manuscript

Robust Stochastic Optimization Made Easy with RSOME

Zhi Chen
Department of Management Sciences, College of Business, City University of Hong Kong, Kowloon Tong, Hong Kong
zhi.chen@cityu.edu.hk

Melvyn Sim, Peng Xiong
Department of Analytics & Operations, NUS Business School, National University of Singapore, 119077, Singapore
dscsimm@nus.edu.sg, bizxio@nus.edu.sg

We present a new distributionally robust optimization model called *robust stochastic optimization* (RSO), which unifies both scenario-tree based stochastic linear optimization and distributionally robust optimization in a practicable framework that can be solved using the state-of-the-art commercial optimization solvers. We also develop a new algebraic modeling package, RSOME to facilitate the implementation of RSO models. The model of uncertainty incorporates both discrete and continuous random variables, typically assumed in scenario-tree based stochastic linear optimization and distributionally robust optimization respectively. To address the non-anticipativity of recourse decisions, we introduce the event-wise recourse adaptations, which integrate the scenario-tree adaptation originating from stochastic linear optimization and the affine adaptation popularized in distributionally robust optimization. Our proposed event-wise ambiguity set is rich enough to capture traditional statistic-based ambiguity sets with convex generalized moments, mixture distribution, ϕ-divergence, Wasserstein (Kantorovich-Rubinstein) metric, and also inspire machine-learning-based ones using techniques such as K-means clustering, and classification and regression trees. Several interesting RSO models, including optimizing over the Hurwicz criterion and two-stage problems over Wasserstein ambiguity sets, are provided.

Key words: stochastic linear optimization, distributionally robust optimization, machine learning.
History: January 12, 2020.

---

1. Introduction

In the era of data analytics, the ubiquity of general purpose deterministic mathematical optimization frameworks such as linear, mixed-integer and conic optimization models, as well as their impact on improving management decision-making, cannot be understated. Algebraic modeling packages and state-of-the-art optimization solvers have been developed on these successful frameworks to facilitate implementation of prescriptive analytics to address a wide variety of real-world problems. Comparatively, frameworks to support generic modeling and optimization under uncertainty, despite their importance, are relatively less established. These frameworks include stochastic linear optimization, robust optimization and more recently, distributionally robust optimization, each of them has its strengths and weaknesses.

Stochastic linear optimization extends the linear optimization framework to minimize the total average cost associated with the optimal *here-and-now* and *wait-and-see* (or recourse) decisions under a known probability distribution (Danzig 1955). For enormous or infinite number of scenarios, we can obtain approximate here-and-now solutions using the sample average approximation (SAA) (Kall and Wallace 1994, Birge and Louveaux 2011, Shapiro and Homem-de Mello 1998, Kleywegt et al. 2002). These approximate solutions are necessarily random and *optimistically biased*, *i.e.*, the actual realized objectives are statistically worse off than those attained by using SAA. Stochastic linear optimization has the versatility of modeling different types of recourse decisions including those with discrete outcomes, albeit at the expense of greater computational effort.

In classical robust optimization (Soyster 1973, Ben-Tal and Nemirovski 1998, El Ghaoui et al. 1998, Bertsimas and Sim 2004, Ben-Tal et al. 2015), the solution is obtained by reformulating the model to a deterministic optimization problem that can be solved using available solvers. The underlying uncertainty is a distribution-free continuous random variable with support confined to a convex *uncertainty set*. The solution obtained via classical robust optimization hedges against the worst-case outcome within the uncertainty set and hence is *pessimistically biased*, *i.e.*, the realized objective value by the robust solution would often be better than the objective value attained by solving the robust optimization problem. To reduce the conservativeness, distributionally robust optimization incorporates an *ambiguity set* of probability distributions and its solution hedges against the worst-case distribution within the ambiguity set (Dupačová 1976, Shapiro and Kleywegt 2002, El Ghaoui et al. 2003, Shapiro and Ahmed 2004, Delage and Ye 2010, Wiesemann et al. 2014). Under an embedded linear optimization framework, both robust optimization and distributionally robust optimization have been extended to address problems with recourse decisions (see, for instance, Ben-Tal et al. 2004, Takriti and Ahmed 2004, Bertsimas et al. 2019b). As these models are generally computationally intractable (Shapiro and Nemirovski 2005, Ben-Tal et al. 2004), approximate solutions are sought by restricting the recourse decisions to affine mappings of the uncertainty and by requiring the recourse decisions to remain feasible almost surely.

In this paper, we introduce a new distributionally robust optimization model that we call *robust stochastic optimization* (RSO). From the methodological perspective, the RSO model unifies both scenario-tree based stochastic linear optimization and distributionally robust optimization in a tractable framework. Specifically, the RSO framework incorporates both discrete and continuous random variables, and introduces the event-wise static and event-wise affine adaptations to address the non-anticipativity of recourse decisions. The equipped event-wise ambiguity set is rich to capture traditional statistic-based ambiguity sets and also opens up new ones based on machine learning techniques such as K-means clustering, and classification and regression trees. We showcase several interesting models that we can integrate in our framework. From the practical perspective, a wide variety of stochastic and disrtributionally robust optimization problems can be effectively prototyped and tested in an algebraic modeling package based on such a unified RSO model. Indeed, we design the RSO framework with the mindset that it can be integrated in a general purpose software that would be accessible to modelers. As a proof of concept, we develop an algebraic modeling package, RSOME (Robust Stochastic Optimization Made Easy) to facilitate modeling of problems under the RSO framework. With its intuitive syntax, RSOME provides a convenient way for modelers to specify RSO models and interfaces with state-of-the-art commercial solvers such as CPLEX, Gurobi, and MOSEK to obtain the optimal solutions to these models.

Our work significantly extends the frameworks of Wiesemann et al. (2014) and Bertsimas et al. (2019b). In terms of modeling uncertainty, the ambiguity sets in these two models do not include discrete random variable and they are special cases of our event-wise ambiguity set. Moreover, their ambiguity sets are chiefly moment-based and do not naturally incorporate statistical-distance-based-information such as Wasserstein metric or ϕ-divergence. We also would like to highlight the difficulties of developing an algebraic modeling toolbox based on the Wiesemann et al. (2014) framework. Among other things, to ensure tractability, the model requires to impose a nesting condition on the confidence sets and a key challenge is to check whether the nesting condition holds. To resolve these issues collectively, we introduce a discrete random variable as part of the event-wise ambiguity set and associate its outcome with confidence sets. An immediate consequence is the assurance of tractability without having to impose additional conditions on the confidence sets such as nesting and disjoint. In terms of recourse adaptation, Wiesemann et al. (2014) do not consider recourse decisions, while the model of Bertsimas et al. (2019b) is based on affine recourse adaptation, which is a special case of our event-wise affine adaptation.

Notations. Boldface uppercase and lowercase characters denote matrices and vectors, respectively. We denote by $[N] \triangleq \{1, 2, \ldots, N\}$ the set of positive running indices up to N. We use $\mathcal{P}_0(\mathbb{R}^I)$ to represent the set of all distributions on $\mathbb{R}^I$. A random variable, $\tilde{\boldsymbol{z}}$ is denoted with a tilde sign and we use $\tilde{\boldsymbol{z}} \sim \mathbb{P}$, $\mathbb{P} \in \mathcal{P}_0(\mathbb{R}^{I_z})$ to define $\tilde{\boldsymbol{z}}$ as an I_z-dimensional random variable with distribution $\mathbb{P}$.

2. Framework for Robust Stochastic Optimization

We now introduce the *robust stochastic optimization* (RSO) model, which combines both scenario-tree based stochastic linear optimization and distributionally robust optimization in a unified framework. The uncertainty associated with the RSO model comprises both discrete and continuous random variables. Specifically, $\tilde{s}$ represents a discrete random scenario taking values in $[S]$, while $\tilde{\boldsymbol{z}}$ represents a continuous random variable with outcomes in $\mathbb{R}^{I_z}$. Conditioning on the realization of a scenario $s \in [S]$, the support set of the random variable $\tilde{\boldsymbol{z}}$ is tractable conic representable and is denoted by $\mathcal{Z}_s$. The joint distribution of $(\tilde{\boldsymbol{z}}, \tilde{s})$ is denoted by $\mathbb{P} \in \mathcal{F}$, where $\mathcal{F}$ is the *ambiguity set* of probability distributions that share some identical distributional information.

We denote by $\boldsymbol{w} \in \mathbb{R}^{J_w}$ the *here-and-now* decision of the RSO model. The recourse decisions depend on the realization of the random variables $\tilde{\boldsymbol{z}}$ and $\tilde{s}$. As it would become clear shortly, to obtain a tractable reformulation, we introduce two types of recourse decisions, which are function maps respectively denoted by $\boldsymbol{x}(s) : [S] \mapsto \mathbb{R}^{J_x}$ and $\boldsymbol{y}(s, \boldsymbol{z}) : [S] \times \mathbb{R}^{I_z} \mapsto \mathbb{R}^{J_y}$. Here, $\boldsymbol{x}(\cdot)$ adapts only to the outcome of the random scenario $\tilde{s}$, while $\boldsymbol{y}(\cdot, \cdot)$ adapts to the outcomes of $\tilde{s}$ and $\tilde{\boldsymbol{z}}$. Similar to most tractable adaptive robust optimization problems, the RSO model requires that for each given scenario $s \in [S]$, the function map $\boldsymbol{y}(s, \boldsymbol{z})$ is affinely dependent on $\boldsymbol{z}$ as follows:

$$\boldsymbol{y}(s, \boldsymbol{z}) \triangleq \boldsymbol{y}^0(s) + \sum_{i \in [I_z]} \boldsymbol{y}^i(s) z_i,$$

where $\boldsymbol{y}^0(s), \ldots, \boldsymbol{y}^{I_z}(s)$ account for the raw decision variables associated with $\boldsymbol{y}(\cdot, \cdot)$ at scenario s.

To characterize the objective function (with index 0) and constraints (with indices $m \in [M]$), we first define the following random variable mappings for all $m \in [M] \cup \{0\}$,

$$\begin{cases} \boldsymbol{a}_m(s, \boldsymbol{z}) \triangleq \boldsymbol{a}_{ms}^0 + \sum_{i \in [I_z]} \boldsymbol{a}_{ms}^i z_i \\ \boldsymbol{b}_m(s, \boldsymbol{z}) \triangleq \boldsymbol{b}_{ms}^0 + \sum_{i \in [I_z]} \boldsymbol{b}_{ms}^i z_i \\ \quad \boldsymbol{c}_m(s) \triangleq \boldsymbol{c}_{ms} \\ d_m(s, \boldsymbol{z}) \triangleq d_{ms}^0 + \sum_{i \in [I_z]} d_{ms}^i z_i \end{cases} \tag{1}$$

for given parameters

$$\boldsymbol{a}_{ms}^i \in \mathbb{R}^{J_w}, \boldsymbol{b}_{ms}^i \in \mathbb{R}^{J_x}, \boldsymbol{c}_{ms} \in \mathbb{R}^{J_y}, d_{ms}^i \in \mathbb{R} \quad \forall i \in [I_z] \cup \{0\},\ s \in [S].$$

The objective function of the RSO model to be minimized,

$$\sup_{\mathbb{P} \in \mathcal{F}} \mathbb{E}_{\mathbb{P}} \left[\boldsymbol{a}_0^\top(\tilde{s}, \tilde{\boldsymbol{z}}) \boldsymbol{w} + \boldsymbol{b}_0^\top(\tilde{s}, \tilde{\boldsymbol{z}}) \boldsymbol{x}(\tilde{s}) + \boldsymbol{c}_0^\top(\tilde{s}) \boldsymbol{y}(\tilde{s}, \tilde{\boldsymbol{z}}) + d_0(\tilde{s}, \tilde{\boldsymbol{z}}) \right],$$

reflects the ambiguity aversion of the decision maker against an ambiguity set $\mathcal{F}$ that we will introduce subsequently. Note that the random variable $\tilde{\boldsymbol{z}} \triangleq (\tilde{\boldsymbol{u}}, \tilde{\boldsymbol{v}})$ includes both the primary I_u-dimensional random variable $\tilde{\boldsymbol{u}}$ and the auxiliary (or lifted) I_v-dimensional random variable $\tilde{\boldsymbol{v}}$ associated with $\tilde{\boldsymbol{u}}$. As in the same spirit of linear optimization models, the provision of the auxiliary random variable $\tilde{\boldsymbol{v}}$ would greatly enhance the modeling power of the RSO model.

There are two types constraints: hard and soft ones, which are respectively associated with the partition $\mathcal{M}_1, \mathcal{M}_2 \subseteq [M]$ of indices of constraints.

The hard constraints (with indices $m \in \mathcal{M}_1$) of the RSO model, which must be satisfied almost surely, are given by the following set of semi-infinite constraints:

$$\boldsymbol{a}_m^\top(s,\boldsymbol{z})\boldsymbol{w} + \boldsymbol{b}_m^\top(s,\boldsymbol{z})\boldsymbol{x}(s) + \boldsymbol{c}_m^\top(s)\boldsymbol{y}(s,\boldsymbol{z}) + d_m(s,\boldsymbol{z}) \leq 0 \quad \forall \boldsymbol{z} \in \mathcal{Z}_s,\ s \in [S],\ m \in \mathcal{M}_1.$$

Observe that for each given scenario $s \in [S]$, the semi-infinite constraint corresponds to the standard linear robust counterpart, which can be transformed to a modest sized constraint system that can be handled by modern solvers. Scenario-tree based stochastic linear optimization is a special case of the RSO model when in the absence of the recourse decision $\boldsymbol{y}(\cdot,\cdot)$. Likewise, adaptive robust optimization is also a special case when $S = 1$.

To enhance the modeling, RSO also supports soft constraints (with indices $m \in \mathcal{M}_2$), which must be satisfied in expectation over all distributions within the ambiguity set $\mathcal{F}$:

$$\sup_{\mathbb{P}\in\mathcal{F}} \mathbb{E}_{\mathbb{P}}\left[\boldsymbol{a}_m^\top(\tilde{s},\tilde{\boldsymbol{z}})\boldsymbol{w} + \boldsymbol{b}_m^\top(\tilde{s},\tilde{\boldsymbol{z}})\boldsymbol{x}(\tilde{s}) + \boldsymbol{c}_m^\top(\tilde{s})\boldsymbol{y}(\tilde{s},\tilde{\boldsymbol{z}}) + d_m(\tilde{s},\tilde{\boldsymbol{z}})\right] \leq 0 \quad \forall m \in \mathcal{M}_2.$$

As in the objective function, soft constraints are evaluated in the expected sense, and hence, they capture the risk neutrality of the decision maker under ambiguity aversion. By introducing additional recourse decisions that are also embedded in the hard constraints, the RSO model is capable to capture risk-averse objective functions or safeguarding constraints; see Section 5.

Apart from hard and soft constraints, for a given scenario s, we can impose additional constraints jointly on the decisions $\boldsymbol{w}$, $\boldsymbol{x}(s)$ and $\boldsymbol{y}^0(s),\ldots,\boldsymbol{y}^{I_z}(s)$. Specifically, we have

$$\boldsymbol{r}(s) \triangleq \left(\boldsymbol{w}, \boldsymbol{x}(s), \boldsymbol{y}^0(s), \ldots, \boldsymbol{y}^{I_z}(s)\right) \in \mathcal{X}_s \quad \forall s \in [S],$$

where the feasible set $\mathcal{X}_s$ may encompass nonlinear constraints such as conic and integral ones.

3. Event-Wise Recourse Adaptations

Stochastic linear optimization and distributionally robust optimization have different approaches for addressing dynamic decision-making, where uncertainty is revealed in stages and recourse decisions taken in different stages should be non-anticipative to uncertainty realization. In distributionally robust optimization, this can be achieved by restricting the dependency of a recourse decision on only a subset of the uncertainty $\tilde{\boldsymbol{z}}$ that has been revealed (see an example in Figure 1). In stark contrast, dynamic modeling in scenario-tree based stochastic linear optimization is more involved and requires enumerating the complete sample paths from the beginning to the end of the decision horizon. In this regard, a scenario represents a sample path and a scenario tree is typically used to showcase sample paths as well as decisions (Høyland and Wallace 2001, Pflug 2001, Heitsch and Römisch 2009). Figure 2 presents the scenario tree for a three-stage problem with five scenarios: in accordance of non-anticipativity, the first-stage decision $\boldsymbol{w}$ is independent of the scenarios; while

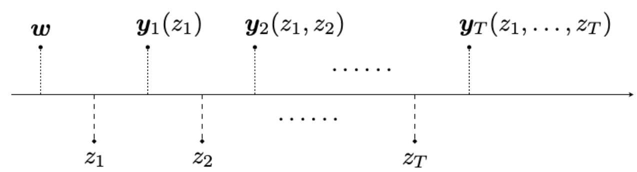

Figure 1 Timeline of a multi-stage problem. Uncertain parameters are revealed in stages as $z_1, \ldots, z_T$. The decision $\boldsymbol{w}$, made before any uncertainty realizes, is non-adaptive to any uncertain parameters. The recourse decision $\boldsymbol{y}_t$, made after observing the uncertainty realization z_t, is adaptive to all revealed $z_1, \ldots, z_t$.

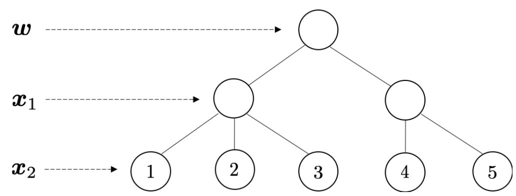

Figure 2 Scenario tree. The recourse decision $\boldsymbol{x}_1(\cdot)$ satisfies $\boldsymbol{x}_1(1) = \boldsymbol{x}_1(2) = \boldsymbol{x}_1(3)$ and $\boldsymbol{x}_1(4) = \boldsymbol{x}_1(5)$, while the recourse decision $\boldsymbol{x}_2(\cdot)$ can take differently depending on the outcome from scenarios $1, 2, \ldots, 5$. The second level of the scenario tree gives a collection of two MECE events $\{1,2,3\}$ and $\{4,5\}$, while the third level gives a collection of five singleton MECE events $\{1\}$, $\{2\}$, $\{3\}$, $\{4\}$, and $\{5\}$. The collection of MECE events associated with a lower level (*e.g.*, the third level) is nested in the collection associated with a higher level (*e.g.*, the second level).

the second-stage decision $\boldsymbol{x}_1(\cdot)$ shall be indifferent among scenarios 1, 2, and 3 and be indifferent between scenarios 4 and 5, and the third stage decision $\boldsymbol{x}_2(\cdot)$ can adapt to scenarios $1, 2, \ldots, 5$. Inspired by these two distinctive approaches, we propose a novel event-wise recourse adaptation scheme that incorporates their essences in terms of modeling dynamic decision-making: the proposed scheme allows the recourse decision to be adaptive to event realization, as in stochastic linear optimization, and affinely adaptive to revealed uncertainty, as in distributionally robust optimization. When there is only one scenario (*i.e.*, event), the event-wise recourse adaptation reduces to the affine recourse adaptation in Bertsimas et al. (2019b).

To formally specify the event-wise adaptation of the recourse decision, $\boldsymbol{x}(\cdot)$, we first define an event $\mathcal{E} \subseteq [S]$ by a subset of scenarios. A partition of scenarios then induces a collection $\mathcal{C}$ of *mutually exclusive* and *collectively exhaustive* (MECE) events. Correspondingly, we define a mapping $\mathcal{H}_{\mathcal{C}} : [S] \mapsto \mathcal{C}$ such that $\mathcal{H}_{\mathcal{C}}(s) = \mathcal{E}$, for which $\mathcal{E}$ is the only event in $\mathcal{C}$ that contains the scenario s. Given a collection $\mathcal{C}$ of MECE events, we define the *event-wise static adaptation*,

$$\mathcal{A}(\mathcal{C}) \triangleq \left\{ x : [S] \mapsto \mathbb{R} \;\middle|\; \begin{array}{l} x(s) = x^{\mathcal{E}}, \ \mathcal{E} = \mathcal{H}_{\mathcal{C}}(s) \\ \text{for some } x^{\mathcal{E}} \in \mathbb{R} \end{array} \right\},$$

which follows along a similar vein as in scenario-tree based stochastic linear optimization. Note that each level of the scenario tree naturally gives a partition of scenarios that induces a collection of MECE events, which in turn is used in the input of event-wise recourse adaptations for recourse decisions associated with that level (*i.e.*, stage); see Figure 2. In Appendix E, we provide a financial planning example to illustrate the use of event-wise static adaptation to formulate a multi-stage stochastic linear optimization problem via scenario tree.

Similarly, for the recourse decision, $\boldsymbol{y}(\cdot,\cdot)$, we define the *event-wise affine adaptation*,

$$\bar{\mathcal{A}}(\mathcal{C},\mathcal{I}) \triangleq \left\{ y : [S]\times\mathbb{R}^{I_z} \mapsto \mathbb{R} \;\middle|\; \begin{array}{l} y(s,\boldsymbol{z}) = y^0(s) + \sum_{i\in\mathcal{I}} y^i(s) z_i \\ \text{for some } y^0, y^i \in \mathcal{A}(\mathcal{C}), i \in \mathcal{I} \end{array} \right\}$$

for a subset $\mathcal{I} \subseteq [I_z]$. As in distributionally robust optimization, the information set $\mathcal{I}$ of the recourse decision $\boldsymbol{y}(\cdot,\cdot)$ taken at a certain stage tracks the indices of revealed uncertainties up to that stage, and $\boldsymbol{y}(\cdot,\cdot)$ is affinely adaptive to those components of $\boldsymbol{z}$ residing in the information set. Along with the collection $\mathcal{C}$ of MECE events, the information set captures the non-anticipativity of $\boldsymbol{y}(\cdot,\cdot)$.

