Mathematical Programming (2022) 195:1069–1105
https://doi.org/10.1007/s10107-021-01724-0

FULL LENGTH PAPER

Series A

Distributionally robust stochastic programs with side information based on trimmings

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

Received: 22 September 2020 / Accepted: 1 October 2021 / Published online: 22 November 2021

Abstract
We consider stochastic programs conditional on some covariate information, where the only knowledge of the possible relationship between the uncertain parameters and the covariates is reduced to a finite data sample of their joint distribution. By exploiting the close link between the notion of *trimmings* of a probability measure and the *partial* mass transportation problem, we construct a data-driven Distributionally Robust Optimization (DRO) framework to hedge the decision against the intrinsic error in the process of inferring conditional information from limited joint data. We show that our approach is computationally as tractable as the standard (without side information) Wasserstein-metric-based DRO and enjoys performance guarantees. Furthermore, our DRO framework can be conveniently used to address data-driven decision-making problems under contaminated samples. Finally, the theoretical results are illustrated using a single-item newsvendor problem and a portfolio allocation problem with side information.

Keywords Distributionally robust optimization · Trimmings · Side information · Partial mass transportation problem · Newsvendor problem · Portfolio optimization

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

✉ Juan M. Morales
juan.morales@uma.es

Adrián Esteban-Pérez
adrianesteban@uma.es

[1] Universidad de Málaga, Andalucía Tech, Departamento de Matemática Aplicada, 29071 Málaga, Spain

1 Introduction

Today's decision makers not only collect observations of the uncertainties directly affecting their decision-making processes, but also gather data about measurable exogenous variables that may have some predictive power on those uncertainties [5]. In Statistics, Operations Research and Machine Learning, these variables are often referred to as *covariates*, *explanatory variables*, *side information* or *features* [39].

In the framework of Optimization Under Uncertainty, the side information acts by changing the probability measure of the uncertainties. In fact, if the joint distribution of the features and the uncertainties were known, this measure change would correspond to conditioning that distribution on the side information given. Unfortunately, in practice, the decision maker only has an incomplete picture of such a joint distribution in the form of a finite data sample. The development of optimization methods capable of exploiting the side information to make improved decisions, in a context of limited knowledge of its explanatory power on the uncertainties, defines the ultimate purpose of the so-called *Prescriptive Stochastic Programming* or *Conditional Stochastic Optimization* paradigm. This paradigm has recently become very popular in the technical literature, see, for instance, [5,7,39] and references therein. More specifically, a data-driven approach to address the newsvendor problem, whereby the decision is explicitly modeled as a parametric function of the features, is proposed in [5]. This approach thus seeks to optimize said function. In contrast, the authors in [7] formulate and formalize the problem of minimizing the conditional expectation cost given the side information, and develop various schemes based on machine learning methods (typically used for regression and prediction) to get data-driven solutions. Their approach is *non-parametric* in the sense that the optimal decision is not constrained to be a member of a certain family of the features' functions. The inspiring work in [7] has been subject to further study and improvement in two principal directions, namely, the design of efficient algorithms to trim down the computational burden of the optimization [16] and the development of strategies to reduce the variance and bias of the decision obtained and its associated cost (the pairing of both interpreted as a statistical estimator). In the latter case, we can cite the work in [12], where they leverage ideas from bootstrapping and machine learning to confer robustness on the decision and acquire asymptotic performance guarantees. Similarly, the authors in [8] and [39] propose regularization procedures to reduce the variance of the data-driven solution to the conditional expectation cost minimization problem which is formalized and studied in [7]. A scheme to robustify the data-driven methods introduced in this work is also proposed in [9] for dynamic decision-making.

A different, but related thrust of research focuses on developing methods to construct predictions specifically tailored to the optimization problem that is to be solved and where those predictions are then used as input information. Essentially, the predictions are intended to yield decisions with a low disappointment or regret. This framework is known in the literature as (smart) *Predict-then-Optimize*, see, e.g., [4,17,18,36], and references therein.

