Check for
updates

Partition-based distributionally robust optimization via optimal transport with order cone constraints

Adrián Esteban-Pérez[1] · **Juan M. Morales[1]**

Abstract

In this paper we wish to tackle stochastic programs affected by ambiguity about the probability law that governs their uncertain parameters. Using optimal transport theory, we construct an ambiguity set that exploits the knowledge about the distribution of the uncertain parameters, which is provided by: (1) sample data and (2) *a-priori* information on the order among the probabilities that the true data-generating distribution assigns to some regions of its support set. This type of order is enforced by means of order cone constraints and can encode a wide range of information on the shape of the probability distribution of the uncertain parameters such as information related to monotonicity or multi-modality. We seek decisions that are distributionally robust. In a number of practical cases, the resulting distributionally robust optimization (DRO) problem can be reformulated as a finite convex problem where the *a-priori* information translates into linear constraints. In addition, our method inherits the finite-sample performance guarantees of the Wasserstein-metric-based DRO approach proposed by Mohajerin Esfahani and Kuhn (Math Program 171(1–2):115–166. https://doi.org/10.1007/s10107-017-1172-1, 2018), while generalizing this and other popular DRO approaches. Finally, we have designed numerical experiments to analyze the performance of our approach with the newsvendor problem and the problem of a strategic firm competing à la Cournot in a market.

Keywords Distributionally robust optimization · Optimal transport · Wasserstein metric · Order cone constraints · Multi-item newsvendor problem · Strategic firm

✉ Juan M. Morales
    juan.morales@uma.es

    Adrián Esteban-Pérez
    adrianesteban@uma.es

[1]   Department of Applied Mathematics, University of Málaga, Málaga 29071, Spain

Mathematics Subject Classification 90C15 Stochastic programming · 90C47 Minimax problems

1 Introduction

Distributionally robust optimization (DRO) is a powerful modeling framework for optimization under uncertainty that emerges from considering that the probability distribution of the problem's uncertain parameters is, in itself, also uncertain. This gives rise to the notion of the *ambiguity set*, that is, a set where the modeler assumes that the true distribution of the problem's uncertain parameters is contained. The goal of DRO is therefore to find the decision maker's choice that is optimal against the worst-case probability distribution within the prescribed ambiguity set. Hence, DRO can be seen as a *marriage* between stochastic programming and robust optimization, working with probability distributions as the former does, while hedging the decision-maker against the worst case as the latter typically aims to do. Since the work of Scarf (1958), many DRO models have been proposed and studied in the technical literature, especially over the last decade, in which DRO has attracted a lot of attention and become very popular in the field of optimization under uncertainty as an alternative to other paradigms. We refer the reader to Keith and Ahner (2021), Rahimian and Mehrotra (2019) for recent surveys on DRO and optimization under uncertainty. Naturally, the construction of the ambiguity set is key to the practical performance of DRO. It is no wonder, therefore, that much effort has been applied to this issue, resulting in several ways to specify and characterize the ambiguity set, namely:

1. *Moment-based approach:* The ambiguity set is defined as the set of all probability distributions whose moments satisfy certain constraints; see, Delage and Ye (2010), Gao and Kleywegt (2017), Liu et al. (2018, 2019), Mehrotra and Papp (2014), Nakao et al. (2017), Xin and Goldberg (2021), Zymler et al. (2013), to name a few.
2. *Dissimilarity-based approach:* The ambiguity set is defined as the set of all probability distributions whose *dissimilarity* to a prescribed distribution (often referred to as the *nominal distribution*) is lower than or equal to a given value. Within this category, the choice of the *dissimilarity* function leads to a wealth of distinct variants:

    (a) *Optimal-transport-based (OTP) approach:* Here, we include the work in Blanchet et al. (2019, 2021), Gao and Kleywegt (2016), Mohajerin Esfahani and Kuhn (2018), Shafieezadeh-Abadeh et al. (2019), among many others, all of which use, as the dissimilarity function, the well-known *Wasserstein* distance, which exhibits some nice statistical convergence properties. Our work is also based on optimal mass transportation and consequently, it falls within this category.
    (b) *φ-divergences-based approach:* This class comprises all those approaches, which use φ-divergences (such as the Kullback-Leibler divergence), for instance, Bayraksan and Love (2015), Ben-Tal et al. (2013), Namkoong and Duchi (2016). We also include in this group the likelihood-based approaches, proposed by Duchi et al. (2021) and Wang et al. (2016).

(c) *Other measures of dissimilarity:* This category includes all other dissimilarity-based procedures for constructing ambiguity sets than those already mentioned, such as those that utilize the family of ζ-structure probability metrics (for example, the total variation metric, the Bounded Lipschitz metric …), see, for example, the work in Rahimian et al. (2019) and Zhao and Guan (2015), and the Prokhorov metric (Erdoğan and Iyengar 2006).

3. *Hypothesis-test-based approach:* The ambiguity set is made up of all those probability distributions which, given a data sample, pass a certain hypothesis test with a prescribed confidence level; see, for example, the work in Bertsimas et al. (2018a, b), Chen et al. (2019).

In the work we present here, we focus on ambiguity sets that are formulated by way of an optimal mass transportation problem. In fact, when the cost function in this problem is a metric, we recover the Wasserstein metric, which is indeed a metric for probability measures. According to Blanchet et al. (2021), Gao and Kleywegt (2016), Mohajerin Esfahani and Kuhn (2018), the Wasserstein metric has nice and interesting properties which make it a good choice in DRO, compared to popular alternative choices such as ϕ-divergences (see Sects. 1.1 and 5.1 in Gao and Kleywegt (2016), and the Introduction in Mohajerin Esfahani and Kuhn (2018) for a comparative analysis). Interestingly, the Wassertein distance offers a powerful theoretical framework to establish rates and guarantees of convergence. Furthermore, the conservatism implied by the ambiguity sets, built by means of the Wasserstein metric can be easily controlled, based on those rates.

Other cost functions can be used in the optimal mass transportation problem, but these do not generally result in a metric, which, most likely, makes it much harder to establish rates of convergence and theoretical guarantees.

However, one disadvantage of using the Wasserstein metric is that the worst-case probability distribution may take the form of a Dirac distribution (Yue et al. 2020), which is implausible in practice. Ambiguity sets containing unrealistic distributions may result in overly conservative solutions, since protection against these implausible distributions may require a decision that is more expensive than actually needed. Consequently, ambiguity sets that are solely based on a Wasserstein ball may lead to excessively costly solutions. In order to reduce the degree of conservatism, the authors in Gao and Kleywegt (2017), Liu (2021), Wang et al. (2018), Yao et al. (2018) consider ambiguity sets that are formulated using the Wasserstein metric in conjunction with moment constraints. Specifying these constraints, however, requires the estimation of the relevant parameters. Moreover, adding second-order moment information leads to semidefinite programs. In fact, as underlined in Liu (2021), although the mixture of moment conditions and the Wasserstein metric allows the decision maker to exclude pathological distributions and results in good out-of-sample performance, only in some special cases, e.g., when the objective function is piecewise linear with respect to the uncertain parameter, can the DRO problem be reformulated as a tractable semidefinite program. For this reason, they propose a method to approximate the solution of DRO problems with ambiguity sets that are based on both moment conditions and the Wasserstein metric.

Our work follows the path of the work in Gao and Kleywegt (2017), Liu (2021), Wang et al. (2018), Yao et al. (2018): In an attempt to avoid overly conservative solutions, we seek to enrich the specification of Wasserstein ambiguity sets with *a-priori* information on the true probability distribution of the problem's uncertain parameters. Nonetheless, unlike the aforementioned approaches, we represent this information in the form of order cone constraints on the probability masses associated with a partition of the sample space. This has the advantage that the inclusion of such *a-priori* information does not jeopardize the computational tractability of the underlying mathematical program. Our main contributions can be summarized as follows:

1. In real-world decision-making problems, it is common to count on qualitative and expert information conveying some sense of order between the probabilities of occurrence of certain events. For instance, in the multi-item newsvendor problem, the experienced decision maker may state that high demand values for a certain item are *more likely* to occur than low ones. This can occur, for example, when the true data-generating probability distribution is known or believed to be a mixture of distributions. In this case, determining the number of partitions in our DRO approach would be equivalent to estimating the number of components of the mixture. Indeed, one should use a number of partitions close to the number of distributions in the mixture. The task of inferring that number and the contribution of each component to the mixture is a relevant and well-known problem in statistics, which falls within the so-called realm of *finite mixture models* (see, Chapter 6 of McLachlan and Peel (2000)). In our approach, however, we assume that part of this inference task has already been done and so some of the inference results are available to the decision maker. Our aim is to exploit this type of qualitative information in the construction of the ambiguity set. Most importantly, our DRO approach protects the decision maker against the ambiguity in this inference process. For this purpose, we propose partitioning the support of the random parameter vector and bestow a partial order on (some of) the probability masses of the resulting subregions. This partial order can be described by a graph, which, in turn, can be associated with a convex cone. Consequently, the partial order can be embedded into the formulation of the ambiguity set in the form of conic constraints. The use of these types of cones is well known in the field of statistical inference with order restrictions (see, Németh and Németh 2016; Silvapulle and Sen 2011).

2. As shown in the numerical tests, this partial order can be leveraged, among other things, to easily encode multi-modality using *linear* constraints, as opposed to other approaches based on semidefinite programming (see, for example, the work in Hanasusanto et al. 2015), with the consequent benefit in terms of computational complexity. The recent papers Chen et al. (2019), Lam and Mottet (2017), Li et al. (2019) consider ambiguity sets with moment and generalized unimodal constraints. Our approach, however, can practically model a wider range of "shapes" beyond unimodality (see Sect. 2.3 for more details).

3. In addition to the order cone constraints on the probability masses linked to the different subregions of the partitioned sample space, these probability masses can also be treated as random, with their probability distribution belonging to a certain ambiguity set. This way, our modeling framework extends the two popular DRO

paradigms proposed by Mohajerin Esfahani and Kuhn (2018), and Bayraksan and Love (2015), Ben-Tal et al. (2013), respectively. Indeed,

- If we consider one partition only, that is, the entire sample space itself, there is no uncertainty about the associated probability mass (which is, evidently, equal to one) and no partial order can be established. If we now use a distance as the transportation cost function, our DRO framework reduces to that of Mohajerin Esfahani and Kuhn (2018).
- On the contrary, in order to get the DRO framework of Bayraksan and Love (2015), Ben-Tal et al. (2013), we just need to (1) consider a number of partitions such that every partition contains a single data point from the sample, (2) assume that the distribution of their probability masses belongs to a ϕ-divergence-based ambiguity set and (3) ignore any other information on the true probability distribution of the problem's uncertain parameters (namely, partial order and ambiguity in the conditional distributions).

For their part, the authors in Chen et al. (2020) have proposed a different ambiguity set that also covers these two DRO approaches as special cases. However, their ambiguity set does not include the DRO framework we propose, as we note later.

4. Under mild assumptions, we provide a tractable reformulation of our proposed DRO framework and show that it enjoys finite sample and asymptotic consistency guarantees.
5. Finally, we numerically illustrate the benefits in having *a-priori* information by comparing our DRO framework with the well-known sample average approximation (SAA) solution and the Wasserstein metric-based approach of Mohajerin Esfahani and Kuhn (2018). To this end, we consider the single and multi-item newsvendor problems and the problem of a strategic firm competing à la Cournot in a market.

The rest of the paper is organized as follows. Section 2 includes some preliminaries to the optimal transport problem, we formulate the proposed DRO approach and present tractable reformulations. Convergence properties and performance guarantees are theoretically discussed in Sect. 3. Section 4 provides the results from numerical experiments. Finally, Sect. 5 concludes the paper.

Notation We use $\overline{\mathbb{R}}$ to denote the extended real line, and adopt the conventions of its associated arithmetic. Furthermore, $\mathbb{R}_+$ denotes the set of non-negative real numbers. We employ lower-case bold face letters to represent vectors and bold face capital letters for matrices. We use $\mathrm{diag}(a_1, \ldots, a_m)$ for a diagonal matrix of size $m \times m$ whose diagonal elements are equal to $a_1, \ldots, a_m$. Moreover, given a matrix $\mathbf{M}$, its transpose matrix will be written as $\mathbf{M}^T$. We define $\mathbf{e}$ as the array with all its components equal to 1. The inner product of two vectors $\mathbf{u}, \mathbf{v}$ (in a certain space) is denoted $\langle \mathbf{u}, \mathbf{v} \rangle = \mathbf{u}^T \mathbf{v}$. Given any norm $\|\cdot\|$ in the Euclidean space (of a given dimension d), the dual norm is defined as $\|\mathbf{u}\|_* = \sup_{\|\mathbf{v}\| \leq 1} \langle \mathbf{u}, \mathbf{v} \rangle$. Given a function $f : \mathbb{R}^d \to \overline{\mathbb{R}}$, we will say that f is a proper function if $f(\mathbf{x}) < +\infty$ for at least one $\mathbf{x}$ and $f(\mathbf{x}) > -\infty$ for all $\mathbf{x} \in \mathbb{R}^d$. Additionally, the convex conjugate function of f, f^*, is defined as $f^*(\mathbf{y}) := \sup_{\mathbf{x} \in \mathbb{R}^d} \langle \mathbf{y}, \mathbf{x} \rangle - f(\mathbf{x})$. It is well known that if f is a proper function, then f^* is also a proper function. Given a set $A \subseteq \mathbb{R}^d$, we denote its relative interior as

relint(A). Similarly, we refer to its interior as int(A). The support function of set A, S_A, is defined as $S_A(\mathbf{b}) := \sup_{\mathbf{a} \in A} \langle \mathbf{b}, \mathbf{a} \rangle$. The dual cone $\mathscr{C}^*$ of a cone $\mathscr{C}$ is given by $\mathscr{C}^* := \{\mathbf{y} \ / \ \langle \mathbf{y}, \mathbf{x} \rangle \geq 0, \ \forall \mathbf{x} \in \mathscr{C}\}$. We use the symbol $\delta_{\boldsymbol{\xi}}$ to represent the Dirac distribution supported on $\boldsymbol{\xi}$. In addition, we reserve the symbol " $\widehat{}$ " for objects which are dependent on the sample data. The symbols $\mathbb{E}$ and $\mathbb{P}$ denote, respectively, "expectation" and "probability." Finally, for the rest of the paper we assume that we always have measurability for those objects, whose expected values we consider.

2 Data-driven distributionally robust optimization model

First, we briefly introduce some concepts from the *optimal transport problem* (also known as the *mass transportation problem*) that are at the core of the development of our DRO framework.

