Bounds for Multistage Mixed-Integer Distributionally Robust Optimization

Güzin Bayraksan∗ Francesca Maggioni† Daniel Faccini‡ Ming Yang§

February 7, 2022

Abstract: Multistage mixed-integer distributionally robust optimization (DRO) forms a class of extremely challenging problems since their size grows exponentially with the number of stages. One way to model the uncertainty in multistage DRO is by creating sets of conditional distributions (the so-called conditional ambiguity sets) on a finite scenario tree and requiring that such distributions remain close to nominal conditional distributions according to some measure of similarity/distance (e.g., φ-divergences or Wasserstein distance). In this paper, new bounding criteria for this class of difficult decision problems are provided through scenario grouping using the ambiguity sets associated with various commonly used φ-divergences and the Wasserstein distance. Our approach does not require any special problem structure such as linearity, convexity, stagewise independence, and so forth. Therefore, while we focus on multistage mixed-integer DRO, our bounds can be applied to a wide range of DRO problems including two-stage and multistage, with or without integer variables, convex or nonconvex, and nested or non-nested formulations. Numerical results on a multistage mixed-integer production problem show the efficiency of the proposed approach through different choices of partition strategies, ambiguity sets, and levels of robustness.

Keywords: Multistage Distributionally Robust Optimization, Bounding, Dynamic Measures of Risk, Phi-Divergences, Wasserstein Distance.

∗Bayraksan.1@osu.edu
Department of Integrated Systems Engineering, The Ohio State University, USA
†Francesca.Maggioni1@unibg.it
Department of Management, Information and Production Engineering, University of Bergamo, Italy
‡Daniel.Faccini@unibg.it
Department of Management, Information and Production Engineering, University of Bergamo, Italy
§Yang.3149@osu.edu
Department of Integrated Systems Engineering, The Ohio State University, USA
1 Introduction

Multistage stochastic programming has been widely used to solve important problems arising in various fields including finance [9], transportation [5], energy [26, 42] and the environment [44], among others. Despite their wide applicability, this class of problems suffers from two main issues. First, traditional models assume that the underlying stochastic process that governs the uncertain parameters is known. This is rarely true in real life. Second, multistage stochastic programs—particularly those involving mixed-integer variables and nonlinear terms—are notoriously difficult to solve. To alleviate the first issue, Distributionally Robust Optimization (DRO) can be used, where the assumed-known distribution is replaced by an ambiguity set of distributions [40]. Unfortunately, the resulting multistage problem is still extremely challenging to solve due to the exponential growth of the problem in the number stages, and can become even more challenging depending on the type of DRO used. Furthermore, many real-world problems lack special structures (e.g., convexity, stagewise independence, binary state variables) that prevent efficient solution algorithms. Therefore, approximation techniques that provide bounds on the optimal value for multistage DRO problems can be very useful in practice. In these situations, easy-to-compute bounds and approximations are desirable.

This paper investigates easy-to-compute lower bounds (LBs) (for minimization problems) through scenario grouping and convolution of risk measures for a class of multistage DRO formed using $\phi$-divergences and Wasserstein distance. Because obtaining LBs for minimization problems is typically more challenging than obtaining upper bounds (UBs), majority of the paper focuses on LBs. UBs, which are briefly discussed, are calculated by fixing some decisions obtained through the LBs, finding a feasible policy, and using the cost of that policy; see, e.g., [20, 23].

Most of the existing literature on DRO focuses on static, two-stage, or chance-constrained settings [33], and there is relatively little work on multistage DRO. Many of these works investigate different ways of forming ambiguity sets in the multistage setting, which can be more complicated relative to the static/two-stage setting [29, 40]. Moment-based [7, 41, 43], nested Wasserstein [27], modified $\chi^2$ distance [28], general $\phi$-divergences [25], $L_\infty$-norm [15], Wasserstein [10] and $\infty$-Wasserstein distance [6] have been examined to form multistage DRO. Most papers assume linear models with continuous decision variables, except for [6, 43], which consider mixed-integer decision variables. Our bounds do not assume any problem structure such as linearity, convexity, and continuity. Majority of
the existing works also focus on solution methods through nested Benders’ decomposition or its sampling-based variant, stochastic dual dynamic programming [10, 15, 25, 28, 43]. Linear decision rules have also been used to approximately solve these problems [6, 7].

The approach presented in this paper divides the sample space into subgroups. The subgroups, being of smaller size, can be solved more efficiently. They can be solved either by the traditional expected-value objective approach or a DRO approach. Then, the optimal values of subgroups can be combined, e.g., in a distributionally robust manner, to form LBs on the optimal value. Turns out, not all combinations of subgroups’ optimal values yield LBs. We provide conditions on ways to combine the optimal values of the subgroups to obtain LBs for ambiguity sets formed via many commonly used $\phi$-divergences and Wasserstein distance. The Wasserstein setting is more complicated because it requires an appropriate distance between subgroups. We define such a distance between subgroups to ensure LBs, and discuss how to apply these bounds in the multistage setting.

**Related work.** Bounding techniques have a rich history in the stochastic programming literature, and these have been successfully applied to traditional multistage stochastic programs with expected-value objectives. For instance, [11] considers two-stage bounds-based distributional approximations for multistage stochastic linear programs (i.e., moment-based approximations derived as solutions to certain generalized moment problems), relaxing the nonanticipativity constraints. Nonanticipativity is regained progressively via a disaggregation procedure. In [12], the authors propose tight UBs and LBs to stochastic convex programs with random right-hand sides. Using a constraint aggregation procedure, a group of stages from the end of the multistage stochastic program are aggregated to form a single stage, and error bounds are developed. In [18], the author elaborates an approximation scheme that integrates stage-aggregation and discretization through coarsening of sigma-algebras to ensure computational tractability, while providing deterministic error bounds.

Bounds for multistage stochastic linear programs via scenario tree decomposition were proposed for the first time in [19], by solving pair subproblems, measuring the quality of the deterministic solution, and introducing rolling horizon measures. In [20], the authors extend the bounding approach of [8, 19, 38] for stochastic multistage mixed-integer linear programs, solving a sequence of group subproblems made by a subset of reference scenarios plus a subset of scenarios from the finite support. They show the monotonicity of the chains of LBs in terms of the cardinality of reference scenarios and of the remaining
scenarios in each subgroup. A scalable bounding framework for general multistage stochastic programs, extending the work of [38], has been investigated in [39]. This framework scales well with problem size and obtains high-quality solutions within a reasonable time frame. Recently, [2] introduced sampling scenario set partition dual bounds for multistage stochastic programs.

An alternative approach to bound the original multistage stochastic program is to construct two approximating trees, a lower tree and an upper tree, the solutions of which lead to UBs and LBs for the optimal value of the original continuous problem. Results in this direction were first obtained by [14], followed by [13, 17]. In [17], barycentric discretizations are adopted in a more general setting for convex multistage stochastic programs with a generalized non-convex dependence on the random variables. In [22], the authors generalize the bounding ideas of [13, 14, 17] to not necessarily Markovian scenario processes and derive valid LBs and UBs for the convex case. They construct new discrete probability measures directly from the simulated data of the whole scenario process based on the concepts of first order and convex order stochastic dominance.

Bounds for risk-averse multistage mixed-integer stochastic programs via scenario tree decomposition were first proposed by [21] and [23]. In particular, [21] considers multistage convex problems with concave risk functional applied to the total cost over the planning horizon. New refinement chains of LBs are constructed, where each bound can be computed by solving sets of group subproblems less complex than the original one, and recalculating the probabilities of each scenario in the group accordingly. A monotonically nondecreasing behavior in the cardinality of scenarios of each subproblem is proved. LBs for replacing the scenario process by its expectation are also considered as well as UBs based on inserting feasible solutions derived from smaller subproblems. In [23], the authors consider a dynamic risk functional in the objective function, formed by a convex combination of mean and Conditional Value-at-Risk (mean-CVaR). LBs by using convolution of mean-CVaRs with different parameters are obtained through various scenario partition strategies, and a solution algorithm for mean-CVaR multistage mixed-integer stochastic problems is provided; see also [24] for algorithmic use of these bounds.

**Summary of contributions and outline.** This paper, to the best of our knowledge for the first time, introduces new LB criteria for multistage DRO through scenario tree decomposition. UBs are also examined. Our work is similar in spirit to [21, 23], but we consider a large class of DRO formed on finite scenario trees, where the ambiguity
sets are constructed using a $\phi$-divergence (e.g., variation distance, Cressie-Read power divergence, $J$-divergence, etc.) or Wasserstein distance on a finite support [10]. We provide conditions on how the optimal values of subgroup problems can be combined to yield LBs. As mentioned earlier, our results do not require any structural properties, and thus they are applicable to a broad class of problems including two- and multistage, with or without integer variables, nested vs. non-nested formulations. Therefore, after recalling background information in Section 2, in Section 3, we first present our results in the two-stage setting and then discuss how to apply these LBs in the multistage setting. We then investigate the effectiveness of the proposed bounds on a multistage mixed-integer production planning problem in Section 4, discussing the insights gained from these experiments. Section 5 concludes the paper and outlines future research directions. We end by noting that the proposed approach has the important advantage to split a given problem into independent scenario groups. This allows to tackle problems for which simple linear relaxations leave large optimality gaps, problems lacking special structure like convexity, stagewise independence, etc. that prevent efficient solution methods, and large-scale multistage problems that are not solvable by commercial solvers.

2 Basic facts and notation

2.1 Multistage DRO

We consider a finite-horizon sequential decision making problem under uncertainty, where decisions are made at discrete stages $t \in \mathcal{T} := \{0, 1, \ldots, T\}$ and $T$ denotes the planning horizon. The decision process begins with initial decision $x_0 \in \mathbb{R}^{n_0}_+ \times \mathbb{Z}^{n_0}_+$ at stage $t = 0$, called the first-stage or here-and-now decision, followed by sequential decisions $x_t \in \mathbb{R}^{n_t}_+ \times \mathbb{Z}^{n_t}_+$ at stages $t \in \mathcal{T} \setminus \{0\}$. The history of the decision process, at a given point in time, is denoted by $x^t := (x_0, x_1, \ldots, x_t)$, $t \in \mathcal{T}$. The uncertainty is described by a random process $\xi := \{\xi_0, \xi_1, \ldots, \xi_T\} \in \mathbb{R}^{d_0} \times \ldots \times \mathbb{R}^{d_T}$ defined on a measurable space $(\Omega, \mathcal{F})$. We assume $\xi_0$ is a constant and $\xi$ is a random parameter evolving as a discrete-time stochastic process with finite support. The filtration associated with $\xi$ is denoted by $\{\emptyset, \Omega\} = \mathcal{F}_0 \subseteq \mathcal{F}_1 \subseteq \ldots \subseteq \mathcal{F}_T = \mathcal{F}$, where $\mathcal{F}_t, t \in \mathcal{T}$—a $\sigma$-subalgebra of $\mathcal{F}$—models the information available so far. Each $\xi_t, t \in \mathcal{T}$, has finite support $\Omega_t \in \mathbb{R}^{d_t}$ and nominal distribution $Q_t$. Support of $\xi$ is given by $\Omega = \bigcap_{t=0}^T \Omega_t$, and the history of the random process up to stage $t$ is denoted by $\xi^t := (\xi_0, \ldots, \xi_t)$, $t \in \mathcal{T}$. 

5
The following represents the nested formulation of a multistage DRO (see [40]):

\[
\min_{x_0 \in X_0(\xi_0)} c_0(x_0, \xi_0) + \max_{P_1 | \xi^0 \in P_1} \mathbb{E}_{P_1 | \xi^0} \left[ \min_{x_1 \in \lambda_1(x_0, \xi_1)} c_1(x_1, \xi_1) + \max_{P_2 | \xi^1 \in P_2} \mathbb{E}_{P_2 | \xi^1} \left[ \ldots + \max_{P_T | \xi^{T-1} \in P_T} \mathbb{E}_{P_T | \xi^{T-1}} \left[ \min_{x_T \in \lambda_T(x_{T-1}, \xi_T)} c_T(x_T, \xi_T) \ldots \right] \right] \right],
\]

(1)

where the mixed-integer first-stage feasibility set is given by \( x_0^0 \subseteq \mathbb{R}^{n_0} \times \mathbb{Z}^{n_0}_+ \) and, for \( t \in T \setminus \{0\} \), \( \lambda_t : \mathbb{R}^{n_t-1} \times \mathbb{Z}^{n_t}_+ \times \mathbb{R} \rightarrow \mathbb{R}^{n_t} \times \mathbb{Z}^{n_t}_+ \) are \( F_t \)-measurable mixed-integer point-to-set mappings. The possibly nonlinear cost functions are given by \( c_0 : \mathbb{R}^{n_0} \times \mathbb{Z}^{n_0}_+ \rightarrow \mathbb{R} \) in the first stage and by \( c_t : \mathbb{R}^{n_t} \times \mathbb{Z}^{n_t}_+ \times \mathbb{R} \rightarrow \mathbb{R} \) in stages \( t \in T \setminus \{0\} \), which are \( F_t \)-measurable. We assume all relevant optimization problems in the paper have finite optimal solutions. Set \( P_{t|\xi^{t-1}} \) denotes the conditional ambiguity set at period \( t \in T \setminus \{0\} \), conditioned on the history \( \xi^{t-1} \), and it is defined as

\[
P_{t|\xi^{t-1}} := \left\{ P_{t|\xi^{t-1}} \in \mathcal{M}(\Omega_t|\xi^{t-1}) : \Delta(P_{t|\xi^{t-1}}, Q_{t|\xi^{t-1}}) \leq \rho_t \right\},
\]

(2)

where \( \rho_t \geq 0 \) is a given radius, also called the level of robustness. Above, \( \mathcal{M}(\Omega_t|\xi^{t-1}) \) represents a class of probability distributions defined on the finite support \( \Omega_t \times \mathbb{R}^{n_t}_+ \), \( Q_{t|\xi^{t-1}} \) denotes the nominal conditional probability measure at stage \( t \), \( t \in T \setminus \{0\} \), conditioned on the history of the process \( \xi^{t-1} \), and \( \Delta(\cdot, \cdot) \) denotes a measure of similarity or distance between \( P_{t|\xi^{t-1}} \) and \( Q_{t|\xi^{t-1}} \). We are interested in building ambiguity sets using existing data via \( \phi \)-divergences and Wasserstein distance, which we recall in the next sections.

Before we do so, let us define notation that is used throughout the paper.

**Scenario tree and nominal probability notation.** Because we assume \( \xi \) has finite support, the information structure can be described in the form of a scenario tree \( \Sigma \) with \( T+1 \) levels (stages). Let \( \Omega_t \) be the set of ordered nodes of the tree \( \Sigma \) at stage \( t \in T \) and let \( \Omega := \Omega_1 \times \ldots \times \Omega_T \). By assumption, we have a discrete number \( |\Omega_t| \) of nodes at each stage \( t \in T \). Each stage-\( t \) node \( n \) is connected to a unique node at stage \( t-1 \), called ancestor and denoted \( a(n) \). Similarly, each stage-\( t \) node \( n \) is connected to nodes at stage \( t+1 \) called successors or children, where \( B(n) \) denotes the set of children nodes of \( n \). With \( q_{a(n),n} \) we denote the conditional nominal probability of the random process at node \( n \) given its history up to the ancestor node \( a(n) \). A scenario \( \omega_i \), \( i = 1, \ldots, |\Omega_T| \) is a path through nodes from the root node at \( t = 0 \) to a leaf node at \( t = T \). We indicate with \( q_{\omega_i} \) the probability of a scenario \( \omega_i \) passing through nodes \( n_0, n_1, \ldots, n_T \) (where \( n_T \).
\[ t = 0, \ldots, T \] represent generic nodes at stage \( t \), defined as \( q_{\omega_i} := q_{n_0,n_1} \cdot q_{n_1,n_2} \cdots q_{n_{T-1},n_T} \).