Our research, in contrast, builds upon Distributionally Robust Optimization (DRO), which is a powerful modeling paradigm to protect the task of decision-making against the ambiguity of the underlying probability distribution of the uncertainty [40].

Nevertheless, the technical literature on the use of DRO to address Prescriptive or Conditional Stochastic Programming problems is still relatively scarce. We highlight papers [9,15,28,31,33,37,38][1], with [38] being a generalization of [37]. In [15], they resort to a scenario-dependent ambiguity set to exploit feature information in a DRO framework. However, their objective is to minimize a joint expectation and consequently, their approach cannot directly handle the Conditional Stochastic Optimization setting we consider here. In [28], the authors deal with a stochastic control problem with time-dependent data. They extend the idea of [29] to a fully dynamic setting and robustify the control policy against the worst-case weight vector that is within a certain χ^2-distance from the one originally given by the Nadaraya-Watson estimator. In the case of [9], the authors propose using the conditional empirical distribution given by a local predictive method as the center of the Wasserstein ball that characterizes the DRO approach in [35]. This proposal, nonetheless, fails to explicitly account for the inference error associated with the local estimation. In [31,33], the authors develop a two-step procedure whereby a regression model between the uncertainty and the features is first estimated and then a distributionally robust decision-making problem is formulated, considering a Wasserstein ball around the empirical distribution of the residuals. Finally, the authors in [38] also consider a Wasserstein-ball ambiguity set as in [9,31,33], but centered at the empirical distribution of the joint data sample of the uncertainty and the features. In addition, they further constrain the ambiguity set by imposing that the worst-case distribution assigns some probability mass to the support of the uncertainty conditional on the values taken on by the features.

Against this background, our main contributions are:

1. *Modeling power:* We develop a general framework to handle prescriptive stochastic programs within the DRO paradigm. Our DRO framework is based on a new class of ambiguity sets that exploit the close and convenient connection between trimmings and the partial mass problem to immunize the decision against the error incurred in the process of inferring *conditional* information from *joint* (limited) data. We also show that our approach serves as a natural framework for the application of DRO in data-driven decision-making under contaminated samples.
2. *Computational tractability:* Our framework is as complex as the Wasserstein-metric-based DRO approach proposed in [35] without side information. Therefore, we extend the mass-transportation approach to the realm of Conditional Stochastic Optimization while preserving its appealing tractability properties.
3. *Theoretical results and performance guarantees:* Leveraging theory from probability trimmings and optimal transport, we show that our DRO model enjoys a finite sample guarantee and is asymptotically consistent.
4. *Numerical results:* We evaluate our DRO approach on the single-item newsvendor problem and the portfolio allocation problem, and compare it with the KNN method described in [7], the robustified KNN proposed in [9], and a KNN followed by the standard Wassertein-distance-based DRO model introduced in [35], as suggested in [9] too. Unlike all these approaches, ours explicitly accounts for the cost impact of the potential error made when inferring conditional information from a joint sample of the uncertainty and the covariates. To this end, we minimize the worst-

[1] The preprints [31,33,37,38] became available online while this paper was under review in this journal.

case cost over a Wasserstein ball of probability measures with *an ambiguous center*.

The rest of the paper is organized as follows. In Sect. 2, we formulate our DRO framework to address decision-making problems under uncertainty in the presence of side information and show that it is as tractable as the standard Wasserstein-metric-based DRO approach developed in [35]. In Sect. 3.1, we deal with the case in which the side information corresponds to an event of known and positive probability and discuss its application to data-driven decision-making under contaminated samples. The situation in which the probability of such an event is positive, but unknown, is treated in Sect. 3.2. Section 3.3 elaborates on the case in which the side information reduces to a specific realization of the feature vector, more precisely, the instance where the side information represents an event of zero probability. Section 4 provides results from numerical experiments and, finally, Sect. 5 concludes the paper.