Intuitively speaking, the optimal transport problem (OTP) centers on the question of how to move masses between two probability distributions in such a way that the transportation cost is minimal. Let P and Q be two probability distributions in a Polish space S such that P is the distribution of mass seen as the origin (i.e. the source) and Q is the distribution of mass seen as the destination (i.e., the sink), and let c be a measurable cost function with $c(x, y)$ representing the cost of moving a unit of mass from location x to location y. The OTP can be stated as follows

$$C(P, Q) = \inf_{\Pi} \left\{ \int c(x, y) \Pi(dx, dy) : \Pi \text{ is a joint distribution} \right.$$
$$\left. \text{with marginals } P \text{ and } Q, \text{ respectively}\right\}$$

We assume that the cost function c is a non-negative jointly convex lower semicontinuous function such that if $x = y$, then $c(x, y) = 0$. In the remainder of the paper we assume that we have existence and uniqueness of the OTP (see, for example, Theorem 4.1 in Villani (2008)).

For more technical details about the assumptions on the cost function in the OTP, we refer to Villani (2008) and Santambrogio (2015). Note that if we choose a distance on S as the cost function (for example, a p-norm, with $p \geq 1$, if S is the Euclidean space $\mathbb{R}^n$), we get the so-called *Wasserstein* metric of order 1, which we represent as $\mathscr{W}(P, Q)$ and which is also known as the Kantorovich metric.

It is well known that this probability distance metrizes the weak convergence property. Furthermore, convergence with respect to the Wasserstein metric of order 1 is equivalent to weak convergence plus convergence of the first moment. Wherever the Wasserstein metric of order 1 is used in this paper, we implicitly consider the set of all probability distributions with finite moment of order 1. Likewise, we refer to the *Wasserstein ball of radius $r \geq 0$ centered at a certain nominal probability distribution P_0*, which we denote by $\mathbb{B}_r(P_0)$, as the set of all probability distributions whose Wasserstein metric of order 1 to P_0 is at most r.

2.1 Formulation of the proposed model

Problem (P) below formulates the data-driven distributionally robust optimization (DDRO) framework we propose.

$$(P) \quad \inf_{\mathbf{x} \in X} \sup_{Q \in \mathcal{Q}} \mathbb{E}_Q \left[f(\mathbf{x}, \boldsymbol{\xi}) \right] \tag{1a}$$

$$\text{s.t.} \quad \mathbb{P}_Q \left[\boldsymbol{\xi} \in \Xi_i \right] = p_i, \forall i \in \mathscr{I} \tag{1b}$$

$$\widetilde{c}(\mathbf{p} - \widehat{\mathbf{p}}) \leq \rho \tag{1c}$$

$$\sum_{i \in \mathscr{I}} p_i C(Q_i, \widehat{Q}_i) \leq \varepsilon \tag{1d}$$

$$Q_i \in \mathcal{Q}_i, \forall i \in \mathscr{I} \tag{1e}$$

$$\mathbf{p} \in \Theta \tag{1f}$$

where $X \subseteq \mathbb{R}^n$ is the set of feasible decisions, $\boldsymbol{\xi} : \Omega \to \Xi \subseteq \mathbb{R}^d$ is a random vector defined on the measurable space $(\Omega, \mathscr{F})$ with σ-algebra $\mathscr{F}$, and $\mathcal{Q}$ is the set of all probability distributions over the measurable space $(\Omega, \mathscr{F})$. Moreover, for each $i \in \mathscr{I}$, Q_i is the conditional distribution of Q given $\boldsymbol{\xi} \in \Xi_i$, that is $Q_i = Q(\boldsymbol{\xi} \, / \, \boldsymbol{\xi} \in \Xi_i) \in \mathcal{Q}_i$, with $\mathcal{Q}_i$ being the set of all conditional probability distributions of Q given $\boldsymbol{\xi} \in \Xi_i$. In this setting, $\mathscr{I}$ is the set of regions Ξ_i with pairwise disjoint interiors into which the support set Ξ is partitioned, that is, $\bigcup_{i \in \mathscr{I}}(\Xi_i) = \Xi$ and $\text{int}(\Xi_i) \bigcap \text{int}(\Xi_j) = \emptyset$, $\forall i, j \in \mathscr{I}, i \neq j$. Furthermore, we assume that $Q^*(\Xi_i \cap \Xi_j) = 0, \forall i, j \in \mathscr{I}, i \neq j$, where Q^* is the true data-generating distribution. This is equivalent to stating that $\{\Xi_i\}_{i \in \mathscr{I}}$ constitutes a Q^*-packing (see a formal definition of this concept in page 50 of Graf and Luschgy (2000)) and will allow us to unequivocally assign samples from Q^* to the partitions $\Xi_i, i \in \mathscr{I}$. Finally, constraint (1c) defines the set of all probability vectors $\mathbf{p}$ that differ from the nominal empirical probability vector $\widehat{\mathbf{p}}$ in at most ρ according to the cost function $\widetilde{c}$. This is a function that quantifies how dissimilar two probability vectors $\mathbf{p}$ and $\mathbf{q}$ are. For this purpose, we require that $\widetilde{c}$ be a non-negative jointly convex lower semicontinuous function such that if $\mathbf{p} = \mathbf{q}$, then $\widetilde{c}(\mathbf{p}, \mathbf{q}) = 0$. As mentioned further on, function $\widetilde{c}$ could, for example, take the form of a norm or a ϕ-divergence. To ease the notation and the formulation, we use $\boldsymbol{\xi}$ to represent either the random vector $\boldsymbol{\xi}(\omega)$, with $\omega \in \Omega$ or an element of $\mathbb{R}^d$. Note that we can consider the probability measure induced by the random vector $\boldsymbol{\xi}$, if we choose the corresponding Borel σ-algebra $\mathscr{B}$ on Ξ. Thus, we can see $\mathcal{Q}$ as a set of probability measures defined over $(\Xi, \mathscr{B})$, so we write $\mathcal{Q} = \mathcal{Q}(\Xi)$. We define the uncertainty set $\mathscr{P}$ for the probability vector $\mathbf{p} \in \mathbb{R}^{|\mathscr{I}|}$, with $|\mathscr{I}|$ being the number of partitions, as the intersection of Θ and the set defined by constraint (1c). The support set Θ, which includes the order cone constraints on the probability masses $\mathbf{p}$, is given by:

$$\Theta = \{\mathbf{p} \in \mathbb{R}^{|\mathscr{I}|} : \langle \mathbf{e}, \mathbf{p} \rangle = 1, \mathbf{p} \in \mathbb{R}_+^{|\mathscr{I}|}, \mathbf{p} \in \mathscr{C}\} \tag{2}$$

where $\mathscr{C}$ is a proper (convex, closed, full and pointed) cone. Hence, Θ is a convex compact set.

In problem (P), ρ and ε are non-negative parameters, to be tuned by the decision maker, which control the size of the ambiguity set defined by Eqs. (1b)–(1f).

We represent this set as $\mathscr{U}_{\rho,\varepsilon}(\widehat{Q})$, where $\widehat{Q}$ is a nominal distribution expressed in terms of $\widehat{\mathbf{p}}$ and $\widehat{Q}_i$ as

$$\widehat{Q} = \sum_{i \in \mathscr{I}} \widehat{p}_i \widehat{Q}_i \tag{3}$$

where

$$\widehat{p}_i = \frac{N_i}{N + |I'|} \tag{4}$$

and

$$\widehat{Q}_i = \frac{1}{N_i} \sum_{j=1}^{N_i} \delta_{\widehat{\xi}_j^i} \tag{5}$$

Additionally,

$I' = \{i \in \mathscr{I}$ such that partition i does not contain any data from the sample$\}$, $\widehat{\xi}_j^i \in \{\widehat{\xi}_1^i, \ldots, \widehat{\xi}_{N_i}^i\}$ and N_i is the number of atoms in region $\varXi_i$. Here we set $N_i = 1$ and $\widehat{\xi}_1^i := \arg \sup_{\xi \in \varXi_i} f(\mathbf{x}, \xi)$ for those $i \in I'$. Implicitly, we assume that this supremum is attained. We remark that this modeling choice protects the decision maker in those cases where there is a total absence of information on the conditional distributions $Q_i, i \in I'$. Indeed, by introducing the "artificial" data point $\widehat{\xi}_1^i := \arg \sup_{\xi \in \varXi_i} f(\mathbf{x}, \xi)$ in a partition $\varXi_i$ with no samples, we are considering the worst-case form that the true conditional distribution Q_i could possibly take, that is, a Dirac distribution supported on $\widehat{\xi}_1^i$.

Finally, we note that the ambiguity set defined by constraints (1b)–(1f) is unequivocally determined by specifying the partitions $\varXi_i, i \in \mathscr{I}$, the nominal distribution $\widehat{Q}$, the budgets ρ and ε, and the order cone constraints $\mathbf{p} \in \mathscr{C}$ in (2). In fact, if these constraints are removed and we set $\rho = \varepsilon = 0$, then we have $p_i = \widehat{p}_i$ and $Q_i = \widehat{Q}_i, \forall i$, and therefore, $Q = \widehat{Q}$.

The following theorem shows that problem (P) can be reformulated as a single-level problem.

Theorem 1 (Reformulation based on strong duality) *For any non-negative values of parameters ε, ρ, problem (P) is equivalent to the following:*

$$(P0) \quad \inf_{\mathbf{x}, \lambda, \boldsymbol{\mu}, \eta \, \widetilde{\mathbf{p}}, \theta, \mathbf{t}} \lambda \rho + \eta + \theta \varepsilon + \lambda \widetilde{c}_{\widehat{\mathbf{p}}}^* \left(\frac{\left(\frac{1}{N_i} \sum_{j=1}^{N_i} t_{i,j} \right)_{i \in \mathscr{I}} + \boldsymbol{\mu} - \eta \mathbf{e} + \widetilde{\mathbf{p}}}{\lambda} \right)$$

$$s.t. \ t_{i,j} \geq \sup_{\xi \in \varXi_i} \left[f(\mathbf{x}, \xi) - \theta c(\xi, \widehat{\xi}_j^i) \right], \ \forall i \in \mathscr{I}, \ j \leq N_i$$

$$\mathbf{x} \in X, \lambda \geq 0, \boldsymbol{\mu} \in \mathbb{R}_+^{|\mathscr{I}|}, \eta \in \mathbb{R}, \ \widetilde{\mathbf{p}} \in \mathscr{C}^*, \theta \geq 0$$

$$t_{i,j} \in \mathbb{R}, \forall i \in \mathscr{I}, j \leq N_i \tag{6}$$

*where $\widetilde{c}_{\widehat{\mathbf{p}}}^{\,*}(\cdot)$ is the convex conjugate function of $\widetilde{c}(\cdot, \widehat{\mathbf{p}})$, with $\widehat{\mathbf{p}}$ fixed, and $\left(\frac{1}{N_i}\sum_{j=1}^{N_i} t_{i,j}\right)_{i \in \mathscr{I}}$
is the vector with the $|\mathscr{I}|$ components $\frac{1}{N_i}\sum_{j=1}^{N_i} t_{i,j}$.*

Proof Recall that we have assumed that regions $\varXi_i$ are disjoint. Thus, using the law
of total probability, we can rewrite problem (P) as follows:

$$\inf_{\mathbf{x}\in X}\ \sup_{\mathbf{p}\in\mathscr{P}} G(\mathbf{x}, \mathbf{p}) \tag{7}$$

where we have considered the subproblem (SP):

$$(\text{SP})\ G(\mathbf{x}, \mathbf{p}) = \sup_{Q_i \in \mathscr{Q}_i, \forall i}\ \sum_{i \in \mathscr{I}} p_i \mathbb{E}_{Q_i}\left[f(\mathbf{x}, \boldsymbol{\xi})\right] \tag{8a}$$

$$\text{s.t.}\ \sum_{i \in \mathscr{I}} p_i C(Q_i, \widehat{Q}_i) \le \varepsilon \tag{8b}$$

The probability distribution $\widehat{Q}_i$ is defined as $\widehat{Q}_i = \frac{1}{N_i}\sum_{j=1}^{N_i}\delta_{\widehat{\boldsymbol{\xi}}_j^i}$, with $\widehat{\boldsymbol{\xi}}_j^i \in$
$\{\widehat{\boldsymbol{\xi}}_1^i, \dots, \widehat{\boldsymbol{\xi}}_{N_i}^i\}$, and N_i being the number of data points in $\varXi_i$.

Note that the structure of problem (7) does not fit in the general ambiguity set
proposed in Chen et al. (2020).

Equivalently, we can recast the subproblem (SP) as

$$(\text{SP}) = \begin{cases} \displaystyle\sup_{Q_i \in \mathscr{Q}_i, \Pi_i, \forall i}\ \sum_{i \in \mathscr{I}} p_i \int_{\varXi_i} f(\mathbf{x}, \boldsymbol{\xi}) Q_i(d\boldsymbol{\xi}) \\[2ex] \qquad \text{s.t.}\qquad \displaystyle\sum_{i \in \mathscr{I}} p_i \int_{\varXi_i^2} c(\boldsymbol{\xi}, \boldsymbol{\xi}') \Pi_i(d\boldsymbol{\xi}, d\boldsymbol{\xi}') \le \varepsilon \\[2ex] \qquad \begin{cases} \forall\, i,\ \Pi_i \text{ is a joint distribution of } \boldsymbol{\xi} \text{ and } \boldsymbol{\xi}' \\ \text{with marginals } Q_i \text{ and } \widehat{Q}_i,\ \text{respectively} \end{cases} \end{cases} \tag{9}$$

$$= \begin{cases} \displaystyle\sup_{\widetilde{Q}_j^i, \forall i \in \mathscr{I}, j \le N_i}\ \sum_{i \in \mathscr{I}}\frac{p_i}{N_i}\sum_{j=1}^{N_i}\int_{\varXi_i} f(\mathbf{x}, \boldsymbol{\xi})\widetilde{Q}_j^i(d\boldsymbol{\xi}) \\[2ex] \qquad \text{s.t.}\qquad \displaystyle\sum_{i \in \mathscr{I}}\frac{p_i}{N_i}\sum_{j=1}^{N_i}\int_{\varXi_i} c(\boldsymbol{\xi}, \widehat{\boldsymbol{\xi}}_j^i)\widetilde{Q}_j^i(d\boldsymbol{\xi}) \le \varepsilon \\[2ex] \qquad \displaystyle\int_{\varXi_i}\widetilde{Q}_j^i(d\boldsymbol{\xi}) = 1,\ \forall i \in \mathscr{I}, j \le N_i \end{cases} \tag{10}$$

where reformulation (10) follows on from the fact that the marginal distribution of $\boldsymbol{\xi}'$
is the discrete uniform distribution supported on points $\widehat{\boldsymbol{\xi}}_j^i$, $j = 1, \dots, N_i$. Thus, Π_i
is completely determined by the conditional distributions $\widetilde{Q}_j^i = \Pi_i(\boldsymbol{\xi}, \boldsymbol{\xi}'|\boldsymbol{\xi}' = \widehat{\boldsymbol{\xi}}_j^i)$,
$\forall i \le N_i$, that is, $\Pi_i(d\boldsymbol{\xi}, d\boldsymbol{\xi}') = \frac{1}{N_i}\sum_{j=1}^{N_i}\delta_{\widehat{\boldsymbol{\xi}}_j^i}(d\boldsymbol{\xi}')\widetilde{Q}_j^i(d\boldsymbol{\xi})$ (Mohajerin Esfahani and
Kuhn 2018).

