbol{\lambda}$ be the primal feasible solution of problem* (12)*, and $\boldsymbol{\tau}$ be the feasible solution to the dual problem* (13)*, then $\boldsymbol{\lambda}$ and $\boldsymbol{\tau}$ are the optimal primal dual pairs for problem* (12) *if and only if the following complementarity conditions hold:*

$$(c_j-\boldsymbol{a}_j^\top\boldsymbol{\tau})\lambda_j=0, \qquad \forall j\in\mathcal{J}. \tag{16}$$

One can therefore confirm optimality by checking if a dual feasible solution of problem (13) satisfies the condition (16). More specifically, given the optimal solution $\boldsymbol{\lambda}$ of RMP, we set

$$\bar{\lambda}_j := \begin{cases} \lambda_j \text{ if } j \in \mathcal{J}', \\ 0 \ \text{ otherwise.} \end{cases}$$

Let $\mathcal{J}^+ := \{j \in \mathcal{J} \big| \bar{\lambda}_j > 0\}$. We then solve the following optimization problem:

$$\text{maximize}\, 0 \tag{17a}$$

$$\text{subject to}\, \boldsymbol{a}_j^\top \boldsymbol{\tau} = c_j, \qquad \forall j \in \mathcal{J}^+, \tag{17b}$$

$$\boldsymbol{a}_j^\top \boldsymbol{\tau} \le c_j, \qquad \forall j \in \mathcal{J} \backslash \mathcal{J}^+, \tag{17c}$$

$$\boldsymbol{\tau} \in \mathbb{R}^N. \tag{17d}$$

If the above problem is feasible, we have that the feasible solution $\boldsymbol{\tau}$ to (17) is also feasible to (13) and satisfies when paired with $\bar{\lambda}$ condition (16). Therefore, $\bar{\lambda}$ is optimal in the FMP.

5. Numerical Study

In this section, we focus on the distributionally robust multi-item newsvendor problem with two types of ambiguity sets. In Section 5.1, we provide some implementation details. We present the computational results for the Wasserstein ambiguity set in Section 5.2 and for the event-wise mean absolute deviation ambiguity set in Section 5.3.

5.1. Implementation Details

In our test examples, following Chen et al. (2020), we set $\underline{u} = 0$ and $\bar{u}_i$ is randomly generated from a uniform distribution on $[0, 100]$ for $i \in \mathcal{I}$. For each $s \in \mathcal{S}$, $\hat{u}_s$ and μ_s are randomly generated from a uniform distribution on $[\underline{u}, \bar{u}]$. We let $c_i = 1$, $i \in \mathcal{I}$, and $d = 50I$. b_i is randomly generated from a uniform distribution on $[0, 5]$ for all $i \in \mathcal{I}$. We vary the Wasserstein sets radius $\theta \in \{1, 2, 5, 10\}$. For the event-wise mean absolute deviation set, $\sigma_s \in \{0, 1, 2, 5, 10\}$ was used for $s \in \mathcal{S}$. The number of items $I \in \{10, 20\}$ and the sample size $S \in \{500, 2000, 8000\}$ were used to generate instances. For each problem size, ten instances were generated to generate statistics about the performance of the different algorithms. Table 1 presents the number of variables, the number of constraints, and the proportion of the number of variables in set $\mathcal{J}_0$ for instances using the Wasserstein ambiguity set and the event-wise mean absolute deviation set.

Table 1 Classes for the standard form LP with Wasserstein ambiguity set and event-wise mean absolute deviation set, which include the total number of variables (# of variables), the total number of constraints (# of constraints) and the proportion of the number of variables in set $\mathcal{J}_0$ over the total number of variables (prop).

I	S	Wasserstein ambiguity set			Mean absolute deviation set		
		# of variables	# of constraints	prop	# of variables	# of constraints	prop
10	500	567,514	231,002	45%	816,513	325,502	62%
	2000	2,270,014	924,002	45%	3,266,013	1,302,002	62%
	8000	9,080,014	3,696,002	45%	13,064,013	5,208,002	62%
20	500	2,132,524	861,002	43%	3,131,523	1,250,502	61%
	2000	8,530,024	3,444,002	43%	12,526,023	5,002,002	61%
	8000	34,120,024	13,776,002	43%	50,104,023	20,008,002	61%

All experiments are conducted on the Cedar cluster of Compute Canada with a single CPU core and 32G memory. The algorithm was implemented in the C programming language using IBM CPLEX solver, version 12.10 callable libraries. For all computations, we set the maximum number of iteration to be 100 and runtime limit to be $7,200$ seconds. For instances that are solved to optimality, we report the average solution time in seconds, and use "-" to denote that none of the ten instances were solved to optimality within the runtime limit. The reduced cost is generally computed within seconds, so we did not present these runtimes in the table.