We also indicate with \( q^n_n \) the nominal probability of node \( n \) at stage \( t \). So, if node \( n \) at stage \( t \) is reachable through node \( n_0 \) at stage 0, node \( n_1 \) at stage 1, \ldots, node \( n_{t-1} \) at stage \( t-1 \), then \( q^n_n := q_{n_0,n_1} \cdot q_{n_1,n_2} \cdots q_{n_{T-1},n_T} \). Moreover, \( \sum_{n \in \Omega_t} q^n_n = 1, \ t \in \mathcal{T} \) and \( \sum_{m \in \mathcal{B}(n)} q_{n,m} = 1, \ n \in \Omega_t, \ t = 0, \ldots, T-1. \)

**Set notation.** We use shorthand notation \([m]\) to denote the set \( \{1,2,\ldots,m\} \).

For simplicity, from here until Section 3.5, we consider two-stage DRO and only point to changes for multistage case. That is, we set \( T = 1 \) in (1), drop \( \xi_0 \) as it is a constant, and let \( \xi_1 \equiv \xi \) be defined on a probability space \((\Omega, \mathcal{F}, Q)\) with sample space \( \Omega := \{\omega_1,\omega_2,\ldots,\omega_{|\Omega|}\} \), \( \sigma \)-algebra \( \mathcal{F} \) and nominal probability \( Q \). The probability of scenario \( \omega_i \in \Omega \) can be specified as \( q_{\omega_i} \geq 0 \) with \( \sum_{i \in \Omega} q_{\omega_i} = 1 \). Similarly, we simply use \( P \) with probability of scenario \( \omega_i \) as \( p_{\omega_i} \geq 0 \) to define ambiguity set (2). So, (2) becomes

\[ \mathcal{P} = \{ P : \Delta(P, Q) \leq \rho, \sum_{i \in \Omega} p_{\omega_i} = 1, p_{\omega_i} \geq 0, \forall \omega_i \in \Omega \}. \]

**2.2 \( \phi \)-divergences**

For this class of ambiguity sets, \( \Delta \) in (2) is given by

\[ \Delta_{\phi}(P, Q) := \sum_{i \in \Omega} q_{\omega_i} \phi \left( \frac{p_{\omega_i}}{q_{\omega_i}} \right), \]

where the convex \( \phi \)-divergence function \( \phi(u) \geq 0 \) takes value 0 when both \( p_{\omega_i}, q_{\omega_i} > 0 \) have the same value; i.e., \( \phi(1) = 0 \). When \( q_{\omega_i} = 0 \), it holds that \( 0 \cdot \phi(p_{\omega_i}/0) = p_{\omega_i} \lim_{u \to \infty} (\phi(u)/u) \) and \( 0 \cdot \phi(0/0) = 0 \). Accordingly, ambiguity set \( \mathcal{P}_{t|t-1} \) in (2) can be built using some of the well-known \( \phi \)-divergences described in Table 1 and Table 2. These include Variation Distance (VD) and \( J \)-divergence, along with two families of \( \phi \)-divergences, namely, the Cressie-Read (CR) power divergence family and the \( \chi \)-divergence family of order \( a > 1 \). CR power divergence family includes some of the most widely used \( \phi \)-divergences as a special case—e.g., the modified \( \chi^2 \) distance and the Kullback-Leibler divergence—when its parameter \( \theta \) takes specific values or when the limit of \( \theta \) tends to 0 or 1. These special cases are listed in Table 2. Equivalence of the well-known divergences in Table 2 and the CR power divergence family in Table 1 is achieved when the radius \( \rho_t \) in the ambiguity set (2) formed via a divergence in Table 2 is set to an adjusted value \( c \cdot \rho_t^{\theta_{CR}} \), where \( \rho_t^{\theta_{CR}} \) is the radius of the CR divergence in Table 1. Values of coefficient \( c \) corresponding to certain \( \theta \) are listed in the last column of Table 2. For example, when
the radius of the modified $\chi^2$ distance, denoted $\rho_{t,m\chi^2}$, equals $2 \cdot \rho_{t,CR}^{\theta=2}$, where $\rho_{t,CR}^{\theta=2}$ represents the radius formed via CR power divergence with $\theta = 2$, the two ambiguity sets are equivalent.

<table>
<thead>
<tr>
<th>Divergence</th>
<th>$\phi(u)$</th>
<th>$\phi(u), u \geq 0$</th>
<th>$\Delta_\phi(P, Q)$</th>
</tr>
</thead>
<tbody>
<tr>
<td>Variation Distance</td>
<td>$\phi_v$</td>
<td>$</td>
<td>u - 1</td>
</tr>
<tr>
<td>Cressie-Read Power Divergence</td>
<td>$\phi_{CR}^{\theta}$</td>
<td>$\frac{1-\theta+q_{\omega_i}-u_{\omega_i}}{\theta(1-\theta)}$, $\theta \neq 0, 1$</td>
<td>$\frac{1-\sum p_{\omega_i}^\theta q_{\omega_i}^{1-\theta}}{\theta(1-\theta)}$, $\theta \neq 0, 1$</td>
</tr>
<tr>
<td>J-Divergence</td>
<td>$\phi_J$</td>
<td>$(u - 1) \log u$</td>
<td>$\sum (p_{\omega_i} - q_{\omega_i}) \log \left(\frac{p_{\omega_i}}{q_{\omega_i}}\right)$</td>
</tr>
<tr>
<td>$\chi$-Divergence of order $a &gt; 1$</td>
<td>$\phi_{\chi}^a$</td>
<td>$</td>
<td>u - 1</td>
</tr>
</tbody>
</table>

Table 1: Common $\phi$-divergences.

<table>
<thead>
<tr>
<th>$\theta$</th>
<th>Corresponding Divergence</th>
<th>$\phi(u)$</th>
<th>$\phi(u), u \geq 0$</th>
<th>$\Delta_\phi(P, Q)$</th>
<th>$\phi_{CR}^{\theta}(u)$</th>
<th>$c$</th>
</tr>
</thead>
<tbody>
<tr>
<td>2</td>
<td>Modified $\chi^2$ Distance</td>
<td>$\phi_{m\chi^2}$</td>
<td>$(u - 1)^2$</td>
<td>$\sum \left</td>
<td>\frac{p_{\omega_i} - q_{\omega_i}}{p_{\omega_i}}\right</td>
<td>^2$</td>
</tr>
<tr>
<td>$\frac{1}{2}$</td>
<td>Hellinger Distance</td>
<td>$\phi_H$</td>
<td>$(\sqrt{u} - 1)^2$</td>
<td>$\sum (\sqrt{p_{\omega_i}} - \sqrt{q_{\omega_i}})^2$</td>
<td>$4(\frac{1}{2} + \frac{1}{2} - \sqrt{u}) = 2(1 - \sqrt{u})^2$</td>
<td>$\frac{1}{2}$</td>
</tr>
<tr>
<td>-1</td>
<td>$\chi^2$ Distance</td>
<td>$\phi_{\chi^2}$</td>
<td>$\frac{1}{2}(u - 1)^2$</td>
<td>$\sum \left</td>
<td>\frac{p_{\omega_i} - q_{\omega_i}}{p_{\omega_i}}\right</td>
<td>^2$</td>
</tr>
<tr>
<td>$\to 1$</td>
<td>Kullback-Leibler Divergence</td>
<td>$\phi_{KL}$</td>
<td>$u \log u - u + 1$</td>
<td>$\sum p_{\omega_i} \log \left(\frac{p_{\omega_i}}{q_{\omega_i}}\right)$</td>
<td>$u(\log u - 1) + 1$</td>
<td>1</td>
</tr>
<tr>
<td>$\to 0$</td>
<td>Burg Entropy</td>
<td>$\phi_B$</td>
<td>$-\log u + u - 1$</td>
<td>$\sum q_{\omega_i} \log \left(\frac{p_{\omega_i}}{q_{\omega_i}}\right)$</td>
<td>$-\log u + u - 1$</td>
<td>1</td>
</tr>
</tbody>
</table>

Table 2: Some special cases of CR power divergence family. Kullback-Leibler divergence and Burg entropy are obtained by taking the limit of $\theta$ to 1 and 0, respectively.

2.3 Wasserstein distance

Let $\eta$ be a random variable on $(\Omega, F, Q)$ taking values $(\eta_{\omega_1}, \ldots, \eta_{|\Omega|})$. We quantify distributions $P$ close to nominal distribution $Q$ on the same support via Wasserstein distance (see [10]), where $\Delta$ in (2) is defined by

$$\Delta_W(P, Q) := \min_{\varepsilon \geq 0} \left\{ \sum_{\omega_i \in \Omega} \sum_{\omega_j \in \Omega} d_{\omega_i, \omega_j} z_{\omega_i, \omega_j} : \sum_{\omega_i \in \Omega} z_{\omega_i, \omega_j} = q_{\omega_j}, \forall j \in |\Omega|, \sum_{\omega_i \in \Omega} z_{\omega_i, \omega_j} = p_{\omega_i}, \forall i \in |\Omega| \right\}$$

with $d_{\omega_i, \omega_j} := ||\eta_{\omega_i} - \eta_{\omega_j}||_\varsigma$ a distance between the two scenarios $\omega_i$ and $\omega_j$ using $\varsigma$-norm (e.g., $\varsigma \in \{1, 2, \infty\}$). Ambiguity set $\mathcal{P}_{t\varsigma^{t-1}}$ in (2) can be built accordingly.

2.4 Relation to risk-averse optimization

Because the ambiguity sets considered in this paper are compact convex subsets of (conditional) probability measures and optimal values are assumed to be real-valued, DRO
is equivalent to Risk-Averse Stochastic Optimization (RASO) with the objective function expressed by a coherent risk measure; see e.g., [3, 34, 40].

Let us recall coherent risk measures. Let $Z := L_\infty(\Omega, \mathcal{F}, Q)$ be the space of bounded and $\mathcal{F}$-measurable random variables with respect to sample space $\Omega$ and probability distribution $Q$, and let $\eta \in Z$ be a random variable taking values $(\eta_{\omega_1}, \ldots, \eta_{\omega_{|\Omega|}})$. First defined by [1], a function $R(\eta) : Z \to \mathbb{R}$ is called a coherent measure of risk if it satisfies the following properties:

1. Convexity: $R(\lambda \eta^1 + (1 - \lambda) \eta^2) \leq \lambda R(\eta^1) + (1 - \lambda) R(\eta^2)$ for all $\eta^1, \eta^2 \in Z$ and $\lambda \in [0, 1]$;
2. Monotonocity: $\eta^1 \geq \eta^2$ implies $R(\eta^1) \geq R(\eta^2)$ for all $\eta^1, \eta^2 \in Z$;
3. Translation Equivariance: $R(\eta + \lambda) = R(\eta) + \lambda$ for all $\lambda \in \mathbb{R}$ and $\eta \in Z$;
4. Positive Homogeneity: $R(\lambda \cdot \eta) = \lambda \cdot R(\eta)$ for all $\lambda > 0$ and $\eta \in Z$.

Coherent measures of risk can be interpreted as worst-case expectations from a compact convex set of probability measures through their dual representation: $R(\eta) := \max_{P \in \mathcal{P}} \mathbb{E}_P[\eta]$. Therefore, it follows that a RASO can be re-written as a DRO

$$\min_{x \in X} \mathcal{R}(c(x, \xi)) := \min_{x \in X} \max_{P \in \mathcal{P}} \mathbb{E}_P[c(x, \xi)].$$

The above conclusion is extended to the multistage setting by recursively using conditional ambiguity sets, which we recall in Section 3.5; see e.g., [35, 36, 40] for nested coherent composite risk measures in the multistage setting.

3 Lower bounds for DRO

The aim of this section is to provide LBs for DRO formed by $\phi$-divergences and the Wasserstein distance. For this purpose, instead of dealing with the whole sample space $\Omega$, whose large cardinality may lead to computational concerns, the LB is achieved by dividing the sample space $\Omega$ into subgroups that can be considered separately and then combining the optimal values of the subgroups using an ambiguity set with possibly another radius. To perform such a division, we consider the approaches presented in [21] and summarized below. For ease of presentation (of probabilities, computations, etc.), below we set each subgroup to have the same cardinality. However, subgroups with different cardinalities are possible.
3.1 Dissecting the scenario tree

We construct a collection of \( m_l \) subsets, each of cardinality \( l \), of the sample space \( \Omega \)

\[
\left( \Omega_1^{(l)}, \Omega_2^{(l)}, \ldots, \Omega_{m_l}^{(l)} \right)
\]

with the property that their union covers the whole space \( \Omega = \bigcup_{g \in [m_l]} \Omega_g^{(l)} \). For each \( \Omega_g^{(l)} \), \( g \in [m_l] \), there corresponds a probability measure \( Q_g^{(l)} \). Therefore

\[
\left( Q_1^{(l)}, Q_2^{(l)}, \ldots, Q_{m_l}^{(l)} \right)
\]

represents a dissection of the probability measure \( Q = \sum_{g \in [m_l]} \pi_g^{(l)} Q_g^{(l)} \) with \( \sum_{g \in [m_l]} \pi_g^{(l)} = 1 \) and \( \pi_g^{(l)} \geq 0 \) for all \( g \in [m_l] \). For instance, when \( l = 1 \), then \( \Omega_1^{(1)} = \{ \omega_g \} \), \( Q_1^{(1)} = \delta_{\omega_g} \), \( g \in [||\Omega||] \), where \( \delta_{\omega_g} \) represents the Dirac measure at scenario \( \omega_g \); hence, subgroup \( g \) has probability one for scenario \( \omega_g \) and probability zero for all other scenarios. Each measure \( Q_g^{(l)} \) in the collection is given by \( Q_g^{(l)} = \sum_{\omega_i \in \Omega_g^{(l)}} (q_{\omega_i})_g^{(l)} \cdot \delta_{\omega_i} \), where \( (q_{\omega_i})_g^{(l)} \) denotes the nominal probability of scenario \( \omega_i \) within subgroup \( g \) with \( \sum_{\omega_i \in \Omega_g^{(l)}} (q_{\omega_i})_g^{(l)} = 1 \). Below, we provide details of the measures \( Q_g^{(l)} \)—and hence details of \( (q_{\omega_i})_g^{(l)} \)—based on different constructions. Collections of subgroups can be constructed principally in two ways: by keeping one or several scenarios fixed in all subsets, or by choosing them disjoint.

**Fixed scenarios.** We first consider the case where one or more scenarios appear in all subsets. Without loss of generality, we assume that the first \( f < l \) scenarios of \( \Omega \) \( (\Omega_f = \{ \omega_1, \ldots, \omega_f \}) \) are fixed. Consequently the number of subgroups with cardinality \( l \) is \( m_l = \frac{||\Omega|| - f}{l-f} \in \mathbb{N} \). Then, the probability measures \( Q_g^{(l)} \) can be calculated as follows:

\[
Q_g^{(l)} := \sum_{\omega_i \in \Omega_f} q_{\omega_i} \cdot \delta_{\omega_i} + \sum_{\omega_i \in \Omega_g^{(l)} \setminus \Omega_f} \frac{q_{\omega_i}}{\pi_g^{(l)}} \cdot \delta_{\omega_i}
\]

with weights \( \pi_g^{(l)} := \frac{\sum_{\omega_i \in \Omega_g^{(l)} \setminus \Omega_f} q_{\omega_i}}{1 - \sum_{\omega_i \in \Omega_f} q_{\omega_i}} \), for all subgroups \( g \in [m_l] \).

**Disjoint partitions.** Alternatively, one may also consider disjoint partitions: \( \Omega = \bigcup_{g \in [m_l]} \Omega_g^{(l)} \) with \( \Omega_{g_1}^{(l)} \cap \Omega_{g_2}^{(l)} = \emptyset \) for \( g_1 \neq g_2 \). Consequently, the number of subgroups with cardinality \( l \) is \( m_l = \frac{||\Omega||}{l} \in \mathbb{N} \). In this case, probability measures \( Q_g^{(l)} \) are given by

\[
Q_g^{(l)} := \sum_{\omega_i \in \Omega_g^{(l)} \pi_g^{(l)}} q_{\omega_i} \cdot \delta_{\omega_i}
\]
with weights \( \pi_g^{(l)} := \sum_{\omega_i \in \Omega_g^{(l)}} q_{\omega_i} \), for all subgroups \( g \in [m_l] \).

In the multistage setting, for every scenario \( \omega_t \in \Omega_g^{(l)} \) passing through nodes \( n_0, n_1, \ldots, n_T \) (with \( n_t, t \in T \) a generic node at stage \( t \)), conditional probabilities are adjusted as follows:

\[
(q_{n_{t-1}, n_t})_g^{(l)} := \frac{\sum_{\omega \in \Omega_g^{(l)}(n_t)} (q_{\omega})_g^{(l)}}{\sum_{\omega \in \Omega_g^{(l)}(n_{t-1})} (q_{\omega})_g^{(l)}} \quad t \in T \setminus \{0\},
\]

where \( \Omega_g^{(l)}(n_t) \) denotes the set of scenarios of subtree \( \Omega_g^{(3)} \) passing through node \( n_t \).

**Example 1.** Figure 1a displays a sample space \( \Omega = \{\omega_i\}_{i=1}^{15} \) with 15 scenarios. We divide it into 7 subsets \( \Omega_g^{(3)} \), \( g \in [7] \) each of them of cardinality \( l = 3 \) with scenario \( \omega_1 \) fixed. That is, \( \Omega_f = \{\omega_1\} \), \( \Omega_1^{(3)} = \{\omega_1, \omega_2, \omega_3\} \), \( \Omega_2^{(3)} = \{\omega_1, \omega_4, \omega_5\} \), and so forth. Assuming equal probability for each scenario, \( i.e., q_{\omega_i} = \frac{1}{15}, i \in [15] \), the probability of fixed scenario \( \omega_1 \) within each scenario subgroup \( g \) is \( (q_{\omega_1})_g^{(3)} = \frac{1}{15} \) and the probability of other two scenarios in the same subgroup is \( (q_{\omega_i})_g^{(3)} = \frac{7}{15} \). The weight of each subgroup is \( \pi_g^{(3)} = \frac{1}{7}, g \in [7] \).

### 3.2 Convolution of risk measures to obtain lower bounds

Given a collection of subsets of the scenario tree, our next step is to solve the resulting subgroup problems and combine them judiciously to form LBs on the optimal value of DRO. Toward this end, we use *convolution* (also referred to as *composition*) of risk measures induced by the considered ambiguity sets [23]. We describe this process next.

Let \( \mathcal{G} \) be the \( \sigma \)-algebra generated by the collection of subsets \( \Omega = \bigcup_{g \in [m_l]} \Omega_g^{(l)} \), where each subset \( \Omega_g^{(l)} \) corresponds to an elementary event of \( \mathcal{G} \). First, we solve each subgroup \( g \) using DRO given in (3) with radius \( \bar{\rho}_g \). This induces a risk measure denoted \( \mathcal{R}_g^{(l)} \) on each subgroup \( g \in [m_l] \), where \( \mathcal{R}_g^{(l)} \) is the corresponding ambiguity set with radius \( \bar{\rho}_g \). Collectively, we denote these radii as \( \bar{\rho} = \{\bar{\rho}_g\}_{g=1}^{m_l} \). Also, combined together, we represent the one-step conditional risk measure as \( \mathcal{R}_{\bar{\rho}}^{\mathcal{G}} \) and the associated ambiguity set as \( \bar{\mathcal{P}}_{\bar{\rho}}^{\mathcal{G}} := \bigcup_{g \in [m_l]} \mathcal{P}_{\bar{\rho}_g}^{(l)} \). Note that \( \bar{\mathcal{P}}_{\bar{\rho}}^{\mathcal{G}} \) can be represented in terms of \( \mathcal{R}_g^{(l)} \), \( g \in [m_l] \).