The mathematical program (10) constitutes a generalized moment problem over the normalized measures $\widetilde{Q}^i_j$, for which strong duality holds (see, for example, Shapiro (2001)). We can, therefore, dualize the ε-budget constraint on the transport cost, thus obtaining:

$$\inf_{\theta \geq 0} \sup_{\widetilde{Q}^i_j, \forall i \in \mathscr{I}, j \leq N_i} \theta \varepsilon + \sum_{i \in \mathscr{I}} \frac{p_i}{N_i} \sum_{j=1}^{N_i} \int_{\Xi_i} \left[f(\mathbf{x}, \boldsymbol{\xi}) - \theta c(\boldsymbol{\xi}, \widehat{\boldsymbol{\xi}}^i_j) \right] \widetilde{Q}^i_j(d\boldsymbol{\xi}) \tag{11}$$

$$\text{s.t.} \int_{\Xi_i} \widetilde{Q}^i_j(d\boldsymbol{\xi}) = 1, \ \forall i \in \mathscr{I}, j \leq N_i \tag{12}$$

$$= \inf_{\theta \geq 0} \theta \varepsilon + \sum_{i \in \mathscr{I}} \frac{p_i}{N_i} \sum_{j=1}^{N_i} \sup_{\widetilde{Q}^i_j} \int_{\Xi_i} \left[f(\mathbf{x}, \boldsymbol{\xi}) - \theta c(\boldsymbol{\xi}, \widehat{\boldsymbol{\xi}}^i_j) \right] \widetilde{Q}^i_j(d\boldsymbol{\xi}) \tag{13}$$

$$\text{s.t.} \int_{\Xi_i} \widetilde{Q}^i_j(d\boldsymbol{\xi}) = 1, \ \forall i \in \mathscr{I}, j \leq N_i \tag{14}$$

$$= \inf_{\theta \geq 0} \theta \varepsilon + \sum_{i \in \mathscr{I}} \frac{p_i}{N_i} \sum_{j=1}^{N_i} \sup_{\boldsymbol{\xi} \in \Xi_i} \left[f(\mathbf{x}, \boldsymbol{\xi}) - \theta c(\boldsymbol{\xi}, \widehat{\boldsymbol{\xi}}^i_j) \right] \tag{15}$$

$$= \inf_{\theta, t_{ij}, \forall i \in \mathscr{I}, j \leq N_i} \theta \varepsilon + \sum_{i \in \mathscr{I}} \frac{p_i}{N_i} \sum_{j=1}^{N_i} t_{i,j} \tag{16}$$

$$\text{s.t.} \ t_{i,j} \geq \sup_{\boldsymbol{\xi} \in \Xi_i} \left[f(\mathbf{x}, \boldsymbol{\xi}) - \theta c(\boldsymbol{\xi}, \widehat{\boldsymbol{\xi}}^i_j) \right], \ \forall i \in \mathscr{I}, \ j \leq N_i \tag{17}$$

$$\theta \geq 0 \tag{18}$$

where the second equality derives from the fact that we can choose a Dirac distribution supported on Ξ_i as $\widetilde{Q}^i_j$.

Now, dualizing the ρ-budget constraint on the transport cost in the inner supremum of problem (7), we obtain:

$$\inf_{\lambda \geq 0} \lambda \rho + \sup_{\mathbf{p} \in \Theta} \left[G(\mathbf{x}, \mathbf{p}) - \lambda \widetilde{c}(\mathbf{p}, \widehat{\mathbf{p}}) \right] \tag{19}$$

Thus,

$$\inf_{\lambda \geq 0} \lambda \rho + \sup_{\mathbf{p} \in \Theta} \left[G(\mathbf{x}, \mathbf{p}) - \lambda \widetilde{c}(\mathbf{p}, \widehat{\mathbf{p}}) \right] \tag{20}$$

$$= \inf_{\lambda \geq 0} \lambda \rho + \sup_{\mathbf{p} \in \Theta} \left[\inf_{\theta \geq 0, (t_{i,j}) \ s.t.(17)} \theta \varepsilon + \sum_{i \in \mathscr{I}} \frac{p_i}{N_i} \sum_{j=1}^{N_i} t_{i,j} - \lambda \widetilde{c}(\mathbf{p}, \widehat{\mathbf{p}}) \right] \tag{21}$$

Since function $\theta\varepsilon + \sum_{i\in\mathscr{I}} \frac{p_i}{N_i} \sum_{j=1}^{N_i} t_{i,j} - \lambda\widetilde{c}(\mathbf{p},\widehat{\mathbf{p}})$ is upper semicontinuous and concave in $\mathbf{p}$ on the compact convex set Θ (recall that $\widetilde{c}$ is nonegative, lower semi-continuous, and convex in $\mathbf{p}$), and linear in θ and $t_{i,j}$ on the convex set defined by $\theta \geq 0$ and (17), we can apply Sion's min-max theorem (Sion 1958) and in this way, interchange the innest infimum with the outer supremum. Then, by merging the two infima, we arrive at

$$\inf_{\lambda\geq 0, \theta\geq 0, (t_{i,j})} \lambda\rho + \theta\varepsilon + \sup_{\mathbf{p}\in\Theta}\left[\sum_{i\in\mathscr{I}} \frac{p_i}{N_i} \sum_{j=1}^{N_i} t_{i,j} - \lambda\widetilde{c}(\mathbf{p},\widehat{\mathbf{p}})\right]$$

$$\text{s.t. } t_{i,j} \geq \sup_{\boldsymbol{\xi}\in\Xi_i}\left[f(\mathbf{x},\boldsymbol{\xi}) - \theta_i c(\boldsymbol{\xi},\widehat{\boldsymbol{\xi}}_j^i)\right], \ \forall i\in\mathscr{I}, \ j\leq N_i$$

We focus now on the inner supremum,

$$\sup_{\mathbf{p}\in\Theta}\left[\left\langle\mathbf{p}, \left(\frac{1}{N_i}\sum_{j=1}^{N_i} t_{i,j}\right)_{i\in\mathscr{I}}\right\rangle - \lambda\widetilde{c}(\mathbf{p},\widehat{\mathbf{p}})\right] \tag{22}$$

where we have written $\sum_{i\in\mathscr{I}} \frac{p_i}{N_i} \sum_{j=1}^{N_i} t_{i,j}$ as $\left\langle\mathbf{p}, \left(\frac{1}{N_i}\sum_{j=1}^{N_i} t_{i,j}\right)_{i\in\mathscr{I}}\right\rangle$. This is a concave maximization problem (be aware that $\langle\mathbf{p},\mathbf{H}(\mathbf{x})\rangle - \lambda\widetilde{c}(\mathbf{p},\widehat{\mathbf{p}})$ is a concave function with respect to $\mathbf{p}$ and Θ is a convex compact set; furthermore, notice that we have $\mathbf{H}(\mathbf{x}) = \left(\frac{1}{N_i}\sum_{j=1}^{N_i} t_{i,j}\right)_{i\in\mathscr{I}}$ in our particular case). Consequently, strong duality holds if a Slater condition is satisfied, that is, if there exists a point $\mathbf{p}^* \in \text{relint}(\mathbb{R}_+^{|\mathscr{I}|})$ such that $\langle\mathbf{e},\mathbf{p}^*\rangle = 1$, and $\mathbf{p}^* \in \text{int}(\mathscr{C})$ (see, for example, Boyd and Vandenberghe 2004). Using a standard duality argument, we dualize the constraints $\mathbf{p} \in \mathbb{R}_+^{|\mathscr{I}|}$, $\langle\mathbf{e},\mathbf{p}\rangle = 1$ and $\mathbf{p} \in \mathscr{C}$, with associated multipliers $\boldsymbol{\mu} \in \mathbb{R}_+^{|\mathscr{I}|}$, $\eta \in \mathbb{R}$ and $\widetilde{\mathbf{p}} \in \mathscr{C}^*$, respectively. Thus, we obtain the following problem:

$$\inf_{\eta\in\mathbb{R}, \boldsymbol{\mu}\in\mathbb{R}_+^{|\mathscr{I}|}, \widetilde{\mathbf{p}}\in\mathscr{C}^*} \sup_{\mathbf{p}}\left\{\left\langle\mathbf{p}, \left(\frac{1}{N_i}\sum_{j=1}^{N_i} t_{i,j}\right)_{i\in\mathscr{I}}\right\rangle - \lambda\widetilde{c}(\mathbf{p},\widehat{\mathbf{p}}) + \langle\boldsymbol{\mu},\mathbf{p}\rangle + \eta(1 - \langle\mathbf{e},\mathbf{p}\rangle) + \langle\widetilde{\mathbf{p}},\mathbf{p}\rangle\right\}$$

$$= \inf_{\eta\in\mathbb{R}, \boldsymbol{\mu}\in\mathbb{R}_+^{|\mathscr{I}|}, \widetilde{\mathbf{p}}\in\mathscr{C}^*} \eta + \sup_{\mathbf{p}}\left\{\left\langle\mathbf{p}, \left(\frac{1}{N_i}\sum_{j=1}^{N_i} t_{i,j}\right)_{i\in\mathscr{I}} + \boldsymbol{\mu} - \eta\mathbf{e} + \widetilde{\mathbf{p}}\right\rangle - \lambda\widetilde{c}(\mathbf{p},\widehat{\mathbf{p}})\right\}$$

$$= \inf_{\eta\in\mathbb{R}, \boldsymbol{\mu}\in\mathbb{R}_+^{|\mathscr{I}|}, \widetilde{\mathbf{p}}\in\mathscr{C}^*} \eta + \lambda\sup_{\mathbf{p}}\left\{\left\langle\mathbf{p}, \frac{\left(\frac{1}{N_i}\sum_{j=1}^{N_i} t_{i,j}\right)_{i\in\mathscr{I}} + \boldsymbol{\mu} - \eta\mathbf{e} + \widetilde{\mathbf{p}}}{\lambda}\right\rangle - \widetilde{c}(\mathbf{p},\widehat{\mathbf{p}})\right\}$$

$$= \inf_{\eta\in\mathbb{R}, \boldsymbol{\mu}\in\mathbb{R}_+^{|\mathscr{I}|}, \widetilde{\mathbf{p}}\in\mathscr{C}^*} \eta + \lambda\widetilde{c}_{\widehat{\mathbf{p}}}^*\left(\frac{\left(\frac{1}{N_i}\sum_{j=1}^{N_i} t_{i,j}\right)_{i\in\mathscr{I}} + \boldsymbol{\mu} - \eta\mathbf{e} + \widetilde{\mathbf{p}}}{\lambda}\right)$$

where $\widetilde{c}_{\widehat{\mathbf{p}}}^*(\cdot)$ is the convex conjugate function of $\widetilde{c}(\cdot,\widehat{\mathbf{p}})$, with $\widehat{\mathbf{p}}$ fixed.

Therefore, problem (7) can be equivalently reformulated as follows:

$$
\text{(P0)} \quad \inf_{\mathbf{x},\lambda,\boldsymbol{\mu},\eta\,\widetilde{\mathbf{p}},\theta,\mathbf{t}} \lambda\rho + \eta + \theta\varepsilon + \lambda\widetilde{c}_{\widehat{\mathbf{p}}}^{*}\left(\frac{\left(\frac{1}{N_i}\sum_{j=1}^{N_i} t_{i,j}\right)_{i\in\mathscr{I}} + \boldsymbol{\mu} - \eta\mathbf{e} + \widetilde{\mathbf{p}}}{\lambda} \right)
$$

$$
\text{s.t. } t_{i,j} \geq \sup_{\boldsymbol{\xi}\in\Xi_i}\left[f(\mathbf{x},\boldsymbol{\xi}) - \theta c(\boldsymbol{\xi},\widehat{\boldsymbol{\xi}}_j^i) \right],\ \forall i\in\mathscr{I},\ j\leq N_i
$$

$$
\mathbf{x}\in X, \lambda\geq 0, \boldsymbol{\mu}\in\mathbb{R}_+^{|\mathscr{I}|}, \eta\in\mathbb{R},\ \widetilde{\mathbf{p}}\in\mathscr{C}^*, \theta\geq 0
$$

$$
t_{i,j}\in\mathbb{R}, \forall i\in\mathscr{I}, j\leq N_i
$$

$\square$

Moreover, in the case that the cost function $\widetilde{c}(\cdot,\cdot)$ is given by a norm, we have $\widetilde{c}_{\widehat{\mathbf{p}}}(\mathbf{p}) = \|\mathbf{p} - \widehat{\mathbf{p}}\|$. The next corollary deals with this particular case.

Corollary 1 *If the cost functions $c(\cdot,\cdot)$ and $\widetilde{c}(\cdot,\cdot)$ are given by norms, then for any non-negative values of parameters ε, ρ, the problem (P) is equivalent to the following problem*

$$
\text{(P1)} \quad \inf_{\mathbf{x},\lambda,\boldsymbol{\mu},\eta\,\widetilde{\mathbf{p}},\theta,\mathbf{t}} \lambda\rho + \eta + \theta\varepsilon + \sum_{i\in\mathscr{I}} \widehat{p}_i\left(\frac{1}{N_i}\sum_{j=1}^{N_i} t_{i,j} + \mu_i - \eta + \widetilde{p}_i \right)
$$

$$
\text{s.t. } t_{i,j} \geq \sup_{\boldsymbol{\xi}\in\Xi_i}\left[f(\mathbf{x},\boldsymbol{\xi}) - \theta\left\|\boldsymbol{\xi} - \widehat{\boldsymbol{\xi}}_j^i\right\| \right], \forall i\in\mathscr{I}, \forall j\leq N_i
$$

$$
\left\| \left(\frac{1}{N_i}\sum_{j=1}^{N_i} t_{i,j} + \mu_i - \eta + \widetilde{p}_i \right)_{i\in\mathscr{I}} \right\|_* \leq \lambda
$$

$$
\mathbf{x}\in X, \lambda\geq 0, \boldsymbol{\mu}\in\mathbb{R}_+^{|\mathscr{I}|}, \eta\in\mathbb{R},\ \widetilde{\mathbf{p}}\in\mathscr{C}^*, \theta\geq 0
$$

$$
t_{i,j}\in\mathbb{R}, \forall i\in\mathscr{I}, \forall j\leq N_i \tag{23}
$$

Proof We use the following Lemma to put problem (P0) in a better shape.