5.2. Computational Results with Wasserstein Ambiguity Set

We now implement the column generation Algorithm 1 to solve the newsvendor problem under the Wasserstein ambiguity set. The following variants are considered for solving the problem in this section:

- CPX refers to the barrier interior point method in CPLEX on problem (12).
- CG_KKT refers to using the CG algorithm (Algorithm 1) with the KKT test-based early-stopping criterion and strategical column selection. The KKT test is also conducted at every iteration.
- CG_KKT* refers to using the CG algorithm (Algorithm 1) with the KKT test-based early-stopping criterion and strategical column selection. The KKT test is conducted only when the objective values in two successive iterations are the same.

- CG_SCS refers to using the CG algorithm (Algorithm 1) with strategical column selection but without the early-stopping criterion.
- CG refers to using the CG algorithm (Algorithm 1) with neither strategical column selection nor the early-stopping criterion.
- CG_KKT_PS refers to the use of CG_KKT but using the primal simplex method to solve the master problem.
- BD refers to using the multi-cut Benders decomposition method (which is described in Appendix A).*

Note that we use the barrier interior point method to solve (RMP) for all the variants except CG_KKT_PS. Table 2 displays the average solution time for solving the problem, the average time spent evaluating the KKT test-based early-stopping criterion, the average number of iterations, and the average sparsity ratio at optimum, where the sparsity ratio for an instance is computed as the number of variables whose optimal value is equal to zero divided by the total number of variables.

We observe from Table 2 that the computation is more expensive as the number of scenarios S increases. However, it seems that the Wasserstein radius θ does not have a significant impact on the computational efficiency. We also observe that CG_KKT yields significantly better performance than the other six methods for most types of instances. In fact, in comparison to CPX, CG_KKT decreases the solution time by an average of about 32%. For $I = 20$ and $S = 8000$, CG_KKT can solve five out of the ten instances to optimality while CPX cannot solve any instance to optimality within the time limit. Moreover, CG_SCS has a significantly worse performance than CG_KKT in terms of the average solution time, and the average number of iterations is at least twice larger than CG_KKT for most types of instances. Only when $I = 20$ and $S = 8000$ can we observe that CG_SCS gives a comparable performance to CG_KKT. Note that CG yields a poor performance when compared with CG_SCS. For most types of instances, it fails to solve to optimality. This confirms that the use of the early-stopping criterion and the column selection strategy proposed in Section 4.1 play a critical role in improving the efficiency of the column generation algorithm. If we compare the CG_KKT and CG_KKT_PS, CG_KKT_PS can only solve 137 of the

*We also implement the multi-cut Benders decomposition method with level regularization that was proposed in Zverovich et al. (2012). We found that this method has a similar performance to BD, thus we did not present it in this section.

Table 2 The average CPU solution time (in seconds) for solving the multi-item DR newsvendor problem with Wasserstein ambiguity set (AvT), and solving the KKT problem (KKT), the average number of iterations (Iters), and the average sparsity ratio (SR) in percent are reported.