Notation. We use $\overline{\mathbb{R}}$ to represent the extended real line, and adopt the conventions of its associated arithmetic. Moreover, $\mathbb{R}_+$ stands for the set of non-negative real numbers. We employ lower-case bold face letters to represent vectors. The inner product of two vectors $\mathbf{u}$, $\mathbf{v}$ is denoted as $\langle \mathbf{u}, \mathbf{v} \rangle = \mathbf{u}^T \mathbf{v}$ and by $\|\mathbf{u}\|$ we denote the norm of the vector $\mathbf{u}$. For a set A, the indicator function $\mathbb{I}_A(\mathbf{a})$ is defined through $\mathbb{I}_A(\mathbf{a}) = 1$ if $\mathbf{a} \in A$; $= 0$ otherwise. The Lebesgue measure in $\mathbb{R}^d$ is denoted as λ^d. We use the symbol $\delta_{\boldsymbol{\xi}}$ to represent the Dirac distribution supported on $\boldsymbol{\xi}$. Additionally, we reserve the symbol "$\widehat{\ }$" for objects which are dependent on the sample data. The K-fold product of a distribution $\mathbb{Q}$ will be denoted as $\mathbb{Q}^K$. Finally, the symbols $\mathbb{E}$ and $\mathbb{P}$ denote, respectively, "expectation" and "probability" (the context will give us the measure under which that expectation or probability is taken).

2 Data-driven distributionally robust optimization with side information

In this paper, we propose a general framework for data-driven distributionally robust optimization with side information that relies on two related tools, namely, the *optimal mass transport theory* and the concept of *trimming of a probability measure*. Next, we introduce some preliminaries that help motivate our proposal. All the proofs that are missing in the main text are compiled in the "Appendix".

2.1 Preliminaries and motivation

Let $\mathbf{x} \in X \subseteq \mathbb{R}^{d_{\mathbf{x}}}$ be the decision variable vector and $\mathbf{y}$, with support set $\Xi_{\mathbf{y}} \subseteq \mathbb{R}^{d_{\mathbf{y}}}$, the random vector that models the uncertainty affecting the value of the decision. Let $\mathbf{z}$, with support set $\Xi_{\mathbf{z}} \subseteq \mathbb{R}^{d_{\mathbf{z}}}$, be the (random) feature vector and denote the objective function to be minimized as $f(\mathbf{x}, \boldsymbol{\xi})$, where $\boldsymbol{\xi} := (\mathbf{z}, \mathbf{y})$.

Given a new piece of information in the form of the event $\boldsymbol{\xi} \in \widetilde{\Xi}$, the decision maker seeks to compute the optimal decision that minimizes the (true) conditional

expected cost:

$$J^* := \inf_{\mathbf{x}\in X} \mathbb{E}_{\mathbb{Q}}\left[f(\mathbf{x},\boldsymbol{\xi}) \;\mid\; \boldsymbol{\xi}\in\widetilde{\Xi}\right] = \inf_{\mathbf{x}\in X} \mathbb{E}_{\mathbb{Q}_{\widetilde{\Xi}}}\left[f(\mathbf{x},\boldsymbol{\xi})\right] \tag{1}$$

where $\mathbb{Q}$ is the true joint distribution of $\boldsymbol{\xi} := (\mathbf{z},\mathbf{y})$ with support set $\Xi \subseteq \mathbb{R}^{d_{\mathbf{z}}+d_{\mathbf{y}}}$ and $\mathbb{Q}_{\widetilde{\Xi}}$ is the associated true distribution of $\boldsymbol{\xi}$ *conditional on* $\boldsymbol{\xi}\in\widetilde{\Xi}$. Hence, we implicitly assume that $\mathbb{Q}_{\widetilde{\Xi}}$ is a regular conditional distribution and that the conditional expectation (1) is well defined.