Armed with the event-wise recourse adaptations, we propose the following RSO framework:

$$\begin{array}{lll}
\min & \sup_{\mathbb{P}\in\mathcal{F}} \mathbb{E}_{\mathbb{P}}\left[\boldsymbol{a}_0^\top(\tilde{s},\tilde{\boldsymbol{z}})\boldsymbol{w} + \boldsymbol{b}_0^\top(\tilde{s},\tilde{\boldsymbol{z}})\boldsymbol{x}(\tilde{s}) + \boldsymbol{c}_0^\top(\tilde{s})\boldsymbol{y}(\tilde{s},\tilde{\boldsymbol{z}}) + d_0(\tilde{s},\tilde{\boldsymbol{z}})\right] & \\
\text{s.t.} & \boldsymbol{a}_m^\top(s,\boldsymbol{z})\boldsymbol{w} + \boldsymbol{b}_m^\top(s,\boldsymbol{z})\boldsymbol{x}(s) + \boldsymbol{c}_m^\top(s)\boldsymbol{y}(s,\boldsymbol{z}) + d_m(s,\boldsymbol{z}) \le 0 & \forall \boldsymbol{z}\in\mathcal{Z}_s,\ s\in[S],\ m\in\mathcal{M}_1 \\
& \sup_{\mathbb{P}\in\mathcal{F}} \mathbb{E}_{\mathbb{P}}\left[\boldsymbol{a}_m^\top(\tilde{s},\tilde{\boldsymbol{z}})\boldsymbol{w} + \boldsymbol{b}_m^\top(\tilde{s},\tilde{\boldsymbol{z}})\boldsymbol{x}(\tilde{s}) + \boldsymbol{c}_m^\top(\tilde{s})\boldsymbol{y}(\tilde{s},\tilde{\boldsymbol{z}}) + d_m(\tilde{s},\tilde{\boldsymbol{z}})\right] \le 0 & \forall m\in\mathcal{M}_2 \\
& (\boldsymbol{w}, \boldsymbol{x}(s), \boldsymbol{y}^0(s), \dots, \boldsymbol{y}^{I_z}(s)) \in \mathcal{X}_s & \forall s\in[S] \\
& x_j \in \mathcal{A}(\mathcal{C}_x^j) & \forall j\in[J_x] \\
& y_j \in \bar{\mathcal{A}}(\mathcal{C}_y^j, \mathcal{I}_y^j) & \forall j\in[J_y],
\end{array} \tag{2}$$

for given $\mathcal{C}_x^j, j\in[J_x]$ and $\mathcal{C}_y^j, j\in[J_y]$ of MECE events, and information index sets $\mathcal{I}_y^j, j\in[J_y]$.

The RSO model (2) includes both static and adaptive multi-stage problems in its presentation: it is static when there is only one MECE event of scenarios and no event-wise affine adaptation, and is dynamic otherwise; it spans across more than two stages when there are uncertainty realizations in more than one stage and recourse decisions taken after each time of uncertainty realization. It has been well known that affine recourse adaptation can be extended to multi-stage problems, by restricting a recourse decision to be selectively dependent on a subset of uncertain parameters which would have been realized when that recourse decision has to be made. When there is only one scenario (*i.e.*, $S = 1$) and no soft constraint (*i.e.*, without the second collection of constraints), the RSO model (2) would recover the general multi-stage problems in Bertsimas et al. (2019b, § 3.2). We refer interested readers to Bertsimas and Thiele (2006), See and Sim (2010), Delage and Iancu (2015) and references therein for more details and examples of modeling and optimizing multi-stage problems using distributionally robust optimization.

For a given scenario $s \in [S]$, the objective function and constraints in the RSO model are bi-affine functions of the underlying decision variable $\boldsymbol{r}(s) \in \mathbb{R}^{J_r}$. Hence, we can write

$$\boldsymbol{a}_m^\top(s,\boldsymbol{z})\boldsymbol{w} + \boldsymbol{b}_m^\top(s,\boldsymbol{z})\boldsymbol{x}(s) + \boldsymbol{c}_m^\top(s)\boldsymbol{y}(s,\boldsymbol{z}) + d_m(s,\boldsymbol{z}) \triangleq \boldsymbol{r}^\top(s)\boldsymbol{G}_m(s)\boldsymbol{z} + h_m(s) \quad \forall m \in [M]\cup\{0\},$$

for parameters $\boldsymbol{G}_m(s) \in \mathbb{R}^{J_r \times I_z}$ and $h_m(s) \in \mathbb{R}$. This relation would enable us to reformulate the hard constraints into deterministic constraint systems using standard robust optimization techniques for the tractable representation of an uncertainty-affected linear inequality; see Theorem 1.3.4 and Chapter 1 in Ben-Tal et al. (2009).

The expansive RSO model can be reformulated as a deterministic optimization problem using our developed algebraic modeling toolbox, RSOME. Before we could do so, we will next introduce the ambiguity set for the objective function and constraints.

4. Event-Wise Ambiguity Set

We propose the event-wise ambiguity set, which is representable in the format

$$\mathcal{F} = \left\{ \mathbb{P} \in \mathcal{P}_0\left(\mathbb{R}^{I_z} \times [S]\right) \;\middle|\; \begin{array}{ll} (\tilde{\boldsymbol{z}}, \tilde{s}) \sim \mathbb{P} & \\ \mathbb{E}_{\mathbb{P}}\left[\tilde{\boldsymbol{z}} \mid \tilde{s} \in \mathcal{E}_k\right] \in \mathcal{Q}_k & \forall k \in [K] \\ \mathbb{P}\left[\tilde{\boldsymbol{z}} \in \mathcal{Z}_s \mid \tilde{s} = s\right] = 1 & \forall s \in [S] \\ \mathbb{P}\left[\tilde{s} = s\right] = p_s & \forall s \in [S] \\ \text{for some } \boldsymbol{p} \in \mathcal{P} & \end{array} \right\} \tag{3}$$

for given events $\mathcal{E}_k, k \in [K]$ and given closed and convex sets $\mathcal{Z}_s, s \in [S]$, $\mathcal{Q}_k, k \in [K]$, and $\mathcal{P} \subseteq \{\boldsymbol{p} \in \mathbb{R}^S_{++} \mid \sum_{s\in[S]} p_s = 1\}$. The random variable $\tilde{s}$ indicates a set of random scenarios whose realization probabilities may be uncertain. For different scenarios, the support of the random variable $\tilde{\boldsymbol{z}}$ could be different, while conditioning on the event realization, the expecation of $\tilde{\boldsymbol{z}}$ can also differ. Quite notably, we can effectively determine the worst-case expectation over the event-wise ambiguity set $\mathcal{F}$ by solving a classical robust optimization problem.

Theorem 1. *Assuming the Slater's condition holds, the worst-case expectation*

$$\sup_{\mathbb{P}\in\mathcal{F}} \mathbb{E}_{\mathbb{P}}\left[\boldsymbol{r}^\top(\tilde{s})\boldsymbol{G}_m(\tilde{s})\tilde{\boldsymbol{z}} + h_m(\tilde{s})\right]$$

is equivalent to the optimal value of the following classical robust optimization problem:

$$\begin{array}{lll} \inf & \gamma & \\ \text{s.t.} & \gamma \geq \boldsymbol{\alpha}^\top\boldsymbol{p} + \sum\limits_{k\in[K]} \boldsymbol{\beta}_k^\top \boldsymbol{\mu}_k & \forall \boldsymbol{p} \in \mathcal{P},\ \dfrac{\boldsymbol{\mu}_k}{\sum_{s\in\mathcal{E}_k} p_s} \in \mathcal{Q}_k,\ k \in [K] \\ & \alpha_s + \sum\limits_{k\in\mathcal{K}_s} \boldsymbol{\beta}_k^\top \boldsymbol{z} \geq \boldsymbol{r}^\top(s)\boldsymbol{G}_m(s)\boldsymbol{z} + h_m(s) & \forall \boldsymbol{z} \in \mathcal{Z}_s,\ s \in [S] \\ & \gamma \in \mathbb{R},\ \boldsymbol{\alpha} \in \mathbb{R}^S,\ \boldsymbol{\beta}_k \in \mathbb{R}^{I_z} & \forall k \in [K], \end{array} \tag{4}$$

where for each $s \in [S]$, $\mathcal{K}_s = \{k \in [K] \mid s \in \mathcal{E}_k\}$.

The tractability of problem (4) depends on the uncertainty sets $\mathcal{Z}_s, s \in [S], \mathcal{Q}_k, k \in [k]$ and $\mathcal{P}$. For practicability, these sets are confined to tractable conic representable sets such as polyhedral or second-order conic representable ones. By using an algebraic modeling toolbox such as RSOME, the classical robust optimization problem is automatically transformed to a polynomial sized linear or second-order conic optimization problem, which can be solved by commercial solvers such as CPLEX, Gurobi and MOSEK. In terms of generalization, the event-wise ambiguity encompasses a wide spectrum of existing ambiguity sets in its intuitive expression and may inspire new ones based on machine learning techniques.

Uncertain Discrete Distribution

The event-wise ambiguity set can naturally specify uncertain discrete distributions as follows:

$$\mathcal{F} = \left\{ \mathbb{P} \in \mathcal{P}_0\left(\mathbb{R}^{I_z} \times [S]\right) \;\middle|\; \begin{array}{ll} (\tilde{\boldsymbol{z}}, \tilde{s}) \sim \mathbb{P} & \\ \mathbb{P}\left[\tilde{\boldsymbol{z}} \in \mathcal{Z}_s \mid \tilde{s} = s\right] = 1 & \forall s \in [S] \\ \mathbb{P}\left[\tilde{s} = s\right] = p_s & \forall s \in [S] \\ \text{for some } \boldsymbol{p} \in \mathcal{P} & \end{array} \right\},$$

where each $\mathcal{Z}_s = \{\hat{\boldsymbol{z}}_s\}$ is a singleton set. As proposed in Ben-Tal et al. (2013), ϕ-divergences can be used to characterize the uncertainty set $\mathcal{P}$ of discrete probability distributions. If the uncertainty set $\mathcal{P}$ is a singleton set, *i.e.*, $\boldsymbol{p}$ is fixed, then the corresponding ambiguity set would also shrinkage to a singleton set containing only the known discrete distribution $\frac{1}{S}\sum_{s \in [S]} p_s \delta_{\hat{\boldsymbol{z}}_s}$.

Generalized-Moment Ambiguity Set

Wiesemann et al. (2014) formally introduce the following generalized-moment ambiguity set that is based on a convex function $\boldsymbol{\phi} : \mathbb{R}^{I_u} \mapsto \mathbb{R}^{I_v}$:

$$\mathcal{G} = \left\{ \mathbb{P} \in \mathcal{P}_0\left(\mathbb{R}^{I_u}\right) \;\middle|\; \begin{array}{l} \tilde{\boldsymbol{u}} \sim \mathbb{P} \\ \mathbb{E}_{\mathbb{P}}\left[\tilde{\boldsymbol{u}}\right] \in \mathcal{Q} \\ \mathbb{E}_{\mathbb{P}}\left[\boldsymbol{\phi}(\tilde{\boldsymbol{u}})\right] \leq \boldsymbol{\sigma} \\ \mathbb{P}\left[\tilde{\boldsymbol{u}} \in \mathcal{U}\right] = 1 \end{array} \right\}.$$

The generalized moments characterized by the convex function $\boldsymbol{\phi}$ can provide useful statistical characterizations of the uncertainty $\tilde{\boldsymbol{u}}$, including (co)-variance, absolute deviation, semi-variance, and expected utility, among others. Based on the lifting and projection theorem (Wiesemann et al. 2014 , Theorem 5), it holds that $\Pi_{\tilde{\boldsymbol{u}}}\mathcal{F} = \mathcal{G}$, where

$$\mathcal{F} = \left\{ \mathbb{P} \in \mathcal{P}_0\left(\mathbb{R}^{I_u + I_v} \times \{1\}\right) \;\middle|\; \begin{array}{l} ((\tilde{\boldsymbol{u}}, \tilde{\boldsymbol{v}}), \tilde{s}) \sim \mathbb{P} \\ \mathbb{E}_{\mathbb{P}}\left[\tilde{\boldsymbol{u}} \mid \tilde{s} = 1\right] \in \mathcal{Q} \\ \mathbb{E}_{\mathbb{P}}\left[\tilde{\boldsymbol{v}} \mid \tilde{s} = 1\right] \leq \boldsymbol{\sigma} \\ \mathbb{P}\left[(\tilde{\boldsymbol{u}}, \tilde{\boldsymbol{v}}) \in \mathcal{Z} \mid \tilde{s} = 1\right] = 1 \\ \mathbb{P}\left[\tilde{s} = 1\right] = 1 \end{array} \right\},$$

where $\mathcal{Z} = \{(\tilde{\boldsymbol{u}}, \tilde{\boldsymbol{v}}) \mid \boldsymbol{u} \in \mathcal{U}, \boldsymbol{v} \geq \boldsymbol{\phi}(\boldsymbol{u})\}$. That is to say, a generalized-moment ambiguity set can be mapped into an event-wise ambiguity set with only one scenario, *i.e.*, $S = 1$.

Wasserstein Ambiguity Set

We consider a data-driven setting as in Mohajerin Esfahani and Kuhn (2018) on the design of a Wasserstein ambiguity set centered around the empirical distribution $\hat{\mathbb{P}} = \frac{1}{S}\sum_{s\in[S]}\delta_{\hat{\boldsymbol{u}}_s}$. Given a tractable distance metric $\rho : \mathbb{R}^{I_u} \times \mathbb{R}^{I_u} \mapsto [0, +\infty)$, the Wasserstein metric (a.k.a Kantorovich-Rubinstein metric) between any two distributions $\mathbb{P}$ and $\hat{\mathbb{P}}$ is defined via an optimization problem:

$$d_W(\mathbb{P}, \hat{\mathbb{P}}) \triangleq \inf_{\mathbb{Q}\in\mathcal{Q}(\mathbb{P},\hat{\mathbb{P}})} \mathbb{E}_{\mathbb{Q}}\left[\rho(\tilde{\boldsymbol{u}}, \tilde{\boldsymbol{u}}^\dagger)\right], \tag{5}$$

where $\tilde{\boldsymbol{u}} \sim \mathbb{P}$, $\tilde{\boldsymbol{u}}^\dagger \sim \hat{\mathbb{P}}$, and $\mathcal{Q}(\mathbb{P}, \hat{\mathbb{P}})$ is the set of all joint probability distrbutions on $\mathbb{R}^{I_u} \times \mathbb{R}^{I_u}$ with marginals $\mathbb{P}$ and $\hat{\mathbb{P}}$. The Wasserstein ambiguity set is then defined by

$$\mathcal{G}_W(\theta) = \left\{ \mathbb{P} \in \mathcal{P}_0(\mathcal{U}) \;\middle|\; \begin{array}{l} \tilde{\boldsymbol{u}} \sim \mathbb{P} \\ d_W(\mathbb{P}, \hat{\mathbb{P}}) \leq \theta \end{array} \right\}, \tag{6}$$

which is a ball of radius $\theta \geq 0$ around $\hat{\mathbb{P}}$. Interestingly, we can provide a new lifted representation of the Wasserstein ambiguity set in the format of an event-wise ambiguity set.

$$\mathcal{F}_W(\theta) = \left\{ \mathbb{P} \in \mathcal{P}_0\left(\mathbb{R}^{I_u+1} \times [S]\right) \;\middle|\; \begin{array}{ll} ((\tilde{\boldsymbol{u}}, \tilde{v}), \tilde{s}) \sim \mathbb{P} & \\ \mathbb{E}_{\mathbb{P}}\left[\tilde{v} \mid \tilde{s} \in [S]\right] \leq \theta & \\ \mathbb{P}\left[(\tilde{\boldsymbol{u}}, \tilde{v}) \in \mathcal{Z}_s \mid \tilde{s} = s\right] = 1 & \forall s \in [S] \\ \mathbb{P}\left[\tilde{s} = s\right] = \frac{1}{S} & \forall s \in [S] \end{array} \right\}, \tag{7}$$

where the primary random variable $\tilde{\boldsymbol{u}}$ and the auxiliary random variable $\tilde{v}$ jointly reside in lifted support sets $\mathcal{Z}_s = \{(\boldsymbol{u}, v) \mid \boldsymbol{u} \in \mathcal{U}, v \geq \rho(\boldsymbol{u}, \hat{\boldsymbol{u}}_s)\}, s \in [S]$ for different scenarios.

Theorem 2. *The Wasserstein ambiguity set $\mathcal{G}_W(\theta)$ is equivalent to the marginal distribution of $\tilde{\boldsymbol{u}}$ under $\mathbb{P}$, for all $\mathbb{P} \in \mathcal{F}_W(\theta)$. That is, for all $\theta \geq 0$, $\mathcal{G}_W(\theta) = \Pi_{\tilde{\boldsymbol{u}}}\mathcal{F}_W(\theta)$.*

Such a lifted representation can be extended to type-p Wasserstein metric for any $p \in [1, \infty]$ and the result about the type-∞ originates from Bertsimas et al. (2018); see Appendix C.

Mixture-Distribution Ambiguity Set

We can use the event-wise ambiguity set to specify a mixture distribution as proposed in Hanasusanto et al. (2015), which is useful, for example, in modeling ambiguous multi-modal distributions.

$$\mathcal{F} = \left\{ \mathbb{P} \in \mathcal{P}_0\left(\mathbb{R}^{I_u+I_v} \times [S]\right) \;\middle|\; \begin{array}{ll} ((\tilde{\boldsymbol{u}}, \tilde{\boldsymbol{v}}), \tilde{s}) \sim \mathbb{P} & \\ \mathbb{E}_{\mathbb{P}}\left[\tilde{\boldsymbol{u}} \mid \tilde{s} = s\right] \in \mathcal{Q}_s & \forall s \in [S] \\ \mathbb{E}_{\mathbb{P}}\left[\tilde{\boldsymbol{v}} \mid \tilde{s} = s\right] \leq \boldsymbol{\sigma}_s & \forall s \in [S] \\ \mathbb{P}\left[(\tilde{\boldsymbol{u}}, \tilde{\boldsymbol{v}}) \in \mathcal{Z}_s \mid \tilde{s} = s\right] = 1 & \forall s \in [S] \\ \mathbb{P}\left[\tilde{s} = s\right] = p_s & \forall s \in [S] \end{array} \right\}, \tag{8}$$

where for each $s \in [S]$, $\mathcal{Z}_s = \{(\tilde{\boldsymbol{u}}, \tilde{\boldsymbol{v}}) \mid \boldsymbol{u} \in \mathcal{U}_s, \boldsymbol{v} \geq \boldsymbol{\phi}(\boldsymbol{u})\}$. Note that any distribution $\mathbb{P} \in \mathcal{F}$ can be written as $\mathbb{P} = \sum_{s\in[S]} p_s \mathbb{P}_s$, where each mixture component $\mathbb{P}_s$ is an ambiguous distribution with support $\mathcal{U}_s$ and moments $\mathbb{E}_{\mathbb{P}_s}[\tilde{\boldsymbol{u}}] \in \mathcal{Q}_s$ and $\mathbb{E}_{\mathbb{P}_s}[\boldsymbol{\phi}(\tilde{\boldsymbol{u}})] \leq \boldsymbol{\sigma}_s$. Hanasusanto et al. (2015) have used the mixture-distribution ambiguity set to model the uncertain demand in the textile apparel industry, which is known for the multimodality and ambiguity.

K-means Ambiguity Set

We can incorporate clustering techniques in machine learning to construct event-wise ambiguity sets directly from data. Given N historical observations $\hat{\boldsymbol{u}}_1, \ldots, \hat{\boldsymbol{u}}_N$ for the primary random variable $\tilde{\boldsymbol{u}}$, we can partition the support set $\mathcal{U}$ into S clusters $\mathcal{U}_s, s \in [S]$ using the K-means clustering (MacQueen et al. 1967, Dubes and Jain 1988), which gives centroids $\hat{\boldsymbol{\mu}}_s, s \in [S]$ of these clusters (see an illustration in Figure 3). Associated with each cluster, we can determine its support set by

$$\mathcal{U}_s = \{\boldsymbol{u} \in \mathcal{U} \mid 2\boldsymbol{u}^\top(\hat{\boldsymbol{\mu}}_r - \hat{\boldsymbol{\mu}}_s) \leq \hat{\boldsymbol{\mu}}_r^\top \hat{\boldsymbol{\mu}}_r - \hat{\boldsymbol{\mu}}_s^\top \hat{\boldsymbol{\mu}}_s \ \forall r \in [S]\},$$

where the hyperplane $2\boldsymbol{u}^\top(\hat{\boldsymbol{\mu}}_r - \hat{\boldsymbol{\mu}}_s) = \hat{\boldsymbol{\mu}}_r^\top \hat{\boldsymbol{\mu}}_r - \hat{\boldsymbol{\mu}}_s^\top \hat{\boldsymbol{\mu}}_s$ corresponds to the perpendicular bisector of $\hat{\boldsymbol{u}}_s$ and $\hat{\boldsymbol{u}}_r$. Let $\mathbb{I}$ denote the indicator function. The weight of a cluster is

$$\hat{p}_s = \frac{1}{N} \sum_{n\in[N]} \mathbb{I}[\hat{\boldsymbol{u}}_n \in \mathcal{U}_s]$$

and the parameters associated with the convex generalized moments are

$$\hat{\boldsymbol{\sigma}}_s = \frac{1}{\hat{p}_s N} \sum_{n\in[N]} \mathbb{I}[\hat{\boldsymbol{u}}_n \in \mathcal{U}_s] \boldsymbol{\phi}(\hat{\boldsymbol{u}}_n).$$

The corresponding K-means ambiguity set is a special mixture-distribution ambiguity set in the form (8) with cluster-wise estimates $\mathcal{Q}_s = \{\hat{\boldsymbol{\mu}}_s\}$, $p_s = \hat{p}_s$, $\boldsymbol{\sigma}_s = \hat{\boldsymbol{\sigma}}_s$, $s \in [S]$. To account for uncertainty in these estimates, we can further specify uncertainty sets for them.