Lemma 1 *Let $\widetilde{c}_{\widehat{\mathbf{p}}}(\mathbf{p}) = \|\mathbf{p} - \widehat{\mathbf{p}}\|$, where $\widehat{\mathbf{p}}\in\mathbb{R}^{|\mathscr{I}|}$ is a fixed vector and $\|\cdot\|$ a norm in $\mathbb{R}^{|\mathscr{I}|}$. Then, it holds that the convex conjugate function of $\widetilde{c}_{\widehat{\mathbf{p}}}(\mathbf{p})$ is as follows*

$$
\widetilde{c}_{\widehat{\mathbf{p}}}^{*}(\mathbf{s}) = \begin{cases} \sum_{i\in\mathscr{I}} \widehat{p}_i s_i & \text{if } \|\mathbf{s}\|_* \leq 1 \\ \infty & \text{if } \|\mathbf{s}\|_* > 1 \end{cases}
$$

Proof The claim of the Lemma follows from Proposition 5.1.4. (vii) and Example 5.1.2 (b) of Lucchetti (2006). $\square$

Therefore, problem (P0) reduces to

$$
\text{(P1)} \quad \inf_{\mathbf{x},\lambda,\boldsymbol{\mu},\eta\,\widetilde{\mathbf{p}},\theta,\mathbf{t}} \lambda\rho + \eta + \theta\varepsilon + \sum_{i\in\mathscr{I}} \widehat{p}_i\left(\frac{1}{N_i}\sum_{j=1}^{N_i} t_{i,j} + \mu_i - \eta + \widetilde{p}_i \right)
$$

$$\text{s.t. } t_{i,j} \geq \sup_{\boldsymbol{\xi} \in \varXi_i} \left[f(\mathbf{x}, \boldsymbol{\xi}) - \theta \left\| \boldsymbol{\xi} - \widehat{\boldsymbol{\xi}}_j^i \right\| \right], \forall i \in \mathscr{I}, \forall j \leq N_i$$

$$\left\| \left(\frac{1}{N_i} \sum_{j=1}^{N_i} t_{i,j} + \mu_i - \eta + \widetilde{p}_i \right)_{i \in \mathscr{I}} \right\|_* \leq \lambda$$

$$\mathbf{x} \in X, \lambda \geq 0, \boldsymbol{\mu} \in \mathbb{R}_+^{|\mathscr{I}|}, \eta \in \mathbb{R}, \ \widetilde{\mathbf{p}} \in \mathscr{C}^*, \theta \geq 0$$

$$t_{i,j} \in \mathbb{R}, \forall i \in \mathscr{I}, \forall j \leq N_i$$

$\square$

Remarks Our data-driven DRO framework (P) can be easily understood as a generalization of other popular DRO approaches. To see this, first we need to remove the order cone constraints on the probabilities associated with each subregion into which the support $\varXi$ has been partitioned, that is, the condition $\mathbf{p} \in \mathscr{C}$, and then proceed as indicated below:

1. If we set $\varepsilon = 0$, $|\mathscr{I}| = N$, with every partition containing a single and different data point from the sample, and use a ϕ-divergence to build the cost function, i.e., $\widetilde{c}_{\widehat{\mathbf{p}}}(\mathbf{p}) = \sum_{i \in \mathscr{I}} \widehat{p}_i \phi \left(\frac{p_i}{\widehat{p}_i} \right)$ and hence, $c_{\widehat{\mathbf{p}}}^*(\mathbf{s}) = \sum_{i \in \mathscr{I}} \widehat{p}_i \phi^*(s_i)$, then our data-driven DRO approach boils down to that of Ben-Tal et al. (2013) and Bayraksan and Love (2015).
2. On the contrary, if we set $|\mathscr{I}| = 1$, c is given by a norm and take $\widetilde{c}_{\widehat{\mathbf{p}}}(\mathbf{p}) = \|\mathbf{p} - \widehat{\mathbf{p}}\|$ (hence $\widetilde{c}_{\widehat{\mathbf{p}}}^*(\mathbf{s}) = \sum_{i \in \mathscr{I}} \widehat{p}_i s_i$ if $\|\mathbf{s}\|_* \leq 1$), we get the model of Mohajerin Esfahani and Kuhn (2018).

Finally, we remark that constraint (6) for each $i \in I'$ is equivalent (under the assumptions we make on the transportation cost function) to $t_{i,1} \geq \sup_{\xi \in \varXi_i} f(x, \xi)$.

2.2 Tractable reformulations

In this section we provide *nice* reformulations of our DRO model (P) under mild assumptions. For this purpose, we make use of the theoretical foundations laid out in Mohajerin Esfahani and Kuhn (2018). Likewise, some extensions to our model, such as the extension to two-stage stochastic programming problems, are omitted here for brevity and because they can be easily derived in a similar way as found in Mohajerin Esfahani and Kuhn (2018) for the data-driven DRO approach they develop.

We start our theoretical development with the following assumption.

Assumption 1 *We consider that $\varXi_i$, for each $i \in \mathscr{I}$, is a closed convex set, and that $f(\mathbf{x}, \boldsymbol{\xi}) := \max_{k \leq K} g_k(\mathbf{x}, \boldsymbol{\xi})$, with g_k, for each $k \leq K$, being a proper, concave and upper semicontinuous function with respect to $\boldsymbol{\xi}$ (for any fixed value of $\mathbf{x} \in X$) and not identically ∞ on $\varXi_i$.*

Theorem 2 below provides a tractable reformulation of problem (P1) as a finite convex problem. For ease of notation, we suppress the dependence on the variable $\mathbf{x}$ (bearing in mind that this dependence occurs through functions g_k, $k \leq K$).

Theorem 2 *If Assumption* 1 *holds and if we choose a norm (in* $\mathbb{R}^d$ *) as the transportation cost function c, then for any values of* ρ *and* ε *, problem (P1) is equivalent to the following finite convex problem:*

$$
(P1') \quad \inf_{\mathbf{x},\lambda,\eta,\boldsymbol{\mu},\widetilde{\mathbf{p}},\mathbf{z}_{ijk},\mathbf{v}_{ijk},\theta,\mathbf{t}} \lambda\rho + \eta + \theta\varepsilon + \sum_{i\in\mathscr{I}} \widehat{p}_i \left(\frac{1}{N_i} \sum_{j=1}^{N_i} t_{i,j} + \mu_i - \eta + \widetilde{p}_i \right)
$$

$$
s.t. \ t_{i,j} \geq [-g_k]^*(\mathbf{z}_{ijk} - \mathbf{v}_{ijk}) + S_{\Xi_i}(\mathbf{v}_{ijk}) - \langle \mathbf{z}_{ijk}, \widehat{\boldsymbol{\xi}}_j^i \rangle
$$

$$
\forall i \in \mathscr{I}, \forall j \leq N_i, \forall k \leq K
$$

$$
\left\| \mathbf{z}_{ijk} \right\|_* \leq \theta, \forall i \in \mathscr{I}, \forall j \leq N_i, \forall k \leq K
$$

$$
\left\| \left(\frac{1}{N_i} \sum_{j=1}^{N_i} t_{i,j} + \mu_i - \eta + \widetilde{p}_i \right)_{i\in\mathscr{I}} \right\|_* \leq \lambda
$$

$$
\mathbf{x} \in X, \lambda \geq 0, \theta \geq 0, \eta \in \mathbb{R}, \boldsymbol{\mu} \in \mathbb{R}_+^{|\mathscr{I}|}, \widetilde{\mathbf{p}} \in \mathscr{C}^*,
$$

$$
\mathbf{z}_{ijk}, \mathbf{v}_{ijk} \in \mathbb{R}^d, \forall i \in \mathscr{I}, \forall j \leq N_i, \forall k \leq K
$$

$$
t_{i,j} \in \mathbb{R}, \forall i \in \mathscr{I}, \forall j \leq N_i
$$

where $[-g_k]^*(\mathbf{z}_{ijk} - \mathbf{v}_{ijk})$ is the conjugate function of $-g_k$ evaluated at $\mathbf{z}_{ijk} - \mathbf{v}_{ijk}$ and S_{Ξ_i} is the support function of Ξ_i.

Proof In essence, the complexity of problem (P1) depends on our ability to reformulate the supremum in constraint (23) in a tractable manner. This is possible under Asummption 1, following similar steps to those in the proof of Theorem 4.2 in Mohajerin Esfahani and Kuhn (2018), to which we refer. ∎

We note that Asummption 1 covers the particular case where functions $g_k, k \leq K$, are affine and, as a result, f is convex piecewise linear. The single-item newsvendor problem, which we illustrate in the first part of Sect. 4, constitutes a popular example of this case.

2.2.1 Separable objective function

Now we extend the results presented above to a class of objective functions which are additively separable with respect to the dimension d. We assume here that $\boldsymbol{\xi} = (\boldsymbol{\xi}_1, \ldots, \boldsymbol{\xi}_d)$, where $\boldsymbol{\xi}_l \in \mathbb{R}^p$, for each $l = 1, \ldots, d$. Furthermore, we consider the separable norm $\|\boldsymbol{\xi}\|_d := \sum_{l=1}^d \|\boldsymbol{\xi}_l\|$ associated with the base norm $\|\cdot\|$ (on $\mathbb{R}^p$). Finally, we assume that the function f is given as follows:

$$
f(\mathbf{x}, \boldsymbol{\xi}) = \sum_{l=1}^d \max_{k \leq K} g_{lk}(\mathbf{x}, \boldsymbol{\xi}_l) \tag{24}
$$

In this case, the complexity of the resulting DRO problem is linear with respect to the number N of samples. The multi-item newsvendor problem, which we illustrate in the second half of Sect. 4, constitutes a popular example of this case.

Theorem 3 *If $f(\mathbf{x}, \boldsymbol{\xi}) = \sum_{l=1}^{d} \max_{k \leq K} g_{lk}(\mathbf{x}, \boldsymbol{\xi}_l)$, $\{g_{lk}\}_{k \leq K}$ satisfy Assumption 1 for all $l \leq d$, and Ξ_i, for each $i \in \mathscr{I}$, is given by the Cartesian product of closed convex sets (that is, $\Xi_i := \prod_{l=1}^{d} D_l^i$, with D_l^i a closed convex set), and if we choose the norm $\|\cdot\|_d$ as the transportation cost function c, then for any values of ρ and ε, problem (P) is equivalent to the following finite convex problem:*

$$(P2) \quad \inf_{\mathbf{x}, \lambda, \eta, \boldsymbol{\mu}, \, \widetilde{\mathbf{p}}, \mathbf{z}_{ijkl}, \mathbf{v}_{ijkl}, \theta, \boldsymbol{\omega}} \lambda \rho + \eta + \theta \varepsilon + \sum_{i \in \mathscr{I}} \widehat{p}_i \left(\frac{1}{N_i} \sum_{j=1}^{N_i} \sum_{l=1}^{d} \omega_{ijl} + \mu_i - \eta + \widetilde{p}_i \right) \tag{25}$$

$$s.t. \ \omega_{ijl} \geq [-g_{lk}]^*(\mathbf{z}_{ijkl} - \mathbf{v}_{ijkl}) + S_{D_l^i}(\mathbf{v}_{ijkl}) - \langle \mathbf{z}_{ijkl}, \widehat{\boldsymbol{\xi}}_{jl}^i \rangle,$$
$$\forall i \in \mathscr{I}, \forall j \leq N_i, \forall k \leq K, \forall l \leq d \tag{26}$$

$$\|\mathbf{z}_{ijkl}\|_* \leq \theta, \forall i \in \mathscr{I}, \forall j \leq N_i, \forall k \leq K, \forall l \leq d \tag{27}$$

$$\left\| \left(\frac{1}{N_i} \sum_{j=1}^{N_i} \sum_{l=1}^{d} \omega_{ijl} + \mu_i - \eta + \widetilde{p}_i \right) \right\|_* \leq \lambda \tag{28}$$

$$\mathbf{x} \in X, \lambda \geq 0, \theta \geq 0, \eta \in \mathbb{R}, \boldsymbol{\mu} \in \mathbb{R}_+^{|\mathscr{I}|}, \ \widetilde{\mathbf{p}} \in \mathscr{C}^*, \tag{29}$$

$$\omega_{ijl} \in \mathbb{R}, \forall i \in \mathscr{I}, \forall j \leq N_i, \forall l \leq d \tag{30}$$

$$\mathbf{z}_{ijkl}, \mathbf{v}_{ijkl} \in \mathbb{R}^p, \forall i \in \mathscr{I}, \forall j \leq N_i, \forall k \leq K, \forall l \leq d \tag{31}$$

Proof The proof runs in a similar way to that of Theorem 6.1 in Mohajerin Esfahani and Kuhn (2018). □

Remarks If the transportation cost function c is not a norm, there are still some cases where the constraint (6) can be reformulated in a tractable way. In general, Eq. (6) can be seen as the robust counterpart of a constraint affected by the random parameter vector $\boldsymbol{\xi}$, with Ξ_i playing the role of the so-called *uncertainty set*. In our case, the tractability of (6) depends on the nature of each set Ξ_i and each function $\alpha_{ij}(\boldsymbol{\xi}) := f(\mathbf{x}, \boldsymbol{\xi}) - \theta c(\boldsymbol{\xi}, \widehat{\boldsymbol{\xi}}_j^i)$. Indeed, suppose that every Ξ_i is a closed convex set, then:

- If the function f is concave in $\boldsymbol{\xi}$, so is each function $\alpha_{ij}(\boldsymbol{\xi})$ (recall that the transportation cost function c is assumed to be convex and that θ is non-negative). As Roos et al. (2018) points out, this is a tractable instance and tractable reformulations of constraint (6) can be obtained using Fenchel duality following the guidelines in Ben-Tal et al. (2015).
- In contrast, the case in which some $\alpha_{ij}(\boldsymbol{\xi})$ are convex is much more challenging and may call for approximation methods such as the one proposed by Roos et al. (2018).