I	S	θ	CPX		CG_KKT				CG_KKT*				CG_SCS			CG			CG_KKT_PS			BD	
			AvT	SR	AvT	KKT	Iter	SR	AvT	KKT	Iter	SR	AvT	Iter	SR	AvT	Iter	SR	AvT	Iter	SR	AvT	Iter
10	500	1	40	74	23	3	4.6	59	27	0.2	5.5	60	46	6.4	60	2,473	857.1	56	157	4.6	63	560	20.7
		2	39	74	27	4	4.8	60	30	0.6	5.7	59	41	6.4	59	2,262	809.6	56	318	4.8	64	878	20.3
		5	38	75	24	3	4.3	58	28	0.8	5.1	57	40	5.8	57	2,216	790.3	55	312	4.3	63	1,289	19.7
		10	39	74	23	3	4.0	55	28	0.4	4.9	54	42	5.9	55	1,936	723.7	55	316	4.0	62	1,781	19.5
	2000	1	259	73	185	14	4.8	59	249	2.3	5.7	59	335	6.5	59	–	1,490.0	–	4,603	4.8	68	3,842[7]	19.7
		2	224	73	197	15	4.8	59	236	2.1	5.7	58	304	6.4	58	–	1,475.7	–	4,592	4.6	73	–	14.0
		5	245	74	168	15	4.5	58	217	5.0	5.4	58	360	6.5	59	–	1,435.4	–	4,879	4.5	72	–	11.1
		10	260	74	146	15	4.1	56	185	5.3	5.0	56	364	6.6	56	–	1,451.6	–	4,271	4.1	72	–	9.4
	8000	1	1,151	72	883	81	4.8	60	1,069	6.2	5.7	58	1,297	6.1	58	–	439.6	–	–	3.0	–	–	7.2
		2	1,131	73	842	79	4.7	59	1,000	4.9	5.5	58	1,269	6.0	58	–	437.1	–	–	3.0	–	–	3.4
		5	1,404	72	937	79	4.6	59	1,063	4.6	5.4	58	1,442	6.0	58	–	370.9	–	–	3.0	–	–	3.0
		10	1,352	72	767	81	4.1	58	921	17.3	5.0	58	1,347	5.9	58	–	348.9	–	–	3.0	–	–	2.8
20	500	1	295	72	269	22	6.4	61	287	8.1	7.4	62	910	12.0	62	5,435[2]	739.1	58	3,701	6.2	73	–	10.4
		2	346	73	265	23	6.4	62	308	10.0	7.4	62	836	11.5	63	4,940[1]	718.7	87	3,942	6.3	75	–	8.4
		5	359	73	243	22	5.9	63	291	11.8	6.9	63	749	10.9	63	4,904[3]	673.5	69	3,755	5.8	72	–	7.4
		10	423	73	251	25	5.8	61	293	12.2	6.8	61	881	11.5	62	4,475[3]	651.1	58	3,879	5.7	76	–	6.7
	2000	1	2,382	72	1,485	87	6.5	66	1,789	15.7	7.3	66	2,831	8.3	72	–	374.5	–	6,351[5]	4.0	86	–	2.6
		2	2,214	73	1,781	117	6.4	69	1,630	57.9	7.0	69	2,708	8.1	72	–	381.4	–	6,395[5]	3.8	86	–	2.5
		5	2,825	73	1,214	87	5.6	65	1,439	26.9	6.5	65	2,582	7.6	69	–	362.7	–	4,742[5]	3.5	87	–	2.0
		10	2,343	74	1,226	127	5.4	63	1,199	22.4	6.1	63	2,826	7.7	68	–	357.7	–	1,694[2]	3.6	87	–	2.0
	8000	1	–		1,733[5]	781	5.0	88	1,615	219.8	5.3	88	1,446[5]	5.0	87	–	108.7	–	–	2.0	–	–	1.3
		2	–		1,169[5]	586	4.0	87	1,426	233.1	4.1	88	1,117[5]	4.0	87	–	82.4	–	–	2.0	–	–	1.0
		5	–		1,221[5]	578	4.0	85	1,410	150.2	4.6	83	1,404[5]	5.0	82	–	82.1	–	–	2.0	–	–	1.0
		10	–		1,537[5]	575	4.0	79	1,671	66.2	4.7	77	1,592[5]	5.0	77	–	75.4	–	–	2.0	–	–	1.0

"–" in column of AvT means no instance can be solved to optimality within the time limit.

"[·]" in column of AvT means the number of instances (among 10 instances) that can be solved to optimality within 7,200 seconds.

240 instances within 2 hours, while CG_KKT can solve 220 instances, which indicates that the interior point method can greatly improve the performance of the column generation. We also provide an alternative way to implement the KKT-based early-stopping criterion referred to as CG_KKT*. Naturally, it saves certain time for solving the KKT test, yet, the performance is slightly worse in terms of the average solution time when compared with CG_KKT. Generally, the solution time for one more iteration is longer than that of extra KKT tests in CG_KKT. Furthermore, the results in Table 2 show the numerical limitations of BD for solving such problems. There are only a few problem sizes that BD can solve to optimality.

In the case of CPX, the average sparsity ratio is about 73%, which means that about 73% of the optimal solution have a value of zero. Thus, this is the observation that initially motivated us to exploit a CG algorithm. In fact, by using our proposed CG algorithm, we can reduce the sparsity ratio by about 12% on average and 20% at most. This seems to confirm that CG can successfully identify the redundant columns and decrease the problem size. Overall, CG_KKT and CG_KKT* are the most effective methods for solving the large-scale newsvendor problem with the Wasserstein ambiguity set.

5.3. Computational Results with Event-Wise Mean Absolute Deviation Set

In this section, we present computational results of the newsvendor problem under the event-wise mean absolute deviation set by utilizing the seven methods described in Section 5.2. Table 3 presents the average solution time for solving the problem, the average time for evaluating the KKT-test based early stopping criterion, the average number of iterations, and the average sparsity ratio, where the sparsity ratio is defined in Section 5.2. We compare the performance of the approaches described in Section 5.2. Note that CG_KKT* has a similar performance to CG_SCS, and the multi-cut Benders decomposition with level regularization has a similar performance to BD, thus we did not present them in Table 3. We also note that the BD method for the newsvendor problem under an event-wise mean absolute deviation set is similar to the one under a Wasserstein ambiguity set (i.e., see Appendix A). Therefore, we will omit the details of the BD method for the problem under an event-wise mean absolute deviation set.