We refer interested readers to a recent work by Perakis et al. (2018) of using data to construct the K-means ambiguity set to address a joint pricing and production problem. In a simpler example presented in Appendix F, we study a three-period portfolio management problem to show how a two-layer K-means ambiguity set can be constructed directly from historical returns. We show that complex RSO models can be effectively formulated with the help of RSOME. A wealth of portfolio literature based on the stochastic programming approach can also address the non-anticipativity of recourse decisions by using cluster heuristics that group samples into different bundles (see, *e.g.*, Bogentoft et al. 2001, Hibiki 2006 and reference therein). We acknowledge that further empirical studies should be done to evaluate and compare their performance in practice, which would benefit from the algebraic modeling package, RSOME in implementing different RSO models.

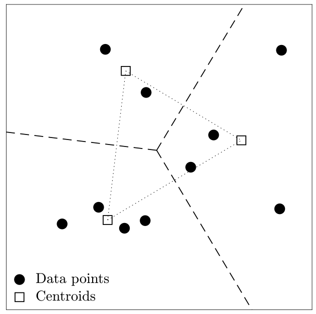

Figure 3 K-means clustering. The distance from a point to its corresponding centroid is not larger than the distance from it to any other centroid. The boundary of a cluster is determined by the boundary of the support set (solid lines) and the perpendicular bisectors of its centroid and centroids of other clusters (dash lines).

Side-Information Ambiguity Set

Recently, Hao et al. (2019) give a novel interpretation of the event-wise ambiguity set to incorporate side information. Suppose that the uncertainty of the primary random variable $\tilde{\boldsymbol{u}}$ is strongly associated with some uncertain side information (*i.e.*, covariates), represented by an auxiliary random variable $\tilde{\boldsymbol{v}}$, and collectively, one accesses to the data $(\hat{\boldsymbol{u}}_1, \hat{\boldsymbol{v}}_1), \cdots, (\hat{\boldsymbol{u}}_N, \hat{\boldsymbol{v}}_N)$. Such side information can be captured in the event-wise ambiguity set by grouping the primary random variable $\tilde{\boldsymbol{u}}$ based on the side information $\tilde{\boldsymbol{v}}$. In particular, a broad range of machine learning techniques, including classification and regression trees, can separate $\hat{\boldsymbol{v}}_i, i \in [N]$ into S scenarios. Correspondingly, one could partition $\hat{\boldsymbol{u}}_i, i \in [N]$ into S scenarios and specify the statistical information of each scenario (see the example in Figure 4). In the original attempt of incorporating side information, Hao et al. (2019) use the weather information as side information to characterize the uncertain demands of taxis and apply multivariate regression tree to obtain the scenarios. By doing so, the authors achieve significant improvement in taxis allocation under demand uncertainty.

5. Modeling Examples

The RSO framework is expansive and encompasses scenario-tree based stochastic linear optimization and distributionally robust optimization models. Although it is based on expectations of bi-affine functions, it can also provide a tight characterization of the worst-case expectations of some classes of quadratic functions known in the literature, including the seminal works of Ben-Tal and Nemirovski (1998) and Tütüncü and Koenig (2004), and extend them to include discrete scenarios (Appendix B). We next provide several examples in our framework, including optimizing over the Hurwicz criterion and models which have both discrete and continuous recourse decisions

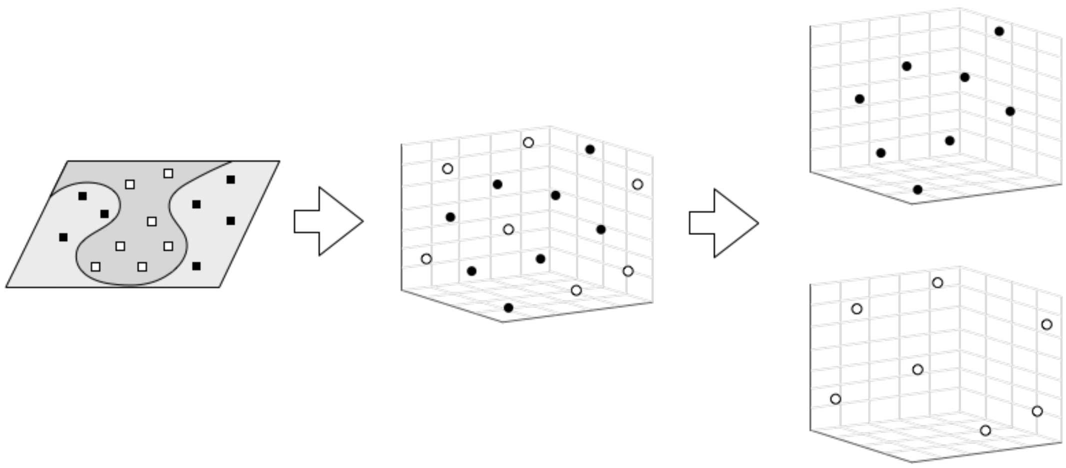

Figure 4 Realizations $\hat{\boldsymbol{v}}_n \in \mathbb{R}^2, n \in [N]$ of the auxiliary random variable are separated into 2 groups (black and white). Correspondingly, realizations $\hat{\boldsymbol{u}}_n \in \mathbb{R}^3, n \in [N]$ of the primary random variable are grouped into 2 scenarios (black and white). The empirical information of these two groups can be different.

and where the uncertainty is characterized using the Wasserstein ambiguity set as well as the K-means ambiguity set (Appendix F)—both are directly constructed from data. We show that an algebraic modeling toolbox such as RSOME could greatly facilitate the implementation without worrying about the tedious reformulation.

Hurwicz Criterion

Hurwicz (1951) is arguably first to propose a decision criterion that articulates the tradeoff between pessimistic and optimistic objectives, which under distributional ambiguity can be formulated as

$$(1-\varphi)\sup_{\mathbb{P}\in\mathcal{F}} \mathbb{E}_{\mathbb{P}}\left[f(\boldsymbol{w},\tilde{\boldsymbol{u}},\tilde{s})\right] + \varphi \inf_{\mathbb{P}\in\mathcal{F}} \mathbb{E}_{\mathbb{P}}\left[f(\boldsymbol{w},\tilde{\boldsymbol{u}},\tilde{s})\right],$$

where the cost function $f(\boldsymbol{w},\boldsymbol{u},s)$ depends on the here-and-now decision $\boldsymbol{w}$, and it is typically convex in $\boldsymbol{u}$ for given $\boldsymbol{w}\in\mathcal{X}$ and scenario $s\in[S]$. Here $\varphi\in[0,1]$ is the level of optimism, with $\varphi=0$ ($\varphi=1$) being the most pessimistic (optimistic) perception of the objective value. In order to obtain a computationally tractable model, we often consider the most pessimistic objective (*i.e.*, $\varphi=0$) because the best-case expectation for the most optimistic objective (*i.e.*, $\varphi=1$) is typically non-convex in its decision $\boldsymbol{w}$. Quite notably, there is a class of ambiguity sets for which the best-case expectation would also be tractable.

Proposition 1. *Consider an event-wise ambiguity set $\mathcal{F}$ in (3) such that for any $\mathbb{P}\in\mathcal{F}$, it satisfies $\mathbb{E}_{\mathbb{P}}[\tilde{\boldsymbol{u}} \mid \tilde{s}=s]=\boldsymbol{\mu}_s$ and $\mathbb{P}[\tilde{s}=s]=p_s$ with known $\boldsymbol{\mu}_s$ and p_s for all $s\in[S]$. Then for any function $g(\boldsymbol{u},s):\mathbb{R}^{I_u}\times[S]\mapsto\mathbb{R}$ that is convex in $\boldsymbol{u}$ for a given $s\in[S]$, we have*

$$\inf_{\mathbb{P}\in\mathcal{F}} \mathbb{E}_{\mathbb{P}}\left[g(\tilde{\boldsymbol{u}},\tilde{s})\right] = \sum_{s\in[S]} p_s g(\boldsymbol{\mu}_s, s).$$

Proof. The proposition follows immediately from Jensen's inequality. □

The mixture-distribution ambiguity set with singleton sets $\mathcal{Q}_s, s\in[S]$ and the K-means ambiguity set fit in this class, for which we can optimize over the Hurwicz criterion

$$\min_{\boldsymbol{w}\in\mathcal{X}}\left\{(1-\varphi)\sup_{\mathbb{P}\in\mathcal{F}}\mathbb{E}_{\mathbb{P}}\left[f(\boldsymbol{w},\tilde{\boldsymbol{u}},\tilde{s})\right]+\varphi\sum_{s\in[S]}p_s f(\boldsymbol{w},\boldsymbol{\mu}_s,s)\right\}$$

by formulating via the RSO framework. Note that it would also be possible to account for scenarios with uncertain probabilities vector $\boldsymbol{p}=(p_s)_{s\in[S]}\subseteq\mathcal{P}$, as long as the uncertainty set $\mathcal{P}$ is a polytope with modest number of extreme points, $\boldsymbol{p}^e, e\in[E]$. In such circumstances, we would solve

$$\min_{e\in[E]}\min_{\boldsymbol{w}\in\mathcal{X}}\left\{(1-\varphi)\sup_{\mathbb{P}\in\mathcal{F}}\mathbb{E}_{\mathbb{P}}\left[f(\boldsymbol{w},\tilde{\boldsymbol{u}},\tilde{s})\right]+\varphi\sum_{s\in[S]}p^e_s f(\boldsymbol{w},\boldsymbol{\mu}_s,s)\right\}.$$

Expectation of Convex and Piecewise Affine Functions

Expectation of convex and piecewise affine functions are commonly encountered in modeling risk aversion based on the utility (Gilboa and Schmeidler 1989) or risk measure including the shortfall risk measure (Föllmer and Schied 2002) and the optimized certainty equivalent (Ben-Tal and Teboulle 2007). We show that by simply introducing a recourse decision $y(\cdot,\cdot)$, we can achieve an equivalent formulation under the RSO framework.

Theorem 3. *The worst-case expectation*

$$\sup_{\mathbb{P}\in\mathcal{F}}\mathbb{E}_{\mathbb{P}}\left[\max_{\ell\in[L]}\left\{\boldsymbol{r}^\top(\tilde{s})\boldsymbol{G}_\ell(\tilde{s})\tilde{\boldsymbol{z}}+h_\ell(\tilde{s})\right\}\right] \tag{9}$$

for a finite index set $[L]$, is equivalent to the following problem

$$\begin{aligned}\min\ &\sup_{\mathbb{P}\in\mathcal{F}}\mathbb{E}_{\mathbb{P}}\left[y(\tilde{s},\tilde{\boldsymbol{z}})\right]\\ \text{s.t.}\ & y(s,\boldsymbol{z})\geq\boldsymbol{r}^\top(s)\boldsymbol{G}_\ell(s)\boldsymbol{z}+h_\ell(s)\quad\forall\boldsymbol{z}\in\mathcal{Z}_s,\ s\in[S],\ \ell\in[L]\\ & y\in\bar{\mathcal{A}}(\bar{\mathcal{C}},[I_z]),\end{aligned} \tag{10}$$

where the collection $\bar{\mathcal{C}}\triangleq\{\{s\}\mid s\in[S]\}$ consists of singleton MECE events.

Expected Utility with Mean-Covariance Ambiguity Sets

Consider the mean-covariance ambiguity set that commonly appears in portfolio management

$$\mathcal{G}(\boldsymbol{\mu},\boldsymbol{\Sigma})=\left\{\mathbb{P}\in\mathcal{P}_0\left(\mathbb{R}^{I_u}\right)\ \middle|\ \begin{array}{l}\tilde{\boldsymbol{u}}\sim\mathbb{P}\\ \mathbb{E}_{\mathbb{P}}[\tilde{\boldsymbol{u}}]=\boldsymbol{\mu}\\ \mathbb{E}_{\mathbb{P}}[(\tilde{\boldsymbol{u}}-\boldsymbol{\mu})(\tilde{\boldsymbol{u}}-\boldsymbol{\mu})^\top]=\boldsymbol{\Sigma}\end{array}\right\}.$$

Here the I_u-dimensional random variable $\tilde{\boldsymbol{u}}$ can be a scalar or a vector and refers to the random return(s) of the risky asset(s). For any utility function $U:\mathbb{R}\mapsto\mathbb{R}$, Popescu (2007) has shown that the robust expected utility of the random weighted sum $\boldsymbol{w}^\top\tilde{\boldsymbol{u}}$ satisfies:

$$\inf_{\mathbb{P}\in\mathcal{G}(\boldsymbol{\mu},\boldsymbol{\Sigma})}\mathbb{E}_{\mathbb{P}}[U(\boldsymbol{w}^\top\tilde{\boldsymbol{u}})]=\inf_{\mathbb{P}\in\mathcal{G}(\boldsymbol{w}^\top\boldsymbol{\mu},\boldsymbol{w}^\top\boldsymbol{\Sigma}\boldsymbol{w})}\mathbb{E}_{\mathbb{P}}[U(\tilde{u})]. \tag{11}$$

This property enables Natarajan et al. (2010) to obtain an attractive computationally tractable second-order cone reformulation when U is concave piecewise affine, which is surprising as a direct duality approach would result in a positive semidefinite program that is much harder to solve. We next recover this result using the RSO framework to obtain the second-order cone reformulation.

Theorem 4. *Given a concave piecewise affine utility function* $U(u)=\min_{\ell\in[L]}\{g_\ell u+h_\ell\}$*, the robust expected utility* $\inf_{\mathbb{P}\in\mathcal{G}(\boldsymbol{\mu},\boldsymbol{\Sigma})}\mathbb{E}_{\mathbb{P}}[U(\boldsymbol{w}^\top\tilde{\boldsymbol{u}})]$ *is equivalent to*

$$\begin{aligned}
\max\ & \inf_{\mathbb{P}\in\mathcal{F}}\mathbb{E}_{\mathbb{P}}[y(\tilde{u},\tilde{v})]\\
\text{s.t.}\ & y(u,v)\le g_\ell(ru+\boldsymbol{w}^\top\boldsymbol{\mu})+h_\ell\quad \forall(u,v)\in\mathcal{Z},\ \ell\in[L]\\
& r\ge\sqrt{\boldsymbol{w}^\top\boldsymbol{\Sigma}\boldsymbol{w}}\\
& r\in\mathbb{R},\ y\in\bar{\mathcal{A}}(\{1\},\{1,2\}),
\end{aligned}$$

where the ambiguity set

$$\mathcal{F}=\left\{\mathbb{P}\in\mathcal{P}_0\left(\mathbb{R}^2\right)\ \middle|\ \begin{array}{l}(\tilde{u},\tilde{v})\sim\mathbb{P}\\ \mathbb{E}_{\mathbb{P}}[\tilde{u}]=0,\ \mathbb{E}_{\mathbb{P}}[\tilde{v}]\le 1\\ \mathbb{P}[(\tilde{u},\tilde{v})\in\mathcal{Z}]=1\end{array}\right\}$$

has only one scenario (i.e., $S=1$*) and takes a support set* $\mathcal{Z}=\{(u,v)\in\mathbb{R}^2\mid v\ge u^2\}$*.*

Expectation of Saddle Functions

The RSO model is primary based on a linear optimization framework, where the objective function and soft constraints are bilinear with respect to the underlying decision variable $\boldsymbol{r}(s)$ and the random variable $\tilde{\boldsymbol{z}}$. With auxiliary decisions and auxiliary random variables, we can also consider saddle functions that are convex with respect to the decision variables and concave with respect to the random variables (see, Ben-Tal et al. 2015). Observe that unlike earlier robust and distributionally robust optimization models, the random variable mappings in (1) include affine relations involving the auxiliary random variable, $\tilde{\boldsymbol{v}}$, which is embedded in $\tilde{\boldsymbol{z}}$. This generality allows us to extend the objective function and soft constraints to saddle functions under the RSO framework.

We consider a saddle function $f(\boldsymbol{r}(s),\boldsymbol{u},s)$ such that for a given scenario s, it is jointly convex with respect to the decision $\boldsymbol{r}(s)\in\mathcal{X}_s$ for any fixed $\boldsymbol{u}\in\mathcal{U}_s$ and jointly concave with respect to $\boldsymbol{u}\in\mathcal{U}_s$ for any fixed $\boldsymbol{r}(s)\in\mathcal{X}_s$ as follows:

$$f(\boldsymbol{r}(s),\boldsymbol{u},s)\triangleq\boldsymbol{\xi}^\top(\boldsymbol{r}(s),s)\boldsymbol{\zeta}(\boldsymbol{u},s)=\sum_{\ell\in[I_v]}\xi_\ell(\boldsymbol{r}(s),s)\zeta_\ell(\boldsymbol{u},s). \tag{12}$$

Here for a given scenario s and the corresponding partition of indices $[I_v] = \mathcal{L}_{s1} \cup \mathcal{L}_{s2} \cup \mathcal{L}_{s3} \cup \mathcal{L}_{s4}$:

- $\xi_\ell(\boldsymbol{r}(s), s)$ is nonnegative and convex in $\boldsymbol{r}(s)$ and $\zeta_\ell(\boldsymbol{u}, s)$ is nonnegative and concave in $\boldsymbol{u}$, for $\ell \in \mathcal{L}_{s1}$;
- $\xi_\ell(\boldsymbol{r}(s), s)$ is nonnegative and affine in $\boldsymbol{r}(s)$ and $\zeta_\ell(\boldsymbol{u}, s)$ is concave in $\boldsymbol{u}$, for $\ell \in \mathcal{L}_{s2}$;
- $\xi_\ell(\boldsymbol{r}(s), s)$ is convex in $\boldsymbol{r}(s)$ and $\zeta_\ell(\boldsymbol{u}, s)$ is nonnegative and affine in $\boldsymbol{u}$, for $\ell \in \mathcal{L}_{s3}$;
- $\xi_\ell(\boldsymbol{r}(s), s)$ is affine in $\boldsymbol{r}(s)$ and $\zeta_\ell(\boldsymbol{u}, s)$ is affine in $\boldsymbol{u}$, for $\ell \in \mathcal{L}_{s4}$.

Theorem 5. *The worst-case expectation* $\sup_{\mathbb{P} \in \mathcal{G}} \mathbb{E}_{\mathbb{P}}[f(\boldsymbol{r}(\tilde{s}), \tilde{\boldsymbol{u}}, \tilde{s})]$ *for the ambiguity set*

$$\mathcal{G} = \left\{ \mathbb{P} \in \mathcal{P}_0\left(\mathbb{R}^{I_u} \times [S]\right) \;\middle|\; \begin{array}{ll} (\tilde{\boldsymbol{u}}, \tilde{s}) \sim \mathbb{P} & \\ \mathbb{E}_{\mathbb{P}}[\tilde{\boldsymbol{u}} \mid \tilde{s} \in \mathcal{E}_k] \in \mathcal{Q}_k & \forall k \in [K] \\ \mathbb{P}[\tilde{\boldsymbol{u}} \in \mathcal{U}_s \mid \tilde{s} = s] = 1 & \forall s \in [S] \\ \mathbb{P}[\tilde{s} = s] = p_s & \forall s \in [S] \\ \textit{for some } \boldsymbol{p} \in \mathcal{P} & \end{array} \right\},$$

is the same as

$$\begin{array}{lll} \min & \sup_{\mathbb{P} \in \mathcal{F}} \mathbb{E}_{\mathbb{P}}[\bar{\boldsymbol{r}}^\top(\tilde{s})\tilde{\boldsymbol{v}}] & \\ \text{s.t.} & \bar{r}_\ell(s) \geq \xi_\ell(\boldsymbol{r}(s), s) & \forall \ell \in \mathcal{L}_{s1} \cup \mathcal{L}_{s3},\ s \in [S] \\ & \bar{r}_\ell(s) = \xi_\ell(\boldsymbol{r}(s), s) & \forall \ell \in \mathcal{L}_{s2} \cup \mathcal{L}_{s4},\ s \in [S] \\ & \bar{\boldsymbol{r}}(s) \in \mathbb{R}^{I_v} & \forall s \in [S], \end{array} \tag{13}$$

for the lifted event-wise ambiguity set

$$\mathcal{F} = \left\{ \mathbb{P} \in \mathcal{P}_0\left(\mathbb{R}^{I_u + I_v} \times [S]\right) \;\middle|\; \begin{array}{ll} ((\tilde{\boldsymbol{u}}, \tilde{\boldsymbol{v}}), \tilde{s}) \sim \mathbb{P} & \\ \mathbb{E}_{\mathbb{P}}[\tilde{\boldsymbol{u}} \mid \tilde{s} \in \mathcal{E}_k] \in \mathcal{Q}_k & \forall k \in [K] \\ \mathbb{P}[(\tilde{\boldsymbol{u}}, \tilde{\boldsymbol{v}}) \in \mathcal{Z}_s \mid \tilde{s} = s] = 1 & \forall s \in [S] \\ \mathbb{P}[\tilde{s} = s] = p_s & \forall s \in [S] \\ \textit{for some } \boldsymbol{p} \in \mathcal{P} & \end{array} \right\} \tag{14}$$

with lifted support sets

$$\mathcal{Z}_s = \left\{ (\boldsymbol{u}, \boldsymbol{v}) \in \mathbb{R}^{I_u + I_v} \;\middle|\; \begin{array}{ll} \boldsymbol{u} \in \mathcal{U}_s & \\ v_\ell \leq \zeta_\ell(\boldsymbol{u}, s) & \forall \ell \in \mathcal{L}_{s1} \cup \mathcal{L}_{s2} \\ v_\ell = \zeta_\ell(\boldsymbol{u}, s) & \forall \ell \in \mathcal{L}_{s3} \cup \mathcal{L}_{s4} \end{array} \right\} \quad \forall s \in [S].$$

Two-Stage Problem with Wasserstein Ambiguity Sets

Optimization models based on the Wasserstein ambiguity set have recently attracted considerable interests from both stochastic programming and robust optimization communities. While most of the existing models are static, dynamic models with the Wasserstein ambiguity set are scarce due to limited solution approaches. We next demonstrate that the RSO framework provides a tractable approximation for two-stage linear optimization problems with the Wasserstein ambiguity set (6), which has the potential to serve the modeling of multi-stage dynamic problems.