Let \( z_g^{(l)} \) be the optimal value of subgroup \( g \in [m_l] \). Consider a \( \mathcal{G} \)-measurable random variable \( \zeta_{LB} \) taking values \( \{z_g^{(l)}\}_{g=1}^{m_l} \) with nominal probabilities \( \pi_g^{(l)} \) for each \( g \in [m_l] \). For instance, in Example 1, \( \zeta_{LB} \) has support \( \{z_1^{(3)}, z_2^{(3)}, \ldots, z_7^{(3)}\} \), consisting of optimal values of the 7 subgroups, each with nominal probability \( \pi_g^{(l)} = \frac{1}{7} \). We create an ambiguity set \( \bar{\mathcal{P}}_{\bar{\rho}}^{\mathcal{G}} \) around this nominal distribution using radius \( \bar{\rho} \). We then aim to obtain a LB on the optimal value of DRO by calculating

\[
\bar{\mathcal{R}}_{\bar{\rho}}^{\mathcal{G}}(\zeta_{LB}) := \max_{\bar{\mathcal{P}}_{\bar{\rho}}^{\mathcal{G}}} \mathbb{E}_{\bar{\mathcal{P}}}[\zeta_{LB}].
\]
This process results in the risk measure \( \tilde{R}_{\tilde{\rho},\rho}(\cdot) := (\tilde{R}_\rho \circ \tilde{R}_\tilde{\rho})^{\mathcal{G}}(\cdot) \), which is the convolution of the one-step conditional risk measure \( \tilde{R}_\rho^{\mathcal{G}} : \mathcal{Z} \rightarrow \mathcal{L}_\infty(\Omega, \mathcal{G}, Q) \) and the risk measure on the collection of subsets \( \tilde{R}_\tilde{\rho}^{\mathcal{G}} : \mathcal{L}_\infty(\Omega, \mathcal{G}, Q) \rightarrow \mathbb{R} \). The ambiguity set corresponding to \( \tilde{R}_{\tilde{\rho},\rho}(\cdot) \) is denoted by \( \tilde{P}_{\tilde{\rho},\rho} \) so that for any \( \eta \in \mathcal{Z} \), \( \tilde{R}_{\tilde{\rho},\rho}(\eta) = \max_{\rho' \in \tilde{P}_{\tilde{\rho},\rho}} \mathbb{E}_{\rho'}[\eta] \) (see [23]). Above, we focused on how to use convolution of risk measures to obtain LBs for DRO. This involves an “optimization” step in \( x \) in (3). For convolution of risk measures on random variables, one can consider a “fixed” \( x \) in (3). In the next two sections, we present LB criteria on general random variables \( \eta \in \mathcal{Z} \). These results are directly applicable to static/two-stage DRO because minimization preserves the LBs. Their application to multistage DRO will be discussed in Section 3.5.

The ambiguity sets mentioned above can be formulated as follows. First, given the subset \( \Omega_g^{(l)} \) for subgroup \( g \), the ambiguity set associated with DRO in (3) using radius \( \tilde{\rho}_g \geq 0 \) inducing risk measure \( \mathcal{R}_{\tilde{\rho}_g}^{(l)} \) on this subgroup is

\[
\mathcal{P}_{\tilde{\rho}_g}^{(l)} := \left\{ \tilde{F}_g^{(l)} : \Delta(\tilde{F}_g^{(l)}, Q_g^{(l)}) \leq \tilde{\rho}_g, \sum_{\omega_i \in \Omega_g^{(l)}} (\tilde{p}_{\omega_i})_g^{(l)} = 1, (\tilde{p}_{\omega_i})_g^{(l)} \geq 0, \forall \omega_i \in \Omega_g^{(l)} \right\},
\]

where \( (\tilde{p}_{\omega_i})_g^{(l)} \) represents the probability \( \tilde{F}_g^{(l)} \) assumes for scenario \( \omega_i \in \Omega_g^{(l)} \). Hence, the ambiguity set corresponding to the one-step conditional risk measure \( \tilde{R}_\rho^{\mathcal{G}} \) is

\[
\tilde{P}_\rho^{\mathcal{G}} := \left\{ P : \Delta(P_g^{(l)}, Q_g^{(l)}) \leq \rho_g, \forall g \in [m], \sum_{\omega_i \in \Omega_g^{(l)}} (p_{\omega_i})_g^{(l)} = 1, \forall g \in [m], (p_{\omega_i})_g^{(l)} \geq 0, \forall \omega_i \in \Omega_g^{(l)}, \forall g \in [m] \right\}.
\]
Above, $\tilde{P} := \{\tilde{P}^{(l)}_{g}\}_{g=1}^{m_l}$. Next, the ambiguity set using radius $\tilde{\rho} \geq 0$ inducing risk measure $\tilde{R}_{\tilde{\rho}}^g$ on the collection of subsets is

$$\tilde{R}_{\tilde{\rho}}^g := \left\{ \tilde{P} : \Delta(\tilde{P}, \check{Q}) \leq \tilde{\rho}, \sum_{g\in[m_l]} \tilde{p}^{(l)}_g = 1, \tilde{p}^{(l)}_g \geq 0, \forall g \in [m_l] \right\},$$

where $\check{Q}$ is the nominal distribution composed of the weights $\pi^{(l)}_g$ detailed in Section 3.1 and $\tilde{p}^{(l)}_g$ represents the probability $\tilde{P}$ assumes for the subgroup $g$ with cardinality $l$. Finally, the ambiguity set corresponding to the convolution $\tilde{R}_{\tilde{\rho}}^g \circ \tilde{R}_{\tilde{\rho}}^{g'}$ with associated risk measure $\tilde{R}_{\tilde{\rho}, \tilde{\rho}}$ becomes

$$\tilde{R}_{\tilde{\rho}, \tilde{\rho}} := \left\{ \tilde{P} : \tilde{p}_{\omega, g} = \tilde{p}^{(l)}_g \cdot (\tilde{p}_{\omega})^{(l)}_g, \forall \omega_i \in \Omega_f, g \in [m_l], \tilde{p}'_{\omega_i} = \sum_{g\in[m_l]} \tilde{p}_{\omega_i, g}, \forall \omega_i \in \Omega_f, \right.$$

$$\text{and } \tilde{p}'_{\omega_i} = \tilde{p}^{(l)}_g \cdot (\tilde{p}_{\omega_i})^{(l)}_g, \forall \omega_i \in \Omega_g^{(l)} \setminus \Omega_f, g \in [m_l], \check{P} \in \tilde{R}_{\tilde{\rho}}^g, \quad \tilde{P} \in \tilde{R}_{\tilde{\rho}}^{g'}, \right\}$$

(4)

Recall $\Omega_f$ denotes the set of fixed scenarios (Section 3.1), and if $\Omega_f = \emptyset$, disjoint partitions are used. For any fixed scenario $\omega_i \in \Omega_f$, its probability $\tilde{p}'_{\omega_i}$ is found by summing up its group probabilities $\tilde{p}'_{\omega_i, g}$ for all subgroups $g \in [m_l]$. For instance, in Example 1, the probability of scenario $\omega_1$ after convolution is found by $\tilde{p}'_{\omega_1} = \sum_{g=1}^{7} \tilde{p}^{(3)}_g \cdot (\tilde{p}_{\omega_1})^{(3)}_g$, whereas for all other scenarios we have $\tilde{p}'_{\omega_2} = \tilde{p}^{(3)}_1 \cdot (\tilde{p}_{\omega_2})^{(3)}_1, \ldots, \tilde{p}'_{\omega_5} = \tilde{p}^{(3)}_2 \cdot (\tilde{p}_{\omega_5})^{(3)}_2, \ldots, \tilde{p}'_{\omega_15} = \tilde{p}^{(3)}_7 \cdot (\tilde{p}_{\omega_{15}})^{(3)}_7$ (see Figure 1b). Notice that in (4) the condition $\sum_{\omega_i \in \Omega} \tilde{p}'_{\omega_i} = 1$ always holds because the respective ambiguity sets $\tilde{R}_{\tilde{\rho}}^g$ and $\tilde{R}_{\tilde{\rho}}^{g'}$ require $\sum_{g\in[m_l]} \tilde{p}^{(l)}_g = 1$ and $\sum_{\omega_i \in \Omega_g^{(l)}} (\tilde{p}_{\omega_i})^{(l)}_g = 1$ for all subgroups $g \in [m_l]$:

$$\sum_{\omega_i \in \Omega} \tilde{p}'_{\omega_i} = \sum_{\omega_i \in \Omega_f} \tilde{p}'_{\omega_i} + \sum_{\omega_i \in (\Omega \setminus \Omega_f)} \tilde{p}'_{\omega_i} = \sum_{\omega_i \in \Omega_f} \sum_{g\in[m_l]} \tilde{p}^{(l)}_g \cdot (\tilde{p}_{\omega_i})^{(l)}_g + \sum_{g\in[m_l]} \sum_{\omega_i \in \Omega_f \cup \Omega_g^{(l)}} \tilde{p}^{(l)}_g \cdot (\tilde{p}_{\omega_i})^{(l)}_g$$

$$= \sum_{g\in[m_l]} \sum_{\omega_i \in \Omega_g^{(l)}} \tilde{p}^{(l)}_g \cdot (\tilde{p}_{\omega_i})^{(l)}_g = \sum_{g\in[m_l]} \left( \sum_{\omega_i \in \Omega_g^{(l)}} \tilde{p}^{(l)}_g \cdot \sum_{\omega_i \in \Omega_f \cup \Omega_g^{(l)}} (\tilde{p}_{\omega_i})^{(l)}_g \right) = \sum_{g\in[m_l]} \left( \sum_{\omega_i \in \Omega_f \cup \Omega_g^{(l)}} \tilde{p}^{(l)}_g \cdot \sum_{\omega_i \in \Omega_g^{(l)}} \tilde{p}^{(l)}_g \cdot 1 \right) = 1,$$

where the complement set with respect to $\Omega$ is denoted by $(\cdot)^C$.

In the rest of this section, we denote the nominal probabilities of the fixed scenarios after dissection as $q_{\omega_i, g} = \pi^{(l)}_g \cdot (q_{\omega_i})^{(l)}_g, \forall \omega_i \in \Omega_f, g \in [m_l]$, where $q_{\omega_i} = \sum_{g\in[m_l]} q_{\omega_i, g}$ for any $\omega_i \in \Omega_f$. We use $\tilde{\rho}_{\text{max}}$ to denote the maximal value of $\tilde{\rho}$ among subgroups $g \in [m_l]$ (i.e., $\tilde{\rho}_{\text{max}} = \max_{g\in[m_l]} \tilde{\rho}_g$). Subscript $\phi^g_{CR}$ is used to represent all relevant ambiguity sets and risk measures induced by CR power divergence family with parameter $\theta$. For instance, $R_{\phi^g_{CR}}$ denotes the risk measure induced by the CR power divergence with ambiguity set.
\( P^{\phi_{CR}(\rho)} \) using radius \( \rho \) of the original DRO, and \( \tilde{R}^{\phi_{CR}(\tilde{\rho}, \rho)} = \left( \tilde{R}^{\phi_{CR}(\tilde{\rho})} \circ \tilde{R}^{\phi_{CR}(\rho)} \right) \) denotes the risk measure after convolution using radii \( \tilde{\rho} \) and \( \rho \), and so forth. Similarly, we use subscripts \( \phi_v, \phi_J, \phi_{\chi}^a \), and \( W \) to denote VD, \( J \)-divergence, \( \chi \)-divergence of order \( a > 1 \) and Wasserstein distance, respectively. These notations are used in the subsequent results and their proofs.

### 3.3 Lower-bound criteria for \( \phi \)-divergences

We now present LB criteria for DRO formed via commonly used \( \phi \)-divergences listed in Table 1 through scenario grouping. We begin with CR power divergence family and present the proof in detail. The LB criteria for the special cases of the CR power divergence family in Table 2 can be acquired from the below result.

**Proposition 1. (LB criteria for CR power divergences).** Consider the convolution formed by CR power divergence with parameter \( \theta \neq 0, 1 \). For \( 0 < \theta < 1 \), suppose the radii \( \rho, \bar{\rho}, \rho, \bar{\rho}' \in [0, \frac{1}{\theta} + \frac{1}{1-\theta}] \), \( g \in [m_l] \). For \( \theta < 0 \) or \( \theta > 1 \), suppose the radii \( \rho, \bar{\rho}, \rho, \bar{\rho'} \geq 0 \) and the sample space \( \Omega \) is dissected by disjoint partitions (i.e., \( \Omega_f = \emptyset \)). If

\[
\begin{align*}
\bar{\rho} + \rho_{\text{max}} &\leq \rho \quad \text{when} \quad \theta \in (0, 1), \\
\bar{\rho} + \rho_{\text{max}} - \theta(1-\theta) \cdot \bar{\rho} \cdot \rho_{\text{max}} &\leq \rho \quad \text{when} \quad \theta < 0 \text{ or } \theta > 1 \text{ and } \Omega_f = \emptyset, \\
\bar{\rho} + \rho_{\text{max}} &\leq \rho \quad \text{when} \quad \theta \to 0 \text{ or } \theta \to 1,
\end{align*}
\]

then \( \tilde{R}^{\phi_{CR}(\tilde{\rho}, \rho)}(\eta) \leq R^{\phi_{CR}(\rho)}(\eta) \) for all \( \eta \in Z \).

**Proof.** Let \( P' \in \tilde{P}^{\phi_{CR}(\tilde{\rho}, \rho)} \). Then there exists \( P \in \tilde{P}^{\phi_{CR}(\tilde{\rho})} \) and \( \bar{P} \in \tilde{P}^{\phi_{CR}(\rho)} \) such that

\[
\sum_{g \in [m_l]} \tilde{p}_g^{(l)} = 1, \quad \text{and} \quad \sum_{\omega_i \in \Omega_g^{(l)}} (\bar{p}_{\omega_i})^{(l)} = 1
\]

and, by the definition of \( \Delta^{\phi_{CR}} \) from Table 1, satisfying

\[
1 - \frac{\sum_{g \in [m_l]} \left( \tilde{p}_g^{(l)} \right)^\theta \left( \pi_g^{(l)} \right)^{1-\theta}}{\theta(1-\theta)} \leq \tilde{\rho}, \quad 1 - \frac{\sum_{\omega_i \in \Omega_g^{(l)}} \left( \bar{p}_{\omega_i}^{(l)} \right)^\theta \left( \bar{q}_{\omega_i}^{(l)} \right)^{1-\theta}}{\theta(1-\theta)} \leq \bar{\rho}, \quad \forall g \in [m_l]. \tag{5}
\]

We now show the steps to find the criteria for \( \Delta^{\phi_{CR}}(P', Q) \leq \rho \):
When \( \theta \in (0, 1) \), writing out \( \Delta_{\phi_{CR}}(P', Q) \), we obtain

\[
1 - \frac{\sum_{\omega_i \in \Omega} \left( p_{\omega_i}' \right)^\theta (q_{\omega_i})^{1-\theta}}{\theta(1-\theta)} = 1 - \frac{\sum_{\omega_i \in \Omega_f} \left( p_{\omega_i}' \right)^\theta (q_{\omega_i})^{1-\theta} - \sum_{\omega_i \in (\Omega_f)^C} \left( p_{\omega_i}' \right)^\theta (q_{\omega_i})^{1-\theta}}{\theta(1-\theta)}
\]

\[
1 - \frac{\sum_{\omega_i \in \Omega_f} \sum_{g \in [m]} \left( \tilde{p}_g \right)^\theta (\bar{\pi}_g (q_{\omega_i})^{(l)})^{1-\theta} - \sum_{g \in [m]} \sum_{\omega_i \in (\Omega_f)^g} \left( \tilde{p}_g \right)^\theta (\bar{\pi}_g (q_{\omega_i})^{(l)})^{1-\theta}}{\theta(1-\theta)}
\]

\[
1 - \frac{\sum_{g \in [m]} \sum_{\omega_i \in (\Omega_f)^g} \left( \tilde{p}_g \right)^\theta (\bar{\pi}_g (q_{\omega_i})^{(l)})^{1-\theta}}{\theta(1-\theta)} + \frac{\sum_{g \in [m]} \left( \tilde{p}_g \right)^\theta (\bar{\pi}_g^{(l)})^{1-\theta} - \sum_{g \in [m]} \sum_{\omega_i \in (\Omega_f)^g} \left( \tilde{p}_g \right)^\theta (\bar{\pi}_g (q_{\omega_i})^{(l)})^{1-\theta}}{\theta(1-\theta)}
\]

\[
\leq \tilde{\rho} + \sum_{g \in [m]} \left[ \left( \tilde{p}_g \right)^\theta (\bar{\pi}_g^{(l)})^{1-\theta} \right] \leq \tilde{\rho} + \sum_{g \in [m]} \left[ \left( \tilde{p}_g \right)^\theta (\bar{\pi}_g^{(l)})^{1-\theta} \right] \cdot \tilde{\rho}_{\max}
\]

\[
= \tilde{\rho} + \tilde{\rho}_{\max} - \theta(1-\theta) \cdot \tilde{\rho}_{\max}
\]

where inequality (6) follows from Hölder’s inequality applied on the fixed scenarios. The first inequality in (8) follows from (5) and the second from definition of \( \tilde{\rho}_{\max} \). Let us denote the right-hand side of (9) as \( A \). Since \( -\theta(1-\theta) \) is negative, \( A \leq \tilde{\rho} + \tilde{\rho}_{\max} \). Therefore, if \( \tilde{\rho} + \tilde{\rho}_{\max} \leq \rho \), we have \( \tilde{\mathcal{P}}_{\phi_{CR}}(\rho, \tilde{\rho}) \subseteq \mathcal{P}_{\phi_{CR}}(\rho) \). This implies that for any \( \eta \in \mathcal{Z} \), we obtain
\[ \tilde{R}_{\phi \theta}^{\text{CR}}(\bar{\rho}, \bar{\rho}) = \max_{P \in \tilde{\mathcal{P}}_{\phi \theta}^{\text{CR}}} \mathbb{E}_P[\eta] \leq \max_{P \in \mathcal{P}_{\phi \theta}^{\text{CR}}} \mathbb{E}_P[\eta] = R_{\phi \theta}^{\text{CR}}(\rho). \]

2. When \( \theta < 0 \) or \( \theta > 1 \), we can no longer apply Hölder’s inequality on the fixed scenarios in (6). However, when \( \Omega_f = \emptyset \) (i.e., disjoint partitions are used), we can directly start from (7) and follow the steps to (9). Since \( -\theta(1-\theta) \) is positive, \( A \leq \tilde{\rho} + \tilde{\rho}_{\text{max}} - \theta(1-\theta) \cdot \tilde{\rho} \cdot \tilde{\rho}_{\text{max}} \) by (5). Therefore, if \( \tilde{\rho} + \tilde{\rho}_{\text{max}} - \theta(1-\theta) \cdot \tilde{\rho} \cdot \tilde{\rho}_{\text{max}} \leq \rho \), the result follows.