From Table 3, we can derive some similar observations as discussed in Table 2. For instance, the average solution time increases as the number of scenarios S increases, and is not significantly different when σ varies from 1 to 10. BD is also the most inefficient approach

Table 3 The average CPU solution time (in seconds) for for solving the multi-item DR newsvendor problem with event-wise mean absolute deviation set (AvT), and solving the KKT problem (KKT), the average number of iterations (Iter), and the average sparsity ratio (SR) in percent are reported.

I	S	σ	CPX		CG_KKT				CG_SCS			CG			CG_KKT_PS			BD	
			AvT	SR	AvT	KKT	Iter	SR	AvT	Iter	SR	AvT	Iter	SR	AvT	Iter	SR	AvT	Iter
10	500	0	2	92.7	2	0.0	1.0	92.9	2	1.0	92.9	1	1.0	92.9	1	1.0	92.9	148	18.3
		1	30	96.6	6	0.6	3.0	94.8	6	3.0	94.8	11	5.3	94.7	7	3.0	94.8	432	20.3
		2	29	96.9	6	0.5	3.0	95.2	6	3.0	95.2	11	5.4	95.2	8	3.0	95.2	438	20.9
		5	30	97.2	6	0.4	3.0	95.6	6	3.0	95.6	12	5.7	95.6	11	3.0	95.7	438	21.4
		10	29	97.3	6	0.4	3.0	95.8	6	3.0	95.8	13	6.3	95.8	7	3.0	95.9	432	21.5
	2000	0	9	92.9	6	0.0	1.0	92.9	6	1.0	92.9	6	1.0	92.9	9	1.0	92.9	455	17.1
		1	200	96.1	37	13.3	3.0	93.8	22	3.1	93.8	92	11.6	93.8	84	3.0	93.9	1,591	18.5
		2	205	96.4	45	19.7	3.0	94.3	23	3.1	94.3	64	8.3	94.3	88	3.0	94.4	1,635	19.1
		5	200	96.7	74	48.5	3.0	94.9	25	3.3	94.9	84	9.7	94.9	115	3.0	94.9	1,795	20.7
		10	197	97.1	67	41.6	3.0	95.6	27	3.3	95.6	100	10.8	95.6	107	3.0	95.6	1,681	20.1
	8000	0	43	92.6	25	0.0	1.0	92.9	28	1.0	92.9	32	1.0	92.9	65	1.0	92.9	**	**
		1	850	96.8	153	9.9	3.0	94.9	152	3.0	94.9	1,148	18.0	94.8	1,255	3.0	95.0	**	**
		2	836	96.9	223	77.0	3.0	95.2	218	3.1	95.2	653	11.7	95.1	1,493	3.0	95.2	**	**
		5	796	97.3	228	71.9	3.1	95.8	186	3.2	95.8	679	10.1	95.7	1,637	3.1	95.9	**	**
		10	807	97.3	228	72.3	3.0	95.9	172	3.1	95.9	808	11.4	95.8	1,486	3.0	95.9	**	**
20	500	0	5	96.3	6	0.0	1.0	96.3	5	1.0	96.3	6	1.0	96.3	6	1.0	96.3	473	31.6
		1	161	98.1	25	2.8	3.0	96.9	22	3.0	96.9	59	8.0	96.9	53	3.0	97.0	4,280	35.4
		2	169	98.3	25	2.8	3.0	97.3	22	3.0	97.3	55	7.6	97.3	86	3.0	97.3	4,174	36.1
		5	169	98.5	26	4.0	3.0	97.6	22	3.1	97.6	67	9.1	97.6	58	3.0	97.6	3,724	33.7
		10	156	98.6	24	2.3	3.0	97.8	22	3.0	97.8	75	10.3	97.8	59	3.0	97.9	3,682	34.9
	2000	0	35	96.2	28	0.0	1.0	96.3	28	1.0	96.3	27	1.0	96.3	35	1.0	96.3	**	**
		1	855	98.3	118	13.7	3.0	97.2	104	3.0	97.2	672	17.3	97.2	261	3.1	97.2	**	**
		2	863	98.4	120	13.7	3.0	97.4	106	3.0	97.4	512	12.8	97.4	322	3.0	97.4	**	**
		5	854	98.5	115	10.7	3.0	97.7	104	3.0	97.7	557	13.0	97.7	358	3.0	97.7	**	**
		10	872	98.6	186	80.9	3.0	97.8	80	3.1	97.8	739	17.3	97.8	360	3.0	97.8	**	**
	8000	0	159	96.3	125	0.0	1.0	96.3	125	1.0	96.3	119	1.0	96.3	250[9]	1.0	96.3	**	**
		1	–	–	1,143	98.1	3.0	97.2	1,045	3.0	97.2	1,410	7.7	97.8	3,524[9]	3.0	97.3	**	**
		2	–	–	1,149	101.4	3.0	97.6	1,048	3.0	97.6	779	7.6	97.8	4,249[9]	3.1	97.6	**	**
		5	–	–	1,074	78.0	3.0	97.7	996	3.0	97.7	703	8.1	97.9	5,220[9]	3.0	97.8	**	**
		10	–	–	1,097	68.5	3.0	97.8	1,028	3.0	97.8	837	7.8	97.9	5,245[8]	3.0	97.9	**	**