In particular, we consider the following second-stage problem given the here-and-now decision $\boldsymbol{w}$ and the realization $\boldsymbol{u}$ of the underlying primary random variable $\tilde{\boldsymbol{u}}$.

$$\begin{aligned} f(\boldsymbol{w},\boldsymbol{u}) = \min\ & \boldsymbol{c}_0^\top \boldsymbol{y} \\ \text{s.t.}\ & \boldsymbol{a}_\ell^\top(\boldsymbol{u})\boldsymbol{w} + \boldsymbol{c}_\ell^\top \boldsymbol{y} \geq d_\ell(\boldsymbol{u}) \quad \forall \ell \in [L] \\ & \boldsymbol{y} \in \mathbb{R}^{J_y}, \end{aligned} \tag{15}$$

where similar to the random variable mappings in (1), for each ℓ^{th} constraint, $\boldsymbol{a}_\ell$ and d_ℓ are affine mappings of the realization of $\tilde{\boldsymbol{u}}$. For any here-and-now decision $\boldsymbol{w}$, we approximate its worst-case expected second-stage cost $\sup_{\mathbb{P}\in\mathcal{F}_W(\theta)} \mathbb{E}_\mathbb{P}[f(\boldsymbol{w},\tilde{\boldsymbol{u}})]$ under the Wasserstein ambiguity set through

$$\begin{aligned} \min\ & \sup_{\mathbb{P}\in\mathcal{F}_W(\theta)} \mathbb{E}_\mathbb{P}\left[\boldsymbol{c}_0^\top \boldsymbol{y}(\tilde{s},\tilde{\boldsymbol{z}})\right] \\ \text{s.t.}\ & \boldsymbol{a}_\ell^\top(\boldsymbol{u})\boldsymbol{w} + \boldsymbol{c}_\ell^\top \boldsymbol{y}(s,\boldsymbol{z}) \geq d_\ell(\boldsymbol{u}) \quad \forall \boldsymbol{z}\in\mathcal{Z}_s,\ s\in[S],\ \ell\in[L] \\ & y_j \in \bar{\mathcal{A}}(\mathcal{C}_j,\mathcal{I}_j) \qquad \forall j\in[J_y], \end{aligned} \tag{16}$$

where $\tilde{\boldsymbol{z}} = (\tilde{\boldsymbol{u}},\tilde{v})$ and where the collections $\mathcal{C}_j, j\in[J_y]$ of MECE events and information index sets $\mathcal{I}_j \subseteq [I_u+1], j\in[J_y]$ jointly control how the recourse decision $\boldsymbol{y}(\cdot,\cdot)$ adapts to $(\tilde{\boldsymbol{u}},\tilde{v})$ and $\tilde{s}$. The optimal $\boldsymbol{w}$ can then be selected by minimizing the sum of the deterministic first-stage cost $\boldsymbol{c}_0^\top\boldsymbol{w}$ and the worst-case expected second-stage cost (16). In Appendix D, we report the performance of this conservative approximation in comparison with *(i)* the computationally expensive exact approach and *(ii)* a state-of-the-art approximation scheme by Hanasusanto and Kuhn (2018).

The event-wise adaptation can be extended to address two-stage problems with the type-∞ Wasserstein metric, and interestingly, it would coincide with the multi-policy approximation (MPA) proposed by Bertsimas et al. (2019a). We refer to Bertsimas et al. (2018) for using generalization of MPA to address multi-stage problems with the type-∞ Wasserstein ambiguity set.

6. Conclusion

The RSO model unifies an important class of scenario-tree based stochastic linear optimization problems and a number of distributionally robust optimization models (based on convex generalized moments, mixture distribution, ϕ-divergence, Wasserstein metric, etc.) that have been considered in isolation to date. As we have demonstrated, the RSO model also opens up to new approaches including those inspired by machine learning techniques. Based on such a unified framework for generic modeling and optimization under uncertainty, we firmly believe that algebraic modeling package such as RSOME would help us in navigating and evaluating the plethora of approaches to address a wide variety of uncertainty-affected optimization problems in practice. We refer readers to the perpetually updated manual of RSOME for more examples that can be modeled in RSO.

Acknowledgments

The authors gratefully acknowledge the Editor-in-Chief David Simchi-Levi, the anonymous associate editor, and two reviewers, whose constructive and valuable comments led to substantial improvements of the paper. The authors also thank Zhaowei Hao, Long He, Zhenyu Hu, Jun Jiang, Brad Sturt and Qinshen Tang for the helpful discussions and feedback. The research of Zhi Chen is funded by the start-up grant for new faculty (project number: 7200646) from the City University of Hong Kong. The research of Melvyn Sim is supported by Singapore Ministry of Education Social Science Research Thematic Grant under MOE2016-SSRTG-059. Any opinions, findings, and conclusions or recommendations expressed in this material are those of the author(s) and do not necessarily reflect the views of the Singapore Ministry of Education.

References

Ben-Tal, Aharon, Dick Den Hertog, Anja De Waegenaere, Bertrand Melenberg, Gijs Rennen. 2013. Robust solutions of optimization problems affected by uncertain probabilities. *Management Science* **59**(2) 341–357.

Ben-Tal, Aharon, Dick Den Hertog, Jean-Philippe Vial. 2015. Deriving robust counterparts of nonlinear uncertain inequalities. *Mathematical Programming* **149**(1-2) 265–299.

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

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

Ben-Tal, Aharon, Arkadi Nemirovski. 1998. Robust convex optimization. *Mathematics of Operations Research* **23**(4) 769–805.

Ben-Tal, Aharon, Arkadi Nemirovski. 2001. *Lectures on modern convex optimization: analysis, algorithms, and engineering applications*. SIAM.

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

Bertsimas, Dimitris, Shtern Shimrit, Sturt Bradley. 2018. A data-driven approach for multi-stage linear optimization. *Available at Optimization Online.*

Bertsimas, Dimitris, Shtern Shimrit, Sturt Bradley. 2019a. Two-stage sample robust optimization. *Available at* `https://arxiv.org/abs/1907.07142`.

Bertsimas, Dimitris, Melvyn Sim. 2004. The price of robustness. *Operations Research* **52**(1) 35–53.

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

Bertsimas, Dimitris, Aurélie Thiele. 2006. A robust optimization approach to inventory theory. *Operations Research* **54**(1) 150–168.

Birge, John R, Francois Louveaux. 2011. *Introduction to stochastic programming*. Springer.

Bogentoft, Erik, H Edwin Romeijn, Stanislav Uryasev. 2001. Asset/liability management for pension funds using CVaR constraints. *The Journal of Risk Finance* **3**(1) 57–71.

Danzig, George B. 1955. Linear programming under uncertainty. *Management Science* **1**(3-4) 197–206.

Delage, Erick, Dan A Iancu. 2015. Robust multistage decision making. *The Operations Research Revolution*. INFORMS, 20–46.

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

Dubes, Richard C, Anil K Jain. 1988. *Algorithms for clustering data*. Prentice Hall.

Dupačová, Jitka. 1976. Minimax stochastic programs with nonconvex nonseparable penalty functions. *Progress in Operations Research* **1** 303–316.

El Ghaoui, Laurent, Maksim Oks, Francois Oustry. 2003. Worst-case value-at-risk and robust portfolio optimization: a conic programming approach. *Operations Research* **51**(4) 543–556.

El Ghaoui, Laurent, Francois Oustry, Hervé Lebret. 1998. Robust solutions to uncertain semidefinite programs. *SIAM Journal on Optimization* **9**(1) 33–52.

Föllmer, Hans, Alexander Schied. 2002. Convex measures of risk and trading constraints. *Finance and Stochastics* **6**(4) 429–447.

Gilboa, Itzhak, David Schmeidler. 1989. Maxmin expected utility with non-unique prior. *Journal of Mathematical Economics* **18**(2) 141–153.

Grant, Michael C, Stephen P Boyd. 2008. Graph implementations for nonsmooth convex programs. *Recent Advances in Learning and Control*. Springer, 95–110.

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

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

Hao, Zhaowei, Long He, Zhenyu Hu, Jun Jiang. 2019. Robust vehicle pre-allocation with uncertain covariates. *Forthcoming in Production and Operation Management.*

Heitsch, Holger, Werner Römisch. 2009. Scenario tree modeling for multistage stochastic programs. *Mathematical Programming* **118**(2) 371–406.

Hibiki, Norio. 2006. Multi-period stochastic optimization models for dynamic asset allocation. *Journal of Banking & Finance* **30**(2) 365–390.

Høyland, Kjetil, Stein W Wallace. 2001. Generating scenario trees for multistage decision problems. *Management Science* **47**(2) 295–307.

Hurwicz, Leonid. 1951. The generalized Bayes minimax principle: a criterion for decision making uncer ucertainty. *Cowles Comm. Discuss. Paper Stat* (355).

Kall, Peter, Stein W Wallace. 1994. *Stochastic programming*. Springer.

Kleywegt, Anton J, Alexander Shapiro, Tito Homem-de Mello. 2002. The sample average approximation method for stochastic discrete optimization. *SIAM Journal on Optimization* **12**(2) 479–502.

Kuhn, Daniel, Peyman Mohajerin Esfahani, Viet Anh Nguyen, Soroosh Shafieezadeh-Abadeh. 2019. Wasserstein distributionally robust optimization: Theory and applications in machine learning. *Forthcoming in INFORMS TutORials in Operations Research.*

MacQueen, James, et al. 1967. Some methods for classification and analysis of multivariate observations. *Proceedings of the fifth Berkeley Symposium on Mathematical Statistics and Probability*, vol. 1. Oakland, CA, USA, 281–297.

Mohajerin Esfahani, Peyman, Daniel Kuhn. 2018. Data-driven distributionally robust optimization using the Wasserstein metric: performance guarantees and tractable reformulations. *Mathematical Programming* **171**(1-2) 1–52.

Natarajan, Karthik, Melvyn Sim, Joline Uichanco. 2010. Tractable robust expected utility and risk models for portfolio optimization. *Mathematical Finance* **20**(4) 695–731

Perakis, Georgia, Melvyn Sim, Qinshen Tang, Peng Xiong. 2018. Joint pricing and production: a fusion of machine learning and robust optimization. *Available at SSRN.*

Pflug, Georg Ch. 2001. Scenario tree generation for multiperiod financial optimization by optimal discretization. *Mathematical Programming* **89**(2) 251–271.

Popescu, Ioana. 2007. Robust mean-covariance solutions for stochastic optimization. *Operations Research* **55**(1) 98–112.

See, Chuen-Teck, Melvyn Sim. 2010. Robust approximation to multiperiod inventory management. *Operations Research* **58**(3) 583–594.

Shapiro, Alexander, Shabbir Ahmed. 2004. On a class of minimax stochastic programs. *SIAM Journal on Optimization* **14**(4) 1237–1249.

Shapiro, Alexander, Tito Homem-de Mello. 1998. A simulation-based approach to two-stage stochastic programming with recourse. *Mathematical Programming* **81**(3) 301–325.

Shapiro, Alexander, Anton Kleywegt. 2002. Minimax analysis of stochastic problems. *Optimization Methods and Software* **17**(3) 523–542.

Shapiro, Alexander, Arkadi Nemirovski. 2005. On complexity of stochastic programming problems. *Continuous Optimization*. Springer, 111–146.

Soyster, Allen L. 1973. Technical note: convex programming with set-inclusive constraints and applications to inexact linear programming. *Operations Research* **21**(5) 1154–1157.

Takriti, Samer, Shabbir Ahmed. 2004. On robust optimization of two-stage systems. *Mathematical Programming* **99**(1) 109–126.

Tütüncü, Reha H, M Koenig. 2004. Robust asset allocation. *Annals of Operations Research* **132**(1-4) 157–187.

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

Zhen, Jianzhe, Daniel Kuhn, Wolfram Wiesemann. 2019. Distributionally robust nonlinear optimization. *Working paper.*

A. Proofs

Proof of Theorem 1. Let $\boldsymbol{\mu} = (\boldsymbol{\mu}_k)_{k\in[K]}$ and $\mathcal{Q} = \{\boldsymbol{\mu} \mid \boldsymbol{\mu}_k \in \mathcal{Q}_k \ \forall k \in [K]\}$. We can re-express

$$\lambda^* = \sup_{\mathbb{P}\in\mathcal{F}} \mathbb{E}_{\mathbb{P}}\left[\boldsymbol{r}^\top(\tilde{s})\boldsymbol{G}_m(\tilde{s})\tilde{\boldsymbol{z}} + h_m(\tilde{s})\right]$$

by $\lambda^* = \sup_{(\boldsymbol{p},\boldsymbol{\mu})\in\mathcal{P}\times\mathcal{Q}} \lambda(\boldsymbol{p},\boldsymbol{\mu})$, where given $(\boldsymbol{p},\boldsymbol{\mu}) \in \mathcal{P}\times\mathcal{Q}$, we define an ambiguity set

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

and correspondingly the worst-case expectation

$$\lambda(\boldsymbol{p},\boldsymbol{\mu}) = \sup_{\mathbb{P}\in\mathcal{F}(\boldsymbol{p},\boldsymbol{\mu})} \mathbb{E}_{\mathbb{P}}\left[\boldsymbol{r}^\top(\tilde{s})\boldsymbol{G}_m(\tilde{s})\tilde{\boldsymbol{z}} + h_m(\tilde{s})\right].$$

Using the law of total probability, we can construct the joint distribution $\mathbb{P}$ of $(\tilde{\boldsymbol{z}},\tilde{s})$ from the marginal distribution $\hat{\mathbb{P}}$ of $\tilde{s}$ supported on $[S]$ and the conditional distributions $\mathbb{P}_s$ of $\tilde{\boldsymbol{z}}$ given $\tilde{s}=s$, $s\in[S]$. In this way, we can reformulate $\lambda(\boldsymbol{p},\boldsymbol{\mu})$ as

$$\begin{aligned} \lambda(\boldsymbol{p},\boldsymbol{\mu}) = \sup\ & \sum_{s\in[S]} p_s \mathbb{E}_{\mathbb{P}_s}\left[\boldsymbol{r}^\top(\tilde{s})\boldsymbol{G}_m(\tilde{s})\tilde{\boldsymbol{z}} + h_m(\tilde{s})\right] & \\ \text{s.t.}\ & \sum_{s\in\mathcal{E}_k} p_s \mathbb{E}_{\mathbb{P}_s}\left[\tilde{\boldsymbol{z}}\right] = q_k\boldsymbol{\mu}_k & \forall k\in[K] \\ & \mathbb{P}_s\left[\tilde{\boldsymbol{z}}\in\mathcal{Z}_s\right] = 1 & \forall s\in[S] \end{aligned}$$

with $q_k = \sum_{s\in\mathcal{E}_k} p_s$, $k\in[K]$. We can express the dual of $\lambda(\boldsymbol{p},\boldsymbol{\mu})$ as

$$\begin{aligned} \lambda_1(\boldsymbol{p},\boldsymbol{\mu}) = \inf\ & \sum_{s\in[S]} \alpha_s + \sum_{k\in[K]} q_k\boldsymbol{\beta}_k^\top\boldsymbol{\mu}_k & \\ \text{s.t.}\ & \alpha_s + p_s\sum_{k\in\mathcal{K}_s} \boldsymbol{\beta}_k^\top\boldsymbol{z} \geq p_s(\boldsymbol{r}^\top(s)\boldsymbol{G}_m(s)\boldsymbol{z} + h_m(s)) & \forall \boldsymbol{z}\in\mathcal{Z}_s,\ s\in[S] \\ & \boldsymbol{\alpha}\in\mathbb{R}^S,\ \boldsymbol{\beta}_k\in\mathbb{R}^{I_z} & \forall k\in[K] \\[1ex] = \inf\ & \boldsymbol{\alpha}^\top\boldsymbol{p} + \sum_{k\in[K]} q_k\boldsymbol{\beta}_k^\top\boldsymbol{\mu}_k & \\ \text{s.t.}\ & \alpha_s + \sum_{k\in\mathcal{K}_s} \boldsymbol{\beta}_k^\top\boldsymbol{z} \geq \boldsymbol{r}^\top(s)\boldsymbol{G}_m(s)\boldsymbol{z} + h_m(s) & \forall \boldsymbol{z}\in\mathcal{Z}_s,\ s\in[S] \\ & \boldsymbol{\alpha}\in\mathbb{R}^S,\ \boldsymbol{\beta}_k\in\mathbb{R}^{I_z} & \forall k\in[K], \end{aligned}$$

where the second equality follows from for all $s\in[S]$, first changing variable from α_s to $p_s\alpha_s$ and then dividing both sides of the constraint by p_s, which is allowed since $\boldsymbol{p}\in\mathcal{P}$ is strictly positive.

By weak duality, $\lambda(\boldsymbol{p},\boldsymbol{\mu}) \leq \lambda_1(\boldsymbol{p},\boldsymbol{\mu})$. By the general min-max theorem, we further observe that

$$\lambda_1^* = \sup_{(\boldsymbol{p},\boldsymbol{\mu})\in\mathcal{P}\times\mathcal{Q}} \lambda_1(\boldsymbol{p},\boldsymbol{\mu}) \leq \lambda_2^*,$$

where

$$
\begin{aligned}
\lambda_2^* = \inf \ & \gamma \\
\text{s.t.} \ & \gamma \geq \boldsymbol{\alpha}^\top \boldsymbol{p} + \sum_{k\in[K]} q_k \boldsymbol{\beta}_k^\top \boldsymbol{\mu}_k && \forall \boldsymbol{p}\in\mathcal{P},\ \boldsymbol{\mu}_k\in\mathcal{Q}_k,\ k\in[K] \\
& \alpha_s + \sum_{k\in\mathcal{K}_s} \boldsymbol{\beta}_k^\top \boldsymbol{z} \geq \boldsymbol{r}^\top(s)\boldsymbol{G}_m(s)\boldsymbol{z} + h_m(s) && \forall \boldsymbol{z}\in\mathcal{Z}_s,\ s\in[S] \\
& \gamma\in\mathbb{R},\ \boldsymbol{\alpha}\in\mathbb{R}^S,\ \boldsymbol{\beta}_k\in\mathbb{R}^{I_z} && \forall k\in[K].
\end{aligned}
\tag{17}
$$

Due to the presence of products of uncertain variables (*e.g.*, $q_k\boldsymbol{\mu}_k$), problem (17) is nonconvex. Since $\boldsymbol{p} > 0$ (and hence $q_k > 0$), an equivalent convex representation can be obtained by changing variables in problem (17) from $q_k\boldsymbol{\mu}_k$ to $\boldsymbol{\mu}_k$ for all $k\in[K]$, which turns out to be problem (4).