An example of $\widetilde{\Xi}$ would be $\widetilde{\Xi} := \{\boldsymbol{\xi}=(\mathbf{z},\mathbf{y})\in\Xi \;:\; \mathbf{z}\in\mathscr{Z}\}$, with $\mathscr{Z}\subseteq\Xi_{\mathbf{z}}$ being an uncertainty set built from the information on the features. We note that this definition includes the case in which $\mathscr{Z}$ reduces to a singleton $\mathbf{z}^*$ representing a particular realization of the features.

Unfortunately, when it comes to solving problem (1), neither the true distribution $\mathbb{Q}$ nor—even less so—the conditional one $\mathbb{Q}_{\widetilde{\Xi}}$ are generally known to the decision maker. Actually, the decision maker typically counts only on a data sample consisting of N observations $\widehat{\boldsymbol{\xi}}_i := (\widehat{\mathbf{z}}_i,\widehat{\mathbf{y}}_i)$ for $i=1,\ldots,N$, which we assume are i.i.d. Therefore, the solution to problem (1) *per se* is, in practice, out of reach and the best the decision maker can do is to approximate the solution to (1) with some (probabilistic) performance guarantees. Within this context, *Distributionally Robust Optimization* (DRO) emerges as a powerful modeling framework to achieve that goal. In brief, the DRO approach aims to find a decision $\mathbf{x}\in X$ that is *robust* against all *conditional* probability distributions that are somehow *plausible* given the information at the decision maker's disposal. This is mathematically stated as follows:

$$\inf_{\mathbf{x}\in X} \sup_{Q_{\widetilde{\Xi}}\in\widehat{\mathscr{U}}_N} \mathbb{E}_{Q_{\widetilde{\Xi}}}\left[f(\mathbf{x},\boldsymbol{\xi})\right] \tag{2}$$

where $\widehat{\mathscr{U}}_N$ is a so-called *ambiguity set* that contains all those plausible conditional distributions. This ambiguity set must be built from the available information on $\boldsymbol{\xi}$, which, in our case, comprises the N observations $\{\widehat{\boldsymbol{\xi}}_i\}_{i=1}^N$. The subscript N in $\widehat{\mathscr{U}}_N$ is intended to underline this issue. Furthermore, the condition $Q_{\widetilde{\Xi}}(\widetilde{\Xi})=1$ for all $Q_{\widetilde{\Xi}}\in\widehat{\mathscr{U}}_N$ is implicit in the construction of that set. In our setup, however, problem (2) poses a major challenge, which has to do with the fact that the observations $\{\widehat{\boldsymbol{\xi}}_i\}_{i=1}^N$ pertain to the true *joint* distribution $\mathbb{Q}$, and *not* to the conditional one $\mathbb{Q}_{\widetilde{\Xi}}$. Consequently, we need to build an ambiguity set $\widehat{\mathscr{U}}_N$ for the plausible *conditional* distributions from the limited joint information on $\mathbb{Q}$ provided by the data $\{\widehat{\boldsymbol{\xi}}_i\}_{i=1}^N$.

At this point, we should note that there are several approaches in the technical literature to handle the conditional stochastic optimization problem (1) for the particular case in which $\widetilde{\Xi}$ is defined as $\widetilde{\Xi} := \{\boldsymbol{\xi}=(\mathbf{z},\mathbf{y})\in\Xi \;:\; \mathbf{z}=\mathbf{z}^*\}$. For example, the authors of [7] approximate (1) by the following conditional estimate

$$\inf_{\mathbf{x}\in X} \sum_{i=1}^{N} w_N^i(\mathbf{z}^*)\, f(\mathbf{x},(\mathbf{z}^*,\widehat{\mathbf{y}}_i)) \tag{3}$$

where $w_N^i(\mathbf{z}^*)$ is a weight function that can be given by various non-parametric machine learning methods such as K-nearest neighbors, kernel regression, CART, and random forests. Formulation (3) can be naturally interpreted as a (conditional) Sample-Average-Approximation (SAA) of problem (1).