"–" in column of AvT means no instance can be solved to optimality within the time limit.
"**" in column of AvT and *Iter* means that all the instances that can not be solved to optimality due to out of memory.
"[·]" in column of AvT means the number of instances (among 10 instances) that can be solved to optimality within 7,200 seconds.

among all other five approaches, it can only solve 50% of instances (i.e. 150 out of 300 instances) to optimality. Besides, CG_SCS and CG_KKT significantly outperform CG and CG_KKT_PS, which further confirms the efficiency of the proposed the column selection strategy and the interior point method. However, there are also some different findings. Unlike in Table 2, CG provides a better performance on the average solution time than CPX for most types of instances. In particular, CG reduces the average solution time by an average of about 40%. For $I=20$ and $S=8000$, CG can solve 31 from 50 instances to optimality, while CPX can only solve 10 instances. It is worth noting that CG_SCS and CG_KKT* slightly outperform CG_KKT. These methods have a similar number of iterations, which seems to indicate that there is no need to use an early stopping technique for the CG under the event-wise mean absolute deviation set. Moreover, the results from Table 3 show that the average sparsity ratio in CPX is about 97%, using CG only reduce this number by on average 1.1%. Although the reduction appears small, it actually reflects a large reduction in problem size. For example, the average sparsity ratio from the instances with $I=10$, $S=500$ and $\sigma=1$ shows that the solutions have an average of about 27,000 non-zero variables and 816,513 total variables for CPX, while for CG the solutions have an average of about 27,000 non-zero variables and 516,636 total variables. The CG scheme effectively removed about 300,000 variables from the FMP.

6. Concluding Remarks

This paper studied the distributionally robust multi-item newsvendor problem with the event-wise ambiguity set with event-wise LDRs proposed in Chen et al. (2020). A column generation-based decomposition scheme and several accelerating strategies are developed to solve the problem. A numerical study is performed to confirm that the techniques developed in this paper can improve the computational efficiency. In particular, the proposed algorithm allows us to identify optimal solutions for several large-scale instances that cannot be solved using CPLEX and the multicut-based Benders decomposition method within a two-hour CPU time limit, and significantly reduces the computational time for other instances. It would be a very interesting direction for future research to investigate how to extend our solution scheme to the multi-stage setting and other practical applications such as inventory control problems.

Acknowledgments

The authors are thankful to Jacques Desrosiers and Jean Bertrand Gauthier for their helpful feedback regarding an early draft of this work. Shanshan Wang was partially supported by the National Natural Science Foundation of China [Grant 71972012] and Groupe d'études et de recherche en analyse des décisions (GERAD). Erick Delage was partially supported by the Canadian Natural Sciences and Engineering Research Council [Grant RGPIN-2016-05208 and 492997-2016] and by the Canada Research Chair program [950-230057]. Finally, this research was enabled in part by support provided by Compute Canada (https://www.computecanada.ca/).

References

Ardestani-Jaafari, A., Delage, E., 2016. Robust optimization of sums of piecewise linear functions with application to inventory problems. Operations Research 64, 474–494.

Bansal, M., Huang, K.L., Mehrotra, S., 2018. Decomposition algorithms for two-stage distributionally robust mixed binary programs. SIAM Journal on Optimization 28, 2360–2383.

Bayraksan, G., Love, D.K., 2015. Data-driven stochastic programming using phi-divergences, in: The Operations Research Revolution. INFORMS, pp. 1–19.

Ben-Tal, A., den Hertog, D., De Waegenaere, A., Melenberg, B., Rennen, G., 2013. Robust solutions of optimization problems affected by uncertain probabilities. Management Science 59, 341–357.

Ben-Tal, A., den Hertog, D., Vial, J.P., 2015. Deriving robust counterparts of nonlinear uncertain inequalities. Mathematical programming 149, 265–299.

Bertsimas, D., Gupta, V., Kallus, N., 2018. Robust sample average approximation. Mathematical Programming 171, 217–282.

Bertsimas, D., Sim, M., Zhang, M., 2019. Adaptive distributionally robust optimization. Management Science 65, 604–618.