Assuming the conic representation of the following system

$$
\begin{cases}
\dfrac{\sum_{s\in\mathcal{E}_k}\boldsymbol{\xi}_s}{\sum_{s\in\mathcal{E}_k}\tau_s} \in \mathcal{Q}_k & \forall k\in[K] \\
\dfrac{\boldsymbol{\xi}_s}{\tau_s} \in \mathcal{Z}_s & \forall s\in[S] \\
\boldsymbol{\tau}\in\mathcal{P}
\end{cases}
\tag{18}
$$

satisfies the Slater's condition (see Theorem 1.4.2 in Ben-Tal and Nemirovski 2001), one can establish strong duality, *i.e.*, $\lambda^* = \lambda_1^* = \lambda_2^*$ and show that problem (4) is solvable (Bertsimas et al. 2019b, Theorem 1). □

Proof of Theorem 2. We consider an ambiguity set without the auxiliary random variable $\tilde{v}$

$$
\bar{\mathcal{G}}_W(\theta) = \left\{ \mathbb{P}\in\mathcal{P}_0\left(\mathbb{R}^{I_u}\times[S]\right) \ \middle|\ \begin{array}{ll} (\tilde{\boldsymbol{u}},\tilde{s})\sim\mathbb{P} & \\ \mathbb{E}_{\mathbb{P}}\left[\rho(\tilde{\boldsymbol{u}},\hat{\boldsymbol{u}}_{\tilde{s}}) \mid \tilde{s}\in[S]\right] \leq \theta & \\ \mathbb{P}\left[\tilde{\boldsymbol{u}}\in\mathcal{U} \mid \tilde{s}=s\right] = 1 & \forall s\in[S] \\ \mathbb{P}\left[\tilde{s}=s\right] = \frac{1}{S} & \forall s\in[S] \end{array} \right\}.
\tag{19}
$$

Since this ambiguity set satisfies $\Pi_{(\tilde{\boldsymbol{u}},\tilde{s})}\mathcal{F}_W(\theta) = \bar{\mathcal{G}}_W(\theta)$ for all $\theta\geq 0$, thus it is sufficient to prove $\Pi_{\tilde{\boldsymbol{u}}}\bar{\mathcal{G}}_W(\theta) = \mathcal{G}_W(\theta)$ for all $\theta\geq 0$.

To this end, we first prove $\mathcal{G}_W(\theta)\subseteq\Pi_{\tilde{\boldsymbol{u}}}\bar{\mathcal{G}}_W(\theta)$. Consider $\tilde{\boldsymbol{u}}\sim\mathbb{P}$ for some $\mathbb{P}\in\mathcal{G}_W(\theta)$. By definition of the Wasserstein ambiguity set $\mathcal{G}_W(\theta)$, there exists a joint distribution $\mathbb{Q}\in\mathcal{P}(\mathbb{P},\hat{\mathbb{P}})$ of $(\tilde{\boldsymbol{u}},\tilde{\boldsymbol{u}}^\dagger)$ such that $\Pi_{\tilde{\boldsymbol{u}}}\mathbb{Q}=\mathbb{P}$, $\Pi_{\tilde{\boldsymbol{u}}^\dagger}\mathbb{Q}=\hat{\mathbb{P}}$, and $\mathbb{E}_{\mathbb{Q}}\left[\rho(\tilde{\boldsymbol{u}},\tilde{\boldsymbol{u}}^\dagger)\right]\leq\theta$. Since we can construct $\mathbb{Q}$ from the marginal distribution $\hat{\mathbb{P}}$ of $\tilde{\boldsymbol{u}}^\dagger$ supported on $\{\hat{\boldsymbol{u}}_1,\ldots,\hat{\boldsymbol{u}}_S\}$ and the conditional distributions $\mathbb{P}_s$ of $\tilde{\boldsymbol{u}}$, given the realization of $\tilde{\boldsymbol{u}}^\dagger$ is $\hat{\boldsymbol{u}}_s$, $s\in[S]$, we have $(\tilde{\boldsymbol{u}},\tilde{\boldsymbol{u}}^\dagger)\sim\frac{1}{S}\sum_{s\in[S]}\mathbb{P}_s\otimes\delta_{\hat{\boldsymbol{u}}_s}$. We can then construct a distribution $\mathbb{Q}'\in\mathcal{P}_0(\mathbb{R}^{I_u}\times[S])$ for the random variable $(\tilde{\boldsymbol{u}},\tilde{s})\sim\mathbb{Q}'$ via $\mathbb{Q}'=\frac{1}{S}\sum_{s\in[S]}\mathbb{P}_s\otimes\delta_s$. Observe that $\mathbb{Q}'\in\bar{\mathcal{G}}_W(\theta)$, hence $\mathcal{G}_W(\theta)\subseteq\Pi_{\tilde{\boldsymbol{u}}}\bar{\mathcal{G}}_W(\theta)$.

To prove $\Pi_{\tilde{\boldsymbol{u}}}\bar{\mathcal{G}}_W(\theta) \subseteq \mathcal{G}_W(\theta)$, we fix any $\mathbb{P} \in \bar{\mathcal{G}}_W(\theta)$ and we write its projection over $\tilde{\boldsymbol{u}}$ as $\Pi_{\tilde{\boldsymbol{u}}}\mathbb{P} = \frac{1}{S}\sum_{s\in[S]}\mathbb{P}_s$, where $\mathbb{P}_s$ is the conditional distribution of $\tilde{\boldsymbol{z}}$ given the outcome of the random scenario is s. We can then construct a joint distribution $\mathbb{Q} = \frac{1}{S}\sum_{s\in[S]}\mathbb{P}_s \otimes \delta_{\hat{\boldsymbol{u}}_s}$ of $(\tilde{\boldsymbol{u}}, \tilde{\boldsymbol{u}}^\dagger)$ that satisfies

$$\mathbb{E}_{\mathbb{Q}}\left[\rho(\tilde{\boldsymbol{u}}, \tilde{\boldsymbol{u}}^\dagger)\right] = \frac{1}{S}\sum_{s\in[S]}\mathbb{E}_{\mathbb{P}_s}[\rho(\tilde{\boldsymbol{u}}, \hat{\boldsymbol{u}}_s)] = \mathbb{E}_{\mathbb{P}}\left[\rho(\tilde{\boldsymbol{u}}, \hat{\boldsymbol{u}}_{\tilde{s}}) \mid \tilde{s} \in [S]\right] \leq \theta.$$

Hence, $\Pi_{\tilde{\boldsymbol{u}}}\mathbb{P} \in \mathcal{G}_W(\theta)$, which gives $\Pi_{\tilde{\boldsymbol{u}}}\bar{\mathcal{G}}_W(\theta) \subseteq \mathcal{G}_W(\theta)$ to conclude $\mathcal{G}_W(\theta) = \Pi_{\tilde{\boldsymbol{u}}}\bar{\mathcal{G}}_W(\theta)$. □

Proof of Theorem 3. Previous derivations in the proof of Theorem 1 implies that *(i)* the worst-case expectation (9) is equivalent to the following problem

$$\begin{array}{lll} \inf & \gamma & \\ \text{s.t.} & \gamma \geq \boldsymbol{\alpha}^\top \boldsymbol{p} + \sum_{k\in[K]} \boldsymbol{\beta}_k^\top \boldsymbol{\mu}_k & \forall \boldsymbol{p} \in \mathcal{P},\ \frac{\boldsymbol{\mu}_k}{\sum_{s\in\mathcal{E}_k} p_s} \in \mathcal{Q}_k,\ k \in [K] \\ & \alpha_s + \sum_{k\in\mathcal{K}_s} \boldsymbol{\beta}_k^\top \boldsymbol{z} \geq \boldsymbol{r}^\top(s)\boldsymbol{G}_\ell(s)\boldsymbol{z} + h_\ell(s) & \forall \boldsymbol{z} \in \mathcal{Z}_s,\ s \in [S],\ \ell \in [L] \\ & \gamma \in \mathbb{R},\ \boldsymbol{\alpha} \in \mathbb{R}^S,\ \boldsymbol{\beta}_k \in \mathbb{R}^{I_z} & \forall k \in [K]; \end{array} \tag{20}$$

and *(ii)* problem (10) is equivalent to

$$\begin{array}{lll} \inf & \gamma & \\ \text{s.t.} & \gamma \geq \boldsymbol{\alpha}^\top \boldsymbol{p} + \sum_{k\in[K]} \boldsymbol{\beta}_k^\top \boldsymbol{\mu}_k & \forall \boldsymbol{p} \in \mathcal{P},\ \frac{\boldsymbol{\mu}_k}{\sum_{s\in\mathcal{E}_k} p_s} \in \mathcal{Q}_k,\ k \in [K] \\ & \alpha_s + \sum_{k\in\mathcal{K}_s} \boldsymbol{\beta}_k^\top \boldsymbol{z} \geq y(s, \boldsymbol{z}) & \forall \boldsymbol{z} \in \mathcal{Z}_s,\ s \in [S] \\ & y(s, \boldsymbol{z}) \geq \boldsymbol{r}^\top(s)\boldsymbol{G}_\ell(s)\boldsymbol{z} + h_\ell(s) & \forall \boldsymbol{z} \in \mathcal{Z}_s,\ s \in [S],\ \ell \in [L] \\ & y \in \bar{\mathcal{A}}(\bar{\mathcal{C}}, [I_z]) & \\ & \gamma \in \mathbb{R},\ \boldsymbol{\alpha} \in \mathbb{R}^S,\ \boldsymbol{\beta}_k \in \mathbb{R}^{I_z} & \forall k \in [K]. \end{array} \tag{21}$$

It is then sufficient to construct a feasible solution to problem (21) from a feasible solution to problem (20) such that the constructive solution yields the same objective. Indeed, given a feasible solution $(\gamma^\dagger, \boldsymbol{\alpha}^\dagger, (\boldsymbol{\beta}_k^\dagger)_{k\in[K]})$ to problem (20), we can construct such a desired solution via:

$$\gamma = \gamma^\dagger,\ \boldsymbol{\alpha} = \boldsymbol{\alpha}^\dagger,\ \boldsymbol{\beta}_k = \boldsymbol{\beta}_k^\dagger\ \forall k \in [K],\ y(s, \boldsymbol{z}) = \alpha_s^\dagger + \sum_{k\in\mathcal{K}_s} (\boldsymbol{\beta}_k^\dagger)^\top \boldsymbol{z}\ \forall s \in [S],$$

for which the recourse decision $y(\cdot, \cdot) \in \bar{\mathcal{A}}(\bar{\mathcal{C}}, [I_z])$. □

Proof of Theorem 4. Using the result of Popescu (2007), we first show that

$$\inf_{\mathbb{P}\in\mathcal{G}(\boldsymbol{w}^\top\boldsymbol{\mu}, \boldsymbol{w}^\top\boldsymbol{\Sigma}\boldsymbol{w})} \mathbb{E}_{\mathbb{P}}[U(\tilde{u})] = \sup_r \left\{ \inf_{\mathbb{P}\in\mathcal{F}} \mathbb{E}_{\mathbb{P}}[U(r\tilde{u} + \boldsymbol{w}^\top\boldsymbol{\mu})] \;\middle|\; r \geq \sqrt{\boldsymbol{w}'\boldsymbol{\Sigma}\boldsymbol{w}} \right\}.$$

By duality, we have

$$\begin{aligned} \inf_{\mathbb{P}\in\mathcal{G}(\boldsymbol{w}^\top\boldsymbol{\mu}, \boldsymbol{w}^\top\boldsymbol{\Sigma}\boldsymbol{w})} \mathbb{E}_{\mathbb{P}}[U(\tilde{u})] &= \inf_{\mathbb{P}\in\mathcal{G}(0, \boldsymbol{w}^\top\boldsymbol{\Sigma}\boldsymbol{w})} \mathbb{E}_{\mathbb{P}}[U(\tilde{u} + \boldsymbol{w}^\top\boldsymbol{\mu})] \\ &= \sup_{\alpha, \beta_1, \beta_2} \left\{ \alpha + \boldsymbol{w}^\top\boldsymbol{\Sigma}\boldsymbol{w} \cdot \beta_2 \;\middle|\; \alpha + \beta_1 u + \beta_2 u^2 \leq U(u + \boldsymbol{w}^\top\boldsymbol{\mu})\ \ \forall u \right\}. \end{aligned}$$

Note that it requires $\beta_2 \le 0$ for the above problem to be feasible, as otherwise the constraint would be violated for some sufficiently large u. Hence, we can further rewrite this problem into

$$\begin{aligned}
&\sup_{\alpha,\beta_1,\beta_2,r}\left\{\alpha + r^2\beta_2 \;\middle|\; \begin{array}{l}\alpha+\beta_1 u+\beta_2 u^2 \le U(u+\boldsymbol{w}^\top\boldsymbol{\mu}) \quad \forall u\\ r \ge \sqrt{\boldsymbol{w}^\top\boldsymbol{\Sigma}\boldsymbol{w}}\end{array}\right\}\\
=\;& \sup_{r}\left\{\inf_{\mathbb{P}\in\mathcal{G}(0,r^2)} \mathbb{E}_{\mathbb{P}}[U(\tilde{u}+\boldsymbol{w}^\top\boldsymbol{\mu})] \;\middle|\; r \ge \sqrt{\boldsymbol{w}^\top\boldsymbol{\Sigma}\boldsymbol{w}}\right\}\\
=\;& \sup_{r}\left\{\inf_{\mathbb{P}\in\mathcal{G}(0,1)} \mathbb{E}_{\mathbb{P}}[U(r\tilde{u}+\boldsymbol{w}^\top\boldsymbol{\mu})] \;\middle|\; r \ge \sqrt{\boldsymbol{w}^\top\boldsymbol{\Sigma}\boldsymbol{w}}\right\}\\
=\;& \sup_{\alpha,\beta_1,\beta_2,r}\left\{\alpha + \beta_2 \;\middle|\; \begin{array}{l}\alpha+\beta_1 u+\beta_2 u^2 \le U(ru+\boldsymbol{w}^\top\boldsymbol{\mu}) \quad \forall u\\ r \ge \sqrt{\boldsymbol{w}^\top\boldsymbol{\Sigma}\boldsymbol{w}}\end{array}\right\}\\
=\;& \sup_{\alpha,\beta_1,\beta_2,r}\left\{\alpha + \beta_2 \;\middle|\; \begin{array}{l}\alpha+\beta_1 u+\beta_2 v \le U(ru+\boldsymbol{w}^\top\boldsymbol{\mu}) \quad \forall (u,v): v \ge u^2\\ \beta_2 \le 0,\; r \ge \sqrt{\boldsymbol{w}^\top\boldsymbol{\Sigma}\boldsymbol{w}}\end{array}\right\}\\
=\;& \sup_{r}\left\{\inf_{\mathbb{P}\in\mathcal{F}} \mathbb{E}_{\mathbb{P}}[U(r\tilde{u}+\boldsymbol{w}^\top\boldsymbol{\mu})] \;\middle|\; r \ge \sqrt{\boldsymbol{w}^\top\boldsymbol{\Sigma}\boldsymbol{w}}\right\}.
\end{aligned}$$

The result then follows by applying Theorem 3. □

Proof of Theorem 5. Observe that for any feasible recourse decision $\bar{\boldsymbol{r}}(\cdot)$ to problem (13), we have

$$\sup_{\mathbb{P}\in\mathcal{F}} \mathbb{E}_{\mathbb{P}}[\bar{\boldsymbol{r}}^\top(\tilde{s})\tilde{\boldsymbol{v}}] = \sup_{\mathbb{P}\in\mathcal{G}} \mathbb{E}_{\mathbb{P}}[\bar{\boldsymbol{r}}^\top(\tilde{s})\boldsymbol{\zeta}(\tilde{\boldsymbol{u}},\tilde{s})].$$

In addition, the optimal $\bar{\boldsymbol{r}}^\star(\cdot)$ to problem (13) satisfies $\bar{r}^\star_\ell(s) = \xi_\ell(\boldsymbol{r}(s),s)$ for all $\ell\in[I_v]$ and $s\in[S]$. Therefore, our claim holds. □

B. Worst-Case Expectation of Quadratic Functions

The RSO framework can be used to provide a tight characterization of the worst-case expectation of some quadratic functions that are known in the literature and extend them to include discrete scenarios. Let $\mathbb{S}^I$ be the space of symmetric matrices in $\mathbb{R}^{I\times I}$. Given $\boldsymbol{X},\boldsymbol{Y}\in\mathbb{S}^I$, we denote by $\boldsymbol{X}\succeq\boldsymbol{Y}$ (resp., $\boldsymbol{X}\succ\boldsymbol{Y}$) to represent $\boldsymbol{X}-\boldsymbol{Y}$ is positive semidefinite (resp., definite), and denote by $\boldsymbol{X}\bullet\boldsymbol{Y}$ as the trace inner product of $\boldsymbol{X},\boldsymbol{Y}$. Special matrices and vectors of the appropriate dimension include $\boldsymbol{O}$, $\boldsymbol{I}$, and $\boldsymbol{0}$, which respectively correspond to the zero matrix, the identity matrix, and the zero vector.

Bi-Convex-Quadratic Function

We explore the following bi-convex-quadratic function as an extension of Ben-Tal and Nemirovski (1998) to include discrete scenarios:

$$g(\boldsymbol{r}(s),\boldsymbol{u},s) \triangleq \boldsymbol{u}^\top\boldsymbol{A}^\top(\boldsymbol{r}(s),s)\boldsymbol{A}(\boldsymbol{r}(s),s)\boldsymbol{u} + 2\boldsymbol{u}^\top\boldsymbol{b}(\boldsymbol{r}(s),s) + c(\boldsymbol{r}(s),s),$$

where given a scenario s, $\boldsymbol{A}(\boldsymbol{r}(s),s), \boldsymbol{b}(\boldsymbol{r}(s),s), c(\boldsymbol{r}(s),s)$ are affine mappings of $\boldsymbol{r}(s)$. The event-wise ambiguity set is given by

$$\mathcal{G} = \left\{ \mathbb{P} \in \mathcal{P}_0\left(\mathbb{R}^{I_u} \times [S]\right) \;\middle|\; \begin{array}{ll} (\tilde{\boldsymbol{u}}, \tilde{s}) \sim \mathbb{P} & \\ \mathbb{E}_{\mathbb{P}}\left[\binom{1}{\tilde{\boldsymbol{u}}}\binom{1}{\tilde{\boldsymbol{u}}}^\top \;\middle|\; \tilde{s} \in \mathcal{E}_k\right] \in \mathcal{Q}_k & \forall k \in [K] \\ \mathbb{P}\left[\binom{1}{\tilde{\boldsymbol{u}}}\binom{1}{\tilde{\boldsymbol{u}}}^\top \in \mathcal{U}_s \;\middle|\; \tilde{s} = s\right] = 1 & \forall s \in [S] \\ \mathbb{P}[\tilde{s} = s] = p_s & \forall s \in [S] \\ \text{for some } \boldsymbol{p} \in \mathcal{P} & \end{array} \right\}.$$

The support set $\mathcal{U}_s$ is general enough to capture the ubiquitous uncertainty set $\{\boldsymbol{u} \mid \boldsymbol{u}^\top \boldsymbol{\Lambda}_s \boldsymbol{u} \leq 1\}$ parameterized by some $\boldsymbol{\Lambda}_s \succ \boldsymbol{0}$, for which we only need to define

$$\mathcal{U}_s = \left\{ \boldsymbol{U} \in \mathbb{S}^{I_u+1} \;\middle|\; \boldsymbol{U} \bullet \begin{pmatrix} -1 & \boldsymbol{0}^\top \\ \boldsymbol{0} & \boldsymbol{\Lambda}_s \end{pmatrix} \leq 0 \right\}. \tag{22}$$

THEOREM 6. *The worst-case expectation*

$$\sup_{\mathbb{P} \in \mathcal{G}} \mathbb{E}_{\mathbb{P}}\left[g(\boldsymbol{r}(\tilde{s}), \tilde{\boldsymbol{u}}, \tilde{s})\right] \tag{23}$$

is bounded from above by

$$\begin{array}{lll} \min & \sup_{\mathbb{P} \in \mathcal{F}} \mathbb{E}_{\mathbb{P}}[\boldsymbol{R}(\tilde{s}) \bullet \tilde{\boldsymbol{Z}}] & \\ \text{s.t.} & \begin{pmatrix} \boldsymbol{R}(s) - \begin{pmatrix} 1 & \boldsymbol{b}^\top(\boldsymbol{r}(s),s) \\ \boldsymbol{b}(\boldsymbol{r}(s),s) & \boldsymbol{O} \end{pmatrix} & \begin{pmatrix} 0 & \boldsymbol{0}^\top \\ \boldsymbol{0} & \boldsymbol{A}^\top(\boldsymbol{r}(s),s) \end{pmatrix} \\ \begin{pmatrix} 0 & \boldsymbol{0}^\top \\ \boldsymbol{0} & \boldsymbol{A}(\boldsymbol{r}(s),s) \end{pmatrix} & \boldsymbol{I} \end{pmatrix} \succeq \boldsymbol{O} & \forall s \in [S] \\ & \boldsymbol{R}(s) \in \mathbb{S}^{I_u+1} & \forall s \in [S], \end{array} \tag{24}$$

where the lifted event-wise ambiguity set

$$\mathcal{F} = \left\{ \mathbb{P} \in \mathcal{P}_0\left(\mathbb{S}^{I_u+1} \times [S]\right) \;\middle|\; \begin{array}{ll} (\tilde{\boldsymbol{Z}}, \tilde{s}) \sim \mathbb{P} & \\ \mathbb{E}_{\mathbb{P}}[\tilde{\boldsymbol{Z}} \mid \tilde{s} \in \mathcal{E}_k] \in \mathcal{Q}_k & \forall k \in [K] \\ \mathbb{P}[\tilde{\boldsymbol{Z}} \in \mathcal{Z}_s \mid \tilde{s} = s] = 1 & \forall s \in [S] \\ \mathbb{P}[\tilde{s} = s] = p_s & \forall s \in [S] \\ \textit{for some } \boldsymbol{p} \in \mathcal{P} & \end{array} \right\}$$

takes lifted support sets $\mathcal{Z}_s = \{\boldsymbol{Z} \in \mathcal{U}_s \mid \boldsymbol{Z} \succeq \boldsymbol{O},\ [\boldsymbol{Z}]_{1,1} = 1\}$, $s \in [S]$. *Moreover, the bound is tight for ellipsoidal support sets defined in (22).*

Proof of Theorem 6. We note that

$$g(\boldsymbol{r}(s), \boldsymbol{u}, s) = \begin{pmatrix} 1 & \boldsymbol{b}^\top(\boldsymbol{r}(s),s) \\ \boldsymbol{b}(\boldsymbol{r}(s),s) & \boldsymbol{A}^\top(\boldsymbol{r}(s),s)\boldsymbol{A}(\boldsymbol{r}(s),s) \end{pmatrix} \bullet \left(\begin{pmatrix} 1 \\ \boldsymbol{u} \end{pmatrix} \begin{pmatrix} 1 \\ \boldsymbol{u} \end{pmatrix}^\top \right).$$

By Schur complement, each positive semidefinite constraint of problem (24) is equivalent to

$$\boldsymbol{R}(s) \succeq \begin{pmatrix} 1 & \boldsymbol{b}^\top(\boldsymbol{r}(s), s) \\ \boldsymbol{b}(\boldsymbol{r}(s), s) & \boldsymbol{A}^\top(\boldsymbol{r}(s), s)\boldsymbol{A}(\boldsymbol{r}(s), s) \end{pmatrix}.$$

Since $\boldsymbol{Z} \in \mathcal{Z}_s$ is positive semidefinite, an optimal $\boldsymbol{R}(s)$ would be

$$\boldsymbol{R}(s) = \begin{pmatrix} 1 & \boldsymbol{b}^\top(\boldsymbol{r}(s), s) \\ \boldsymbol{b}(\boldsymbol{r}(s), s) & \boldsymbol{A}^\top(\boldsymbol{r}(s), s)\boldsymbol{A}(\boldsymbol{r}(s), s) \end{pmatrix}.$$

Observe that the ambiguity set $\mathcal{F}$ coincides with $\mathcal{G}$ if every support set $\mathcal{Z}_s$ is replaced by $\bar{\mathcal{Z}}_s = \{\boldsymbol{Z} \in \mathcal{U}_s \mid \boldsymbol{Z} \succeq \boldsymbol{O},\ [\boldsymbol{Z}]_{1,1} = 1,\ \mathrm{rank}(\boldsymbol{Z}) = 1\}$, which however, would lead to a harder problem to solve due to the rank constraint. Since $\bar{\mathcal{Z}}_s \subseteq \mathcal{Z}_s$, we obtain the conservative upper bound.