The authors in [8] extend the work in [7] to accommodate the setting in which the outcome of the uncertainty $\mathbf{y}$ may be contingent on the taken decision $\mathbf{x}$. For this purpose, they work with an enriched data set comprising observations of the uncertainty $\mathbf{y}$, the decision $\mathbf{x}$ and the covariates $\mathbf{z}$, and allow the weights in (3) to depend on $\mathbf{x}$ too. Besides, they add terms to the objective function of (3) to penalize estimates of its variance and bias. The case in which the weight function (3) is given by the Nadaraya-Watson (NW) kernel regression estimator is considered in [29,39]. In [39], in addition, they leverage techniques from moderate deviations theory to design a regularization scheme that reduces the optimistic bias of the NW approximation and to provide insight into its out-of-sample performance. The work in [12] focuses on conditional estimators (3) where the weights are provided by the NW or KNN method. They use DRO, based on the relative entropy distance for discrete distributions to get decisions from (3) that perform well on a large portion of resamples *bootstraped* from the empirical distribution of the available data set.

Finally, the authors in [9] provide a robustified version of the conditional estimator (3), which takes the following form

$$\inf_{\mathbf{x}\in X} \sum_{i=1}^{N} w_N^i(\mathbf{z}^*) \sup_{\mathbf{y}\in\mathscr{U}_N^i} \left[f(\mathbf{x}, (\mathbf{z}^*, \mathbf{y})) \right] \tag{4}$$

where $\mathscr{U}_N^i := \{\mathbf{y} \in \Xi_{\mathbf{y}} : \|\mathbf{y} - \widehat{\mathbf{y}}_i\|_p \leq \varepsilon_N\}$. This problem can be seen as a robust SAA method capable of exploiting side information and has also been used in [10,11].

In our case, however, we follow a different path to address the conditional stochastic optimization problem (1) by way of (2). More precisely, we leverage the notion of *trimmings of a distribution* and the related theory of *partial mass transportation*.

2.2 The partial mass transportation problem and trimmings

This section introduces some concepts about trimmings and the partial mass transportation problem that help us construct the ambiguity set $\widehat{\mathscr{U}}_N$ in (2) from the sample data $\{\widehat{\boldsymbol{\xi}}_i\}_{i=1}^N$. For simplicity, we restrict ourselves to probability measures defined in $\mathbb{R}^d$.

If $\mathbb{Q}(\widetilde{\Xi}) = \alpha > 0$ (our analysis, though, will also cover the case $\alpha = 0$ later in Sect. 3.3), problem (1) can be recast as

$$J^* := \inf_{\mathbf{x}\in X} \frac{1}{\alpha} \mathbb{E}_{\mathbb{Q}} \left[f(\mathbf{x}, \boldsymbol{\xi}) \mathbb{I}_{\widetilde{\Xi}}(\boldsymbol{\xi}) \right] \tag{5}$$

which only requires that $\mathbb{E}_{\mathbb{Q}} \left[|f(\mathbf{x}, \boldsymbol{\xi}) \mathbb{I}_{\widetilde{\Xi}}(\boldsymbol{\xi})| \right] < \infty$ for all $\mathbf{x} \in X$ (see [27, Eq. 6.2]).

Now we introduce the notion of a *trimming* of a distribution, which is at the core of our proposed DRO framework.

Definition 1 ($(1-\alpha)$-trimmings, Definition 1.1 from [6]) Given $0 \leq \alpha \leq 1$ and probability measures $P, Q \in \mathbb{R}^d$, we say that Q is an $(1-\alpha)$-*trimming* of P if Q is absolutely continuous with respect to P, and the Radon-Nikodym derivative satisfies $\frac{dQ}{dP} \leq \frac{1}{\alpha}$. The *set of all* $(1-\alpha)$-*trimmings* (or trimming set of level $1-\alpha$) of P will be denoted by $\mathscr{R}_{1-\alpha}(P)$.