3. When \( \theta \to 1 \), the CR power divergence is equivalent to Kullback-Leibler divergence. Detailed proof is provided in Appendix.

4. When \( \theta \to 0 \), the proof is similar to the \( \theta \to 1 \) case and hence omitted. \( \Box \)

Appendix provides a detailed proof of Kullback-Leibler divergence (i.e., \( \theta \to 1 \) limit case). Appendix and proof of Proposition 1 reveal that when \( \theta \in (0, 1) \) or in the limit cases of Kullback-Leibler divergence (\( \theta \to 1 \)) and Burg entropy (\( \theta \to 0 \)), the sample space \( \Omega \) can be dissected in any way, either using disjoint partitions or fixed scenarios. However, when \( \theta < 0 \) or \( \theta > 1 \), the above result is valid for disjoint partitions.

LB criteria for other \( \phi \)-divergences in Table 1 can be obtained using a similar proof technique. Below, we provide the results and relegate the proofs to Appendix.

**Proposition 2. (LB criterion for variation distance)**. Consider the convolution formed by variation distance, where radii \( \rho, \tilde{\rho}, \tilde{\rho} \in [0, 2], g \in [m] \). If

\[ \tilde{\rho} \cdot \tilde{\rho}_{\text{max}} + \tilde{\rho} + \tilde{\rho}_{\text{max}} \leq \rho, \]

then \( \tilde{R}_{\phi \theta}^{\text{CR}}(\bar{\rho}, \bar{\rho})(\eta) \leq R_{\phi \theta}^{\text{CR}}(\rho)(\eta) \) for all \( \eta \in Z \).

**Proposition 3. (LB criterion for J-divergence)**. Consider the convolution formed by J-divergence, where radii \( \rho, \tilde{\rho}, \tilde{\rho} \geq 0, g \in [m] \). If

\[ \tilde{\rho} + \tilde{\rho}_{\text{max}} \leq \rho, \]

then \( \tilde{R}_{\phi \theta}^{\text{CR}}(\bar{\rho}, \bar{\rho})(\eta) \leq R_{\phi \theta}^{\text{CR}}(\rho)(\eta) \) for all \( \eta \in Z \).

**Proposition 4. (LB criterion for \( \chi \)-divergence of order \( a > 1 \))**. Consider the convolution formed by \( \chi \)-divergence of order \( a > 1 \), where radii \( \rho, \tilde{\rho}, \tilde{\rho} \geq 0, g \in [m] \) and the sample space \( \Omega \) is dissected by disjoint partitions (i.e., \( \Omega_f = \emptyset \)). If

\[ \left[ \left( \tilde{\rho} \right)^{\frac{a}{2}} + \left( \tilde{\rho}_{\text{max}} \right)^{\frac{a}{2}} + \left( \tilde{\rho} \cdot \tilde{\rho}_{\text{max}} \right)^{\frac{a}{2}} \right]^{\frac{2}{a}} \leq \rho, \]

16
then \( \tilde{R}_{\phi}(\tilde{\rho}, \rho)(\eta) \leq R_{\phi}(\rho)(\eta) \) for all \( \eta \in \mathbb{Z} \).

Note that Proposition 4 is a general result that applies to all values of \( a > 1 \). For certain values of \( a \), a tighter inequality might be available (e.g., when \( a = 2 \)).

### 3.4 Lower-bound criterion for Wasserstein distance

We now provide a LB criterion for Wasserstein distance through scenario decomposition. The main idea is the same: to find criteria that guarantee the convoluted ambiguity set being a subset of the ambiguity set of the original problem. Recall that Wasserstein distance needs a distance \( d_{\omega_i, \omega_j} \) between any two scenarios \( \omega_i \) and \( \omega_j \). To apply it to scenario groups, we need a distance between subgroups as well. We provide such a distance between subgroups and a criterion for radii \( \bar{\rho}, \tilde{\rho} \) to ensure LBs below.

**Proposition 5. (LB criterion for Wasserstein distance).** Consider the convolution of DRO formed by Wasserstein distance, where radii \( \rho, \bar{\rho}, \tilde{\rho} \geq 0 \), \( g \in [m] \). Let the distance between scenario groups be defined as

\[
d_{g_1, g_2} := \begin{cases} \max_{\omega_i \in \Omega(g_1), \omega_j \in \Omega(g_2)} \{ d_{\omega_i, \omega_j} \} & \text{when } g_1 \neq g_2 \\ 0 & \text{when } g_1 = g_2, \end{cases} \tag{10}
\]

where \( g_1, g_2 \in [m] \). If \( \bar{\rho} + \tilde{\rho}_{\text{max}} \leq \rho \),

then \( \tilde{R}_W(\bar{\rho}, \rho)(\eta) \leq R_W(\rho)(\eta) \) for all \( \eta \in \mathbb{Z} \).

**Proof.** First assume \( \Omega \) is dissected using disjoint partitions (i.e., \( \Omega_f = \emptyset \)). Given \( P' \in P_w(\bar{\rho}, \rho) \) formed after scenario grouping and convolution, the Wasserstein distance between \( P' \) and \( Q \) for the original problem can be written as

\[
\Delta_W(P', Q) := \min_{z \geq 0} \left\{ \sum_{\omega_i \in \Omega} \sum_{\omega_j \in \Omega(g)} d_{\omega_i, \omega_j} z_{\omega_i, \omega_j} : \sum_{\omega_i \in \Omega} z_{\omega_i, \omega_j} = q_{\omega_j}, \forall \omega_j \in \Omega \right\}, \tag{11}
\]

and the Wasserstein distance for ambiguity set \( \tilde{P}_W(\bar{\rho}) \) can be written as

\[
\Delta_W(\tilde{P}, \bar{Q}) := \min_{y \geq 0} \left\{ \sum_{g_1 \in [m]} \sum_{g_2 \in [m]} d_{g_1, g_2} y_{g_1, g_2} : \sum_{g_1 \in [m]} y_{g_1, g_2} = \tilde{p}^{(g)}_{g_2}, \forall g_2 \in [m], \sum_{g_2 \in [m]} y_{g_1, g_2} = \tilde{p}^{(g)}_{g_1}, \forall g_1 \in [m] \right\}. \tag{12}
\]
Similarly, for each subgroup \( g \in [m_l] \), we have

\[
\Delta_W(P(l), Q(l)) := \min_{x \geq 0} \left\{ \sum_{\omega_i \in \Omega_g(l)} \sum_{\omega_j \in \Omega_g(l)} d_{\omega_i, \omega_j} (x_{\omega_i, \omega_j})_g : \sum_{\omega_i \in \Omega_g(l)} (x_{\omega_i, \omega_j})_g = (q_{\omega_j})_g^l, \omega_j \in \Omega_g(l), \right. \\
\left. \sum_{\omega_i \in \Omega_g(l)} (x_{\omega_i, \omega_j})_g = (\tilde{p}_{\omega_i})_g^l, \omega_i \in \Omega_g(l) \right\}.
\] (13)

Let us now define one way to obtain \( z_{\omega_i, \omega_j} \) in (11) by using the transportation decisions \( y_{g1, g2} \) and \( (x_{\omega_i, \omega_j})_g \) in (12) and (13), respectively:

\[
\begin{align*}
  z_{\omega_i, \omega_j} &= y_{g, g} (x_{\omega_i, \omega_j})_g, && \forall \omega_i, \omega_j \in \Omega_g(l), \forall g \in [m_l] \quad (14) \\
  \sum_{\omega_j \in \Omega_{g2}(l)} z_{\omega_i, \omega_j} &= y_{g1, g2} \sum_{\omega_j \in \Omega_{g2}(l)} (x_{\omega_i, \omega_j})_g, && \forall \omega_j \in \Omega_{g2}(l), \forall g2 \in [m_l] \quad (15) \\
  \sum_{\omega_i \in \Omega_{g1}(l)} z_{\omega_i, \omega_j} &= y_{g1, g2} \sum_{\omega_i \in \Omega_{g1}(l)} (x_{\omega_i, \omega_j})_g, && \forall \omega_i \in \Omega_{g1}(l), \forall g1 \in [m_l]. \quad (16)
\end{align*}
\]

Above, we also assume \( z \geq 0 \). (This is automatically satisfied for within-subgroup \( z_{\omega_i, \omega_j} \) by (14) given \( y \geq 0 \) and \( x \geq 0 \).) With the transformation above, we can show that the constraints in (11) of the Wasserstein distance \( \Delta_W(P', Q) \) are all satisfied, even though \( z_{\omega_i, \omega_j} \) formed through (14)–(16) may not be optimal to \( \Delta_W(P', Q) \). For instance, the first set of constraints in (11) for all \( \omega_j \in \Omega \) (or equivalently all \( \omega_j \in \Omega_g(l), g \in [m_l] \)) are satisfied by (15) and the first sets of constraints in (12) and (13):

\[
\begin{align*}
  \sum_{\omega_i \in \Omega} z_{\omega_i, \omega_j} &= \sum_{g1 \in [m_l]} \sum_{\omega_i \in \Omega_{g1}(l)} z_{\omega_i, \omega_j} = \sum_{g1 \in [m_l]} \left( \sum_{\omega_i \in \Omega_{g1}(l)} (y_{g1, g} \sum_{\omega_j \in \Omega_{g2}(l)} (x_{\omega_i, \omega_j})_g) \right) \\
  &= \left( \sum_{g1 \in [m_l]} y_{g1, g} \right) \left( \sum_{\omega_i \in \Omega_{g1}(l)} (x_{\omega_i, \omega_j})_g \right) = \pi_{g1}^l (q_{\omega_j})_g^l = q_{\omega_j},
\end{align*}
\]

where \( g \) denotes the subgroup scenario \( \omega_j \) belongs to. The second set of constraints in (11) can be shown similarly by using (16) and the second sets of constraints in (12) and (13). Hence, all feasible solutions to constraints in (12) and (13) are also feasible to the constraints in (11).

Let \( Z, Y, \) and \( X \) denote the feasible regions given by the constraints in (11), (12), and (13), respectively, each supplemented with their nonnegativity constraints \( z \geq 0, y \geq 0, \) and \( x \geq 0 \). We now show steps to find criteria for \( \Delta_W(P', Q) \leq \rho \):

\[
\Delta_W(P', Q) = \min_{x \in \mathbb{Z}} \sum_{\omega_i \in \Omega} \sum_{\omega_j \in \Omega} d_{\omega_i, \omega_j} z_{\omega_i, \omega_j}.
\]
\[
= \min_{z \in \mathbb{Z}} \sum_{g \in [m]} \sum_{\omega_j \in \Omega_g} \left( \sum_{\omega_i \in \Omega_g} d_{\omega_i, \omega_j} z_{\omega_i, \omega_j} + \sum_{\omega_j \in (\Omega_g')^c} d_{\omega_i, \omega_j} z_{\omega_i, \omega_j} \right)
\]

\[
\leq \min_{y \in \mathbb{Z}} \sum_{x \in \mathbb{X}} \sum_{g \in [m]} \left( \sum_{\omega_i \in \Omega_g} \sum_{\omega_j \in \Omega_g'} d_{\omega_i, \omega_j} y_{g, g}(x_{\omega_i, \omega_j}) + \sum_{\omega_j \in (\Omega_g')^c} d_{\omega_i, \omega_j} z_{\omega_i, \omega_j} \right) \tag{a}
\]

\[
\leq \min_{y \in \mathbb{Z}} \sum_{x \in \mathbb{X}} \sum_{g \in [m]} \left( \sum_{\omega_i \in \Omega_g} \sum_{\omega_j \in \Omega_g'} d_{\omega_i, \omega_j} y_{g, g}(x_{\omega_i, \omega_j}) + \sum_{g \in [m]} \sum_{g \in [m]} d_{g, g_2} y_{g, g_2} \right) \tag{b}
\]

\[
\leq \min_{y \in \mathbb{Z}} \sum_{x \in \mathbb{X}} \sum_{g \in [m]} \left( \sum_{\omega_i \in \Omega_g} \sum_{\omega_j \in \Omega_g'} d_{\omega_i, \omega_j} y_{g, g}(x_{\omega_i, \omega_j}) + \sum_{g \in [m]} \sum_{g \in [m]} d_{g, g_2} y_{g, g_2} \right) \tag{c}
\]

\[
= \sum_{g \in [m]} \left( \pi_g^{(l)} \sum_{x \in \mathbb{X}} \sum_{\omega_i \in \Omega_g} d_{\omega_i, \omega_j}(x_{\omega_i, \omega_j}) \right) + \min_{y \in \mathbb{Z}} \sum_{g \in [m]} \sum_{g \in [m]} d_{g, g_2} y_{g, g_2} \tag{d}
\]

\[
= \sum_{g \in [m]} \pi_g^{(l)} \Delta W(\bar{P}_g^{(l)}, Q_g^{(l)}) + \Delta W(\bar{P}, \bar{Q}) \leq \sum_{g \in [m]} \pi_g^{(l)} \cdot \bar{r}_{g} + \bar{p} \leq \bar{p} + \bar{p}_{\max}. \tag{f}
\]

where (a) follows from (14). Note that this is an inequality because decisions \( z \) obtained through this transformation may not be optimal. Inequality (b) follows (10). By summing over \( \omega_j \in \Omega_g \) on both sides of (15), we can show \( y_{g, g_2} = \sum_{\omega_i \in \Omega_g} \sum_{\omega_j \in \Omega_g} z_{\omega_i, \omega_j}, \forall g, g_2 \in [m] \). Then (c) follows. Inequality (d) follows from \( y_{g, g} \leq \pi_g^{(l)} \) (see (12)). Equality (e) is due to the fact that the resulting problem is separable. Finally, the equality in (f) follows from the definition of \( \Delta W(\bar{P}_g^{(l)}, Q_g^{(l)}) \) and \( \Delta W(\bar{P}, \bar{Q}) \), and the inequality in (f) follows by construction. Therefore, similar to the statement at the end of the proof of part 1 of Proposition 1, when \( \theta \in (0, 1) \), if \( \bar{p} + \bar{p}_{\max} \leq p \), then \( \bar{R}_W(\bar{p}, \bar{r})(\eta) \leq \bar{R}_W(r)(\eta) \) for all \( \eta \in \mathbb{Z} \).