We next show that the bound is tight for ellipsoidal uncertainty sets defined in (22). After using Theorem 1 to reformulate problem (23), we need to deal with the following robust counterpart.

$$\alpha_s \geq \boldsymbol{\Phi}_s \bullet \left(\begin{pmatrix} 1 \\ \boldsymbol{u} \end{pmatrix} \begin{pmatrix} 1 \\ \boldsymbol{u} \end{pmatrix}^\top \right) \quad \forall \boldsymbol{u}^\top \boldsymbol{\Lambda}_s \boldsymbol{u} \leq 1$$

for some $\alpha_s \in \mathbb{R}, \boldsymbol{\Phi}_s \in \mathbb{S}^{I_u+1}$, which by S-lemma, is equivalent to

$$\begin{pmatrix} \alpha_s & \boldsymbol{0}^\top \\ \boldsymbol{0} & \boldsymbol{0} \end{pmatrix} + \delta_s \begin{pmatrix} -1 & \boldsymbol{0}^\top \\ \boldsymbol{0} & \boldsymbol{\Lambda}_s \end{pmatrix} \succeq \boldsymbol{\Phi}_s,$$

for some $\delta_s \geq 0$. On the other hand, the robust counterpart in the reformulation of problem (24)

$$\alpha_s \geq \boldsymbol{\Phi}_s \bullet \boldsymbol{Z} \quad \forall \boldsymbol{Z} \bullet \begin{pmatrix} -1 & \boldsymbol{0}^\top \\ \boldsymbol{0} & \boldsymbol{\Lambda}_s \end{pmatrix} \leq 0,\ [\boldsymbol{Z}]_{1,1} = 1,\ \boldsymbol{Z} \succeq \boldsymbol{O}$$

is equivalent to

$$\begin{pmatrix} \tau_s & \boldsymbol{0}^\top \\ \boldsymbol{0} & \boldsymbol{0} \end{pmatrix} + \delta_s \begin{pmatrix} -1 & \boldsymbol{0}^\top \\ \boldsymbol{0} & \boldsymbol{\Lambda}_s \end{pmatrix} \succeq \boldsymbol{\Phi}_s,$$

for some $\tau_s \leq \alpha_s$ and $\delta_s \geq 0$, for which we can replace τ_s with α_s without affecting its feasibility. This establishes the desired tight bound for ellipsoidal uncertainty sets. □

Bi-Conic-Quadratic Function

We can also extend the bi-conic-quadratic function considered in Ben-Tal and Nemirovski (1998) to include discrete scenarios as follows:

$$h(\boldsymbol{r}(s), \boldsymbol{u}, s) \triangleq \|\boldsymbol{A}(\boldsymbol{r}(s), s)\boldsymbol{u} + \boldsymbol{b}(\boldsymbol{r}(s), s)\|_2,$$

where given s, $\boldsymbol{A}(\boldsymbol{r}(s), s), \boldsymbol{b}(\boldsymbol{r}(s), s)$ are affine mappings of $\boldsymbol{r}(s)$. The event-wise ambiguity set takes

$$\mathcal{G} = \left\{ \mathbb{P} \in \mathcal{P}_0\left(\mathbb{R}^{I_u} \times [S]\right) \;\middle|\; \begin{array}{ll} (\tilde{\boldsymbol{u}}, \tilde{s}) \sim \mathbb{P} & \\ \mathbb{P}\left[\binom{1}{\tilde{\boldsymbol{u}}}\binom{1}{\tilde{\boldsymbol{u}}}^\top \in \mathcal{U}_s \;\middle|\; \tilde{s} = s\right] = 1 & \forall s \in [S] \\ \mathbb{P}[\tilde{s} = s] = p_s & \forall s \in [S] \\ \text{for some } \boldsymbol{p} \in \mathcal{P} & \end{array} \right\}.$$

THEOREM 7. *The worst-case expectation*

$$\sup_{\mathbb{P}\in\mathcal{G}} \mathbb{E}_{\mathbb{P}}\left[h(\boldsymbol{r}(\tilde{s}), \tilde{\boldsymbol{u}}, \tilde{s})\right]$$

is bounded from above by

$$\begin{array}{lll} \min & \sup_{\mathbb{P}\in\mathcal{F}} \mathbb{E}_{\mathbb{P}}[x(\tilde{s})] & \\ \text{s.t.} & x(s) \geq \boldsymbol{R}(s) \bullet \boldsymbol{Z} & \forall \boldsymbol{Z} \in \mathcal{Z}_s,\ s \in [S] \\ & \begin{pmatrix} \boldsymbol{R}(s) & \begin{pmatrix} \boldsymbol{b}^\top(\boldsymbol{r}(s), s) \\ \boldsymbol{A}^\top(\boldsymbol{r}(s), s) \end{pmatrix} \\ (\boldsymbol{b}(\boldsymbol{r}(s), s)\ \ \boldsymbol{A}(\boldsymbol{r}(s), s)) & x(s)\boldsymbol{I} \end{pmatrix} \succeq \boldsymbol{O} & \forall s \in [S] \\ & \boldsymbol{R}(s) \in \mathbb{S}^{I_u+1} & \forall s \in [S] \\ & x \in \mathcal{A}(\bar{\mathcal{C}}), & \end{array} \tag{25}$$

where $\bar{\mathcal{C}} \triangleq \{\{s\} \mid s \in [S]\}$ *and where the lifted event-wise ambiguity set*

$$\mathcal{F} = \left\{ \mathbb{P} \in \mathcal{P}_0\left(\mathbb{S}^{I_u+1} \times [S]\right) \left| \begin{array}{ll} (\tilde{\boldsymbol{Z}}, \tilde{s}) \sim \mathbb{P} & \\ \mathbb{P}[\tilde{\boldsymbol{Z}} \in \mathcal{Z}_s \mid \tilde{s} = s] = 1 & \forall s \in [S] \\ \mathbb{P}[\tilde{s} = s] = p_s & \forall s \in [S] \\ \text{for some } \boldsymbol{p} \in \mathcal{P} & \end{array} \right.\right\}$$

takes lifted support sets $\mathcal{Z}_s = \{\boldsymbol{Z} \in \mathcal{U}_s \mid \boldsymbol{Z} \succeq \boldsymbol{O},\ [\boldsymbol{Z}]_{1,1} = 1\}$, $s \in [S]$*. Moreover, the bound is tight for ellipsoidal uncertainty sets defined in* (22)*.*

Proof of Theorem 7. Since the ambiguity set does not contain any expectation constraint, we can obtain a tractable reformulation by replacing $h(\boldsymbol{r}(s), \boldsymbol{u}, s)$ with a recourse variable $x(s)$ and imposing the following constraint (see reformulation in Theorem 1):

$$x^2(s) \geq h^2(\boldsymbol{r}(s), \boldsymbol{u}, s) \quad \forall \begin{pmatrix} 1 \\ \boldsymbol{u} \end{pmatrix} \begin{pmatrix} 1 \\ \boldsymbol{u} \end{pmatrix}^\top \in \mathcal{U}_s, s \in [S].$$

We next discuss how such a constraint can be specified in problem (25). Observe that

$$h^2(\boldsymbol{r}(s), \boldsymbol{u}, s) = \left(\begin{pmatrix} \boldsymbol{b}^\top(\boldsymbol{r}(s), s) \\ \boldsymbol{A}^\top(\boldsymbol{r}(s), s) \end{pmatrix} (\boldsymbol{b}(\boldsymbol{r}(s), s)\ \ \boldsymbol{A}(\boldsymbol{r}(s), s)) \right) \bullet \left(\begin{pmatrix} 1 \\ \boldsymbol{u} \end{pmatrix} \begin{pmatrix} 1 \\ \boldsymbol{u} \end{pmatrix}^\top \right).$$

By Schur complement, each positive semidefinite constraint of problem (24) is equivalent to

$$x(s)\boldsymbol{R}(s) \succeq \begin{pmatrix} \boldsymbol{b}^\top(\boldsymbol{r}(s), s) \\ \boldsymbol{A}^\top(\boldsymbol{r}(s), s) \end{pmatrix} (\boldsymbol{b}(\boldsymbol{r}(s), s)\ \ \boldsymbol{A}(\boldsymbol{r}(s), s)).$$

Since $\boldsymbol{Z} \in \mathcal{Z}_s$ is positive semidefinite and $x(s) \geq 0$, an optimal $\boldsymbol{R}(s)$ would be

$$x(s)\boldsymbol{R}(s) = \begin{pmatrix} \boldsymbol{b}^\top(\boldsymbol{r}(s), s) \\ \boldsymbol{A}^\top(\boldsymbol{r}(s), s) \end{pmatrix} (\boldsymbol{b}(\boldsymbol{r}(s), s)\ \ \boldsymbol{A}(\boldsymbol{r}(s), s)).$$

The rest of the proof follows similarly as in the proof of Theorem 6. □

Affine-Quadractic Function

As an extension of Tütüncü and Koenig (2004), we consider a saddle function that is convex quadratic with respect to the decision variable and that is affine with respect to $\boldsymbol{z}$:

$$g(\boldsymbol{r}(s), \boldsymbol{z}, s) \triangleq \boldsymbol{r}^\top(s)\boldsymbol{H}(s, \boldsymbol{z})\boldsymbol{r}(s) + \boldsymbol{r}^\top(s)\boldsymbol{G}(s)\boldsymbol{z} + h(s), \tag{26}$$

where given a scenario s, $\boldsymbol{H}(s, \boldsymbol{z})$ is an affine mapping of $\boldsymbol{z}$ and $\mathcal{Z}_s \subseteq \{\boldsymbol{z} \mid \boldsymbol{H}(\boldsymbol{z}, s) \succeq \boldsymbol{O}\}$. Introducing auxiliary variables $\boldsymbol{R}(s) \in \mathbb{S}^{I_v}$, $s \in [S]$ and using the Schur complement, the robust expectation $\sup_{\mathbb{P}\in\mathcal{F}} \mathbb{E}_{\mathbb{P}}\left[g(\boldsymbol{r}(\tilde{s}), \tilde{\boldsymbol{z}}, \tilde{s})\right]$ is the same as

$$\begin{array}{lll} \min & \sup_{\mathbb{P}\in\mathcal{F}} \mathbb{E}_{\mathbb{P}}\left[\boldsymbol{R}(\tilde{s}) \bullet \boldsymbol{H}(\tilde{s}, \tilde{\boldsymbol{z}}) + \boldsymbol{r}^\top(\tilde{s})\boldsymbol{G}(\tilde{s})\tilde{\boldsymbol{z}} + h(\tilde{s})\right] & \\ \text{s.t.} & \begin{pmatrix} 1 & \boldsymbol{r}^\top(s) \\ \boldsymbol{r}(s) & \boldsymbol{R}(s) \end{pmatrix} \succeq \boldsymbol{0} & \forall s \in [S] \\ & \boldsymbol{R}(s) \in \mathbb{S}^{I_v} & \forall s \in [S], \end{array}$$

which falls within the RSO framework.

C. Representation of Wasserstein Ambiguity Sets

For $p \in [1, \infty)$, the type-p Wasserstein metric between two distributions $\mathbb{P}$ and $\hat{\mathbb{P}}$ for a given distance metric ρ is defined as

$$d_W^p(\mathbb{P}, \hat{\mathbb{P}}) \triangleq \inf_{\mathbb{Q}\in\mathcal{Q}(\mathbb{P},\hat{\mathbb{P}})} \left(\mathbb{E}_{\mathbb{Q}}\left[\rho^p(\tilde{\boldsymbol{u}}, \tilde{\boldsymbol{u}}^\dagger)\right]\right)^{\frac{1}{p}}.$$

Correspondingly, the type-p Wasserstein ambiguity set is defined by

$$\mathcal{G}_W^p(\theta) = \left\{ \mathbb{P} \in \mathcal{P}_0(\mathcal{U}) \left| \begin{array}{l} \tilde{\boldsymbol{u}} \sim \mathbb{P} \\ d_W^p(\mathbb{P}, \hat{\mathbb{P}}) \leq \theta \end{array} \right. \right\}.$$

Consider another distance metric ρ^p and the corresponding type-1 Wasserstein metric $\bar{d}_W(\mathbb{P}, \hat{\mathbb{P}})$ between $\mathbb{P}$ and $\hat{\mathbb{P}}$ which is determined by

$$\bar{d}_W(\mathbb{P}, \hat{\mathbb{P}}) \triangleq \inf_{\mathbb{Q}\in\mathcal{Q}(\mathbb{P},\hat{\mathbb{P}})} \mathbb{E}_{\mathbb{Q}}\left[\rho^p(\tilde{\boldsymbol{u}}, \tilde{\boldsymbol{u}}^\dagger)\right].$$

We then have

$$\mathcal{G}_W^p(\theta) = \left\{ \mathbb{P} \in \mathcal{P}_0(\mathcal{U}) \left| \begin{array}{l} \tilde{\boldsymbol{u}} \sim \mathbb{P} \\ \bar{d}_W(\mathbb{P}, \hat{\mathbb{P}}) \leq \theta^p \end{array} \right. \right\}.$$

Equivalently, for $p \in [1, \infty)$, the type-p Wasserstein ambiguity set of radius θ can be re-interpreted as a type-1 Wasserstein ambiguity set of radius θ^p where the type-1 Wasserstein metric between $\mathbb{P}$ and $\hat{\mathbb{P}}$ is $\bar{d}_W(\mathbb{P}, \hat{\mathbb{P}})$. From this perspective, we can directly use Theorem 2 to represent the type-p Wasserstein ambiguity set $\mathcal{G}_W^p(\theta)$ in the format of an event-wise ambiguity set.

For $p=\infty$, the type-∞ Wasserstein metric between two distributions $\mathbb{P}$ and $\hat{\mathbb{P}}$ is defined as

$$d_W^\infty(\mathbb{P},\hat{\mathbb{P}}) \triangleq \inf_{\mathbb{Q}\in\mathcal{Q}(\mathbb{P},\hat{\mathbb{P}})} \mathbb{Q}\text{-ess}\sup_{\mathcal{U}\times\mathcal{U}} \rho(\tilde{\boldsymbol{u}},\tilde{\boldsymbol{u}}^\dagger),$$

where the essential supremum of the joint distribution $\mathbb{Q}$ is defined by

$$\mathbb{Q}\text{-ess}\sup_{\mathcal{U}\times\mathcal{U}} \rho(\tilde{\boldsymbol{u}},\tilde{\boldsymbol{u}}^\dagger) = \inf\{M:\ \mathbb{Q}\left[\rho(\tilde{\boldsymbol{u}}-\tilde{\boldsymbol{u}}^\dagger)>M\right]=0\}.$$

Bertsimas et al. (2018) show that a distribution $\mathbb{P}$ in the type-∞ Wasserstein ambiguity set

$$\mathcal{G}_W^\infty(\theta) = \left\{ \mathbb{P}\in\mathcal{P}_0(\mathcal{U}) \;\middle|\; \begin{array}{l} \tilde{\boldsymbol{u}}\sim\mathbb{P} \\ d_W^\infty(\mathbb{P},\hat{\mathbb{P}})\le\theta \end{array} \right\}$$

is indeed a mixture distribution $\mathbb{P}=\frac{1}{S}\sum_{s\in[S]}\mathbb{P}_s$ consisting of ambiguous components such that for every $s\in[S]$, $\mathbb{P}_s\in\mathcal{P}_0(\mathcal{U})$ and $\mathbb{P}_s[\rho(\tilde{\boldsymbol{u}},\hat{\boldsymbol{u}}_s)\le\theta]=1$. Therefore, one can represent the type-∞ Wasserstein ambiguity set using the following mixture-distribution ambiguity set

$$\mathcal{F}_W^\infty(\theta) = \left\{ \mathbb{P}\in\mathcal{P}_0\left(\mathbb{R}^{I_u}\times[S]\right) \;\middle|\; \begin{array}{ll} (\tilde{\boldsymbol{u}},\tilde{s})\sim\mathbb{P} & \\ \mathbb{P}\left[\tilde{\boldsymbol{u}}\in\mathcal{U},\rho(\tilde{\boldsymbol{u}},\hat{\boldsymbol{u}}_s)\le\theta \mid \tilde{s}=s\right]=1 & \forall s\in[S] \\ \mathbb{P}\left[\tilde{s}=s\right]=\frac{1}{S} & \forall s\in[S] \end{array} \right\},$$

which is an event-wise ambiguity set satisfying $\mathcal{G}_W^\infty(\theta)=\Pi_{\tilde{\boldsymbol{u}}}\mathcal{F}_W^\infty(\theta)$ for all $\theta\ge 0$.

In recent independent works, based on a generalized 'primal-worst equals dual-best' duality scheme, Kuhn et al. (2019) and Zhen et al. (2019) provide convex reformulations for many distributionally robust optimization problems with type-p Wasserstein metric for $p\in[1,\infty]$.

D. Computational Experiments with Wasserstein Ambiguity Sets

We focus on two-stage linear optimization problems with the data-driven Wasserstein ambiguity set of type-1 in the form (7), given some past observations $\hat{\boldsymbol{u}}_1,\ldots,\hat{\boldsymbol{u}}_S$ of the uncertainty.