In any case, we need to compute convex conjugate functions, which is, in itself, a complicated problem in general. For assistance in this regard, one may resort to symbolic computation in order to get closed formulas for convex conjugate functions (see, for example, Borwein and Hamilton 2009).

2.3 Order cone constraints

To account for *a-priori* knowledge about the probability distribution of the random parameter vector ξ (for example, the decision maker may have some information about the shape of this distribution), we propose to convey this knowledge using order constraints on the probability masses p_i associated with each subregion $\varXi_i$ into which the support $\varXi$ of ξ is partitioned. These order constraints are based on order cones, which, in turn, can be represented in the form of graphs.

We can build order cones from graphs that allow for the comparison of all probabilities p_i. In that case, we say that the graph, and the associated cone, establish a *total order*. If, on the contrary, the graph only allows some of those probabilities to be compared, we talk about *partial order*. For more details about order cones we refer the reader to Németh and Németh ([2016](#)).

Below, we present some common choices of order cones.

– *Simple order cone (monotonicity):*

$$\mathscr{C} = \{p \in \mathbb{R}^{|\mathscr{I}|} : p_1 \geq \ldots \geq p_{|\mathscr{I}|}\}$$

– *Tree order cone:*

$$\mathscr{C} = \{p \in \mathbb{R}^{|\mathscr{I}|} : p_i \geq p_{|\mathscr{I}|}, \ i = 1, \ldots, |\mathscr{I}| - 1\}$$

– *Star-shaped cone (decrease on average):*

$$\mathscr{C} = \left\{p \in \mathbb{R}^{|\mathscr{I}|} : p_1 \geq \frac{p_1 + p_2}{2} \geq \ldots \geq \frac{p_1 + \ldots + p_{|\mathscr{I}|}}{|\mathscr{I}|}\right\}$$

– *Umbrella cone (unimodality):*

$$\mathscr{C} = \{p \in \mathbb{R}^{|\mathscr{I}|} : p_1 \leq p_2 \leq \ldots \leq p_m \geq p_{m+1} \geq \ldots \geq p_{|\mathscr{I}|}\}$$

An order cone is a *polyhedral convex cone* and as such, can be algebraically expressed in the form $\mathscr{C} = \{\mathbf{p} \in \mathbb{R}^{|\mathscr{I}|} : \mathbf{A}\mathbf{p} \geq 0\}$, with $\mathbf{A}$ being a matrix of appropriate dimensions. Its dual $\mathscr{C}^*$ can, therefore, be easily computed as $\mathscr{C}^* = \{\widetilde{\mathbf{p}} = \mathbf{A}^T \boldsymbol{\nu} : \boldsymbol{\nu} \geq \mathbf{0}\}$ (see, for instance, Corollary 3.12.9 in Silvapulle and Sen ([2011](#))). Notwithstanding, our DRO approach can be equally applied under other types of support sets, as long as the problem ([22](#)) admits a strong dual (we refer the interested reader to Ben-Tal et al. ([2013](#)) for a list of types of support sets under which strong duality holds).

As compared to other approaches available in the technical literature, order cones provide a straightforward way of encoding modality information in the ambiguity set of the DRO problem. For instance, Hanasusanto et al. ([2015](#)) indirectly introduces multi-modality information by imposing first and second moment conditions on the different ambiguous components of a mixture with known weights. Their approach, however, results in a semidefinite program. Unlike Hanasusanto et al. ([2015](#)), the authors in Li et al. ([2019](#)) explicitly incorporate modality information into their ambiguity set through moment and generalized unimodal constraints. Nonetheless, they still need

to solve a semidefinite program and their DRO approach overlooks the data-driven nature of those constraints. In Chen et al. (2019), they construct an ambiguity set made up of those absolutely continuous probability distributions whose density function is bounded by some bands with a certain confidence level. Their approach can be used to impose monotonicity or unimodality of the probability distributions, but can only be applied to the univariate case.

Beyond modality, the order cone constraints on the partition probabilities that characterize our DRO approach equip the decision maker with a versatile and intuitive framework to exploit information on the shape of the ambiguous probability distribution. For example, as we do in the numerical experiments in Sect. 4, we can construct an order cone that constrains the ratios among the partition probabilities, which can be seen as a discrete approximation of encoding "derivative" information on the ambiguous probability distribution (if this admits a density function). Likewise, other order cones could be used to bestow some sense of "convexity" on this distribution.

3 On convergence and out-of-sample performance guarantees

In this section, we show that our DRO approach (P) naturally inherits the convergence and performance guarantees of that introduced in Mohajerin Esfahani and Kuhn (2018). For this purpose, we first need to recall some terminology and concepts from this paper to which we will resort later on. Throughout this section, we denote the training data sample (that is, the sample path sequence) as $\widehat{\varXi}_N := \{\widehat{\boldsymbol{\xi}}^i\}_{i=1}^N \subseteq \varXi$. Following Mohajerin Esfahani and Kuhn (2018), $\widehat{\varXi}_N$ can be seen as a random vector governed by the probability distribution $\mathbb{P}^N := Q^* \times \cdots \times Q^*$ (N times) supported on $\varXi^N$ (with the respective product σ-algebra).

In the remainder of this paper, we will denote the optimization problem associated with problem (P) under the true probability distribution Q^* as (P*) (that is, the problem defined as $J^* := \inf_{\mathbf{x} \in X} \mathbb{E}_{Q^*}[f(\mathbf{x}, \boldsymbol{\xi})]$). We then say that a *data-driven solution* for problem (P*) is a feasible solution $\widehat{\mathbf{x}}_N \in X$ which is constructed from the sample data. Furthermore, the *out-of-sample performance* of a data-driven solution $\widehat{\mathbf{x}}_N$ is defined as $\mathbb{E}_{Q^*}[f(\widehat{\mathbf{x}}_N, \boldsymbol{\xi})]$.

In line with Mohajerin Esfahani and Kuhn (2018), given a data-driven solution $\widehat{\mathbf{x}}_N$, a *finite sample guarantee* is a relation in the form

$$\mathbb{P}^N\Big[\widehat{\varXi}_N \ : \ \mathbb{E}_{Q^*}[f(\widehat{\mathbf{x}}_N, \boldsymbol{\xi})] \leq \widehat{J}_N\Big] \geq 1 - \beta \tag{32}$$

where $\widehat{J}_N$ is a *certificate* for the out-of-sample performance of $\widehat{\mathbf{x}}_N$ (i.e., an upper bound that is generally contingent on the training dataset), $\beta \in (0, 1)$ is a *significance parameter* with respect to the distribution $\mathbb{P}^N$, on which both $\widehat{\mathbf{x}}_N$ and $\widehat{J}_N$ depend. Moreover, we refer to the probability on the left-hand side of (32) as the *reliability* of $(\widehat{\mathbf{x}}_N, \widehat{J}_N)$.

Ideally, we strive to develop a method capable of identifying a highly reliable data-driven solution with a certificate as low as possible.

The data-driven DRO approach that we propose in this paper to address problem (P*) accounts for the uncertainty about the true data-generating distribution Q^*, while taking advantage of some a-priori *order* information that the decision maker may have on some probabilities induced by Q^* over a partition of the support set Ξ. Below, we claim that the pair $(\widehat{\mathbf{x}}_N, \widehat{J}_N)$ provided by our distributionally robust optimization problem (P) features performance guarantees in line with those discussed in Mohajerin Esfahani and Kuhn (2018). More specifically, for a suitable choice of the ambiguity set, the optimal value $\widehat{J}_N$ of problem (P) constitutes a certificate of the type (32) providing a confidence level $1 - \beta$ on the out-of-sample performance of the data-driven solution $\widehat{\mathbf{x}}_N$. This can be formally stated under some assumptions about the underlying true conditional probability distributions.

To this end, we first provide probabilistic guarantees on the partition probabilities $p_i, \forall i \leq |\mathscr{I}|$. In this vein, note that the empirical probability $\widehat{p}_i$, defined as in Eq. (4), can be modeled as a binomial distribution with success probability p_i^*, divided by the total number of trials. Consequently, by the Strong Law of Large Numbers (SLLN), $\widehat{p}_i$ converges to p_i^* almost surely.

Now suppose that we choose a ϕ-divergence as $\widetilde{c}$, where ϕ is a twice continuously differentiable function around 1 with $\phi''(1) > 0$. Then, take $\beta_p > 0$. If we choose as ρ the value

$$\rho(\beta_p) := (\phi''(1)/(2N))\chi^2_{|\mathscr{I}|-1,1-\beta_p} \tag{33}$$

we get a confidence set of level $1 - \beta_p$ on the true partition probabilities $\mathbf{p}^*$ (see Ben-Tal et al. 2013 and Bayraksan and Love 2015).

If, alternatively, we choose the total variation distance as $\widetilde{c}$, we can use Eq. (19) in Guo and Xu (2019) to take ρ as

$$\rho(\beta_p) := (|\mathscr{I}|/\sqrt{N})(2 + \sqrt{2\log(|\mathscr{I}|/\beta_p)}) \tag{34}$$

and obtain a confidence set of level $1 - \beta_p$ on $\mathbf{p}^*$.

Next we establish a concentration tail inequality of the probability weighted Wasserstein metric of order 1 between each conditional distribution and its respective true conditional distribution. For this purpose, we first need to make the following assumption:

Assumption 2 (Light-tailed Conditional Distributions) *For each* $i \in \mathscr{I}$, *there exist* $a_i, \gamma_i \in \mathbb{R}$, *with* $a_i > 1$ *and* $\gamma_i > 0$ *such that*

$$\mathbb{E}_{Q_i^*}\big[\exp(\gamma_i \|\boldsymbol{\xi}\|^{a_i})\big] = \int_{\Xi_i} \exp(\gamma_i \|\boldsymbol{\xi}\|^{a_i}) \, Q_i^*(\mathrm{d}\boldsymbol{\xi}) < \infty. \tag{35}$$

The following theorem provides a tail concentration inequality for the weighted sum of the Wasserstein metrics of order 1 between the true and empirical conditional distributions.

Theorem 4 (Concentration Inequality for the Conditional Distributions) *If Assumption 2 holds, for each* $i \in \mathscr{I}$, *given* $\beta_i \in (0, 1]$ *we have that* $\forall N_i \geq 1$, $\dim(\boldsymbol{\xi}) \neq 2$

and for all $\varepsilon > \sum_{i \in \mathscr{I}} p_i \varepsilon_{N_i}(\beta_i)$, for any values $p_i, i \in \mathscr{I}$ such that $p_i \geq 0$ and $\sum_{i \in \mathscr{I}} p_i = 1$, the following holds

$$\mathbb{P}\left[\sum_{i \in \mathscr{I}} p_i \mathscr{W}(Q_i^*, \widehat{Q}_i) \leq \varepsilon\right] \geq 1 - \sum_{i \in \mathscr{I}} \beta_i \tag{36}$$

where

$$\varepsilon_{N_i}(\beta_i) := \begin{cases} \left(\dfrac{\log(B_i \beta_i^{-1})}{C_i N_i}\right)^{1/\max\{\dim(\boldsymbol{\xi}), 2\}} & \text{if } N_i \geq \dfrac{\log(B_i \beta_i^{-1})}{C_i}, \\ \left(\dfrac{\log(B_i \beta_i^{-1})}{C_i N_i}\right)^{1/a_i} & \text{if } N_i < \dfrac{\log(B_i \beta_i^{-1})}{C_i}. \end{cases} \tag{37}$$

Proof Given Assumption 3, for all $i \in \mathscr{I}$, we deduce from Theorem 2 in Fournier and Guillin (2015) that

$$\mathbb{P}\left[\mathscr{W}(Q_i^*, \widehat{Q}_i) \leq \varepsilon_{N_i}(\beta_i)\right] \geq 1 - \beta_i.$$

Thus, we have that

$$\mathbb{P}\left[\sum_{i \in \mathscr{I}} p_i \mathscr{W}(Q_i^*, \widehat{Q}_i) \leq \sum_{i \in \mathscr{I}} p_i \varepsilon_{N_i}(\beta_i)\right]$$

$$\geq \mathbb{P}\left[\bigcap_{i \in \mathscr{I}} \left(p_i \mathscr{W}(Q_i^*, \widehat{Q}_i) \leq p_i \varepsilon_{N_i}(\beta_i)\right)\right] \tag{38}$$

$$= 1 - \mathbb{P}\left[\bigcup_{i \in \mathscr{I}} \left(p_i \mathscr{W}(Q_i^*, \widehat{Q}_i) > p_i \varepsilon_{N_i}(\beta_i)\right)\right] \tag{39}$$

$$\geq 1 - \sum_{i \in \mathscr{I}} \mathbb{P}\left[p_i \mathscr{W}(Q_i^*, \widehat{Q}_i) > p_i \varepsilon_{N_i}(\beta_i)\right] \tag{40}$$

$$\geq 1 - \sum_{i \in \mathscr{I}} \beta_i \tag{41}$$

Theorem 4 sets the probabilistic bound $\sum_{i \in \mathscr{I}} p_i \varepsilon_{N_i}(\beta_i)$ on the weighted Wasserstein metric of order 1 between each conditional distribution and its respective true conditional distribution, with at least confidence level $1 - \sum_{i \in \mathscr{I}} \beta_i$. We remark that, if the partitions are compact, stronger results like those in Theorem 2 of Ji and Lejeune (2020) could be used to choose the radii of the Wasserstein balls. More specifically, the result in Theorem 2 of Ji and Lejeune (2020) depends on the diameter of the compact support set (i.e., the maximum distance between two elements of that set). The result stated in our theorem, in contrast, is valid for unbounded partitions, as it only requires the true conditional distribution associated with each partition be light-tailed. The next

theorem states the finite-sample guarantee performance of the proposed DRO method we develop in this paper: □

Theorem 5 (Finite sample guarantee) *Suppose that Assumption 2 holds and that we have chosen as ρ the value given by Eqs.* (33) *or* (34). *Then, the finite sample guarantee* (32) *holds with at least confidence level* $(1 - \beta_p)(1 - \sum_{i \in \mathscr{I}} \beta_i)$.