We now consider the case when \( \Omega \) is dissected using fixed scenarios (\( \Omega_f \neq \emptyset \)). Recall in (4), for any fixed scenario \( \omega_i \in \Omega_f \), we have \( p'_{\omega_i, g} = \bar{p}_g^{(l)}(\bar{p}_{\omega_i})^{(l)}_g, \ g \in [m] \) and \( p'_{\omega_i} = \sum_{g \in [m]} p'_{\omega_i, g} \). Hence splitting the two constraints in (11), we have

\[
\sum_{\omega_i \in \Omega} z_{\omega_i, \omega_j} = q_{\omega_j} = \sum_{g \in [m]} q_{\omega_j, g} = \sum_{g \in [m]} \pi_g^{(l)}(q_{\omega_j})^{(l)}_g, \quad \omega_j \in \Omega_f, \tag{17}
\]

\[
\sum_{\omega_i \in \Omega} z_{\omega_i, \omega_j} = \pi_g^{(l)}(q_{\omega_j})^{(l)}_g, \quad \omega_j \in (\Omega_f)^c, \tag{18}
\]

\[
\sum_{\omega_i \in \Omega} p'_{\omega_i} = \sum_{g \in [m]} p'_{\omega_i, g} = \sum_{g \in [m]} \bar{p}_g^{(l)}(\bar{p}_{\omega_i})^{(l)}_g, \quad \omega_i \in \Omega_f, \tag{19}
\]

\[
\sum_{\omega_i \in \Omega} p'_{\omega_i} = \bar{p}_g^{(l)}(\bar{p}_{\omega_i})^{(l)}_g, \quad \omega_i \in (\Omega_f)^c. \tag{20}
\]
Define a finite expanded space \( \tilde{\Omega} := \{ \omega_{1(1)}, \omega_{1(2)}, \ldots, \omega_{1(m_1)}, \omega_{2(1)}, \omega_{2(2)}, \ldots, \omega_{2(m_2)}, \ldots, \omega_{1(f)}, \omega_{2(f)}, \ldots, \omega_{f+1}, \omega_{f+2}, \ldots, \omega_{|\Omega|} \} \), where the fixed scenarios \( \omega_i \in \Omega_f = \{ \omega_1, \ldots, \omega_f \} \) in different subgroups are considered to have different “atoms” and the rest of the scenarios \( \omega_i \in (\Omega_f)^C \) are left as before. We again use the same subgroups \( \Omega_g^{(l)} \) but on the expanded space \( \tilde{\Omega} \). Then, we have the following Wasserstein distance on \( \tilde{\Omega} \):

\[
\tilde{\Delta}_W(P', Q) := \min_{\tilde{z} \geq 0} \left\{ \sum_{\omega_i \in \Omega} \sum_{\omega_j \in \tilde{\Omega}} d_{\omega_i, \omega_j} \tilde{z}_{\omega_i, \omega_j} : \sum_{\omega_i \in \Omega} \tilde{z}_{\omega_i, \omega_j} = \pi_g^{(l)} (q_g)_g, \forall \omega_j \in \tilde{\Omega}, \right. \\
\left. \sum_{\omega_j \in \tilde{\Omega}} \tilde{z}_{\omega_i, \omega_j} = \tilde{p}_g^{(l)} (\bar{p}_g)_g, \forall \omega_i \in \tilde{\Omega} \right\},
\]

(21)

For any \( \tilde{z} \geq 0 \) feasible to Wasserstein distance \( \tilde{\Delta}_W(P', Q) \)—including the optimal \( \tilde{z} \)—on the expanded space \( \tilde{\Omega} \), we can generate a feasible solution \( z \geq 0 \) to (17)–(20). First, for any non-fixed scenarios \( \omega_i, \omega_j \in (\Omega_f)^C \), we set \( z_{\omega_i, \omega_j} = \tilde{z}_{\omega_i, \omega_j} \) and observe the constraints in (21) are the same as (18) and (20). Next, for any fixed scenario \( \omega_i \in \Omega_f \) and non-fixed scenario \( \omega_j \in (\Omega_f)^C \), we set (i) \( z_{\omega_i, \omega_j} = \sum_{g_1 \in [m_1]} \sum_{g_2 \in [m_1]} \tilde{z}_{\omega_i, (g_1), (g_2)} \cdot \sum_{g \in [m]} \tilde{z}_{(g_1), (g_2)} \), (ii) \( z_{\omega_i, \omega_j} = \sum_{g \in [m]} \tilde{z}_{(g_1), (g_2)} \), and (iii) \( z_{\omega_j, \omega_i} = \sum_{g \in [m]} \tilde{z}_{(g_1), (g_2)} \). Then (17) and (19) are also satisfied. Furthermore, with this \( z \), the objective functions of two Wasserstein distances coincide: \( \sum_{\omega_i \in \Omega} \sum_{\omega_j \in \tilde{\Omega}} d_{\omega_i, \omega_j} z_{\omega_i, \omega_j} = \sum_{\omega_i \in \tilde{\Omega}} \sum_{\omega_j \in \tilde{\Omega}} d_{\omega_i, \omega_j} \tilde{z}_{\omega_i, \omega_j} \) because any distance involving \( \omega_i (g) \in \tilde{\Omega} \) is equivalent to distance involving \( \omega_i \in \Omega_f \), e.g., \( d_{\omega_i (g_1), \omega_i (g_2)} = 0 \) for all \( \omega_i \in \Omega_f, g_1, g_2 \in [m_1] \). As a result, \( \Delta_W(P', Q) \leq \tilde{\Delta}_W(P', Q) \). Observe \( \tilde{\Delta}_W(P', Q) \) is obtained as “disjoint” partitions on \( \tilde{\Omega} \). Then, following similar steps to the disjoint partition, we show \( \tilde{\Delta}_W(P', Q) \leq \bar{p} + \rho_{\text{max}} \). Therefore, if \( \bar{p} + \rho_{\text{max}} \leq \rho \) the result follows.

**Remark.** Although the above propositions use \( \rho_{\text{max}} \), these results can also be obtained using the individual \( \tilde{\rho}_g \) values for each scenario group \( g \in [m_1] \). For instance, the condition for Wasserstein would be \( \bar{p} + \sum_{g \in [m_1]} \tilde{\rho}_g^{(l)} \cdot \tilde{\rho}_g \leq \rho \) and the result for \( J \)-divergence would be \( \bar{p} + \sum_{g \in [m_1]} \tilde{\rho}_g^{(l)} \cdot \bar{\rho}_g \leq \rho \). In our numerical results, we use the same value for each group, i.e., \( \bar{\rho}_g = \bar{\rho}_{\text{max}} \) for all \( g \in [m_1] \).

### 3.5 Lower bounds for multistage problems

We now extend the LBs obtained in Sections 3.3 and 3.4 to multistage DRO. Due to the correspondence between DRO and RASO, here we focus on a RASO formulation and present our results and proofs using the properties of conditional coherent risk measures.

We first recall the results in [35]. For a multistage decision horizon with stages \( t \in T \) let \( Z_t := \mathcal{L}_\infty(\Omega_t, \tilde{\mathcal{F}}_t, Q_t) \). The mapping \( \mathcal{R}^{\tilde{\mathcal{F}}_{t+1}, \mathcal{F}_t} : Z_{t+1} \rightarrow Z_t \) is called one-step conditional
risk measure if it satisfies the properties presented in Section 2.4 for corresponding spaces $Z_t$ and $Z_{t+1}$ for all $t \in \{0, \ldots, T-1\}$. The risk involved in a sequence of random variables $\eta_t \in Z_t$, $t \in T$ adapted to the filtration $\mathcal{F}_t$, $t \in T$ can be evaluated by a time-consistent dynamic risk measure $\mathcal{R}_\rho$ induced by a measure of similarity between distributions $\Delta$ using radii $\rho := (\rho_1, \ldots, \rho_T)$, defined as follows:

$$
\mathcal{R}_\rho(\eta_0, \ldots, \eta_T) := \eta_0 + \mathcal{R}_{\rho_1}^{\mathcal{F}_1|\mathcal{F}_0}(\eta_1 + \mathcal{R}_{\rho_2}^{\mathcal{F}_2|\mathcal{F}_1}(\eta_2 + \ldots + \mathcal{R}_{\rho_T}^{\mathcal{F}_T|\mathcal{F}_{T-1}}(\eta_T))).
$$  (22)

In general, it is not necessary to use the same $\Delta$ at each stage of the problem, but we do so in this paper for simplicity. Also, by changing the radii $\rho_t$ we can choose how close we remain to the nominal distributions at different stages. Setting $\eta_t := c_t(x_t, \xi_t)$ at stages $t \in T$ and using (22), the multistage RASO problem can be formulated as

$$
\min_{x_0 \in X_0(\xi_0)} c_0(x_0, \xi_0) + \mathcal{R}_{\rho_1} \left( Q_1(x_0, \xi_0(\eta_0)) \right)
$$  (23)

where

$$
Q_1(x_0, \xi_0(\eta_0)) := \min_{x_t \in X_t(x_{t-1}, \xi_t), t \in T \setminus \{0\}} \mathcal{R}_{\rho_2}^{\ldots, \rho_T} \left( c_1(x_1, \xi_1), \ldots, c_T(x_T, \xi_T) \right).
$$  (24)

Let $x_0^*$ and $z^*$ be an optimal first-stage solution and the optimal value of (23)–(24), respectively.

Our first approach, which we refer to as first-level LB, is formed as follows. Consider the collection of subsets $\Omega = \bigcup_{g \in [m]} \Omega_g^{(l)}$ and its induced $\sigma$-algebra $\mathcal{G}$. On each subgroup $g \in [m]$, we solve problem (23)–(24) with sample space $\Omega_g^{(l)}$ where $\mathcal{R}_{\rho_1}$ is replaced by $\mathcal{R}_{\rho_g}^{(l)}$ and let $z_g^{*(l)}$ be its optimal value. Also let $\zeta_{LB} := \{z_g^{*(l)}\}_{g=1}^{m}$ be a $\mathcal{G}$-measurable random variable with probabilities $\pi_g^{(l)}$, $\forall g \in [m_1]$. A LB on $z^*$ can be obtained by applying the LB risk measure $\tilde{\mathcal{R}}_{\rho}^{\mathcal{G}}$ introduced in the previous section to $\zeta_{LB}$, hence computing $\tilde{\mathcal{R}}_{\rho}^{\mathcal{G}}(\zeta_{LB})$.

There is no need to make any changes to the radii from stage 2 to stage $T$; hence the name ‘first-level’ LB. Observe that the scenario tree can be dissected at finer partitions than just the first stage (e.g., a single scenario at every stage $t \in T$), but the convolution is performed only at the first stage.

**Proposition 6. (First-level LB for multistage DRO).** Given problem (23)–(24), assume the risk measure at each stage is induced by a $\phi$-divergence. Consider the risk measure $\tilde{\mathcal{R}}_{\rho}^{\mathcal{G}} : \mathcal{L}_\infty(\Omega, \mathcal{G}, Q) \to \mathbb{R}$ to combine the subgroups and the one-step conditional risk measure on the subgroups $\tilde{\mathcal{R}}_{\rho}^{\mathcal{F}|\mathcal{G}} : \mathcal{L}_\infty(\Omega, \mathcal{F}, Q) \to \mathcal{L}_\infty(\Omega, \mathcal{G}, Q)$ (i.e., $\mathcal{R}_{\rho_g}^{(l)}$ :
sets at subsequent stages are not modified. Furthermore, all the subgroups involved in solution for each subgroup problem

Proof. If \( x^*_g \) is an optimal first-stage solution of (23)—(24), then it is a feasible first-stage solution for each subgroup problem \( g \in [m_l] \). Thus, we have \( c_0(x^*_g, \xi_0) + \tilde{R}_\rho^{(l)}(Q_1(x^*_g, \xi^0)) ) \geq z_g^{(l)} \), \( \forall g \in [m_l] \), or equivalently \( c_0(x^*_g, \xi_0) + \tilde{R}_\rho^{(l)}(Q_1(x^*_g, \xi^0)) ) \geq \zeta_{LB} \). Both sides of this inequality are \( \mathcal{G} \)-measurable, and since \( \tilde{R}_\rho^{(l)} \) is a coherent risk measure that satisfies the monotonicity property (see Section 2.4), we obtain \( \tilde{R}_\rho^{(l)} \left(c_0(x^*_g, \xi_0) + \tilde{R}_\rho^{(l)}(Q_1(x^*_g, \xi^0)) \right) \geq \tilde{R}_\rho^{(l)}(\zeta_{LB}) \)). We can now apply translation equivariance property (see Section 2.4) to the left-hand side of above inequality to get \( \tilde{R}_\rho^{(l)} \left(c_0(x^*_g, \xi_0) + Q_1(x^*_g, \xi^0) \right) \geq \tilde{R}_\rho^{(l)}(\zeta_{LB}) \)). Since the criteria from Propositions 1-4 are satisfied, we obtain \( R_{\rho_l}(c_0(x^*_g, \xi_0) + Q_1(x^*_g, \xi^0)) \geq \tilde{R}_\rho^{(l)}(\zeta_{LB}) \)). Using once more the translation equivariance property, we reach the result \( z^* = c_0(x^*_g, \xi_0) + R_{\rho_l}(Q_1(x^*_g, \xi^0)) \) \( \geq \tilde{R}_\rho^{(l)}(\zeta_{LB}) \)). See also [23].

Remark: The approach just described is also applicable for the case where the risk measure is applied to the whole scenario cost as a time-inconsistent objective function, given as \( R_\rho^{\text{whole}}(\eta_0, \ldots, \eta_T) := \eta_0 + R_\rho(\eta_1 + \eta_2 + \ldots + \eta_T) \).

The first-level bounding scheme cannot be applied in general to the Wasserstein distance (except for a special case discussed below) because we do not have a distance between subgroups in the multistage setting. Thus, we introduce the following definition for a specific way to dissect the scenario tree, following its structure. This dissection reduces the computation of the distance between subgroups in the multistage setting to a recursive application of the distance in the two-stage setting.

Definition 1. (Dissected up to stage \( \tau \)). Let \( \Omega_{\tau,g}^{(l)} \) denote set of nodes of subgroup \( \Omega_g^{(l)} \) at stage \( t \in \mathcal{T} \). A scenario tree \( \mathcal{T} \) is dissected up to stage \( \tau \in \mathcal{T} \setminus \{0\} \) if

1. for every subgroup \( \Omega_g^{(l)} \), \( g \in [m_l] \) all nodes at stage \( \tau \) of that subgroup have the same ancestor, i.e., \( a(n') = a(n''), n', n'' \in \Omega_{g}^{(l)} \), for all subgroups \( g \in [m_l] \);
2. all the children of stage-\( \tau \) (\( \tau < T \)) nodes belong to the same subgroup, i.e., \( \mathcal{B}(n) \subseteq \Omega_{\tau+1,g}^{(l)} \), for all stage-\( \tau \) nodes \( n \in \Omega_{\tau,g}^{(l)} \), \( \tau \neq T \) for all subgroups \( g \in [m_l] \);
3. subgroups are disjoint, i.e., \( \Omega_{\tau,g_1}^{(l)} \cap \Omega_{\tau,g_2}^{(l)} = \emptyset \), for any two \( g_1, g_2 \in [m_l] \), \( g_1 \neq g_2 \).

According to Definition 1, we split the ambiguity sets up to stage \( \tau \) while the ambiguity sets at subsequent stages are not modified. Furthermore, all the subgroups involved in
the splitting of a given ambiguity set at stage \( \tau \) share the same single path up to stage \( \tau - 1 \). For example, in Figure 2b, the scenario tree depicted in Figure 2a is dissected up to stage \( \tau = 2 \). Consequently, subgroups \( \Omega^{(2)}_1 \) and \( \Omega^{(2)}_2 \) share the same nodes 1 and 2 up to stage \( \tau - 1 = 1 \). Similarly, for the subgroups \( \Omega^{(2)}_3 \) and \( \Omega^{(2)}_4 \). Furthermore, according to Definition 1, a tree dissected up to stage \( \tau \in T \setminus \{0, 1\} \) forms a refinement of a tree dissected up to stage \( \tau - 1 \). Compare, for instance, the dissections in Figure 2b and Figure 2c. Similarly, dissection up to stage \( \tau = 1 \) forms a refinement of the tree \( T \).

Notice that Proposition 6 can also be applied to the Wasserstein distance case if the scenario tree is dissected up to stage \( \tau = 1 \) according to Definition 1. In this situation, we are able to compute distances among subgroups because the ambiguity sets of the subsequent stages are unchanged; so it essentially behaves like a two-stage problem at \( \tau = 1 \) (see (23)–(24), where \( Q_1(x_0, \xi^0) \) is viewed as the optimal value function of the second stage). For the other situations, we propose the following multi-level LB scheme, which works for both \( \phi \)-divergences and the Wasserstein distance.

The multi-level LB scheme works as follows. First, the value of \( \tau \) plays an important role. Any stage-\( t \) radii for \( t \leq \tau \) can be changed to \( \bar{\rho}_1, \ldots, \bar{\rho}_\tau \) for subgroup problems, but all radii for stages \( t > \tau \) are kept the same as the original problem \( \rho_{\tau+1}, \ldots, \rho_T \). Then, the optimal values of the subgroups are combined in a nested fashion, following the structure of the tree, using the convolution method described in Section 3.2 starting from stage \( \tau \) up to stage 1 with radii \( \bar{\rho}_\tau, \ldots, \bar{\rho}_1 \). Before we present the theoretical result, let us first illustrate it with an example.

**Example 2.** Consider the scenario tree \( \Xi \) in Figure 2a, which is dissected up to stage \( \tau = 2 \) in Figure 2b. Viewing subgroups \( \Omega^{(2)}_1 \) and \( \Omega^{(2)}_2 \) in Figure 2b as a dissection of the first subset \( \Omega^{(4)}_1 \) in Figure 2c, we apply the convolution method in Section 3.2 using \( \bar{\rho}_2 \) and \( \bar{\rho}_2 \). If these radii satisfy the criteria presented in Sections 3.3 and 3.4, we obtain a LB on the optimal value of the first subset \( \Omega^{(4)}_1 \) in Figure 2c: \( z^{(4)}_1 \leq z^{(4)}_1 \). In the case of Wasserstein distance, the distance between the subgroups \( \Omega^{(2)}_1 \) and \( \Omega^{(2)}_2 \) is obtained by \( \max\{d_{46}=d_{64}, d_{47}=d_{74}, d_{56}=d_{65}, d_{57}=d_{75}\} \), where \( d_{ij} \) denotes the distance between \( \xi_i \) and \( \xi_j \). The same procedure is performed on the other two subgroups \( \Omega^{(2)}_3 \) and \( \Omega^{(2)}_4 \), obtaining a LB for the second subset \( \Omega^{(4)}_2 \) in Figure 2c: \( z^{(4)}_2 \leq z^{(4)}_2 \). Now, instead of \( z^{(4)}_1, z^{(4)}_2 \), using their LBs \( \hat{z}^{(4)}_1, \hat{z}^{(4)}_2 \), the convolution is applied once again to the subsets \( \Omega^{(4)}_1 \) and \( \Omega^{(4)}_2 \) in Figure 2c, which form a dissection of the tree \( \Xi \) in Figure 2a. Here, the distance between subgroups \( \Omega^{(4)}_1 \) and \( \Omega^{(4)}_2 \) is given by \( d_{23} = d_{32} \). Figure 2d shows
Proposition 7. (Multi-level LB for multistage DRO). Let scenario tree $\mathcal{T}$ be dissected up to stage $\tau$ in subgroups $\Omega_g^{(l)}$, $g \in [m_l]$, according to Definition 1. Let $z^*_g^{(l)}$, $g \in [m_l]$ be the optimal values of problem (23)–(24) with sample space $\Omega_g^{(l)}$, where $\rho_t$ is replaced by $\tilde{\rho}_{t,g}$, $t = 1, \ldots, \tau$ and $\tilde{\rho}_{t,max} := \max_{g \in [m_l]} \tilde{\rho}_{t,g}$. Also let $\zeta_{LB} := \{z_g^{(l)}\}_{g=1}^{m_l}$ be a $\mathcal{G}$-measurable random variable with probabilities $\pi_g^{(l)}$, $g \in [m_l]$. If $\tilde{\rho}_{t,g} = \rho_t$, $g \in [m_l]$, for all $t = \tau + 1, \ldots, T$ and $\tilde{\rho}_{t}, \tilde{\rho}_{t,max}$ satisfy the corresponding criteria from one of the Propositions 1-5 with respect to $\rho_t$ for all $t = 1, \ldots, \tau$, then

$$
\tilde{\mathcal{R}}_{\tilde{\rho}_{t}}^{\mathcal{F}_t | \mathcal{F}_0} \left( \tilde{\mathcal{R}}_{\tilde{\rho}_{t}}^{\mathcal{F}_2 | \mathcal{F}_1} \left( \cdots \left( \tilde{\mathcal{R}}_{\tilde{\rho}_{t}}^{\mathcal{G} | \mathcal{F}_{\tau-1}} \left( \zeta_{LB} \right) \right) \right) \right) \leq z^* ,
$$

with $\tilde{\mathcal{R}}_{\tilde{\rho}_{t}}^{\mathcal{G} | \mathcal{F}_{\tau-1}} \left( \zeta_{LB} \right) := \left\{ \tilde{\mathcal{R}}_{\tilde{\rho}_{t}}^{\mathcal{G} | \mathcal{F}_{\tau-1}} \left( \zeta_{LB} \right) \right\}_{g=1}^{\Omega_{\tau-1}}$ and $\zeta_{LB} := \{z_g^{(l)}\}_{g=1}^{m_l} \subseteq \Omega_{\tau-1}$, with $k > l$ and $\Omega_s^{(k)}$, $s \in [[\Omega_{\tau-1}]]$ dissection of the scenario tree $\mathcal{T}$ up to stage $\tau - 1$.