Multi-Item Newsvendor Problem

We consider a multi-item newsvendor problem with I_u different items. For each item i ($i\in[I_u]$), its unit selling price and ordering cost are denoted by p_i and c_i, respectively. Under a total budget d, the decision maker decides the ordering quantity w_i of each item before its random demand $\tilde{u}_i$ is observed. Once the demand realizes, the selling quantity of each item is decided as $\min\{w_i,u_i\}$. The decision maker maximizes the worst-case expected operating revenue by solving

$$\begin{array}{rl} \max & \inf_{\mathbb{P}\in\mathcal{F}_W(\theta)} \mathbb{E}_{\mathbb{P}}\left[\sum_{i\in[I_u]} p_i\min\{w_i,\tilde{u}_i\}\right] \\ \text{s.t.} & \boldsymbol{c}^\top\boldsymbol{w}=d,\ \boldsymbol{w}\ge\boldsymbol{0}, \end{array}$$

which can be recast as a minimization problem,

$$\begin{aligned} \min\ & -\boldsymbol{p}^\top \boldsymbol{w} + \sup_{\mathbb{P}\in\mathcal{F}_W(\theta)} \mathbb{E}_{\mathbb{P}}\left[\sum_{i\in[I_u]} p_i(w_i-\tilde{u}_i)^+\right] \\ \text{s.t.}\ & \boldsymbol{c}^\top\boldsymbol{w} = d,\ \boldsymbol{w}\geq \boldsymbol{0}. \end{aligned} \tag{27}$$

In the objective function, $\sum_{i\in[I_u]} p_i(w_i-u_i)^+ = \max_{\mathcal{J}\subseteq[I_u]}\sum_{j\in\mathcal{J}} p_j(w_j-u_j)$ is convex and piecewise affine involving 2^{I_u} pieces. Thus by Theorem 3, problem (27) can be exactly solved by

$$\begin{aligned} \lambda^\star = \min\ & -\boldsymbol{p}^\top \boldsymbol{w} + \sup_{\mathbb{P}\in\mathcal{F}_W(\theta)} \mathbb{E}_{\mathbb{P}}\left[y(\tilde{s},\tilde{\boldsymbol{z}})\right] \\ \text{s.t.}\ & y(s,\boldsymbol{z}) \geq \sum_{j\in\mathcal{J}} p_j(w_j-u_j) \qquad \forall \boldsymbol{z}\in\mathcal{Z}_s,\ s\in[S],\ \mathcal{J}\subseteq[I_u] \\ & \boldsymbol{c}^\top\boldsymbol{w} = d,\ \boldsymbol{w}\geq\boldsymbol{0} \\ & y\in\bar{\mathcal{A}}(\bar{\mathcal{C}},[I_u+1]), \end{aligned} \tag{28}$$

where we introduce a recourse variable $y(\cdot,\cdot)$ following the event-wise affine adaptation with the collection $\bar{\mathcal{C}} \triangleq \{\{s\}\mid s\in[S]\}$. Problem size of this exact approach however, increases exponentially in the number of items. Alternatively, we can obtain an upper bound by solving an RSO problem:

$$\begin{aligned} \lambda = \min\ & -\boldsymbol{p}^\top \boldsymbol{x} + \sup_{\mathbb{P}\in\mathcal{F}_W(\theta)} \mathbb{E}_{\mathbb{P}}\left[\boldsymbol{p}^\top\boldsymbol{y}(\tilde{s},\tilde{\boldsymbol{z}})\right] \\ \text{s.t.}\ & \boldsymbol{y}(s,\boldsymbol{z})\geq\boldsymbol{0} \qquad \forall \boldsymbol{z}\in\mathcal{Z}_s,\ s\in[S] \\ & \boldsymbol{y}(s,\boldsymbol{z})\geq \boldsymbol{w}-\boldsymbol{u} \qquad \forall \boldsymbol{z}\in\mathcal{Z}_s,\ s\in[S] \\ & \boldsymbol{c}^\top\boldsymbol{w} = d,\ \boldsymbol{w}\geq\boldsymbol{0} \\ & y_i\in\bar{\mathcal{A}}(\mathcal{C},\mathcal{I}) \qquad \forall i\in[I_u], \end{aligned} \tag{29}$$

where we control how the recourse decision $\boldsymbol{y}(\cdot,\cdot)$ adapts to $(\tilde{\boldsymbol{u}},\tilde{v})$ and $\tilde{s}$ through choosing the collection $\mathcal{C}$ of MECE events and the information index set $\mathcal{I}$.

We consider $I_u\in\{5,7\}$ and $S\in\{5,10,20,50\}$. The random demand belongs to a support set $\mathcal{U}=[\boldsymbol{0},\bar{\boldsymbol{u}}]$, and we use the Euclidean norm $\|\cdot\|_2$ as the distance metric. In each instance, we randomly generate the upper bound $\bar{\boldsymbol{u}}$ from a uniform distribution on $[0,100]^{I_u}$. Subsequently, past observations are randomly generated from the uniform distribution on $[\boldsymbol{0},\bar{\boldsymbol{u}}]$. We set $c_i=1$, $i\in[I_u]$ and $b=50I_u$, and we generate $\boldsymbol{p}$ from a uniform distribution on $[0,5]^{I_u}$. For different choices of θ, we run 100 random instances and compare the performance of different cases of event-wise recourse adaptation against the exact approach:

- case 1: $\boldsymbol{y}(\cdot,\cdot)$ adapts on $\tilde{\boldsymbol{u}},\tilde{v},\tilde{s}$, *i.e.*, $\mathcal{C}=\{\{s\}\mid s\in[S]\}$ and $\mathcal{I}=[I_u+1]$;
- case 2: $\boldsymbol{y}(\cdot,\cdot)$ adapts only on $\tilde{\boldsymbol{u}},\tilde{v}$, *i.e.*, $\mathcal{C}=\{\{1,\ldots,S\}\}$ and $\mathcal{I}=[I_u+1]$;
- case 3: $\boldsymbol{y}(\cdot,\cdot)$ adapts only on $\tilde{\boldsymbol{u}}$, *i.e.*, $\mathcal{C}=\{\{1,\ldots,S\}\}$ and $\mathcal{I}=[I_u]$.

Case 1 corresponds to the full event-wise affine adaptation. We turn off the event-wise adaptation in case 2, while in case 3 we further deprive the recourse decision $\boldsymbol{y}(\cdot,\cdot)$ of the affine adaptation on

		θ = 1			θ = 2			θ = 5			θ = 10			θ = 20		
S	5	< 0.1	2.0	8.7	< 0.1	3.4	8.8	0.1	5.4	9.2	0.2	7.0	9.9	0.5	9.4	10.5
S	10	< 0.1	4.5	11.6	< 0.1	6.4	11.9	0.2	8.6	12.3	0.3	11.0	12.7	1.0	12.6	13.0
S	20	< 0.1	4.8	5.7	< 0.1	5.3	5.8	< 0.1	5.6	6.1	0.1	6.4	6.5	0.3	7.9	7.9
S	50	< 0.1	6.8	7.9	< 0.1	7.4	8.0	< 0.1	8.1	8.4	0.1	9.2	9.2	0.3	11.3	11.3

Table 1 5 items: 90-th percentile optimality gaps (in %) of case 1 (left), case 2 (middle), and case 3 (right).

		θ = 1			θ = 2			θ = 5			θ = 10			θ = 20		
S	5	< 0.1	1.3	8.9	0.1	2.3	9.0	0.2	3.8	9.3	0.3	5.6	9.8	0.6	8.2	10.5
S	10	< 0.1	2.5	6.2	0.1	3.5	6.2	0.1	5.0	6.4	0.2	6.0	6.6	0.5	7.6	7.7
S	20	< 0.1	4.8	6.7	0.1	5.7	6.8	0.1	6.7	6.9	0.2	7.3	7.3	0.6	7.9	7.9
S	50	0.1	5.5	6.1	0.1	5.8	6.1	0.2	6.3	6.3	0.2	6.3	6.6	0.5	7.6	7.6

Table 2 7 items: 90-th percentile optimality gaps (in %) of case 1 (left), case 2 (middle), and case 3 (right).

the auxiliary random variable $\tilde{v}$. For each case, we consider the following relative gap between the objective value using the event-wise recourse adaptation and the exact optimal objective value:

$$\frac{\lambda^\star - \lambda}{\lambda^\star} \times 100\%.$$

Results for $I_u = 5$ and $I_u = 7$ are summarized in Table 1 and Table 2, respectively. With *(i)* the notion of event-wise adaptation and *(ii)* the inclusion of auxiliary random variable $\tilde{v}$, the full event-wise affine adaptation could provide a reasonably good conservative approximation to the exact approach; while excluding either *(i)* or *(ii)* may lead to a more conservative approximation.

We evaluate the scalability of the full event-wise affine adaptation (case 1), the affine adaptation without event-wise dependence (case 2), and the exact approach, by comparing their computation times and limits for different pairs of problem sizes. For the exact approach, the computer runs out of memory when the number of items exceeds 10 and the number of samples exceeds 5 (see Table 3), which is not practically favorable. In contrast, we are able to obtain a conservative solution via the full event-wise affine adaptation with modest computational effort. Quite interestingly, the event-wise adaptation that plays the key role in delivering the less conservative approximation seems to require only a little extra computational effort (see Table 4).

(S, I_u)						
(5, 5)	(10, 5)	(20, 5)	(50, 5)	(100, 5)	(200, 5)	(5, 10)
0.1	1.1	0.2	0.5	1.3	9.4	9.8

Table 3 Computation times of the exact approach. We report only those (S, I_u) pairs for which the exact approach were solved.

		I_u					
		5	10	15	20	25	30
S	5	< 0.1	< 0.1	0.2	0.5	0.7	1.1
	10	< 0.1	0.1	0.3	0.6	2.4	3.7
	20	< 0.1	0.3	0.7	1.3	2.2	3.8
	50	0.2	0.8	1.8	3.7	6.8	10.9
	100	0.4	1.8	5.2	10.3	17.8	31.0
	200	0.5	2.0	5.5	11.0	19.1	32.8
	300	0.8	3.1	7.8	16.5	32.3	49.6
	500	1.4	5.7	16.6	36.6	—	—
	800	2.5	14.8	—	—	—	—

		I_u					
		5	10	15	20	25	30
S	5	< 0.1	0.2	0.2	0.2	0.4	0.5
	10	< 0.1	< 0.1	0.2	0.3	0.7	2.3
	20	< 0.1	0.1	0.3	0.6	1.3	1.9
	50	0.1	0.4	0.8	1.9	3.0	4.4
	100	0.2	1.1	2.7	5.1	9.2	10.5
	200	0.4	1.7	5.2	15.8	30.7	25.3
	300	0.8	3.2	8.6	14.2	36.1	48.4
	500	1.1	5.3	13.8	—	—	—
	800	4.3	13.8	—	—	—	—

Table 4 Computation times and limits of the affine approximation without event-wise adaptation (left) and the full event-wise affine adaptation (right). The symbol "—" indicates "out of memory".

The following code segment shows how to implement the full event-wise affine adaptation for the multi-item newsvendor problem with Wasserstein ambiguity sets in RSOME.

```
% I: number of items
% S: number of past observations
% theta: radius
% Gamma: total budget
% cost (price): cost (price) parameters
% ubar: upper bound of demand
% U = (u_1, ..., u_S): past realizations

% Create RSOME model
model = rsome('newsvendor');

% Define random variables
u = model.random;                        % random demand
v = model.random;                        % auxiliary random variable
P = model.ambiguity(S);                  % create ambiguity set
```

```
% Define support sets for scenarios
for s = 1:S
    P(s).suppset(0 <= u, u <= ubar, norm(u - U(:,s)) <= v);
end

% Define probabilities for scenarios
pr = P.prob;
P.probset(pr == 1/S);

% Define event-wise expectation
P.exptset(expect(u) <= theta);

% Declare Warsserstein ambiguity set
model.with(P);

% Define decision variables
w = model.decision(I,1);
y = model.decision(I,1);

% Define event-wise adaptation
for s = 1:S
    y.evtadapt(s);
end

% Define affine adaptation
y.affadapt(u);
y.affadapt(v);

% Define objective function
model.min(-price'*w - expect(price'*y));

% Define constraints
model.append(y >= 0);
model.append(y >= w - u);
model.append(w >= 0);
model.append(cost'*w == Gamma);

% Solution
model.solve;
```

Experiment of Hanasusanto and Kuhn (2018)

We benchmark the RSO model against a state-of-art approximation scheme proposed by Hanasusanto and Kuhn (2018). Particularly, we repeat their experiment using the same set-ups.

Consider the second-stage problem of the form

$$f(\boldsymbol{u}) = \min\{\boldsymbol{e}^\top \boldsymbol{y} \mid \boldsymbol{y} \geq \boldsymbol{0}, \boldsymbol{y} \geq \boldsymbol{A}\boldsymbol{u} - \boldsymbol{b}\}, \tag{30}$$

where $\boldsymbol{e}$ is a vector of ones. The problem does not have any here-and-now decision $\boldsymbol{w}$ and assumes that the random variable $\tilde{\boldsymbol{u}}$ resides in a box $\mathcal{U} = [0,1]^{I_u}$. Under the distance metric $\rho(\boldsymbol{u}, \boldsymbol{u}^\dagger) =$

$\|\boldsymbol{u}-\boldsymbol{u}^\dagger\|_2^2$, Hanasusanto and Kuhn (2018) have shown that the worst-case expectation

$$\sup_{\mathbb{P}\in\mathcal{F}_W(\theta)} \mathbb{E}_{\mathbb{P}}\left[f(\tilde{\boldsymbol{u}})\right] \tag{31}$$

amounts exactly to the optimal value of the following copositive program.

$$\begin{array}{ll}
\inf & \displaystyle \frac{1}{S}\sum_{s\in[S]}\left(\alpha_s+\overline{\boldsymbol{q}}^\top\boldsymbol{\psi}_s-\beta\|\hat{\boldsymbol{u}}_s\|_2^2+\sum_{\ell\in[I_u+J_y]}\phi_{s\ell}\overline{q}_\ell^2\right)+\beta\theta^2 \\
\text{s.t.} & \begin{pmatrix} \beta\boldsymbol{I}+\overline{\boldsymbol{Q}}^\top\mathrm{diag}(\boldsymbol{\phi}_s)\overline{\boldsymbol{Q}} & -\frac{1}{2}\overline{\boldsymbol{T}}^\top-\overline{\boldsymbol{Q}}^\top\mathrm{diag}(\boldsymbol{\phi}_s)\overline{\boldsymbol{W}}^\top & -\beta\hat{\boldsymbol{u}}_s-\frac{1}{2}\overline{\boldsymbol{Q}}^\top\boldsymbol{\psi}_s \\ -\frac{1}{2}\overline{\boldsymbol{T}}-\overline{\boldsymbol{W}}\mathrm{diag}(\boldsymbol{\phi}_s)\overline{\boldsymbol{Q}} & \overline{\boldsymbol{W}}\mathrm{diag}(\boldsymbol{\phi}_s)\overline{\boldsymbol{W}}^\top & \frac{1}{2}(\overline{\boldsymbol{W}}\boldsymbol{\phi}_s-\overline{\boldsymbol{h}}) \\ -\beta\hat{\boldsymbol{u}}_s^\top-\frac{1}{2}\boldsymbol{\psi}_s^\top\overline{\boldsymbol{Q}} & \frac{1}{2}(\overline{\boldsymbol{W}}\boldsymbol{\phi}_s-\overline{\boldsymbol{h}})^\top & \alpha_s \end{pmatrix}\succeq_{K_{\mathrm{cop}}}\boldsymbol{O}\quad\forall s\in[S] \\
& \boldsymbol{\alpha}\in\mathbb{R}^S,\ \beta\in\mathbb{R}_+,\ \boldsymbol{\psi}_s,\boldsymbol{\phi}_s\in\mathbb{R}^{I_u+J_y}\qquad\qquad\forall s\in[S].
\end{array} \tag{32}$$

Here, $K_{\mathrm{cop}}=\{\boldsymbol{M}\in\mathbb{S}^K\mid\boldsymbol{x}^\top\boldsymbol{M}\boldsymbol{x}\geq 0\ \forall\boldsymbol{x}\geq\boldsymbol{0}\}$ is the copositive cone,

$$\overline{\boldsymbol{Q}}=\begin{pmatrix}\boldsymbol{O}\\ \boldsymbol{I}\end{pmatrix},\ \overline{\boldsymbol{q}}=\begin{pmatrix}\boldsymbol{e}\\ -\boldsymbol{e}\end{pmatrix},\ \overline{\boldsymbol{T}}=\begin{pmatrix}\boldsymbol{A}\\ \boldsymbol{O}\end{pmatrix},\ \overline{\boldsymbol{h}}=\begin{pmatrix}-\boldsymbol{b}\\ \boldsymbol{0}\end{pmatrix},\ \overline{\boldsymbol{W}}=\begin{pmatrix}\boldsymbol{W} & \boldsymbol{O}\\ \boldsymbol{O} & -\boldsymbol{I}\end{pmatrix}\ \text{with}\ \ \boldsymbol{W}=\begin{pmatrix}\boldsymbol{I}\\ \boldsymbol{I}\end{pmatrix},$$

and $\boldsymbol{O}$, $\boldsymbol{I}$, $\boldsymbol{0}$ and $\boldsymbol{e}_i$ respectively correspond to the zero matrix, the identity matrix, the zero vector and the i-th standard unit basis, all of which are of the appropriate dimension. Because the copostive program (32) is generally intractable, Hanasusanto and Kuhn (2018) adopt a conservative K_0-approximation by replacing the copostive cone K_{cop} with

$$K_0=\{\boldsymbol{M}\in\mathbb{S}^K\mid\boldsymbol{M}=\boldsymbol{P}+\boldsymbol{N},\boldsymbol{P}\succeq\boldsymbol{O},\boldsymbol{N}\geq\boldsymbol{O}\}\subseteq K_{\mathrm{cop}},$$

which leads problem (32) to a semidefinite program.

We run numerical tests for different pairs of the uncertainty dimension I_u and the sample size S, and for each pair, we use the same set-ups as in Hanasusanto and Kuhn (2018) to generate 100 random instances. The Wasserstein radius is set to $\theta=1/S$. The dimension J_y of the recourse decision is sampled uniformly at random from $\{1,2,\ldots,\lceil\log(I_u+1)\rceil\}$, $\boldsymbol{A}$ is sampled uniformly from $[0,1]^{J_y\times I_u}$, and $\boldsymbol{b}$ is sampled uniformly from $[0,\boldsymbol{e}^\top\boldsymbol{A}_{1:}]\times\cdots\times[0,\boldsymbol{e}^\top\boldsymbol{A}_{J_y:}]$. Here, $\boldsymbol{A}_{1:}$ stands for the first row of $\boldsymbol{A}$ and so forth. Lastly, we generate independent training samples from the uniform distribution on $[0,1]^{I_u}$. We evaluate the worst-case expectation (31) approximately by using *(i)* the K_0-approximation and *(ii)* the following full event-wise affine adaptation:

$$\begin{array}{lll}
\min & \displaystyle\sup_{\mathbb{P}\in\mathcal{F}_W(\theta)}\mathbb{E}_{\mathbb{P}}\left[\boldsymbol{e}^\top\boldsymbol{y}(\tilde{s},\tilde{\boldsymbol{z}})\right] & \\
\text{s.t.} & \boldsymbol{y}(s,\boldsymbol{z})\geq\boldsymbol{0} & \forall\boldsymbol{z}\in\mathcal{Z}_s,\ s\in[S] \\
& \boldsymbol{y}(s,\boldsymbol{z})\geq\boldsymbol{A}\boldsymbol{u}-\boldsymbol{b} & \forall\boldsymbol{z}\in\mathcal{Z}_s,\ s\in[S] \\
& y_j\in\bar{\mathcal{A}}(\{\{1\},\ldots,\{S\}\},[I_u+1]) & \forall j\in[J_y],
\end{array}$$

		I_u									
		1		2		4		8		16	
S	5	0.3	<0.1	0.3	<0.1	0.3	<0.1	0.5	<0.1	2.3	<0.1
	10	0.4	<0.1	0.4	<0.1	0.5	<0.1	0.9	<0.1	5.1	<0.1
	20	0.6	<0.1	0.7	<0.1	0.8	<0.1	1.8	<0.1	11.6	<0.1
	40	1.1	<0.1	1.3	<0.1	1.6	<0.1	3.5	<0.1	23.1	0.1
	80	2.3	<0.1	2.5	<0.1	3.2	<0.1	7.8	0.1	51.1	0.1
	160	4.5	<0.1	5.1	<0.1	7.0	0.1	18.0	0.2	118.0	0.3
	320	9.2	0.1	10.8	0.1	15.5	0.2	45.4	0.3	281.5	0.6
	640	19.7	0.1	26.9	0.2	43.9	0.3	141.5	1.0	684.3	2.3

Table 5 Computation times (in seconds) of K_0-approximation (left) and event-wise affine adaptation (right).

where for each $s \in [S]$, $\mathcal{Z}_s = \{(\boldsymbol{u}, v) \mid \boldsymbol{u} \in [0,1]^{I_u}, v \geq \|\boldsymbol{u} - \hat{\boldsymbol{u}}_s\|_2^2\}$.

Quite surprisingly, for all pairs of problem sizes, the solutions of both approximation approaches coincide for all 100 randomly generated instances. Unfortunately, we are not able to give a formal proof for this observation. Nevertheless, this observation supports that our proposed event-wise affine adaptation delivers solutions with competitive approximation quality as the state-of-the-art approximation scheme by Hanasusanto and Kuhn (2018). We report in Table 5 the average computation times of both approaches. In terms of computation efficiency, the event-wise affine adaptation outperforms because it leads to a second-order cone approximation to problem (32).

We note that the K_0-approximation by Hanasusanto and Kuhn (2018) also works when the cost vector of the second-stage problem (30) is affinely affected by the uncertainty, while our event-wise affine adaptation does not. On the other hand, the event-wise affine adaptation works with more general distance metrics and more general support sets that are not necessarily polyhedral (in the current experiment, the support set is a box $[0,1]^{I_u}$), while the K_0-approximation does not.

E. Multi-Stage Stochastic Financial Planning Problem

We adopt a financial planning problem from Birge and Louveaux (2011) to illustrate how to incorporate the scenario tree approach in the RSO framework.