Bixby, R.E., Gregory, J.W., Lustig, I.J., Marsten, R.E., Shanno, D.F., 1992. Very large-scale linear programming: A case study in combining interior point and simplex methods. Operations Research 40, 885–897.

Chen, Y., Sun, H., Xu, H., 2021. Decomposition and discrete approximation methods for solving two-stage distributionally robust optimization problems. Computational Optimization and Applications 78, 205–238.

Chen, Z., Sim, M., Xiong, P., 2020. Robust stochastic optimization made easy with rsome. Management Science 66, 3329–3339.

Costa, L., Contardo, C., Desaulniers, G., 2019. Exact branch-price-and-cut algorithms for vehicle routing. Transportation Science 53, 946–985.

Delage, E., Saif, A., 2021. The value of randomized solutions in mixed-integer distributionally robust optimization problems. INFORMS Journal on Computing .

Delage, E., Ye, Y., 2010. Distributionally robust optimization under moment uncertainty with application to data-driven problems. Operations Research 58, 595–612.

Desaulniers, G., Desrosiers, J., Solomon, M.M., 2002. Accelerating strategies in column generation methods for vehicle routing and crew scheduling problems, in: Essays and surveys in metaheuristics. Springer, pp. 309–324.

Desrochers, M., Soumis, F., 1989. A column generation approach to the urban transit crew scheduling problem. Transportation science 23, 1–13.

Donohue, K.L., 2000. Efficient supply contracts for fashion goods with forecast updating and two production modes. Management Science 46, 1397–1411.

Du Merle, O., Villeneuve, D., Desrosiers, J., Hansen, P., 1999. Stabilized column generation. Discrete Mathematics 194, 229–237.

Dupacová, J., 2008. Stochastic programming: minimax approach. Encyclopedia of Optimization , 3778–3782.

Gallego, G., Moon, I., 1993. The distribution free newsboy problem: review and extensions. Journal of the Operational Research Society 44, 825–834.

Gamboa, C.A., Valladão, D.M., Street, A., Homem-de Mello, T., 2021. Decomposition methods for wasserstein-based data-driven distributionally robust problems. Operations Research Letters 49, 696–702.

Gao, R., Kleywegt, A.J., 2016. Distributionally robust stochastic optimization with wasserstein distance. arXiv preprint arXiv:1604.02199 .

Gauvin, C., Delage, E., Gendreau, M., 2017. Decision rule approximations for the risk averse reservoir management problem. European Journal of Operational Research 261, 317–336.

Goh, J., Sim, M., 2010. Distributionally robust optimization and its tractable approximations. Operations Research 58, 902–917.

Hanasusanto, G.A., Kuhn, D., Wallace, S.W., Zymler, S., 2015. Distributionally robust multi-item newsvendor problems with multimodal demand distributions. Mathematical Programming 152, 1–32.

Kamburowski, J., 2015. On the distribution-free newsboy problem with some non-skewed demands. Operations Research Letters 43, 165–171.

Kuhn, D., Wiesemann, W., Georghiou, A., 2011. Primal and dual linear decision rules in stochastic and robust optimization. Mathematical Programming 130, 177–209.

Lübbecke, M.E., Desrosiers, J., 2005. Selected topics in column generation. Operations Research 53, 1007–1023.

Luo, F., Mehrotra, S., 2019. Decomposition algorithm for distributionally robust optimization using wasserstein metric with an application to a class of regression models. European Journal of Operational Research 278, 20–35.

Mehrotra, S., Papp, D., 2014. A cutting surface algorithm for semi-infinite convex programming with an application to moment robust optimization. SIAM Journal on Optimization 24, 1670–1697.

Mohajerin Esfahani, P., Kuhn, D., 2018. Data-driven distributionally robust optimization using the wasserstein metric: Performance guarantees and tractable reformulations. Mathematical Programming 171, 115–166.

Namkoong, H., Duchi, J.C., 2016. Stochastic gradient methods for distributionally robust optimization with f-divergences., in: NIPS, pp. 2208–2216.

Natarajan, K., Sim, M., Uichanco, J., 2018. Asymmetry and ambiguity in newsvendor models. Management Science 64, 3146–3167.

Postek, K., Ben-Tal, A., Den Hertog, D., Melenberg, B., 2018. Robust optimization with ambiguous stochastic constraints under mean and dispersion information. Operations Research 66, 814–833.

Postek, K., den Hertog, D., 2016. Multistage adjustable robust mixed-integer optimization via iterative splitting of the uncertainty set. INFORMS Journal on Computing 28, 553–574.

Qin, Y., Wang, R., Vakharia, A.J., Chen, Y., Seref, M.M., 2011. The newsvendor problem: Review and directions for future research. European Journal of Operational Research 213, 361–374.