Proof. Recall $\Omega_g^{(l)}$, $g \in [m_l]$ is a dissection of the scenario tree $\mathcal{T}$ up to stage $\tau$ and $\Omega_{\tau-1}$ denotes the set of nodes in the original tree $\mathcal{T}$ at stage $\tau - 1$. Let $\Omega_s^{(k)}$, $s \in [[\Omega_{\tau-1}]]$ be a dissection of the scenario tree $\mathcal{T}$ up to stage $\tau - 1$. Fix an $s$. Let $x_{i,s}^*$, $i = 0, \ldots, \tau - 1$, the
optimal solutions of subgroup \( \Omega_s^{(k)} \), be given. Then, they form a feasible solution for each subgroup problem \( \Omega_g^{(l)} \subseteq \Omega_s^{(k)} \) at stages \( 1, \ldots, \tau - 1 \). Thus, for all \( \Omega_g^{(l)} \subseteq \Omega_s^{(k)} \),

\[
c_0(x_0^s, \xi_0) + R_{\tilde{\rho}_s}^{(l)} \left( c_1(x_1^s, \xi_1) + R_{\tilde{\rho}_s}^{(l)} \left( \ldots + R_{\tilde{\rho}_s}^{(l)}(Q_{\tau-1}(x_{\tau-1}^s, \xi_{\tau-1})) \right) \right) \geq z_g^{(l)}
\]

with \( Q_{\tau}(x_{\tau-1}^s, \xi_{\tau-1}) := \min \{ R_{\rho_{\tau+1}, \ldots, \rho_T}(c_T(x_T, \xi_T)) : x_T \in X_T(x_{\tau-1}^s, \xi_{\tau-1}) \} \). Equivalently defining \( \zeta_s^{(l)} : = \{ z_g^{(l)} \} \Omega_g^{(l)} \subseteq \Omega_s^{(k)} \), \( \tilde{\rho}_s^* := \{ \tilde{\rho}_g \} \Omega_s^{(k)} \subseteq \Omega_s^{(k)} \) and \( \mathcal{H}_s^s \) the \( \sigma \)-algebra induced by the collection of subsets \( \Omega_s^{(k)} = \bigcup \Omega_g^{(l)} \subseteq \Omega_s^{(k)} \), we have

\[
c_0(x_0^s, \xi_0) + \tilde{R}_{\tilde{\rho}_s}^{(k)} \left( c_1(x_1^s, \xi_1) + \tilde{R}_{\tilde{\rho}_s}^{(k)} \left( \ldots + \tilde{R}_{\tilde{\rho}_s}^{(k)}(Q_{\tau-1}(x_{\tau-1}^s, \xi_{\tau-1})) \right) \right) \geq \zeta_s^{(k)} L_B,
\]

where we removed notation like \( \mathcal{F}_{\tau-1} | \mathcal{H}_s^s \) on \( \tilde{R}^{(k)} \) for simplicity. Both sides of the inequality are \( \mathcal{H}_s^s \)-measurable. Then, since \( \tilde{R}_{\tilde{\rho}^*}^{\mathcal{H}_s^s} \) is a coherent risk measure that satisfies monotonicity, we obtain

\[
\tilde{R}_{\tilde{\rho}^*}^{\mathcal{H}_s^s} \left( c_0(x_0^s, \xi_0) + \tilde{R}_{\tilde{\rho}_s}^{(k)} \left( c_1(x_1^s, \xi_1) + \tilde{R}_{\tilde{\rho}_s}^{(k)} \left( \ldots + \tilde{R}_{\tilde{\rho}_s}^{(k)}(Q_{\tau-1}(x_{\tau-1}^s, \xi_{\tau-1})) \right) \right) \right) \geq \tilde{R}_{\tilde{\rho}^*}^{\mathcal{H}_s^s}(\zeta_s^{(k)} L_B).
\]

Since by hypothesis \( \tilde{\rho}_s, \tilde{\rho}_s \cdot \max \) satisfy the criteria from one of the the Propositions 1-5 with respect to \( \rho_s \), by translation equivariance, and since by Definition 1 there is a single path up to stage \( \tau - 1 \), we get

\[
z_s^{(k)} = \sum_{i=1}^{\tau-1} c_i(x_i^s, \xi_i) + R_{\tilde{\rho}_s}^{(k)} \left( Q_{\tau-1}(x_{\tau-1}^s, \xi_{\tau-1}) \right) \geq \tilde{R}_{\tilde{\rho}_s}^{\mathcal{H}_s^s}(\zeta_s^{(k)} L_B) := \tilde{z}_s^{(k)}.
\]

Repeating for all \( s \in \{ \Omega_{\tau-1} \} \), we obtain

\[
\begin{bmatrix} z_s^{(k)} \end{bmatrix}^T \geq \begin{bmatrix} \tilde{z}_s^{(k)} \end{bmatrix}^T = \tilde{R}_{\tilde{\rho}_s}^{\mathcal{H}_s^s}(\zeta_s^{(k)} L_B),
\]

where \([ \cdot ]^T\) denotes transpose. Let \( \Omega_{\tau-1}^{(j)} d \in \{ \Omega_{\tau-2} \} \) be a dissection of the scenario tree \( \mathcal{S} \) up to stage \( \tau - 2 \). Let \( x_i^{s,d}, i = 0, \ldots, \tau - 2 \), the optimal solutions of subgroup \( \Omega_{\tau-1}^{(j)} d \), be given. Then, they form a feasible solution for each subgroup problem \( \Omega_{\tau-1}^{(j)} d \). Following the steps above and defining \( \zeta_{d,s}^{(j)} = \{ z_s^{(j)} \} \Omega_s^{(k)} \subseteq \Omega_d^{(j)} \) and \( \mathcal{H}_s^s \) the \( \sigma \)-algebra induced by the collection of subsets \( \Omega_{\tau-1}^{(j)} d = \bigcup \Omega_s^{(k)} \subseteq \Omega_d^{(j)} \), we reach

\[
z_d^{(j)} = \sum_{i=0}^{\tau-2} c_i(x_i^{s,d}, \xi_i) + R_{\tilde{\rho}_{\tau-1}}^{(j)} \left( Q_{\tau-1}(x_{\tau-2}^{s,d}, \xi_{\tau-2}) \right) \geq \tilde{R}_{\tilde{\rho}_{\tau-1}}^{\mathcal{H}_s^s}(\zeta_{d,s}^{(j)} L_B) := \tilde{z}_d^{(j)}.
\]

Repeating for all \( d \in \{ \Omega_{\tau-2} \} \), we obtain

\[
\begin{bmatrix} z_1^{(j)} \end{bmatrix}^T \geq \begin{bmatrix} \tilde{z}_1^{(j)} \end{bmatrix}^T = 25
\]
Repeating the same procedure going backward for other \( \tau - 2 \) times, the result follows.

Remark: When the multi-level LB scheme is applied to the Wasserstein case, the distance at time \( \tau \) between groups \( g_1, g_2 \) is computed as follows:

\[
d_{g_1, g_2} = \begin{cases} 
\max_{i \in \Omega_{r,g_1}, k \in \Omega_{r,g_2}} \{d_{ik}\} & \text{when } g_1 \neq g_2 \\
0 & \text{when } g_1 = g_2,
\end{cases}
\]

where \( g_1 \) and \( g_2 \) are chosen such that \( \forall n_1 \in \Omega_{r,g_1}^{(l)}, n_2 \in \Omega_{r,g_2}^{(l)} : a(n_1) = a(n_2) \) and \( d_{ik} \) is the distance between nodes \( i \) and \( k \). Analogously, going backward for any \( t = \tau - 1, \ldots, 1 \).

3.6 Upper bounds for multistage problems

Finding an UB of an optimization problem is of critical importance when an optimal solution is not available. In general, UBs are obtained by constraining some decision variables to be equal to pre-determined fixed values. In this paper, UBs are obtained by using optimal solutions of scenario subproblems. Using the procedure described before, we solve each single scenario group \( \Omega_{r}^{(l)}, g \in [\Omega_T] \) obtaining \((\hat{x}_0, g, \ldots, \hat{x}_T, g)\) as its optimal solution. Let \( UB_{g}, t \in T \) be the optimal value of the original problem (1) where the variables up to stage \( t \) are set to \( \hat{x}_{i,g} \) for \( i = 0, \ldots, t \). From an algorithmic perspective, this approach requires us to solve problems of smaller dimension than the original one. The best available UB is obtained by taking the minimum value of \( UB_{g} \) over all \( g \in [\Omega_T] \), i.e., \( UB := \min_{g \in [\Omega_T]} UB_{g} \). See [20] for the formal definition and the proof.

4 Case study: a multistage production problem

4.1 Problem Description

To show the effectiveness of the proposed approach, we consider a mixed-integer variant of the inventory management problem introduced in [22], which now includes binary variables to indicate machinery start-ups. The problem can be summarized as follows. Consider a single product inventory system, comprised of a warehouse and a factory equipped with production machinery. At each time step \( t = 0, \ldots, T - 1 \), production can be performed by starting up machinery. Random demands coming from customers have to be satisfied from the existing inventory. If the random demand exceeds the stock, it will be satisfied by
rapid orders from a different source that come at a higher price. The goal is to minimize the total costs of the factory for the entire planning period.

We assume that the distribution of the scenario process is described by a six-stage \((T = 5)\) scenario tree with 5 branches from the root, 4 from each of the second stage nodes, and 3 from each of the third, fourth, fifth stages nodes, resulting in \(|\Omega_T| = 5 \times 4 \times 3 \times 3 \times 3 = 540\) scenarios and 806 nodes. The scenario tree and the nominal distribution are generated as a dependent process across stages; see [21] for details.

The problem formulation is similar to the one presented in [22], except for (i) the binary variables at each node of the scenario tree indicating machine start-ups, (ii) the corresponding start-up costs, and (iii) the typical changes in the constraints that ensure no production takes place if machinery is not started in that period. Below we note changes from [22] in the data used. The value of the demand at the root node \((n = 1)\) is \(\xi_1 = 65\).

At each period \(t = 0, 1, 2, 3, 4\), the maximal production capacity of the factory is 567 units and the machinery start-up cost is \(k_t = 75\). The initial inventory is 10, the final value of the inventory is 2 per unit, and the values of production price \(c_t\), selling price \(s_t\), inventory holding cost \(h_t\) and procurement cost \(b_t\) at time period \(t\) are presented in Table 3.

<table>
<thead>
<tr>
<th>(t)</th>
<th>0</th>
<th>1</th>
<th>2</th>
<th>3</th>
<th>4</th>
<th>5</th>
</tr>
</thead>
<tbody>
<tr>
<td>(c_t)</td>
<td>3.5</td>
<td>3.6</td>
<td>2.3</td>
<td>2.8</td>
<td>3</td>
<td>-</td>
</tr>
<tr>
<td>(s_t)</td>
<td>-</td>
<td>10.7</td>
<td>10.5</td>
<td>10.9</td>
<td>10.6</td>
<td>10</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>(t)</th>
<th>0</th>
<th>1</th>
<th>2</th>
<th>3</th>
<th>4</th>
<th>5</th>
</tr>
</thead>
<tbody>
<tr>
<td>(h_t)</td>
<td>2</td>
<td>1.9</td>
<td>2.1</td>
<td>2.2</td>
<td>2.1</td>
<td>-</td>
</tr>
<tr>
<td>(b_t)</td>
<td>-</td>
<td>4</td>
<td>3.1</td>
<td>4.9</td>
<td>7</td>
<td>7.5</td>
</tr>
</tbody>
</table>

Table 3: Production price \(c_t\), selling price \(s_t\), holding cost \(h_t\) from time \(t\) to time \(t + 1\), and procurement cost \(b_t\) for extra stock from another retailer at time \(t\).

### 4.2 Computation of bounds

This section presents computational results on a DRO version of the production problem described above, considering different ambiguity sets using \(\phi\)-divergences (VD and the modified \(\chi^2\) distance) and the Wasserstein distance. All the considered multistage DRO are implemented in a nested fashion, and we use the same value of the radii \(\rho_t = 0.50\) over all stages \(t \in T \setminus \{0\}\). The problems derived from our case study were solved under AMPL environment using the CPLEX solver 12.8.0.0. Computations have been performed on a 64-bit machine with 8 GB of RAM and a 1.8 GHz Intel i7 processor.

#### 4.2.1 Variation distance case

Table 4 lists the LBs obtained by applying Proposition 6 (first-level LB) using VD. We choose subgroups \(\Omega_{l}^{(l)}\) to be disjoint with \(l = 1, 3, 9, 27, 54, 108\). The instance \(l = 540\)
refers to the original problem, which we report as a benchmark. The first group of bounds \((l = 108, m_l = 5)\) has been obtained by solving \(m_l = 5\) subproblems, each composed of \(l = 108\) consecutive scenarios. The subgroups mainly follow the structure of the scenario tree. Only at \(l = 54\), we split the 5 scenarios at \(t = 1\) individually and group the 4 scenarios at \(t = 2\) two by two consecutively. When \(l = 1\) and \(m_l = 540\), each subgroup forms a deterministic problem with only one scenario. According to Proposition 2, we set LBs (column \(LB\)) 4 times faster overall computation time.

Results show monotonic increases in CPU time per subproblem with both the dimension \(l\) and \(z\). From numerical results of Table 4 obtained considering disjoint subgroups, we observe that the tightest bounds are achieved for greater values of \(\bar{\rho}\) and \(l\), at the cost of increasing CPU running times. Indeed, overall, the best calculated LB is given by \(-1528.74\) (at \(\bar{\rho} = 0.50, \rho_{\max} = 0.00\) when \(l = 108\) and \(m_l = 5\)), while the worst LB is given by \(-1836.90\) (at \(\bar{\rho} = 0.00, \rho_{\max} = 0.50\) when \(l = 1\) and \(m_l = 540\), which is the partition into atoms). Results show monotonic increases in CPU time per subproblem with both the dimension of each subproblem (cardinality \(l\)) and the values of \(\bar{\rho}\). These results also show that very high-quality LBs can be obtained, saving considerable time with respect to the original DRO problem. For example, when \(l = 54\) a %GAP = \(-0.50\%\) can be achieved with about 4 times faster overall computation time.