Proof The claim follows from Theorem 4 and Eqs. (33) and (34), which imply that $\mathbb{P}(Q^* \in \mathscr{U}_{\rho,\varepsilon}(\widehat{Q})) \geq (1 - \beta_p)(1 - \sum_{i \in \mathscr{I}} \beta_i)$. Hence,

$$\mathbb{E}_{Q^*}[f(\widehat{\mathbf{x}}_N, \boldsymbol{\xi})] \leq \sup_{Q \in \mathscr{U}_{\rho,\varepsilon}(\widehat{Q})} \mathbb{E}_Q[f(\widehat{\mathbf{x}}_N, \boldsymbol{\xi})] = \widehat{J}_N$$

with probability at least $(1 - \beta_p)(1 - \sum_{i \in \mathscr{I}} \beta_i)$. □

Remarks In practice, proper values for ε and ρ are set by way of data-driven procedures like bootstrapping or cross-validation, as we illustrate in the numerical experiments in Sect. 4.2 (see also Chen and Paschalidis 2018; Cisneros-Velarde et al. 2020; Mohajerin Esfahani and Kuhn 2018; Shafieezadeh-Abadeh et al. 2019; and Xie 2021 for more examples). These procedures allow the decision maker to tune those parameters as a function of sample size N in order to get reliable decisions without giving up too much on out-of-sample performance. Following this line, and as noted in Remark 5 in Kuhn et al. (2019), the requirement to include the true distribution inside the ambiguity set is only a sufficient, but *not necessary* condition to ensure a finite sample guarantee. Indeed, this guarantee can be sustained even if the parameters of the ambiguity set are reduced below the lowest values for which the ambiguity set represents a confidence set for the true distribution.

Furthermore, recall that the partition probabilities **p** belong to the support set Θ defined by the order cone constraints. Since we assume that these constraints are coherent with the true distribution Q^*, we do not need to explore those probability measures Q in the Wasserstein ball $\mathbb{B}_{\rho_N(\beta)}$ that do not comply with them. Consider, for example, the case in which the worst-case distribution in the ball $\mathbb{B}_{\rho_N(\beta)}$ does not satisfy the order cone constraints. One could expect, therefore, that, in practice, our approach could benefit from this fact to produce a data-driven solution $\widehat{\mathbf{x}}_N$ as reliable as that given by the method of Mohajerin Esfahani and Kuhn (2018), but with a tighter certificate $\widehat{J}_N$. This is precisely what we observe in the numerical experiments that we present below.

We conclude this section with some remarks on the convergence and asymptotic consistency of our DRO approach: We have that, as the number N of samples grows to infinity,

$$(\widehat{\mathbf{x}}_N, \widehat{J}_N) \to (\mathbf{x}^*, J^*) \tag{42}$$

where $\mathbf{x}^*$ (resp. J^*) is an optimizer (resp. the optimal solution value) of problem (P*).

Indeed, assume that Theorem 3.6 in Mohajerin Esfahani and Kuhn (2018) holds, then take a confidence level $1 - \beta$, and choose ε and ρ by way of Theorem 4 and Eqs. (33) (or (34)), respectively. When N grows to infinity, we have, on the one hand, that the conditional distributions converge (in the Wasserstein metric) to their

respective true conditional distributions and the probability weights converge a.s. by the SLLN to their respective true values. Therefore, both ε and ρ tend to zero as N increases to infinity. Consequently, our ambiguity set only contains the empirical distribution $\widehat{Q}_N$, which converges almost surely to the true distribution Q^*.

4 Numerical experiments

The following simulation experiments are designed to provide additional insights into the performance guarantees of our proposed distributionally robust optimization scheme with order cone constraints. For this purpose, we consider two test instances: the (single and multi-item) newsvendor problem and the problem of a strategic firm competing à la Cournot in a market. These two problems have been intentionally selected, because they are qualitatively different when addressed by the standard Wasserstein-metric-based DRO approach proposed in Mohajerin Esfahani and Kuhn (2018). In effect, the former features an objective function $f(\mathbf{x}, \boldsymbol{\xi})$ whose Lipschitz constant with respect to $\boldsymbol{\xi}$ is independent of the decision $\mathbf{x}$. Consequently, as per Remark 6.7 in Mohajerin Esfahani and Kuhn (2018), the standard Wasserstein-metric-based DRO approach renders the same minimizer for this problem as the sample average approximation, whenever the support of the uncertainty $\boldsymbol{\xi}$ is unbounded. This is, in contrast, not true for the problem of a strategic firm competing à la Cournot in a market, which is characterized by an objective function with a Lipschitz constant over $\boldsymbol{\xi}$ that is a function of $\mathbf{x}$. This allows us to highlight the differences of our approach with regard to Mohajerin Esfahani and Kuhn (2018) in two distinct settings.

All the numerical experiments have been implemented in Python. The optimization problems have been built using Pyomo (2019) and solved with CPLEX 12.10 CPLEX Optimizer-IBM (2019) on a PC with Windows 10 and a CPU Intel (R) Core i7-8550U clocking at 1.80 GHz and with 8 GB of RAM. The statistical methods that have been used for the numerical experiments have been coded by means of the module Scikit-learn (see Pedregosa et al. 2011). In what follows we provide some implementation details regarding the proposed model. The numerical experiments have been designed under the following assumptions:

1. *A-priori information*. Given a fixed and known partition of the sample space $\varXi$, we can construct an order cone that is consistent with the true probability distribution. That is, the probability masses that the true distribution assigns to each partition verify the order cone constraints. In practice, this a-priori information is determined by the nature of the problem and the random phenomena, and is assumed to be known by the decision maker based on experience and expert knowledge. Furthermore, in the case that the decision maker has no full certainty about the a-priori information, s/he may resort to statistical hypothesis testing to assess the confidence that the partition probabilities belong to a given order cone (see, for instance, Bhattacharya (1997) and references therein).
   In our numerical experiments, we specifically apply the following approach: Given a fixed number of partitions (later we explain how the partition set is obtained), we consider that the decision maker knows a total order between the probability

masses associated with each of the regions into which the sample space $\varXi$ is split. Furthermore, s/he also knows their ratios approximately, within a certain tolerance (which, in the subsequent experiments, we set to 0.1).

For instance, suppose we have three partitions with (true) probability masses of $p_1^* = 0.6$, $p_2^* = 0.3$ and $p_3^* = 0.1$. The decision maker only knows their relative ratios with a tolerance error of 0.1, that is:

$$p_1 \geq (0.6/0.3 - 0.1)\,p_2$$
$$p_2 \geq (0.3/0.1 - 0.1)\,p_3$$

This way, we get the following order cone constraints:

$$p_1 \geq 1.9\,p_2$$
$$p_2 \geq 2.9\,p_3$$

2. *Support set $\varXi$.* The support set is the Cartesian product of closed intervals (that is, an hypercube, whose size is indicated in each example) and, therefore, is a closed convex set.

3. *True distribution.* For simulation and analysis, the data-generating distribution is approximated by a certain number of data points (15,000 in the newsvendor setting and 10,000 in the problem of the Cournot producer) drawn from a mixture of three normal distributions, whose characteristics are specified in each of the two examples we consider in the following subsections. Furthermore, those data points that fall outside the support set $\varXi$ are discarded.

4. *Construction of partitions $\varXi_i$, $i = 1, \ldots, |\mathscr{I}|$:* In order to construct the partitions, we proceed as follows:

   (a) *Clustering phase:* Firstly, we employ the K-means clustering technique to group the total number of data points that approximate the true data distribution into K clusters. The number K of clusters is decided upon using the well-known Elbow's method (see, for example, Dangeti 2017). It is based on the value of the average distortion produced by different values of K. If K increases, the average distortion will decrease and the improvement in average distortion will diminish. The value of K at which the improvement in distortion decreases the most is called the *elbow*. At this value of K, we should stop dividing the data into further clusters and choose this value as the number of clusters. In addition, we assign a label to identify each of the K clusters. In all the numerical experiments that are presented next, the true data-generating distribution is constructed as a mixture of three (univariate or multivariate) normal distributions. We assume that the decision maker has a good estimate of the number of components of this mixture and thus, we consider, for example, four clusters, i.e., $K = 4$.

   (b) *Decision-tree classifier phase:* Once all the clusters have been labelled, we use the aforementioned total number of data points to train a decision-tree multi-classifier with a maximum number of leafs equal to K. The tree will be then used to allocate new data points into one of the K clusters, which, in effect, is equivalent to having a partition of the support set in K disjoint regions.

5. *Comparative analysis:* We compare three different data-driven approaches to address the solution to problem $\inf_{\mathbf{x} \in X} \ \mathbb{E}_{Q^*}\left[f(\mathbf{x}, \boldsymbol{\xi})\right]$, namely, our approach (DROC), the one of Mohajerin Esfahani and Kuhn (2018) (DROW) and the sample average approximation (SAA). Recall that we denote $x^* \in arg\,min_{\mathbf{x} \in X} \mathbb{E}_{Q^*}\left[f(\mathbf{x}, \boldsymbol{\xi})\right]$ and $J^* = \mathbb{E}_{Q^*}\left[f(\mathbf{x}^*, \boldsymbol{\xi})\right]$, which, in practice, are unknown to the decision maker, but, for analysis purposes, we estimate using the total number of data points that approximate the true data-generating distribution. Moreover, in all numerical experiments, we consider the 1-norm as the functions c and $\widetilde{c}$. To compare the three data-driven approaches we consider, we use two performance metrics, specifically, the *out-of-sample performance* of the data-driven solution (which we also refer to as its *actual expected cost*) and its *out-of-sample disappointment*. The former is given by $\mathbb{E}_{Q^*}\left[f(\widehat{\mathbf{x}}_N^m, \boldsymbol{\xi})\right]$, while the latter is calculated as $J^* - \widehat{J}_N^m$, where $m = \{\text{DROC}, \text{DROW}, \text{SAA}\}$ and $\widehat{J}_N^m$ is the objective function value yielded by the data-driven optimization problem solved by method m. We stress that a negative out-of-sample disappointment represents a favourable outcome. As $\mathbb{E}_{Q^*}\left[f(\widehat{\mathbf{x}}_N^m, \boldsymbol{\xi})\right]$ and $\widehat{J}_N^m$ are random variables (they are direct functions of the sample data), we conduct a certain number of runs, each with an independent sample of size N. This way we can provide (visual) estimates of the expected value and variability of the out-of-sample performance and disappointment for several values of the sample size N. These estimates are illustrated in the form of box plots in a series of figures. In these figures, the dotted black horizontal line corresponds to either solution x^* or to its associated optimal cost J^* with complete information (i.e., without ambiguity about the true data distribution).

For the sole purpose of conducting a comparison as fairly as possible, parameters ε and ρ in both DROC and DROW are tuned so that the underlying true distribution of the data belongs to the corresponding ambiguity set with, at least, a pre-fixed confidence level of probability. In the case of the newsvendor examples, we guarantee this by trial and error for simplicity. In practice, however, these parameters should be calibrated by way of a (statistical) procedure that uses the data available to the decision maker, for example, through cross-validation or bootstrapping. We follow this approach in the problem of the Cournot producer. Finally, we stress that, in our approach, caution should be exercised when selecting ε and ρ, as they should be such that problem (P) has at least one feasible solution. This is not guaranteed in the case that the empirical distribution $\widehat{Q}$ does not satisfy the order cone constraints on the probability masses associated with each subregion $\varXi_i$ of the support set $\varXi$. Intuitively, in this case, optimization problem (P) must have enough "budget" (i.e., ε and ρ must be high enough) to "transport" the empirical distribution to another one that complies with the a-priori information. In other words, the ambiguity set of problem (P) must be sufficiently large to contain at least one probability distribution that assigns probability masses verifying the order cone constraints to the partitions.

4.1 Newsvendor problems

In this section, we illustrate the theoretical results of our paper on the popular *newsvendor* problem (also known as the *newsboy problem*). Many extensions and variants of this problem have been considered since it was first posed in the 50s (see, for example, the work in Gallego and Moon (1993), Choi (2012), Andersson et al. (2013), Pando et al. (2014), and references therein). According to Pando et al. (2013),

The newsboy problem is probably the most studied stochastic inventory model in inventory control theory and the one with most extensions in recent years. This problem reflects many real-life situations and is often used to aid decision making in both manufacturing and retailing. It is particularly important for items with significant demand uncertainty and large over-stocking and under-stocking costs.

4.1.1 The single-item newsvendor problem

In the single-item newsvendor model, the decision maker has to plan the inventory level for a certain product before the random demand ξ for that product is realized, facing both holding and backorder costs. The newsvendor problem can be formulated as

$$\inf_{x \geq 0} \mathbb{E}_Q[h(x - \xi)^+ + b(\xi - x)^+]$$

where x is the order quantity, and $b, h > 0$ are the unit holding cost and the unit backorder cost, respectively. Here we have assumed that $h = 4$ and $b = 2$.

The demand for the item (unknown to the decision maker) is assumed to follow a mixture (with weights $\omega_1 = 0.1$, $\omega_2 = 0.35$ and $\omega_3 = 0.55$) of the three normal distributions $\mathcal{N}_1(0.2, 0.05)$, $\mathcal{N}_1(0.5, 0.1)$, and $\mathcal{N}_1(0.8, 0.05)$, truncated over the unit interval [0, 1]. Figure 1a provides a visual illustration of the resulting mixture. Recall that, in the numerical experiments that follow, we have used 15,000 samples drawn from this mixture of Gaussian distributions to approximate the true distribution of the item demand and to partition its support set [0, 1] into four regions, based on the two-phase procedure we have previously described. In fact, what we show in Fig. 1a is the histogram of those 15,000 data points and its corresponding kernel density estimate.

For the sole purpose of conducting a comparison as fairly as possible, parameters ε and ρ in both DROC and DROW are tuned so that the underlying true distribution of the data belongs to the corresponding ambiguity set with at least 95% of probability. We check whether this condition holds or not *a posteriori* (by trial and error), by counting the number of runs (out of the one thousand we perform) for which the out-of-sample disappointment is negative.

The values we have used for the parameters ε and ρ in DROC and DROW are collated in Table 1. We insist that these parameters have been adjusted so that at most 50 out of the 1000 runs we have conducted for each sample size N deliver a positive out-of-sample disappointment (that is, to achieve and maintain a similar level of reliability for the data-driven solutions given by DROC and DROW). As expected, therefore, the values of both ε and ρ decrease as the sample size N grows.