As extreme cases, we have that for $\alpha = 1$, $\mathscr{R}_0(P)$ is just P, while, for $\alpha = 0$, $\mathscr{R}_1(P)$ is the set of all probability measures absolutely continuous with respect to P. Given a probability P on $\mathbb{R}^d$, if $\alpha_1 \leq \alpha_2$, then $\mathscr{R}_{1-\alpha_2}(P) \subset \mathscr{R}_{1-\alpha_1}(P)$. Especially useful is the fact that a trimming set is a convex set, which is, besides, compact under the topology of weak convergence. We refer the reader to [3, Proposition 2.7] for other interesting properties about the set $\mathscr{R}_{1-\alpha}(P)$.

Consider now the following minimization problem:

$$\inf_{Q \in \mathscr{R}_{1-\alpha}(P)} D(Q, R) \tag{6}$$

where D is a probability metric.

Problem (6) is known as the $(D, 1-\alpha)-$*partial (or incomplete) mass problem* [6]. While there is a variety of probability metrics we could choose from to play the role of D in (6), here we work with the space $\mathscr{P}_p(\mathbb{R}^d)$ of probability distributions supported on $\mathbb{R}^d$ with finite p-th moment and restrict ourselves to the $p-$Wasserstein metric, $\mathscr{W}_p$, for its tractability and theoretical advantages. In such a case (i.e., when $D = \mathscr{W}_p$), problem (6) is referred to as a partial mass *transportation* problem and interpolates between the classical optimal mass transportation problem (when $\alpha = 1$) and the random quantization problem (when $\alpha = 0$).

Intuitively, the partial optimal transport problem goes as follows. We have an excess of offer of a certain quantity of mass at origin (supply) and a mass that needs to be satisfied at destination (demand), so that it is not necessary to serve all the mass (demand$= \alpha \times$supply). In other words, some $(1-\alpha)$-fraction of the mass at origin can be left non-served. The goal is to perform this task at the cheapest transportation cost. If we represent the demand at destination by a target probability distribution R, we can model the supply at origin as $\frac{P}{\alpha}$, where P is another probability distribution and the mass required at destination is α times the mass at origin. This way, a *partial optimal transportation plan* is a probability measure Π on $\mathbb{R}^d \times \mathbb{R}^d$ with first marginal in $\mathscr{R}_{1-\alpha}(P)$ and with second marginal equal to R, which solves the following cost minimization problem:

$$\mathscr{W}_p(\mathscr{R}_{1-\alpha}(P), R) := \min_{Q \in \mathscr{R}_{1-\alpha}(P)} \mathscr{W}_p(Q, R)$$

The following lemma allows us to characterize the connection between the joint distribution $\mathbb{Q}$ and the conditional distribution $\mathbb{Q}_{\widetilde{\Xi}}$ in problem (1) above in terms of the partial mass problem.

Lemma 1 *Let Q be a probability on $\mathbb{R}^d$ such that $Q(\widetilde{\Xi}) = \alpha > 0$ and let $Q_{\widetilde{\Xi}}$ be the Q-conditional probability distribution given the event $\boldsymbol{\xi} \in \widetilde{\Xi}$. Also, assume*

that for a given probability metric D, $\mathscr{R}_{1-\alpha}(Q)$ is closed for D over an appropiate set of probability distributions. Then, $Q_{\widetilde{\Xi}}$ is the unique distribution that satisfies $Q_{\widetilde{\Xi}}(\widetilde{\Xi}) = 1$ and $D\left(\mathscr{R}_{1-\alpha}(Q), Q_{\widetilde{\Xi}}\right) = 0$.