<table>
<thead>
<tr>
<th>(l)</th>
<th>(m_l)</th>
<th>(\bar{\rho})</th>
<th>(\rho_{\max})</th>
<th>(LB)</th>
<th>CPU time overall</th>
<th>CPU time per subpr.</th>
<th>%GAP</th>
</tr>
</thead>
<tbody>
<tr>
<td>108</td>
<td>1</td>
<td>0.00</td>
<td>0.00</td>
<td>-1528.74</td>
<td>14.255</td>
<td>14.255</td>
<td>-1.79%</td>
</tr>
<tr>
<td>108</td>
<td>2</td>
<td>0.25</td>
<td>0.00</td>
<td>-1930.46</td>
<td>7.791</td>
<td>11.920</td>
<td>-5.84%</td>
</tr>
<tr>
<td>108</td>
<td>3</td>
<td>0.50</td>
<td>0.00</td>
<td>-2206.74</td>
<td>1.808</td>
<td>1.808</td>
<td>-2.94%</td>
</tr>
<tr>
<td>108</td>
<td>5</td>
<td>0.00</td>
<td>0.50</td>
<td>-1544.96</td>
<td>4.669</td>
<td>1.994</td>
<td>-1.79%</td>
</tr>
<tr>
<td>54</td>
<td>10</td>
<td>0.00</td>
<td>0.50</td>
<td>-1391.61</td>
<td>4.284</td>
<td>0.328</td>
<td>-1.43%</td>
</tr>
<tr>
<td>54</td>
<td>10</td>
<td>0.25</td>
<td>0.50</td>
<td>-1571.46</td>
<td>3.418</td>
<td>0.348</td>
<td>-2.23%</td>
</tr>
<tr>
<td>27</td>
<td>20</td>
<td>0.00</td>
<td>0.50</td>
<td>-1531.07</td>
<td>4.564</td>
<td>0.358</td>
<td>-1.50%</td>
</tr>
<tr>
<td>27</td>
<td>20</td>
<td>0.25</td>
<td>0.50</td>
<td>-1575.28</td>
<td>2.344</td>
<td>0.117</td>
<td>-1.90%</td>
</tr>
<tr>
<td>27</td>
<td>20</td>
<td>0.50</td>
<td>0.50</td>
<td>-1544.94</td>
<td>2.850</td>
<td>0.115</td>
<td>-1.34%</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>(l)</th>
<th>(m_l)</th>
<th>(\bar{\rho})</th>
<th>(\rho_{\max})</th>
<th>(LB)</th>
<th>CPU time overall</th>
<th>CPU time per subpr.</th>
<th>%GAP</th>
</tr>
</thead>
<tbody>
<tr>
<td>60</td>
<td>0</td>
<td>0.00</td>
<td>0.50</td>
<td>-1703.05</td>
<td>3.266</td>
<td>0.054</td>
<td>-11.79%</td>
</tr>
<tr>
<td>60</td>
<td>0</td>
<td>0.25</td>
<td>0.50</td>
<td>-1692.46</td>
<td>3.025</td>
<td>0.050</td>
<td>-7.15%</td>
</tr>
<tr>
<td>60</td>
<td>0</td>
<td>0.50</td>
<td>0.50</td>
<td>-1681.98</td>
<td>4.109</td>
<td>0.068</td>
<td>-3.84%</td>
</tr>
<tr>
<td>154</td>
<td>0</td>
<td>0.00</td>
<td>0.50</td>
<td>-1836.90</td>
<td>8.484</td>
<td>0.194</td>
<td>-16.90%</td>
</tr>
<tr>
<td>154</td>
<td>0</td>
<td>0.25</td>
<td>0.50</td>
<td>-1835.67</td>
<td>9.061</td>
<td>0.206</td>
<td>-10.76%</td>
</tr>
<tr>
<td>154</td>
<td>0</td>
<td>0.50</td>
<td>0.50</td>
<td>-1822.65</td>
<td>13.359</td>
<td>0.674</td>
<td>-6.54%</td>
</tr>
<tr>
<td>154</td>
<td>0</td>
<td>0.75</td>
<td>0.50</td>
<td>-1836.90</td>
<td>20.836</td>
<td>0.308</td>
<td>-20.17%</td>
</tr>
<tr>
<td>154</td>
<td>0</td>
<td>0.00</td>
<td>0.50</td>
<td>-1066.70</td>
<td>27.141</td>
<td>0.050</td>
<td>-8.74%</td>
</tr>
</tbody>
</table>

Table 4: Collections of LBs with disjoint subsets obtained by applying Proposition 6 (first-level LB) to the multistage inventory problem with VD.

Table 5 provides detailed results obtained by keeping the worst scenario \((\omega_1)\) fixed in all subsets \(\Omega_{\bar{\rho}}^{(1)}\). VD focuses on a convex combination of CVaR and the worst case [16, 31]. So when the worst-case scenario is fixed at each subgroup, we may get better LBs. The cardinality \(l\) of each subproblem has been chosen to have the ratio \(m_l = \frac{540-1}{l-1} \in \mathbb{N}\) and,
specifically, we consider the values \( m_l \in \{2, 8, 12, 50, 78\} \). The combinations \((\bar{\rho}, \bar{\rho}_{\text{max}})\) are chosen as for the disjoint case. Results show that fixing the worst scenario improves the quality of the LB for those partitions with small cardinalities \( l \) and greater values of \( \bar{\rho}_{\text{max}} \) (see for instance \( l = 1, 3, 9 \) in Table 4 and \( l = 2, 8 \) in Table 5). It is also interesting to notice that when the worst scenario is fixed, although the tightest bounds are still obtained for greater values of \( l \), tighter bounds are obtained by setting progressively smaller value of \( \bar{\rho} \) and larger \( \bar{\rho}_{\text{max}} \) (see, for instance, \( l = 50 \) and \( l = 78 \) in Table 5).

<table>
<thead>
<tr>
<th>( l )</th>
<th>( m_l )</th>
<th>( \bar{\rho} )</th>
<th>( \bar{\rho}_{\text{max}} )</th>
<th>CPU time overall</th>
<th>CPU time per edge</th>
<th>%GAP</th>
</tr>
</thead>
<tbody>
<tr>
<td>12</td>
<td>49</td>
<td>0.00</td>
<td>0.00</td>
<td>1674.96</td>
<td>4.508</td>
<td>0.092</td>
</tr>
<tr>
<td>8</td>
<td>77</td>
<td>0.00</td>
<td>0.00</td>
<td>1656.94</td>
<td>5.708</td>
<td>0.114</td>
</tr>
<tr>
<td>2</td>
<td>539</td>
<td>0.00</td>
<td>0.00</td>
<td>1650.21</td>
<td>39.590</td>
<td>0.073</td>
</tr>
</tbody>
</table>

Table 5: Collections of LBs obtained by keeping the worst scenario \((\omega_l)\) fixed in all subsets and applying Proposition 6 (first-level LB) to the multistage inventory problem with VD.

### 4.2.2 Modified \( \chi^2 \) case

In Table 6, we construct collections of LBs applying Proposition 6 (first-level LB) to the multistage inventory problem with modified \( \chi^2 \) distance. Subsets \( \Omega^{(l)}_g \) are chosen to be disjoint, following the structure of the scenario tree with \( l = 1, 3, 9, 27, 54, 108 \) as before. According to Proposition 1 we set \( \bar{\rho}_g = \bar{\rho}_{\text{max}}, g \in [m_l] \) and choose the combinations \((\bar{\rho}, \bar{\rho}_{\text{max}})\) with \( \bar{\rho} \in \{0.00, 0.25, 0.50\} \) and \( \bar{\rho}_{\text{max}} = \frac{m_l - \bar{\rho}}{1 + \bar{\rho}} \). The overall problem, i.e., the full tree with 540 scenarios, was unsolvable within a time limit of 86400 CPU seconds (or 24 hours) and values of percentage deviations (%GAP) with respect to the optimal value could not be explicitly computed. Therefore, to measure the quality of the obtained LB, a new optimality gap is computed as follows: \( \%\text{GAP}^* = \frac{LB^* - LB}{LB^*} \cdot 100 \), with \( LB^* \) representing the best observed LB. Being a problem too large to be solved exactly within the prespecified time limit, the bounding methodology proposed in this paper is particularly helpful. It is worth noting that when \( l = 108 \) the solver also could not solve the subproblems within the time limit, and therefore \( l = 108 \) results are not reported for ease of presentation.

From the numerical results given in Table 6, we observe that regardless of the cardinality \( l \) of each subproblem, the best strategy to get tighter LBs is to set the cardinality \( l \) and \( \bar{\rho} \) as large as possible. Indeed, overall, the best calculated LB is given by \(-1497.95 \) (at \( \bar{\rho} = 0.50, \bar{\rho}_{\text{max}} = 0.00 \) when \( l = 54 \) and \( m_l = 10 \)), although it requires considerable effort.
in terms of CPU time (25395.048 sec.s overall). A drastic reduction in computation time can be obtained by using smaller subgroups of cardinality \( l = 27 \) without sacrificing the quality of the LB too much. Again by setting \( \tilde{\rho} = 0.50, \tilde{\rho}_{\text{max}} = 0.00 \), a LB within 1.82% of the best LB is obtained in approximately 23.5 times faster overall computation time. On the other hand, the worst LB is given by \(-1836.90 \) (at \( \tilde{\rho} = 0.00, \tilde{\rho}_{\text{max}} = 0.50 \) when \( l = 1 \) and \( m_l = 540 \) in just 211.031 CPU seconds. Results also show monotonic increases in CPU time per subproblem with both the dimension \( l \) of each subproblem and the values of \( \tilde{\rho} \).

<table>
<thead>
<tr>
<th>( l )</th>
<th>( m_l )</th>
<th>( \tilde{\rho} )</th>
<th>( \tilde{\rho}_{\text{max}} )</th>
<th>LB</th>
<th>CPU time overall</th>
<th>CPU time per subprob</th>
<th>%GAP</th>
</tr>
</thead>
<tbody>
<tr>
<td>54</td>
<td>10</td>
<td>0.25</td>
<td>0.25</td>
<td>0.25</td>
<td>0.25</td>
<td>0.25</td>
<td>0.25</td>
</tr>
<tr>
<td>27</td>
<td>20</td>
<td>0.50</td>
<td>0.00</td>
<td>0.50</td>
<td>0.00</td>
<td>0.50</td>
<td>0.00</td>
</tr>
</tbody>
</table>

Table 6: Collections of LBs with disjoint subsets obtained by applying Proposition 6 (first-level LB) to the multistage inventory problem with mod. \( \chi^2 \) (time limit = 86400 CPU sec.s).

We now apply the bounding scheme proposed in Proposition 7 (multi-level LB) to the multistage inventory problem with modified \( \chi^2 \) distance. The results are reported in Table 7. Following this partition method, regardless of the cardinality \( l \) of each subproblem and given \( \tau \), the last stage where the scenario tree is partitioned, the best strategy to get tighter LBs is to set \( \tilde{\rho}_l = \rho_l, \tau = 1, \ldots, \tau - 1 \) (which are therefore the only results we report, for ease of exposition). We allow, instead, changes in \( \tilde{\rho}_\tau \) and \( \tilde{\rho}_{\tau,\text{max}} \) taking values 0.00, 0.25 and 0.50. Overall, the best LB is given by \(-1504.41 \), obtained by setting \( \tilde{\rho}_2 = 0.25 \) and \( \tilde{\rho}_{2,\text{max}} = 0.20 \) when \( l = 54, m_l = 10 \) and \( \tau = 2 \). Comparing these results with bounds of Table 6, we conclude that the multi-level bounding technique becomes particularly useful by allowing to get tighter LBs as partitions progressively contain smaller-dimensional subproblems. For instance, at lower values of \( l \), better LBs are obtained with similar overall computation times.

Given that the problem was too large to be solved exactly, to evaluate the quality of obtained LBs, we resort to the upper bounding methodology described in Section 3.6. The only UBs we were able to compute were \( UB^3 = -1453.29 \) (within 72131.000 CPU seconds) and \( UB^3 = -1411.91 \) (within 125.844 CPU seconds). Although \( UB^3 \) performs better than \( UB^4 \), it requires a considerable larger computational effort. All the other UBs (\( UB^3, \ t = 0,1,2 \)) went out of memory because the number of fixed variables was not enough to reduce the dimension of the scenario tree to a computationally tractable
In Table 8, we construct collections of LBs by applying Proposition 7 (multi-level LB) for the Wasserstein distance. Given \( \tau \), the last stage where the tree is dissected, according to Proposition 7 we set \( \bar{\rho}_t = \rho_t, t = 1, \ldots, \tau - 1 \) and choose the combinations \((\bar{\rho}_t, \tilde{\rho}_{\tau, \text{max}})\) with \(\tilde{\rho}_t \in \{0.00, 0.25, 0.50\}\) and \(\tilde{\rho}_{\tau, \text{max}} = \rho_{\tau} - \bar{\rho}_t\). For \( t = \tau + 1, \ldots, T \) we set \(\tilde{\rho}_1 = 0.00\), \(g \in [m_l]\). When \( \tau = 1 \) we work directly with \((\tilde{\rho}_1, \tilde{\rho}_{1, \text{max}})\), see for instance Table 8 with

<table>
<thead>
<tr>
<th>( l )</th>
<th>( m_l )</th>
<th>( \tau )</th>
<th>( (\bar{\rho}<em>t)</em>{t=1}^{\tau-1} )</th>
<th>( (\tilde{\rho}<em>{\tau, \text{max}})</em>{t=1}^{\tau-1} )</th>
<th>( \tilde{\rho}_t )</th>
<th>( \tilde{\rho}_{\tau, \text{max}} )</th>
<th>( \bar{\rho}_t )</th>
<th>( LB )</th>
<th>CPU time overall</th>
<th>CPU time per subpr.</th>
<th>%GAP*</th>
</tr>
</thead>
<tbody>
<tr>
<td>54</td>
<td>10</td>
<td>2</td>
<td>0.50</td>
<td>0.00</td>
<td>0.00</td>
<td>0.50</td>
<td>-1586.64</td>
<td>2731.188</td>
<td>2731.119</td>
<td>-2.35%</td>
<td></td>
</tr>
<tr>
<td>27</td>
<td>20</td>
<td>2</td>
<td>0.50</td>
<td>0.00</td>
<td>0.00</td>
<td>0.50</td>
<td>-1571.74</td>
<td>1002.295</td>
<td>1002.115</td>
<td>-1.93%</td>
<td></td>
</tr>
<tr>
<td>9</td>
<td>60</td>
<td>3</td>
<td>0.50</td>
<td>0.00</td>
<td>0.25</td>
<td>0.25</td>
<td>-1538.46</td>
<td>1571.034</td>
<td>1591.847</td>
<td>-4.70%</td>
<td></td>
</tr>
<tr>
<td>3</td>
<td>180</td>
<td>4</td>
<td>0.50</td>
<td>0.00</td>
<td>0.00</td>
<td>0.50</td>
<td>-1503.95</td>
<td>1583.381</td>
<td>1583.381</td>
<td>-0.95%</td>
<td></td>
</tr>
<tr>
<td>1</td>
<td>540</td>
<td>5</td>
<td>0.50</td>
<td>0.00</td>
<td>0.25</td>
<td>0.25</td>
<td>-1538.21</td>
<td>1001.105</td>
<td>1001.105</td>
<td>-0.89%</td>
<td></td>
</tr>
</tbody>
</table>

Table 8: Collections of LBs with disjoint subsets obtained by applying Proposition 7 (multi-level LB) to the multistage inventory problem with Wasserstein distance.

### 4.2.3 Wasserstein distance case

In Table 8, we construct collections of LBs by applying Proposition 7 (multi-level LB) for the Wasserstein distance. Given \( \tau \), the last stage where the tree is dissected, according to Proposition 7 we set \( \bar{\rho}_t = \rho_t, t = 1, \ldots, \tau - 1 \) and choose the combinations \((\bar{\rho}_t, \tilde{\rho}_{\tau, \text{max}})\) with \(\tilde{\rho}_t \in \{0.00, 0.25, 0.50\}\) and \(\tilde{\rho}_{\tau, \text{max}} = \rho_{\tau} - \bar{\rho}_t\). For \( t = \tau + 1, \ldots, T \) we set \(\tilde{\rho}_1 = 0.00\), \(g \in [m_l]\). When \( \tau = 1 \) we work directly with \((\tilde{\rho}_1, \tilde{\rho}_{1, \text{max}})\), see for instance Table 8 with

<table>
<thead>
<tr>
<th>( l )</th>
<th>( m_l )</th>
<th>( \tau )</th>
<th>( (\bar{\rho}<em>t)</em>{t=1}^{\tau-1} )</th>
<th>( (\tilde{\rho}<em>{\tau, \text{max}})</em>{t=1}^{\tau-1} )</th>
<th>( \tilde{\rho}_t )</th>
<th>( \tilde{\rho}_{\tau, \text{max}} )</th>
<th>( \bar{\rho}_t )</th>
<th>( LB )</th>
<th>CPU time overall</th>
<th>CPU time per subpr.</th>
<th>%GAP*</th>
</tr>
</thead>
<tbody>
<tr>
<td>54</td>
<td>10</td>
<td>2</td>
<td>0.50</td>
<td>0.00</td>
<td>0.00</td>
<td>0.50</td>
<td>-1586.64</td>
<td>2731.188</td>
<td>2731.119</td>
<td>-2.35%</td>
<td></td>
</tr>
<tr>
<td>27</td>
<td>20</td>
<td>2</td>
<td>0.50</td>
<td>0.00</td>
<td>0.00</td>
<td>0.50</td>
<td>-1571.74</td>
<td>1002.295</td>
<td>1002.115</td>
<td>-1.93%</td>
<td></td>
</tr>
<tr>
<td>9</td>
<td>60</td>
<td>3</td>
<td>0.50</td>
<td>0.00</td>
<td>0.25</td>
<td>0.25</td>
<td>-1538.46</td>
<td>1571.034</td>
<td>1591.847</td>
<td>-4.70%</td>
<td></td>
</tr>
<tr>
<td>3</td>
<td>180</td>
<td>4</td>
<td>0.50</td>
<td>0.00</td>
<td>0.00</td>
<td>0.50</td>
<td>-1503.95</td>
<td>1583.381</td>
<td>1583.381</td>
<td>-0.95%</td>
<td></td>
</tr>
<tr>
<td>1</td>
<td>540</td>
<td>5</td>
<td>0.50</td>
<td>0.00</td>
<td>0.25</td>
<td>0.25</td>
<td>-1538.21</td>
<td>1001.105</td>
<td>1001.105</td>
<td>-0.89%</td>
<td></td>
</tr>
</tbody>
</table>

Table 8: Collections of LBs with disjoint subsets obtained by applying Proposition 7 (multi-level LB) to the multistage inventory problem with Wasserstein distance.
\( l = 108 \) and \( m_{l} = 5 \). The overall problem, i.e., the full tree with 540 scenarios, is solved within 85.810 seconds and with optimal value \( z^* = -1706.63 \).