Table 1 Single-item newsvendor problem: values for parameters ε, ρ in DROC and ρ in DROW

N	DROC		DROW
	ε	ρ	ρ
2	0.9	0.9	1
5	0.8	0.8	0.9
10	0.7	0.7	0.8
20	0.4	0.6	0.6
50	0.15	0.25	0.4
100	0.1	0.2	0.25
200	0.01	0.15	0.05

Figure 1b–d show the box plots corresponding to the order quantity, the out-of-sample disappointment and the actual expected cost delivered by each of the considered data-driven approaches for various sample sizes. The shaded areas have been obtained by joining the whiskers of the box plots, while the associated solid lines link their medians. Interestingly, whereas the medians of the order quantity estimators provided by SAA are very close to the optimal one x^*, their high variability results in (large) disappointment with very high probability. On the contrary, the median of the order quantity delivered by DROW is significantly far from the optimal one (with complete information) for small sample sizes, but it manages to keep the out-of-sample disappointment below zero in return. To do so, however, DROW tends to produce costly (overconservative) solutions on average, as inferred from their actual expected cost in Fig. 1d. In plain words, DROW pays quite a lot to ensure a highly reliable/robust order quantity. The proposed approach DDRO, however, is able to leverage the a-priori information on the partition probabilities $(p_i)_{i=1}^{|\mathscr{I}|}$ to substantially reduce the cost to pay for reliable data-driven solutions, especially for small sample sizes. Intuitively, this information enables DROC to identify highly reliable solutions that are myopically deemed as non-reliable and, therefore, discarded by DROW. Logically, this is contingent on the quality of the a-priori information that is supplied to DROC in the form of order cone constraints on $(p_i)_{i=1}^{|\mathscr{I}|}$.
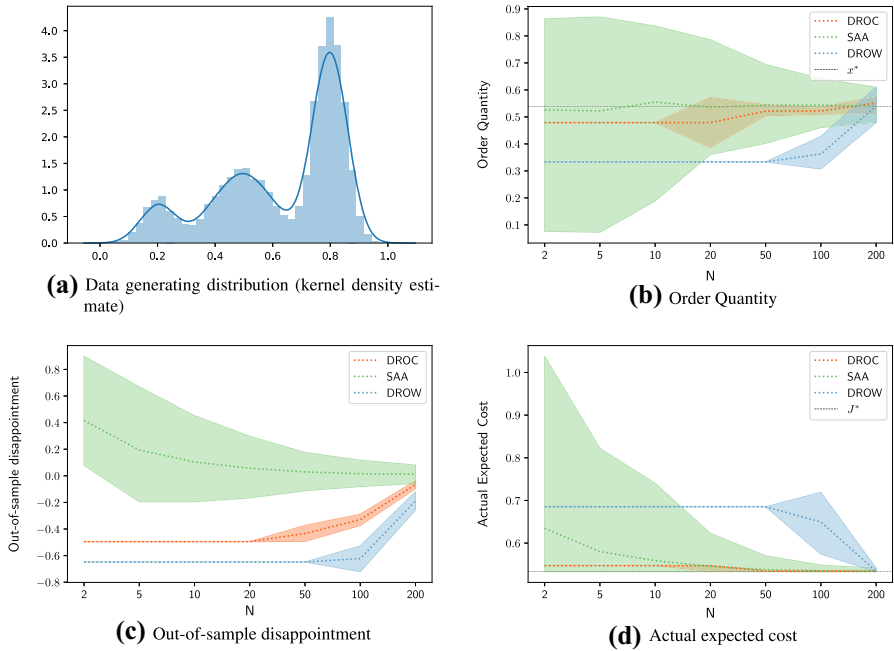
4.1.2 The multi-item newsvendor problem

In this section, we carry out an analysis similar to that of Sect. 4.1.1, but for the multi-item newsvendor problem, which can be formulated as follows:

$$\inf_{\mathbf{x} \geq \mathbf{0}} \mathbb{E}_Q \sum_{l=1}^{d} [h_l(x_l - \xi_l)^+ + b_l(\xi_l - x_l)^+]$$

where x_l is the order quantity for the l-th item, Q is the joint probability distribution governing the demands for the d items, and $b_l, h_l > 0$ are the unit holding cost and the unit backorder cost for the l-th item, respectively.

To illustrate our approach in a higher dimensional setting, we consider twenty items, i.e., $d = 20$.

(a) Data generating distribution (kernel density estimate)

(b) Order Quantity

(c) Out-of-sample disappointment

(d) Actual expected cost

Fig. 1 Single-item newsvendor problem: (Approximate) true data-generating distribution, order quantity and performance metrics

Table 2 Multi-item newsvendor problem: values for parameters ε, ρ in DROC and ρ in DROW

N	DROC		DROW
	ε	ρ	ρ
2	5	2	60
5	5	2	50
10	4.5	1.5	40
20	4	1	20
50	2.5	0.6	10
100	1.75	0.5	8
200	1.25	0.35	4

We consider the following parameters: $h_1 = \cdots = h_{10} = 2$, $h_{11} = \cdots = h_{20} = 4$, $b_1 = \cdots = b_{10} = 4$; and $b_{11} = \cdots = b_{20} = 2$. The demands for the twenty items are assumed to follow a mixture of three multivariate normal distributions $\mathcal{N}_{20}(\boldsymbol{\mu}_1, \Sigma_1)$, $\mathcal{N}_{20}(\boldsymbol{\mu}_2, \Sigma_2)$, and $\mathcal{N}_{20}(\boldsymbol{\mu}_3, \Sigma_3)$, where $\boldsymbol{\mu}_1 = [3, \ldots, 3] \in \mathbb{R}^{20}$, $\Sigma_1 = \mathrm{diag}(1, \ldots, 1) \in \mathbb{R}^{20 \times 20}$; $\boldsymbol{\mu}_2 = [5, \ldots, 5] \in \mathbb{R}^{20}$, $\Sigma_2 = \mathrm{diag}(0.5, \ldots, 0.5) \in \mathbb{R}^{20 \times 20}$; and $\boldsymbol{\mu}_3 = [7, \ldots, 7] \in \mathbb{R}^{20}$, $\Sigma_3 = \mathrm{diag}(0.1, \ldots, 0.1) \in \mathbb{R}^{20 \times 20}$. The weights of the mixture are $\omega_1 = 0.1$, $\omega_2 = 0.65$ and $\omega_3 = 0.25$, respectively. Furthermore, the mixture has been truncated on the hypercube $[0, 10]^{20}$.

The values we have used for the parameters ε and ρ in DROC and DROW are collated in Table 2.
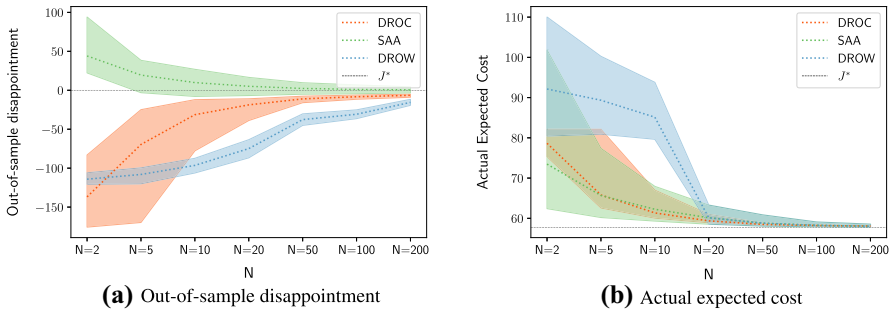
Fig. 2 Multi-item newsvendor problem: Performance metrics

Again, for a meaningful and fair comparison, these parameters have been tuned by trial and error in such a way that at most 50 out of the 1000 runs we have carried out for each sample size N yield a positive out-of-sample disappointment. The values for the parameters, which we need to this end, diminish as we gain more information (i.e., as the sample size N grows). Note that, for small sample sizes, for which the available data provide very little information about their true distribution, a great deal of robustness is required to produce highly reliable data-driven solutions. Consequently, it is little wonder that the selected values for ρ in DROC are equal to two, which is the maximum value that the total variation distance between P and $\widehat{P}$ can take on.

In the same fashion as in the case of the previous example of the single-item newsvendor problem, Fig. 2a and b show, for various sample sizes, the box plots pertaining to the out-of-sample disappointment and the actual expected cost associated with each of the considered data-driven approaches, in that order. The results conveyed by these figures confirm our initial conclusions: The ability of our approach DROC to exploit a-priori knowledge of the order among some partition probabilities permits identifying solutions that perform noticeably better out of sample with the same level of confidence. We underline that, in terms of the out-of-sample disappointment, the decision maker seeks a data-driven method m that renders an estimate $\widehat{J}_N^m$ that results in a positive surprise (i.e., negative disappointment) with a high probability, but that is as close as possible to the cost with full information J^*. Consequently, the large negative out-of-sample disappointment that the solutions given by DROW feature can be attributed to its over-conservativeness.

In terms of computational time, solving DROC for this instance of the multi-item newsvendor problem, with 20 items, four partitions and a sample size of 200, takes less than a second with CPLEX 12.10 running on a Windows 10 PC with a CPU Intel (R) Core i7-8550U clocking at 1.80 GHz and 8 GB of RAM.

4.2 The problem of a strategic firm competing à la Cournot in a market

Next we consider the problem of a strategic firm competing à la Cournot in a market for an undifferentiated product. This could be the case of, for instance, the electricity market (see, e.g., (Gabriel et al. 2012, Ch. 3) and Ruiz et al. 2008). Suppose the firm can produce up to one per-unit amount of product at a cost given by $a_2 x^2 + a_1 x + a_0$,

where x is the per-unit amount of product eventually produced and a_0, a_1 and a_2 are *known* parameters taking values in $\mathbb{R}^+$. Furthermore, assume an inverse residual demand function in the form $\lambda = \alpha - \beta x$, where λ is the market clearing price for the product, and $\alpha, \beta \in \mathbb{R}^+$ are *unknown and uncertain* parameters. The firm seeks, therefore, to minimize its cost $(a_2 x^2 + a_1 x + a_0) - \lambda x$ subject to $x \in [0, 1]$. After some basic manipulation, the problem of the firm can be posed as

$$\inf_{x \in [0,1]} \mathbb{E}_Q[(-x)\xi + x^2]$$

where $\xi = \frac{\alpha - a_1}{\beta + a_2}$.

The most interesting feature of this example is that, unlike in the aforementioned newsvendor problems, the Lipschitz constant of the objective function $f(x, \xi) := (-x)\xi + x^2$ with respect to $\boldsymbol{\xi}$ is dependent on the decision variable $\mathbf{x}$.

We consider that ξ follows a (true) probability distribution given by 10,000 points sampled from a mixture of three Gaussian distributions with variances all equal to 0.3 and means $\mu_1 = 0, \mu_2 = 1.2$ and $\mu_3 = 2.5$. The weights of the mixture are $\omega_1 = 0.5, \omega_2 = 0.2$ and $\omega = 30.3$. Furthermore, the mixture has been truncated (over the interval $[-1.8, 3]$. Figure 3a plots the kernel estimate of the data-generating distribution.

As in the previous experiments, we have divided the support $[-1.8, 3]$ into four partitions, using the procedure described at the beginning of Sect. 4. However, in a different way to what we did in the newsvendor examples, here we select parameters ε and ρ following a procedure that solely relies on the available data, similarly to what is done in Mohajerin Esfahani and Kuhn (2018). Essentially, given a desired confidence level $(1 - \beta)$ for the finite-sample guarantee (set to 0.85 in our numerical experiments), we need to *estimate*, using the data sample available only, the parameters ε and ρ that deliver, at least, this confidence level while yielding the best out-of-sample performance. To this end, we use bootstrapping. The estimator of those parameters is denoted as $param_N^m(\beta)$, underlining that the number and type of parameters to be estimated depend on the method m. The estimation procedure is carried out as follows for each sample of size N (in this experiment, we consider 300 independent data samples for each size N):

1. We construct *kboot* resamples of size N (with replacement), each playing the role of a different training dataset. Moreover, take those data points that have not been resampled to form a validation dataset (one per resample of size N). In our experiments below, we have considered *kboot* = 50.
2. For each resample $k = 1, \ldots, kboot$ and each candidate value for *param*, get a DRO solution from method j with parameter (or pair of paramaters) *param* on the k-th resample. The resulting optimal decision is denoted as $\widehat{x}_N^{j,k}(param)$ and its associated objective value as $\widehat{J}_N^{j,k}(param)$. Subsequently, we compute the out-of-sample performance $J(\widehat{x}_N^{j,k}(param))$ of the data-driven solution $\widehat{x}_N^{j,k}(param)$ over the k-th validation dataset.
3. From among the candidate values for *param* such that $\widehat{J}_N^{j,k}(param)$ exceeds the value $J(\widehat{x}_N^{j,k}(param))$ in at least $(1 - \beta) \times kboot$ different resamples, take the

(a) Data generating distribution (kernel density estimate)

(b) Optimal solution

(c) Out-of-sample disappointment

(d) Actual expected cost
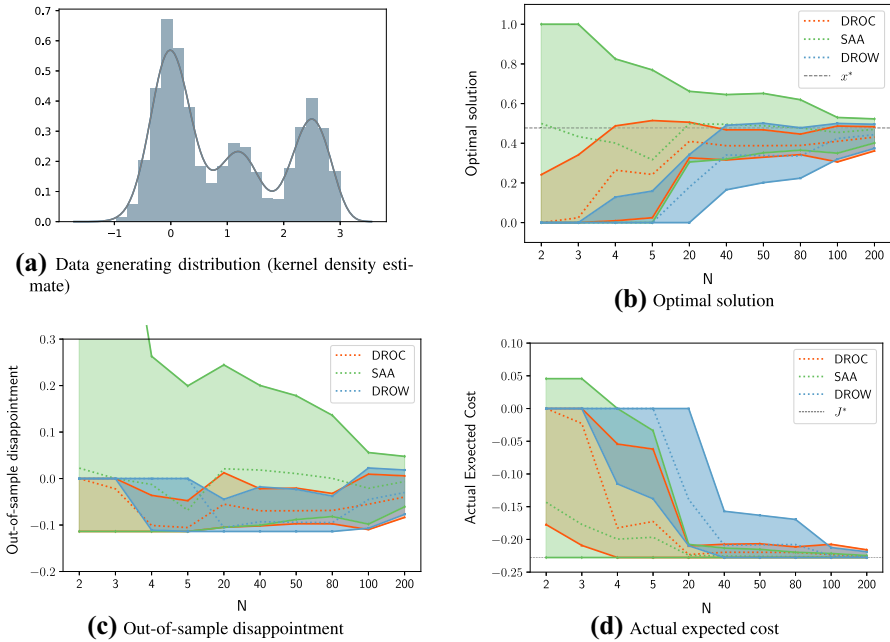
Fig. 3 Strategic firm problem: (Approximate) true data-generating distribution, optimal solution and performance metrics

one with the lowest $\frac{\sum_{k=1}^{kboot} J(\widehat{x}_N^{j,k}(param))}{kboot}$ (that is, with the highest out-of-sample performance averaged over the *kboot* resamples).

4. Finally, compute the solution given by method j with parameter $param_N^{\beta,j}$, $\widehat{x}_N^j :=$ $\widehat{x}_N^j(param_N^{\beta,j})$ and the respective certificate $\widehat{J}_N^j := \widehat{J}_N^j(param_N^{\beta,j})$.