Rahimian, H., Bayraksan, G., Homem-de Mello, T., 2019. Controlling risk and demand ambiguity in newsvendor models. European Journal of Operational Research 279, 854–868.

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

Rousseau, L.M., Gendreau, M., Feillet, D., 2007. Interior point stabilization for column generation. Operations Research Letters 35, 660–668.

Saif, A., Delage, E., 2021. Data-driven distributionally robust capacitated facility location problem. European Journal of Operational Research 291, 995–1007.

Scarf, H., 1957. A Min-max Solution of an Inventory Problem. Arrow K, Karlin S, Scarf H, eds. Studies in the Mathematical Theory of Inventory and Production, (Stanford University Press, Stanford, CA).

Shapiro, A., Dentcheva, D., Ruszczyński, A., 2014. Lectures on stochastic programming: modeling and theory. SIAM.

Vanderbeck, F., 2000. On dantzig-wolfe decomposition in integer programming and ways to perform branching in a branch-and-price algorithm. Operations Research 48, 111–128.

Wang, S., Li, J., Mehrotra, S., 2021. A solution approach to distributionally robust chance-constrained assignment problems. forthcoming INFORMS Journal on Optimization, pre-print available at http://www.optimization-online.org/DB_FILE/2019/05/7207.pdf .

Wiesemann, W., Kuhn, D., Sim, M., 2014. Distributionally robust convex optimization. Operations Research 62, 1358–1376.

Zverovich, V., Fábián, C.I., Ellison, E.F., Mitra, G., 2012. A computational study of a solver system for processing two-stage stochastic lps with enhanced benders decomposition. Mathematical Programming Computation 4, 211–238.

Appendix A: A Multicut-based Benders Decomposition

In this appendix, we present a multicut-based Benders decomposition method for solving the resulting problems. Specifically, here we focus on the problem (10) under a Wasserstein ambiguity set. We remark that a similar procedure can also be applied to the problem (11) under an event-wise mean absolute deviation set. For the sake of simplicity, we omit it.

Given the solution $(\bar{\boldsymbol{x}}, \bar{\beta})$, we solve the following subproblem for $s \in \mathcal{S}$:

$$h_s(\bar{\boldsymbol{x}}, \bar{\beta}) = \text{minimize } \frac{1}{S}\alpha_s \tag{18a}$$

$$\text{subject to } \alpha_s \geq \sum_{i\in\mathcal{I}} b_i y_i^0(s) + \left(\sum_{i\in\mathcal{I}} b_i \boldsymbol{y}_i^u(s) - \boldsymbol{\eta}_s^1 + \boldsymbol{\eta}_s^2\right)^\top \hat{\boldsymbol{u}}_s + \bar{\boldsymbol{u}}^\top \boldsymbol{\eta}_s^1 - \underline{\boldsymbol{u}}^\top \boldsymbol{\eta}_s^2, \tag{18b}$$

$$w_s^1 + \sum_{i\in\mathcal{I}} b_i y_i^v(s) - \bar{\beta} \leq 0, \tag{18c}$$

$$\sum_{i\in\mathcal{I}} b_i y_{ij}^u(s) - \eta_{js}^1 + \eta_{js}^2 \leq w_s^1, \qquad \forall j \in \mathcal{I}, \tag{18d}$$

$$-\sum_{i\in\mathcal{I}} b_i y_{ij}^u(s) + \eta_{js}^1 - \eta_{js}^2 \leq w_s^1, \qquad \forall j \in \mathcal{I}, \tag{18e}$$

$$y_i^0(s) \geq \bar{x}_i - \hat{u}_{is} + \left(-\boldsymbol{y}_i^u(s) - \boldsymbol{\eta}_{is}^3 + \boldsymbol{\eta}_{is}^4\right)^\top \hat{\boldsymbol{u}}_s + \bar{\boldsymbol{u}}^\top \boldsymbol{\eta}_{is}^3 - \underline{\boldsymbol{u}}^\top \boldsymbol{\eta}_{is}^4, \qquad \forall i \in \mathcal{I}, \tag{18f}$$

$$w_{is}^2 - y_i^v(s) \leq 0, \qquad \forall i \in \mathcal{I}, \tag{18g}$$

$$-y_{ij}^u(s) - \eta_{ijs}^3 + \eta_{ijs}^4 - e_{ij} \leq w_{is}^2, \qquad \forall i,j \in \mathcal{I}, \tag{18h}$$

$$y_{ij}^u(s) + \eta_{ijs}^3 - \eta_{ijs}^4 + e_{ij} \leq w_{is}^2, \qquad \forall i,j \in \mathcal{I}, \tag{18i}$$