From the results in Table 8, we observe that, overall, the best calculated LB is given by \(-1711.93\) (obtained by setting \( \bar{\bar{\rho}}_1 = 0.50 \), \( \bar{\rho}_{1,max} = 0.00 \) when \( l = 108 \) and \( m_{l} = 5 \)). On the contrary, the worst LB is given by \(-1808.79\) (at \( \bar{\bar{\rho}}_5 = 0.00 \), \( \bar{\rho}_{5,max} = 0.50 \) when \( l = 1 \) and \( m_{l} = 540 \)). Results still show monotonic increases in CPU time per subproblem with both the dimension \( l \) of each subproblem and the values of \( \bar{\rho}_t \) for this problem. We observe that improvements on LBs as either \( l \) or \( \bar{\rho}_t \) increases, while still present, are less pronounced in this case.

### 4.3 Discussion

Some insights gained from the numerical experiments are as follows. Generally, the greater the number of scenarios per subproblem, the sharper the obtained LBs. For the first-level bounding scheme, when subtrees have a single node at time \( t = 1 \), it is best not to waste any of the robustness budget \( \rho_1 \) on \( \bar{\rho}_{\text{max}} \). This is because, for the subproblem at time \( t = 1 \) there is not really an ambiguity set of distributions with a single scenario. Thus, in this case, the best LBs are obtained by setting \( \bar{\rho}_{\text{max}} = 0 \) and using the largest possible value of \( \bar{\rho} \). This can be seen, e.g., in Table 4 and Table 6 where the 5 scenarios of the original tree at stage \( t = 1 \) are always split in a single-scenario manner. More generally, when subtrees have multiple nodes at time \( t = 1 \), numerical results show that more importance should be assigned to \( \bar{\rho} \) at the expense of \( \bar{\rho}_{\text{max}} \) as the cardinality \( l \) of subgroups decreases (e.g., Table 5 when \( l = 2, 8, 12 \)), while progressively more importance should be assigned to \( \bar{\rho}_{\text{max}} \) at the expense of \( \bar{\rho} \) as the cardinality \( l \) of subgroups increases (e.g., Table 5 when \( l = 50, 78 \)).

Similarly, for the multi-level bounding scheme, when subtrees have a single node at time \( \tau \), it is best not to waste any of the robustness budget \( \rho_{\tau} \) on \( \bar{\rho}_{\tau,\text{max}} \) and use the largest possible value of \( \bar{\rho}_{\tau} \). This can be seen, e.g., in Table 7 and Table 8 when \( l = 1, 3, 9, 27, 108 \). Contrariwise, when there are multiple nodes at time \( \tau \), more importance should be assigned to \( \bar{\rho}_{\tau,\text{max}} \) at the expense of \( \bar{\rho}_{\tau} \) (see, e.g., Table 7 and Table 8 when \( l = 54 \)). For dissections with smaller cardinality \( l \), the multi-level bounding scheme described in Proposition 7 appears to be more effective than the first-level bounding scheme described in Proposition 6, which is instead more useful when the number of scenarios in each subgroup is larger. At smaller cardinalities \( l \) (hence larger values of \( \tau \)), the multi-level scheme seems to provide
more opportunities to improve the LB, instead of only once as in the first-level scheme.

We observe some gains in fixing a worst-case scenario using VD at small cardinalities \( l \) and higher values of \( \bar{\rho}_{max} \), with a slight increase in computation time due to having slightly larger subproblems to solve. This strategy can be useful, e.g., when using \( \phi \)-divergences that can pop scenarios [3]. Divergences that can pop scenarios (like VD and the CR power divergence with \( \theta < 1 \) and \( \theta \to 0 \)) can make the worst-case scenarios to have positive probabilities even if they have a nominal probability of zero. Thus, in large-scale versions of such problems, when small cardinalities \( l \) are needed due to computational bottleneck, it may be possible to obtain better LBs by fixing the worst-case scenarios. Finally, even though we empirically observed monotonicity of LBs in the subgroups’ cardinality \( l \) for fixed values of \( (\bar{\rho}, \bar{\rho}_{max}) \), we found cases (not shown here for brevity) where monotonicity in \( l \) is not satisfied. This is in contrast to the risk-neutral stochastic optimization setting, where the monotonicity of the LBs in \( l \) is guaranteed [20].

5 Conclusions

In this paper new LB criteria for multistage mixed-integer DRO problems—formed by creating ambiguity sets associated with various commonly used \( \phi \)-divergences and the Wasserstein distance on a finite scenario tree—are derived. Conditions on the way the scenario tree is dissected and the convolutions are formed to ensure a LB on the optimal value are established. The scenario tree can be dissected either by disjoint partitions or by fixing certain scenarios in each subgroup, except for CR power divergences with \( \theta < 0 \) and \( \theta > 1 \) and \( \chi \)-divergence of order \( a > 1 \), for which the results are established under disjoint partitions. Two ways to implement these results in the multistage setting are devised: first-level and multi-level. A comparison with classical UBs shows the effectiveness of the proposed LB criteria. Our results do not require any structural properties, and thus they are applicable to a broad class of problems. Numerical results on a multistage production problem show that high-quality LBs can be obtained with a small computation time using the proposed bounding methodology.

Future work could include combining the proposed bounds with sampling-based bounds (e.g., [30]) or using them within solution algorithms (e.g., [24]). Devising new hybrid sampling-based algorithms that could utilize the proposed bounds to be used within Stochastic Dual Dynamic Programming (SDDP) type algorithms (e.g., [10, 28]) merit further research. It would also be interesting to investigate the concept of ineffective and
effective scenarios, defined in [31, 32], in this context. Ineffective scenarios can be removed from the problem without altering the optimal value. Therefore, if such scenarios can be identified, these might significantly speed up the proposed bounds. Effective scenarios might be used as fixed scenarios on each subgroup to improve the bounds. Another area of future research includes dissecting the scenario tree in subgroups with different cardinalities and examining optimal/desirable partitioning strategies (e.g., [37]). Finally, it should be highlighted that the proposed approach has the important advantage of dividing a given problem into subproblems, the solution of which are independent from one another and might be easily parallelized. Such parallel implementations might significantly decrease running times and therefore merit further computational research.

References


Appendix

Proof of LB criterion for Kullback-Leibler divergence (Part of Proposition 1)

Recall from Section 2.2 that Kullback-Leibler (KL) divergence is a special case of the CR power divergence family when the parameter of this family \( \theta \to 1 \). We now prove the LB criteria \( \bar{\rho} + \bar{\rho}_{\text{max}} \leq \rho \) for KL divergence in Proposition 1 by directly working with the ambiguity set of KL divergence. We use subscript \( \phi_{KL} \) to denote all ambiguity sets \( \mathcal{P} \) and divergences \( \Delta \) specific to KL.

Proof. Let \( P' \in \tilde{P}_{\phi_{KL} (\bar{\rho}, \bar{\rho})} \). Then there exists \( \tilde{P} \in \tilde{P}_{\phi_{KL} (\bar{\rho})} \) and \( \tilde{P} \in \tilde{P}_{\phi_{KL} (\bar{\rho})} \) such that

\[
\sum_{g \in [m]} \left[ \bar{p}^{(l)}_{g} \log \left( \frac{\bar{p}^{(l)}_{g}}{\bar{p}^{(l)}_{g}} \right) \right] \leq \bar{\rho}, \quad \sum_{g \in [m]} \bar{p}^{(l)}_{g} = 1 \quad \text{and} \quad \sum_{\omega \in \Omega^{(l)}_{g}} \left( \bar{p}^{(l)}_{\omega} \right) \log \left( \frac{\bar{p}^{(l)}_{\omega}}{\bar{p}^{(l)}_{\omega}} \right) \leq \bar{\rho},
\]

\[
\sum_{\omega \in \Omega^{(l)}_{g}} \left( \bar{p}^{(l)}_{\omega} \right) = 1 \quad \text{for all subgroups} \ g \in [m] \text{ with}
\]

\[
\Delta_{\phi_{KL}} (P', Q) = \sum_{\omega \in \Omega} \left[ p'_{\omega} \log \left( \frac{p'_{\omega}}{q_{\omega}} \right) \right] = \sum_{\omega \in \Omega} \left[ p'_{\omega} \log \left( \frac{p'_{\omega}}{q_{\omega}} \right) \right] + \sum_{\omega \notin \Omega} \left[ p'_{\omega} \log \left( \frac{p'_{\omega}}{q_{\omega}} \right) \right]
\]

\[
= \sum_{\omega \in \Omega} \log \left( \frac{\bar{p}^{(l)}_{g}}{\bar{p}^{(l)}_{g}} \right) + \sum_{\omega \notin \Omega} \left[ p'_{\omega} \log \left( \frac{p'_{\omega}}{q_{\omega}} \right) \right]
\]

\[
\leq \sum_{g \in [m]} \left[ \bar{p}^{(l)}_{g} \log \left( \frac{\bar{p}^{(l)}_{g}}{\bar{p}^{(l)}_{g}} \right) \right] + \sum_{\omega \in \Omega} \left[ p'_{\omega} \log \left( \frac{p'_{\omega}}{q_{\omega}} \right) \right]
\]

\[
= \sum_{g \in [m]} \left[ \bar{p}^{(l)}_{g} \log \left( \frac{\bar{p}^{(l)}_{g}}{\bar{p}^{(l)}_{g}} \right) \right] + \sum_{g \in [m]} \left[ \bar{p}^{(l)}_{g} \log \left( \frac{\bar{p}^{(l)}_{g}}{\bar{p}^{(l)}_{g}} \right) \right]
\]

\[
\leq \sum_{g \in [m]} \left[ \bar{p}^{(l)}_{g} \log \left( \frac{\bar{p}^{(l)}_{g}}{\bar{p}^{(l)}_{g}} \right) \right] + \sum_{g \in [m]} \left[ \bar{p}^{(l)}_{g} \log \left( \frac{\bar{p}^{(l)}_{g}}{\bar{p}^{(l)}_{g}} \right) \right]
\]

where the first inequality (on the third line) follows from the log sum inequality applied to fixed scenarios \( \omega_{i} \in \Omega_{f} \) and the last set of inequalities follow from the facts that \( \Delta_{\phi_{KL}} (\tilde{P}^{(l)}_{g}, \tilde{Q}^{(l)}_{g}) \leq \bar{\rho}_{g} \) for all subgroups \( g \in [m] \), the definition of \( \bar{\rho}_{\text{max}} \), \( \Delta_{\phi_{KL}} (\tilde{P}, \tilde{Q}) \leq \bar{\rho} \),
and $\sum_{\omega_i \in \Omega_g} (\bar{p}_{\omega_i})_g^{(l)} = \sum_{g \in [m_l]} \bar{p}_g^{(l)} = 1$. Therefore, if $\bar{\rho} + \bar{\rho}_{\text{max}} \leq \rho$, the result follows. \hfill \Box

Proof of Proposition 2 (LB criterion for variation distance)

Proof. Let $P' \in \bar{P}_{\phi_v}(\bar{\rho}, \bar{\rho})$. Then there exists $\bar{P} \in \bar{P}_{\phi_v}(\bar{\rho})$ and $\bar{P} \in \bar{P}_{\phi_v}(\bar{\rho})$ such that $\sum_{g \in [m_l]} |\bar{p}_g^{(l)} - \bar{\rho}_g^{(l)}| = 1$ and $\sum_{\omega_i \in \Omega_g} (|\bar{p}_{\omega_i}^{(l)} - (q_{\omega_i})_g^{(l)}|) = \bar{\rho}_g$, $\sum_{\omega_i \in \Omega_g} (|\bar{p}_{\omega_i}^{(l)} - (q_{\omega_i})_g^{(l)}|) = 1$ with

$$\Delta_{\phi_v}(P', Q) = \sum_{\omega_i \in \Omega} |p_{\omega_i} - q_{\omega_i}| = \sum_{\omega_i \in \Omega} \sum_{g \in [m_l]} |p_{\omega_i,g} - q_{\omega_i,g}| + \sum_{\omega_i \in \Omega} |p_{\omega_i} - q_{\omega_i}|$$

$$\leq \sum_{\omega_i \in \Omega} \sum_{g \in [m_l]} \sum_{\omega_i \in \Omega_g} (|\bar{p}_{\omega_i}^{(l)} - (q_{\omega_i})_g^{(l)}| + |\bar{p}_{\omega_i}^{(l)} - (q_{\omega_i})_g^{(l)}| + |\bar{p}_{\omega_i}^{(l)} - (q_{\omega_i})_g^{(l)}|)$$

$$= \sum_{\omega_i \in \Omega} \sum_{\omega_i \in \Omega_g} (|\bar{p}_{\omega_i}^{(l)} - (q_{\omega_i})_g^{(l)}| + \sum_{\omega_i \in \Omega_g} (|\bar{p}_{\omega_i}^{(l)} - (q_{\omega_i})_g^{(l)}| + \sum_{\omega_i \in \Omega_g} (|\bar{p}_{\omega_i}^{(l)} - (q_{\omega_i})_g^{(l)}|)$$

$$\leq \sum_{\omega_i \in \Omega} \sum_{\omega_i \in \Omega_g} (|\bar{p}_{\omega_i}^{(l)} - (q_{\omega_i})_g^{(l)}| + \bar{\rho} + \sum_{\omega_i \in \Omega_g} (|\bar{p}_{\omega_i}^{(l)} - (q_{\omega_i})_g^{(l)}|) \leq \bar{\rho} + \bar{\rho}_{\text{max}} \leq \rho.$$  

where the equality on the second line above follows from the way $p_{\omega_i,g}, q_{\omega_i,g}$ for fixed and $p_{\omega_i}, q_{\omega_i}$ for non-fixed scenarios are defined in set (4) and in Sections 3.1 and 3.2.

The inequality on the third line follows from, for any numbers $a, b, c, d$, that we have $|ac - bd| = |(a - b)(c - d) + (a - b)d + b(c - d)| \leq |(a - b)||c - d| + |(a - b)d + b(c - d)|$.

The final set of inequalities follow from the facts that $\Delta_{\phi_v}(P_g^{(l)}, Q_g^{(l)}) \leq \bar{\rho}_g$ for all subgroups $g \in [m_l]$, the definition of $\bar{\rho}_{\text{max}}$, $\Delta_{\phi_v}(\bar{P}, \bar{Q}) \leq \bar{\rho}$, and $\sum_{\omega_i \in \Omega_g} (q_{\omega_i})_g^{(l)} = \sum_{g \in [m_l]} \bar{p}_g^{(l)} = 1$. Therefore, if $\bar{\rho} \cdot \bar{\rho}_{\text{max}} + \bar{\rho} + \bar{\rho}_{\text{max}} \leq \rho$ the result follows. \hfill \Box

Sketch of Proof of Proposition 3 (LB criterion for J-divergence)

J-divergence is the sum of KL divergence and Burg entropy [4], and Burg entropy is similar to the KL divergence with $q_{\omega_i}$ and $p_{\omega_i}$ exchanged (see Table 1). Therefore, the proof of Burg entropy follows by first splitting the J-divergence as sum of KL divergence and Burg entropy and then following along similar lines as the proof for KL divergence in Appendix. Hence, it is skipped for brevity.

Proof of Proposition 4 (LB criterion for $\chi$-divergence of $a > 1$)

Proof. Let $x_g^{(l)} = 1 - \frac{\bar{p}_g^{(l)}}{\bar{\rho}_g}$ and $(y_{\omega_i})_g^{(l)} = 1 - \frac{(q_{\omega_i})_g^{(l)}}{(q_{\omega_i})_g^{(l)}}$. Let $P' \in \bar{P}_{\phi_v}(\bar{\rho}, \bar{\rho})$. Then there exists $\bar{P} \in \bar{P}_{\phi_v}(\bar{\rho})$ and $\bar{P} \in \bar{P}_{\phi_v}(\bar{\rho})$ such that $\sum_{g \in [m_l]} \bar{p}_g^{(l)} |x_g^{(l)}|^a \leq \bar{\rho}$ and $\sum_{g \in [m_l]} \bar{p}_g^{(l)} |(y_{\omega_i})_g^{(l)}|^a \leq \bar{\rho}_g$ for all subgroups $g \in [m_l]$. Since the scenario tree $\Omega$ is dissected using disjoint partitions (i.e., $\Omega_f = \emptyset$), we have $q_{\omega_i} = \pi_g^{(l)} (q_{\omega_i})_g^{(l)}$, for all $\omega_i \in \Omega_g^{(l)}$ and $g \in [m_l]$. Then,
\[ \Delta_{\phi^e}(P', Q) = \sum_{g \in [m]} \sum_{\omega \in \Omega_g^{(l)}} q_{\omega} \left| 1 - \frac{p_{\omega}^{(l)}(\tilde{\rho}_{\omega})^{(l)}}{\pi_g^{(l)}} \right|^a \leq \sum_{g \in [m]} \sum_{\omega \in \Omega_g^{(l)}} q_{\omega} \left| \phi_g^{(l)} + (\omega_g^{(l)} - \omega_g^{(l)}) \right|^a \]

where the first set of inequalities follow from the facts that \( \Delta \), where the first inequality (on the third line) follows from Minkowski inequality and the last set of inequalities follow from Minkowski inequality and the fact that \( \Delta_{\phi^e}(\tilde{P}, \tilde{Q}) \leq \tilde{\rho} \) and \( \Delta_{\phi^e}(\tilde{P}_g^{(l)}, Q_g^{(l)}) \leq \tilde{\rho}_g \) for all subgroups \( g \in [m] \), the definition of \( \tilde{\rho}_{\max} \), and \( \sum_{\omega \in \Omega_g^{(l)}} (q_{\omega})^{(l)} = \sum_{g \in [m]} \pi_g^{(l)} = 1 \). Therefore, if \( (\tilde{\rho})^{\frac{a}{2}} + (\tilde{\rho}_{\max})^{\frac{a}{2}} + (\tilde{\rho} \cdot \tilde{\rho}_{\max})^{\frac{a}{2}} \leq \rho \) the result follows. \( \square \)