As for the newsvendor examples, Fig. 3b–d show, for various sample sizes, the box plots pertaining to the optimal decision, the out-of-sample disappointment and the actual expected cost associated with each of the considered data-driven approaches, in that order. Once again, the results conveyed by these figures confirm our previous conclusions: Our approach DROC is able to leverage a-priori knowledge of the order among some partition probabilities to deliver solutions that perform significantly better out of sample for the same level of confidence. Furthermore, we see that the decision computed by the proposed method DROC converges to the true optimal solution (with complete information) faster than the solutions provided by the other methods.

5 Conclusions

In this paper, we have presented a novel framework for data-driven distributionally robust optimization (DRO) based on optimal transport theory in combination with order cone constraints to leverage *a-priori* information on the true data-generating distribution. Motivated by the reported over-conservativeness of the traditional DRO

approach based on the Wasserstein metric, we have formulated an ambiguity set able to incorporate information about the order among the probabilities that the true distribution of the problem's uncertain parameters assigns to some subregions of its support set. Our approach can accomodate a wide range of shape information (such as that related to monotonicity or multi-modality) in a practical and intuitive way. Moreover, under mild assumptions, the resulting distributionally robust optimization problem can be, in fact, reformulated as a finite convex problem where the a-priori information (expressed through the order cone constraints) are cast as linear constraints as opposed to the more computationally challenging formulations that exist in the literature. Furthermore, our approach is supported by theoretical performance guarantees and is capable of turning the provided information into solutions with increased reliability and improved performance, as illustrated by the numerical experiments we have prepared based on the well-known newsvendor problem and the problem of a strategic firm competing á la Cournot in a market for a homogeneous product.

Declaration

Conflict of interest The authors declare that they have no conflict of interest.

References

Andersson J, Jörnsten K, Nonås SL, Sandal L, Ubøe J (2013) A maximum entropy approach to the newsvendor problem with partial information. Eur J Oper Res 228(1):190–200. https://doi.org/10.1016/j.ejor.2013.01.031

Bayraksan G, Love DK (2015) Data-driven stochastic programming using phi-divergences. In: The operations research revolution, pp 1–19. INFORMS. https://doi.org/10.1287/educ.2015.0134

Ben-Tal A, den Hertog D, De Waegenaere A, Melenberg B, Rennen G (2013) Robust solutions of optimization problems affected by uncertain probabilities. Manag Sci 59(2):341–357. https://doi.org/10.1287/mnsc.1120.1641

Ben-Tal A, den Hertog D, Vial JP (2015) Deriving robust counterparts of nonlinear uncertain inequalities. Math Program 149(1–2):265–299. https://doi.org/10.1007/s10107-014-0750-8

Bertsimas D, Gupta V, Kallus N (2018) Data-driven robust optimization. Math Program 167(2):235–292. https://doi.org/10.1007/s10107-017-1125-8

Bertsimas D, Gupta V, Kallus N (2018) Robust sample average approximation. Math Program 171(1–2):217–282. https://doi.org/10.1007/s10107-017-1174-z

Bhattacharya B (1997) Testing multinomial parameters under order restrictions. Commun Stat Theory Methods 26(8):1839–1865. https://doi.org/10.1080/03610929708832017

Blanchet J, Kang Y, Murthy K, Zhang F (2019) Data-driven optimal transport cost selection for distributionally robust optimization. In: 2019 winter simulation conference (WSC), pp 3740–3751. https://doi.org/10.1109/WSC40007.2019.9004785

Blanchet J, Kang Y, Zhang F, He F, Hu Z (2021) Doubly robust data-driven distributionally robust optimization. In: Applied modeling techniques and data analysis 1, pp 75–90. Wiley. https://doi.org/10.1002/9781119821588.ch4

Borwein JM, Hamilton CH (2009) Symbolic Fenchel conjugation. Math Program 116(1–2):17–35. https://doi.org/10.1007/s10107-007-0134-4

Boyd S, Vandenberghe L (2004) Convex optimization. https://doi.org/10.1017/CBO9780511804441

Chen R, Paschalidis IC (2018) A robust learning approach for regression models based on distributionally robust optimization. J Mach Learn Res 19(13):1–48

Chen X, Lin Q, Xu G (2019) Distributionally robust optimization with confidence bands for probability density functions. arXiv:1901.02169

Chen Z, Sim M, Xiong P (2020) Robust stochastic optimization made easy with RSOME. Manag Sci 66(8):3329–3339. https://doi.org/10.1287/mnsc.2020.3603

Choi TM (ed) (2012) Handbook of Newsvendor problems, international series in operations research & management science, vol 176. Springer, New York. https://doi.org/10.1007/978-1-4614-3600-3

Cisneros-Velarde P, Petersen A, Oh SY (2020) Distributionally robust formulation and model selection for the graphical lasso. In: Chiappa S, Calandra R (eds) Proceedings of the twenty third international conference on artificial intelligence and statistics, proceedings of machine learning research, vol 108, pp 756–765. PMLR. http://proceedings.mlr.press/v108/cisneros20a.html

CPLEX Optimizer-IBM (2019). https://www.ibm.com/analytics/cplex-optimizer

Dangeti P (2017) Statistics for machine learning: techniques for exploring supervised, unsupervised, and reinforcement learning models with Python and R. Packt Publishing

Delage E, Ye Y (2010) Distributionally robust optimization under moment uncertainty with application to data-driven problems. Oper Res 58(3):595–612. https://doi.org/10.1287/opre.1090.0741

Duchi JC, Glynn PW, Namkoong H (2021) Statistics of robust optimization: a generalized empirical likelihood approach. Math Oper Res. https://doi.org/10.1287/moor.2020.1085

Erdoğan E, Iyengar G (2006) Ambiguous chance constrained problems and robust optimization. Math Program 107(1–2):37–61. https://doi.org/10.1007/s10107-005-0678-0

Fournier N, Guillin A (2015) On the rate of convergence in Wasserstein distance of the empirical measure. Prob Theory Relat Fields 162(3–4):707–738. https://doi.org/10.1007/s00440-014-0583-7

Gabriel SA, Conejo AJ, Fuller JD, Hobbs BF, Ruiz C (2012) Complementarity modeling in energy markets, vol 180. Springer, Berlin

Gallego G, Moon I (1993) The distribution free newsboy problem: review and extensions. J Oper Res Soc 44(8):825. https://doi.org/10.2307/2583894

Gao R, Kleywegt AJ (2016) Distributionally robust stochastic optimization with Wasserstein distance. arXiv:1604.02199

Gao R, Kleywegt AJ (2017) Distributionally robust stochastic optimization with dependence structure. arXiv:1701.04200

Graf S, Luschgy H (2000) Foundations of quantization for probability distributions, lecture notes in mathematics, vol 1730. Springer, Berlin

Guo S, Xu H (2019) Distributionally robust shortfall risk optimization model and its approximation. Math Program 174(1):473–498. https://doi.org/10.1007/s10107-018-1307-z

Hanasusanto GA, Kuhn D, Wallace SW, Zymler S (2015) Distributionally robust multi-item newsvendor problems with multimodal demand distributions. Math Program 152(1–2):1–32. https://doi.org/10.1007/s10107-014-0776-y

Ji R, Lejeune MA (2020) Data-driven optimization of reward-risk ratio measures. INFORMS J Comput. https://doi.org/10.1287/ijoc.2020.1002

Keith AJ, Ahner DK (2021) A survey of decision making and optimization under uncertainty. Ann Oper Res 300(2):319–353. https://doi.org/10.1007/s10479-019-03431-8

Kuhn D, Esfahani PM, Nguyen VA, Shafieezadeh-Abadeh S (2019) Wasserstein distributionally robust optimization: theory and applications in machine learning. In: Operations research & management science in the age of analytics, pp 130–166. INFORMS. https://doi.org/10.1287/educ.2019.0198

Lam H, Mottet C (2017) Tail analysis without parametric models: a worst-case perspective. Oper Res 65(6):1696–1711. https://doi.org/10.1287/opre.2017.1643

Li B, Jiang R, Mathieu JL (2019) Ambiguous risk constraints with moment and unimodality information. Math Program 173(1–2):151–192. https://doi.org/10.1007/s10107-017-1212-x

Liu Y (2021) Discrete approximation scheme in distributionally robust optimization. Numer Math Theory Methods Appl 14(2):285–320. https://doi.org/10.4208/nmtma.OA-2020-0125

Liu J, Chen Z, Lisser A, Xu Z (2019) Closed-form optimal portfolios of distributionally robust mean-CVaR problems with unknown mean and variance. Appl Math Optim 79(3):671–693. https://doi.org/10.1007/s00245-017-9452-y

Liu Q, Wu J, Xiao X, Zhang L (2018) A note on distributionally robust optimization under moment uncertainty. J Numer Math 26(3):141–150. https://doi.org/10.1515/jnma-2017-0020

Lucchetti R (2006) Convexity and well-posed problems. CMS books in mathematics. Springer, New York. https://doi.org/10.1007/0-387-31082-7

McLachlan G, Peel D (2000) Finite mixture models. Wiley series in probability and statistics. Wiley, Hoboken. https://doi.org/10.1002/0471721182

Mehrotra S, Papp D (2014) A cutting surface algorithm for semi-infinite convex programming with an application to moment robust optimization. SIAM J Optim 24(4):1670–1697. https://doi.org/10.1137/130925013

Mohajerin Esfahani P, Kuhn D (2018) Data-driven distributionally robust optimization using the Wasserstein metric: performance guarantees and tractable reformulations. Math Program 171(1–2):115–166. https://doi.org/10.1007/s10107-017-1172-1

Nakao H, Shen S, Chen Z (2017) Network design in scarce data environment using moment-based distributionally robust optimization. Comput Oper Res 88:44–57. https://doi.org/10.1016/j.cor.2017.07.002

Namkoong H, Duchi JC (2016) Stochastic gradient methods for distributionally robust optimization with f-divergences. In: Lee DD, Sugiyama M, Luxburg UV, Guyon I, Garnett R (eds) Advances in neural information processing systems, vol 29. Curran Associates Inc, Red Hook, pp 2208–2216

Németh A, Németh S (2016) Isotonic regression and isotonic projection. Linear Algebra Appl 494:80–89. https://doi.org/10.1016/j.laa.2016.01.006

Pando V, San-José LA, García-Laguna J, Sicilia J (2013) A newsboy problem with an emergency order under a general backorder rate function. Omega 41(6):1020–1028. https://doi.org/10.1016/j.omega.2013.01.003

Pando V, San-José LA, García-Laguna J, Sicilia J (2014) Some general properties for the newsboy problem with an extraordinary order. TOP 22(2):674–693. https://doi.org/10.1007/s11750-013-0287-7

Pedregosa F, Varoquaux G, Gramfort A, Michel V, Thirion B, Grisel O, Blondel M, Prettenhofer P, Weiss R, Dubourg V, Vanderplas J, Passos A, Cournapeau D, Brucher M, Perrot M, Duchesnay E (2011) Scikit-learn: machine learning in python. J Mach Learn Res 12:2825–2830

Pyomo (2019). http://www.pyomo.org/

Rahimian H, Bayraksan G, Homem-de Mello T (2019) Identifying effective scenarios in distributionally robust stochastic programs with total variation distance. Math Program 173(1–2):393–430. https://doi.org/10.1007/s10107-017-1224-6

Rahimian H, Mehrotra S (2019) Distributionally robust optimization: a review. arXiv:1908.05659

Roos E, den Hertog D, Ben-Tal A, de Ruiter F, Zhen J (2018) Approximation of hard uncertain convex inequalities. http://www.optimization-online.org/DB_HTML/2018/06/6679.html

Ruiz C, Conejo A, García-Bertrand R (2008) Some analytical results pertaining to Cournot models for short-term electricity markets. Electr Power Syst Res 78(10):1672–1678. https://doi.org/10.1016/j.epsr.2008.02.008

Santambrogio F (2015) Optimal transport for applied mathematicians. Birkäuser, Basel

Scarf H (1958) A min-max solution of an inventory problem. Studies in the mathematical theory of inventory and production. http://ci.nii.ac.jp/naid/10021308995/en/

Shafieezadeh-Abadeh S, Kuhn D, Esfahani PM (2019) Regularization via mass transportation. J Mach Learn Res 20(103):1–68

Shapiro A (2001) On duality theory of conic linear problems. In: Goberna MÁ, López MA (eds) Semi-infinite programming, nonconvex optimization and its applications. Springer, New York, pp 135–165. https://doi.org/10.1007/978-1-4757-3403-4_7

Silvapulle MJ, Sen PK (2011) Constrained statistical inference: inequality, order, and shape restrictions. Wiley series in probability and statistics. Wiley, Hoboken. https://doi.org/10.1002/9781118165614

Sion M (1958) On general minimax theorems. Pac J Math 8(1):171–176. https://doi.org/10.2140/pjm.1958.8.171

Villani C (2008) Optimal transport: old and new. Grundlehren der mathematischen Wissenschaften. Springer, Berlin

Wang C, Gao R, Qiu F, Wang J, Xin L (2018) Risk-based distributionally robust optimal power flow with dynamic line rating. IEEE Trans Power Syst 33(6):6074–6086. https://doi.org/10.1109/TPWRS.2018.2844356

Wang Z, Glynn PW, Ye Y (2016) Likelihood robust optimization for data-driven problems. Comput Manag Sci 13(2):241–261. https://doi.org/10.1007/s10287-015-0240-3

Xie W (2021) On distributionally robust chance constrained programs with Wasserstein distance. Math Program 186(1–2):115–155. https://doi.org/10.1007/s10107-019-01445-5

Xin L, Goldberg DA (2021) Time (in)consistency of multistage distributionally robust inventory models with moment constraints. Eur J Oper Res 289(3):1127–1141. https://doi.org/10.1016/j.ejor.2020.07.041

Yao L, Wang X, Duan C, Wu X, Zhang W (2018) Risk-based distributionally robust energy and reserve dispatch with Wasserstein-moment metric. In: 2018 IEEE power & energy society general meeting (PESGM), pp 1–5. IEEE. https://doi.org/10.1109/PESGM.2018.8586296

Yue MC, Kuhn D, Wiesemann W (2020) On linear optimization over Wasserstein balls. arXiv:2004.07162

Zhao C, Guan Y (2015) Data-driven risk-averse two-stage stochastic program with ζ-structure probability metrics. http://www.optimization-online.org/DB_FILE/2015/07/5014.pdf

Zymler S, Kuhn D, Rustem B (2013) Distributionally robust joint chance constraints with second-order moment information. Math Program 137(1–2):167–198. https://doi.org/10.1007/s10107-011-0494-7