At the beginning of the first stage in this multi-stage problem, the decision maker allocates the wealth d into two possible investment types, stocks (S) and bonds (B). Eight possible scenarios may occur, which corresponds to independent and equal likelihoods of having high returns of 1.25 for stocks and 1.14 for bonds, or low returns of 1.06 for stocks and 1.12 for bonds over subsequent

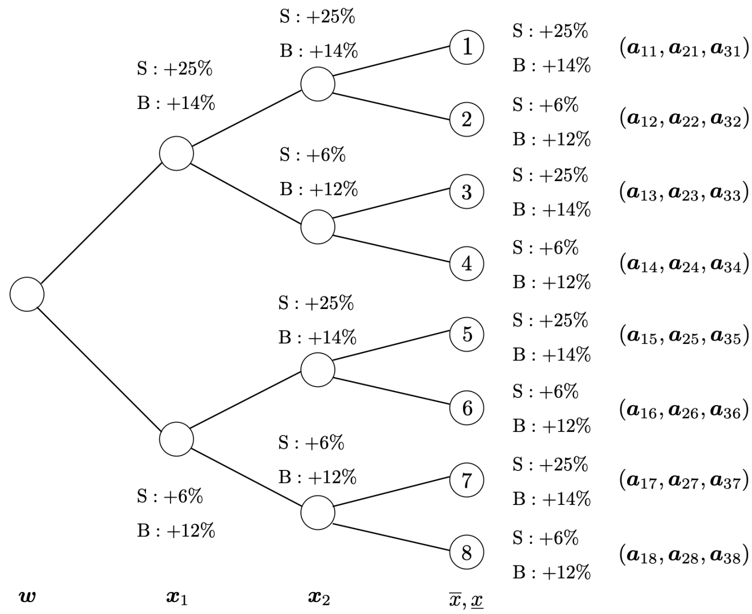

Figure 5 Scenario tree of the financial planning problem.

stages (see Figure 5). Hence we can construct the following singleton ambiguity set of the known discrete distribution of uncertain returns over all stages.

$$\mathcal{F} = \left\{ \mathbb{P} \in \mathcal{P}_0\left(\mathbb{R}^2 \times \mathbb{R}^2 \times \mathbb{R}^2 \times [S]\right) \;\middle|\; \begin{array}{ll} (\tilde{\boldsymbol{z}}_1, \tilde{\boldsymbol{z}}_2, \tilde{\boldsymbol{z}}_3, \tilde{s}) \sim \mathbb{P} & \\ \mathbb{P}\left[(\tilde{\boldsymbol{z}}_1, \tilde{\boldsymbol{z}}_2, \tilde{\boldsymbol{z}}_3) \in \mathcal{Z}_s \mid \tilde{s} = s\right] = 1 & \forall s \in [S] \\ \mathbb{P}\left[\tilde{s} = s\right] = \frac{1}{S} & \forall s \in [S] \end{array} \right\},$$

where $S = 8$ and the singleton support sets $\mathcal{Z}_s = \{(\boldsymbol{a}_{1s}, \boldsymbol{a}_{2s}, \boldsymbol{a}_{3s})\}, s \in [S]$ are determined by $\boldsymbol{a}_{1s} = (1.25, 1.14)$ for $s \in \{1,2,3,4\}$, $\boldsymbol{a}_{2s} = (1.25, 1.14)$ for $s \in \{1,2,5,6\}$, $\boldsymbol{a}_{3s} = (1.25, 1.14)$ for $s \in \{1,3,5,7\}$, and $\boldsymbol{a}_{is} = (1.06, 1.12)$ otherwise.

The decision maker evaluates the difference between the final return r and a prescribed target τ based on a concave and piecewise affine utility function that takes $U(r-\tau) = r - \tau$ if $r \geq \tau$ and $U(r-\tau) = 4(r-\tau)$ otherwise. The initial investment decisions $\boldsymbol{w}$, made before the first stage returns of stocks and bonds realize, must be indifferent among all eight scenarios. The rebalanced investment decison $\boldsymbol{x}_1$, made after the first stage returns realize but before the second stage returns realize, shall be indifferent among scenarios $\{1,2,3,4\}$ and indifferent among scenarios $\{5,6,7,8\}$. Similarly, $\boldsymbol{x}_2$ is indifferent between scenarios $\{1,2\}$ as well as between scenarios $\{3,4\}$, $\{5,6\}$, and $\{7,8\}$. Finally, the nonnegative auxiliary recourse decisions $\overline{x}$ and $\underline{x}$, respectively standing for the excess above or shortfall below the target, are adaptive to revealed uncertainties and thus can be different across the eight scenarios. In all, we can formulate the RSO model as follows:

$$
\begin{array}{lll}
\max & \inf_{\mathbb{P}\in\mathcal{F}} \mathbb{E}_{\mathbb{P}}\left[\overline{x}(\tilde{s}) - 4\underline{x}(\tilde{s})\right] & \\
\text{s.t.} & w_1, w_2 \geq 0,\ w_1 + w_2 = d & \\
& x_{11}(s) + x_{12}(s) - \boldsymbol{z}_1^\top \boldsymbol{w} = 0 & \forall \boldsymbol{z}\in\mathcal{Z}_s, s\in[S] \\
& x_{21}(s) + x_{22}(s) - \boldsymbol{z}_2^\top \boldsymbol{x}_1(s) = 0 & \forall \boldsymbol{z}\in\mathcal{Z}_s, s\in[S] \\
& \boldsymbol{z}_3^\top \boldsymbol{x}_2(s) - \overline{x}(s) + \underline{x}(s) = \tau & \forall \boldsymbol{z}\in\mathcal{Z}_s, s\in[S] \\
& x_{11}(s), x_{12}(s), x_{21}(s), x_{22}(s), \overline{x}(s), \underline{x}(s) \geq 0 & \forall s\in[S] \\
& x_{11}, x_{12} \in \mathcal{A}(\{\{1,2,3,4\},\{5,6,7,8\}\}) & \\
& x_{21}, x_{22} \in \mathcal{A}(\{\{1,2\},\{3,4\},\{5,6\},\{7,8\}\}) & \\
& \overline{x}, \underline{x} \in \mathcal{A}(\{\{1\},\{2\},\{3\},\{4\},\{5\},\{6\},\{7\},\{8\}\}). &
\end{array}
$$

F. Portfolio Management with K-means Adaptive Rebalancing

We consider a three-period portfolio allocation and rebalancing problem to minimize the investment risk at the last period taking into account of transaction costs. At the beginning of the first period, we decide the number of shares $w_i \geq 0$ of stock $i \in [I_u]$ to invest at price a_i, incurring a transaction cost $b_i w_i$. The price of stock i in the second period is $\tilde{a}_i^1 \triangleq a_i(\tilde{u}_i^1 + 1)$, where $\tilde{u}_i^1$ is the corresponding return. Subsequently, for each stock i, we rebalance its shares to $x_i \geq 0$, which incurs a transaction cost $b_i|x_i - w_i|$. In the last period, the price of stock i is $\tilde{a}_i^2 \triangleq a_i(\tilde{u}_i^2 + 1)$, where $\tilde{u}_i^2$ is the third period return with respect to the first period price. The effective portfolio return at the last period, taking into account of the total transaction costs, amounts to

$$
\boldsymbol{w}^\top \tilde{\boldsymbol{a}}^1 - \boldsymbol{w}^\top \boldsymbol{a} + \boldsymbol{x}^\top \tilde{\boldsymbol{a}}^2 - \boldsymbol{x}^\top \tilde{\boldsymbol{a}}^1 - \boldsymbol{b}^\top(\boldsymbol{w} + |\boldsymbol{x} - \boldsymbol{w}|) = \boldsymbol{w}^\top \boldsymbol{A}\tilde{\boldsymbol{u}}^1 + \boldsymbol{x}^\top \boldsymbol{A}(\tilde{\boldsymbol{u}}^2 - \tilde{\boldsymbol{u}}^1) - \boldsymbol{b}^\top(\boldsymbol{w} + |\boldsymbol{x} - \boldsymbol{w}|),
$$

where $\boldsymbol{A} = \operatorname{diag}(\boldsymbol{a})$ and the operator $|\cdot|$ takes the absolute value component-wise.

Ideally, the rebalancing decision $\boldsymbol{x}$ should only depend on the realization of $\tilde{\boldsymbol{u}}^1$. However, this would lead to an intractable problem. Instead, we propose an alternative K-means adaptive approach, where the recourse decision $\boldsymbol{x}(\tilde{s})$ depends on the random scenario $\tilde{s}$ that is associated with the realization of $\tilde{\boldsymbol{u}}^1$. In particular, using the available historical returns $\{(\hat{\boldsymbol{u}}_n^1, \hat{\boldsymbol{u}}_n^2)\}_{n\in[N]}$, we construct a two-layer K-means ambiguity set by first partitioning $\{\hat{\boldsymbol{u}}_n^1\}_{n\in[N]}$ into K_1 clusters, each of which we then further partition into K_2 clusters based on a subset of $\{\hat{\boldsymbol{u}}_n^2\}_{n\in[N]}$ that are affiliated with this specific first-layer cluster. As a result, we obtain a total number of $S = K_1 K_2$ scenarios, each of which corresponds to a unique cluster determined by the first and second layers; see an illustration in Figure 2. For each of the first-layer cluster $k \in [K_1]$, we denote by $\mathcal{E}_k \subseteq [S]$ as the set of scenarios associated with that cluster. Correspondingly, we define $\kappa(s) \in [K_1]$ as the specific

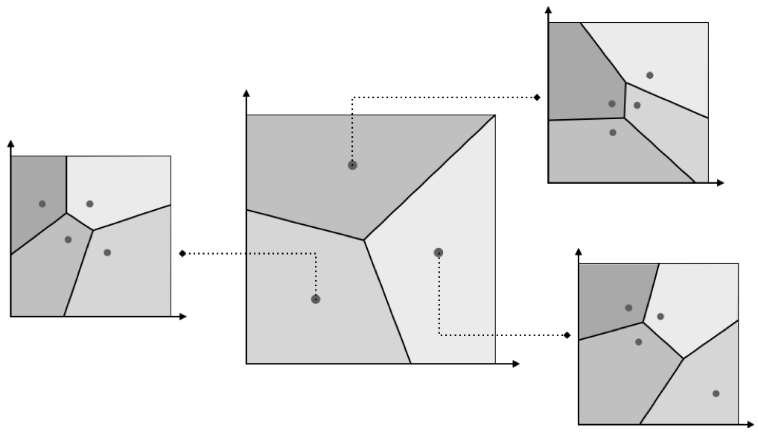

Figure 6 Two-layer K-means clustering. There are 3 clusters for $\{\hat{\boldsymbol{u}}_n^1\}_{n\in[N]}$ in the first layer, and a subset of $\{\hat{\boldsymbol{u}}_n^2\}_{n\in[N]}$ affiliated with each of these clusters is further partitioned into 4 clusters in the second layer. In total, we have 12 distinctive clusters.

first-layer cluster that the scenario s affiliates with. Observe that $\mathcal{C} = \{\mathcal{E}_k \mid k \in [K_1]\}$ is a collection of MECE events. In this way, we obtain the two-layer K-means ambiguity set

$$\mathcal{F} = \left\{ \mathbb{P} \in \mathcal{P}_0\left(\mathbb{R}^{2I_u+2I_v} \times [S]\right) \left| \begin{array}{ll} ((\tilde{\boldsymbol{u}}^1, \tilde{\boldsymbol{u}}^2, \tilde{\boldsymbol{v}}^1, \tilde{\boldsymbol{v}}^2), \tilde{s}) \sim \mathbb{P} & \\ \mathbb{E}_{\mathbb{P}}\left[\tilde{\boldsymbol{u}}^1 \mid \tilde{s} \in \mathcal{E}_k\right] = \hat{\boldsymbol{\mu}}_k^1 & \forall k \in [K_1] \\ \mathbb{E}_{\mathbb{P}}\left[\tilde{\boldsymbol{v}}^1 \mid \tilde{s} \in \mathcal{E}_k\right] \le \hat{\boldsymbol{\sigma}}_k^1 & \forall k \in [K_1] \\ \mathbb{E}_{\mathbb{P}}\left[\tilde{\boldsymbol{u}}^2 \mid \tilde{s} = s\right] = \hat{\boldsymbol{\mu}}_s^2 & \forall s \in [S] \\ \mathbb{E}_{\mathbb{P}}\left[\tilde{\boldsymbol{v}}^2 \mid \tilde{s} = s\right] \le \hat{\boldsymbol{\sigma}}_s^2 & \forall s \in [S] \\ \mathbb{P}\left[(\tilde{\boldsymbol{u}}^1, \tilde{\boldsymbol{u}}^2, \tilde{\boldsymbol{v}}^1, \tilde{\boldsymbol{v}}^2) \in \mathcal{Z}_s \mid \tilde{s} = s\right] = 1 & \forall s \in [S] \\ \mathbb{P}\left[\tilde{s} = s\right] = p_s & \forall s \in [S] \end{array} \right. \right\},$$

where for each $s \in [S]$, the cluster-wise support set is determined by

$$\mathcal{Z}_s = \{(\boldsymbol{u}^1, \boldsymbol{u}^2, \boldsymbol{v}^1, \boldsymbol{v}^2) \mid \boldsymbol{u}^1 \in \mathcal{U}_{\kappa(s)}^1, \boldsymbol{u}^2 \in \mathcal{U}_s^2, \boldsymbol{v}^1 \ge \boldsymbol{\phi}(\boldsymbol{u}^1), \boldsymbol{v}^2 \ge \boldsymbol{\phi}(\boldsymbol{u}^2)\}.$$

The objective function evaluates the worst-case conditional value-at-risk (CVaR) of the final return at a pre-specified risk threshold $\varepsilon \in (0,1)$.

$$\sup_{\mathbb{P}\in\mathcal{F}} \mathbb{P}\text{-CVaR}_{\varepsilon}\left(\boldsymbol{w}^\top \boldsymbol{A}\tilde{\boldsymbol{u}}^1 + \boldsymbol{x}^\top(\tilde{s})\boldsymbol{A}(\tilde{\boldsymbol{u}}^2 - \tilde{\boldsymbol{u}}^1) - \boldsymbol{b}^\top(\boldsymbol{w} + |\boldsymbol{x}(\tilde{s}) - \boldsymbol{w}|)\right),$$

which using now standard techniques, can be rewritten as

$$\min_{\delta}\ \delta + \frac{1}{\varepsilon}\sup_{\mathbb{P}\in\mathcal{F}} \mathbb{E}_{\mathbb{P}}\left[\left(\boldsymbol{x}^\top(\tilde{s})\boldsymbol{A}(\tilde{\boldsymbol{u}}^1 - \tilde{\boldsymbol{u}}^2) - \boldsymbol{w}^\top\boldsymbol{A}\tilde{\boldsymbol{u}}^1 + \boldsymbol{b}^\top(\boldsymbol{w} + |\boldsymbol{x}(\tilde{s}) - \boldsymbol{w}|) - \delta\right)^+\right].$$

Using Theorem 3, we formulate the RSO model for this portfolio optimization problem as follows:

$$
\begin{aligned}
\min\ & \delta + \frac{1}{\varepsilon}\sup_{\mathbb{P}\in\mathcal{F}} \mathbb{E}_{\mathbb{P}}\left[y(\tilde{s},\tilde{\boldsymbol{z}})\right] & \\
\text{s.t.}\ & y(s,\boldsymbol{z}) \geq 0 & \forall \boldsymbol{z}\in\mathcal{Z}_s,\ s\in[S] \\
& y(s,\boldsymbol{z}) \geq \boldsymbol{x}^\top(s)\boldsymbol{A}(\boldsymbol{u}^1-\boldsymbol{u}^2) - \boldsymbol{w}^\top\boldsymbol{A}\boldsymbol{u}^1 + \boldsymbol{b}^\top(\boldsymbol{w}+\bar{\boldsymbol{x}}(s)) - \delta & \forall \boldsymbol{z}\in\mathcal{Z}_s,\ s\in[S] \\
& \bar{\boldsymbol{x}}(s) \geq \boldsymbol{x}(s) - \boldsymbol{w} & \forall s\in[S] \\
& \bar{\boldsymbol{x}}(s) \geq \boldsymbol{w} - \boldsymbol{x}(s) & \forall s\in[S] \\
& \boldsymbol{b}^\top\bar{\boldsymbol{x}}(s) \leq \eta & \forall s\in[S] \\
& \boldsymbol{x}(s) \geq \boldsymbol{0} & \forall s\in[S] \\
& \boldsymbol{a}^\top\boldsymbol{w} = d & \\
& \boldsymbol{w} \geq \boldsymbol{0} & \\
& x_i, \bar{x}_i \in \mathcal{A}(\mathcal{C}) & \forall i\in[I_u] \\
& y \in \bar{\mathcal{A}}(\bar{\mathcal{C}}, [2I_u + 2I_v]) &
\end{aligned}
$$

where $\tilde{\boldsymbol{z}} \triangleq (\tilde{\boldsymbol{u}}^1, \tilde{\boldsymbol{u}}^2, \tilde{\boldsymbol{v}}^1, \tilde{\boldsymbol{v}}^2)$ and $\bar{\mathcal{C}} \triangleq \{\{s\} \mid s\in[S]\}$ consists of singleton MECE events. For simplicity, we impose a limit η on the transaction cost to prohibit over rebalancing in the second period, and d is the initial allocation budget. We next provide the sample code in RSOME to elucidate the intuitive implementation of the RSO model via an algebraic modeling language.

Sample Code for K-means Adaptive Rebalancing

We assume $\mathcal{E}_k = \{(k-1)K_2+1,\ldots,kK_2\} \subseteq [S]$ for all $k\in[K_1]$. Correspondingly, $\kappa(s) = \lceil \frac{s}{K_2} \rceil$ for all $s\in[S]$. We take a convex function $\boldsymbol{\phi}$ that specifies the mean absolute deviation of each random return within a particular cluster. Hence for each $s\in[S]$, the cluster-wise support set is given by

$$
\mathcal{Z}_s = \{(\boldsymbol{u}^1,\boldsymbol{u}^2,\boldsymbol{v}^1,\boldsymbol{v}^2) \mid \boldsymbol{D}^1_{\kappa(s)}\boldsymbol{u}^1 \leq \boldsymbol{f}^1_{\kappa(s)}, \boldsymbol{D}^2_s\boldsymbol{u}^2 \leq \boldsymbol{f}^2_s, \boldsymbol{v}^1 \geq |\boldsymbol{u}^1 - \hat{\boldsymbol{\mu}}^1_{\kappa(s)}|, \boldsymbol{v}^2 \geq |\boldsymbol{u}^2 - \hat{\boldsymbol{\mu}}^2_s|\},
$$

where each cluster is in fact a polyhedron and where $|\cdot|$ applies component-wise. The estimates $\{\boldsymbol{D}^1_k\}_{k\in[K_1]}$, $\{\boldsymbol{f}^1_k\}_{k\in[K_1]}$, $\{\hat{\boldsymbol{\mu}}^1_k\}_{k\in[K_1]}$, $\{\hat{\boldsymbol{\sigma}}^1_k\}_{k\in[K_1]}$ are contained in MATLAB cells D1, f1, mu1, sigma1, and similarly, $\{\boldsymbol{D}^2_s\}_{s\in[S]}$, $\{\boldsymbol{f}^2_s\}_{s\in[S]}$, $\{\hat{\boldsymbol{\mu}}^2_s\}_{s\in[S]}$, $\{\hat{\boldsymbol{\sigma}}^2_s\}_{s\in[S]}$ are contained in D2, f2, mu2, sigma2.

```
% I: number of stocks
% K1: number of first-layer clusters
% K2: number of second-layer clusters
% ps: probabilities of clusters
% epsilon: risk threshold
% a,b,d,eta: parameters

% Create RSOME model
model = rsome('portfolio');

% Define random variables
```

```
u = model.random(I,2);                     % random demand
v = model.random(I,2);                     % auxiliary random variable
P = model.ambiguity(K1*K2);                % create ambiguity set

% Define support sets for scenarios
for s = 1:K1*K2
    P(s).suppset(D1{ceil(s/K2)}*u(:,1) <= f1{ceil(s/K2)}, ...
                 D2{s}*u(:,2) <= f2{s}, ...
                 v(:,1) >= abs(u(:,1) - mu1{ceil(s/K2)}), ...
                 v(:,2) >= abs(u(:,2) - mu2{s}));
end

% Define probabilities for scenarios
pr = P.prob;
P.probset(pr == ps);

% Define event-wise expectation
for k = 1:K1
    P((k-1)*K2+1:k*K2).exptset(expect(u(:,1)) == mu1{k}, ...
                               expect(v(:,1)) <= sigma1{k});
end
for s = 1:K1*K2
    P(s).exptset(expect(u(:,2)) == mu2{s}, ...
                 expect(v(:,2)) <= sigma2{s});
end

% Declare K-means ambiguity set
model.with(P);

% Define decision variables
w = model.decision(I,1);
x = model.decision(I,1);
xbar = model.decision(I,1);
y = model.decision;
delta = model.decision;

% Define event-wise adaptation
for k = 1:K1
    x.evtadapt((k-1)*K2+1:k*K2);
    xbar.evtadapt((k-1)*K2+1:k*K2);
end
for s = 1:K1*K2
    y.evtadapt(s);
end

% Define affine adaptation
y.affadapt(u);
y.affadapt(v);

% Define objective function
model.min(delta + expect((1/epsilon)*y));
```

```
% Define constraints
model.append(y >= 0);
model.append(y >= x'*diag(a)*(u(:, 1)- u(:, 2)) ...
            - w'*diag(a)*u(:, 1) + b'*(w + xbar) - delta);
model.append(xbar >= abs(x - w));
model.append(b'*xbar <= eta);
model.append(x >= 0);
model.append(a'*w == d);
model.append(w >= 0);

% Solution
model.solve;
```

G. Endnote

All mathematical programs in numerical experiments are solved using MOSEK on an Intel Core (TM) @ 3.40 GHz with 8GB RAM. The semidefinite program related to the K_0-approximation is implemented using the CVX interface (Grant and Boyd 2008), while remaining models are implemented using our developped algebraic modeling package RSOME (available at `https://www.rsomerso.com/`).