$$y_i^0(s) \geq \left(-\boldsymbol{y}_i^u(s) - \boldsymbol{\eta}_{is}^5 + \boldsymbol{\eta}_{is}^6\right)^\top \hat{\boldsymbol{u}}_s + \bar{\boldsymbol{u}}^\top \boldsymbol{\eta}_{is}^5 - \underline{\boldsymbol{u}}^\top \boldsymbol{\eta}_{is}^6, \qquad \forall i \in \mathcal{I}, \tag{18j}$$

$$w_{is}^3 - y_i^v(s) \leq 0, \qquad \forall i \in \mathcal{I}, \tag{18k}$$

$$-y_{ij}^u(s) - \eta_{ijs}^5 + \eta_{ijs}^6 \leq w_{is}^3, \qquad \forall i,j \in \mathcal{I}, \tag{18l}$$

$$y_{ij}^u(s) + \eta_{ijs}^5 - \eta_{ijs}^6 \leq w_{is}^3, \qquad \forall i,j \in \mathcal{I}, \tag{18m}$$

$$\boldsymbol{\eta}_s^1, \boldsymbol{\eta}_s^2, \boldsymbol{\eta}_{is}^3, \boldsymbol{\eta}_{is}^4, \boldsymbol{\eta}_{is}^5, \boldsymbol{\eta}_{is}^6 \geq 0, \qquad \forall i \in \mathcal{I}. \tag{18n}$$

Let $\boldsymbol{\tau}^1$-$\boldsymbol{\tau}^{12}$ be the dual variables of problem (18), then the optimality cut is as follows:

$$\rho_s \geq -\tau_s^2 \beta + \sum_{i\in\mathcal{I}} \tau_{si}^5 x_i - \sum_{i\in\mathcal{I}} \tau_{si}^5 \hat{u}_{is} - \sum_{i\in\mathcal{I}} \tau_{sii}^7 + \sum_{i\in\mathcal{I}} \tau_{sii}^8. \tag{19}$$

The master problem is defined as follows:

$$\text{minimize } -\boldsymbol{b}^\top \boldsymbol{x} + \theta\beta + \sum_{s\in\mathcal{S}} \rho_s \tag{20a}$$

$$\text{subject to } \boldsymbol{c}^\top \boldsymbol{x} = d, \tag{20b}$$

$$\rho_s \geq -\tau_s^{2k} \beta + \sum_{i\in\mathcal{I}} \tau_{si}^{5k} x_i - \sum_{i\in\mathcal{I}} \tau_{si}^{5k} \hat{u}_{is} - \sum_{i\in\mathcal{I}} \tau_{sii}^{7k} + \sum_{i\in\mathcal{I}} \tau_{sii}^{8k}, \qquad k \in [K_s], s \in \mathcal{S}, \tag{20c}$$

$$\boldsymbol{x}, \beta \geq 0. \tag{20d}$$

We can obtain a lower bound (LB) by solving the master problem (20), and obtain an upper bound (UB) through,

$$-\boldsymbol{b}^\top \bar{\boldsymbol{x}} + \bar{\beta}\theta + \sum_{s\in\mathcal{S}} h_s(\bar{\boldsymbol{x}}, \bar{\beta}). \tag{21}$$

Finally, Algorithm 2 describes the procedure of our implementation of the Benders decomposition method.

Algorithm 2: Benders Decomposition Implementation

Input A tolerance $\epsilon \geq 0$ and maximum run time *stoptime*
Initialize The number of iteration $\ell = 0$, $LB = -\infty$, $UB = +\infty$, $K_s = 0$ for $s \in \mathcal{S}$.
while ($runtime \leq stoptime$ and $UB - LB > \epsilon$)
    Set $\ell = \ell + 1$, solve the master problem (20).
    Record optimal solution $(\boldsymbol{x}^\ell, \beta^\ell, \boldsymbol{\rho}^\ell)$ and optimal objective $lobj^\ell$.
    Update $LB = lobj^\ell$.
    Fix $\bar{\boldsymbol{x}} := \boldsymbol{x}^\ell$, $\bar{\beta} := \beta^\ell$.
    for $s \in \mathcal{S}$
        Solve the subproblem (18).
        Obtain $h_s(\bar{\boldsymbol{x}}, \bar{\beta})$ and $\boldsymbol{\tau}_s^*$.
        if $\rho_s < h_s(\bar{\boldsymbol{x}}, \bar{\beta})$
            Let $K_s = K_s + 1$ and $\boldsymbol{\tau}_s^{K_s} = \boldsymbol{\tau}_s^*$.
        end if
    end for
    Update $UB = \min\{UB, uobj^\ell\}$ where $uobj^\ell$ is as described in (21).
end while
return UB and the corresponding optimal solution.