By way of Lemma (1), we can reformulate Problem (1) as follows:

$$\inf_{\mathbf{x}\in X} \sup_{Q_{\widetilde{\Xi}}} \mathbb{E}_{Q_{\widetilde{\Xi}}}\left[f(\mathbf{x}, \boldsymbol{\xi})\right] \tag{7a}$$

$$\text{s.t. } \mathscr{W}_p^p(\mathscr{R}_{1-\alpha}(\mathbb{Q}), Q_{\widetilde{\Xi}}) = 0 \tag{7b}$$

$$Q_{\widetilde{\Xi}}(\widetilde{\Xi}) = 1 \tag{7c}$$

which now presents a form which is much more suited to our purpose, that is, to get to the DRO-type of problem (2) we propose. The change, nonetheless, has been essentially cosmetic, because problem (7) still relies on the true *joint* distribution $\mathbb{Q}$ and therefore, is of no use in practice as it stands right now. To make it practical, we need to rewrite it not in terms of the unknown $\mathbb{Q}$, but in terms of the information available to the decision maker, i.e., the sample data $\{\widehat{\boldsymbol{\xi}}_i\}_{i=1}^N$. For that purpose, it seems sensible and natural to replace $\mathbb{Q}$ in (7b) with its best approximation taken directly from the data, namely, the empirical measure of the sample, $\widehat{\mathbb{Q}}_N$. Logically, to accommodate the approximation, we will need to introduce a *budget* $\widetilde{\rho}$ in equation (7b), that is,

$$\text{(P)} \quad \inf_{\mathbf{x}\in X} \sup_{Q_{\widetilde{\Xi}}} \mathbb{E}_{Q_{\widetilde{\Xi}}}\left[f(\mathbf{x}, \boldsymbol{\xi})\right] \tag{8a}$$

$$\text{s.t. } \mathscr{W}_p^p(\mathscr{R}_{1-\alpha}(\widehat{\mathbb{Q}}_N), Q_{\widetilde{\Xi}}) \leq \widetilde{\rho} \tag{8b}$$

$$Q_{\widetilde{\Xi}}(\widetilde{\Xi}) = 1 \tag{8c}$$

Hereinafter we will use $\widehat{\mathscr{U}}_N(\alpha, \widetilde{\rho})$ to denote the ambiguity set defined by constraints (8b)–(8c). Under certain conditions, this uncertainty set enjoys nice topological properties, as we state in [19, Proposition EC.2].

Now we define what we call the *minimum transportation budget*, which plays an important role in the selection of budget $\widetilde{\rho}$ in problem (P).

Definition 2 (Minimum transportation budget) Given $\alpha > 0$ in problem (P), the *minimum transportation budget*, which we denote as $\underline{\epsilon}_{N\alpha}$, is the p-Wasserstein distance between the set $\mathscr{P}_p(\widetilde{\Xi})$ and the $(1-\alpha)$-trimming of the empirical distribution $\widehat{\mathbb{Q}}_N$ that is the *closest* to that set, i.e., $\inf\{\mathscr{W}_p(P, Q) : P \in \mathscr{R}_{1-\alpha}(\widehat{\mathbb{Q}}_N),\ Q \in \mathscr{P}_p(\widetilde{\Xi})\}$, which is given by

$$\underline{\epsilon}_{N\alpha} = \left(\frac{1}{N\alpha}\sum_{k=1}^{\lfloor N\alpha\rfloor} \text{dist}(\boldsymbol{\xi}_{k:N}, \widetilde{\Xi})^p + \left(1 - \frac{\lfloor N\alpha\rfloor}{N\alpha}\right)\text{dist}(\boldsymbol{\xi}_{\lceil N\alpha\rceil:N}, \widetilde{\Xi})^p\right)^{\frac{1}{p}} \tag{9}$$

where $\boldsymbol{\xi}_{k:N}$ is the *k-th* nearest data point from the sample to set $\widetilde{\Xi}$ and $\text{dist}(\boldsymbol{\xi}_j, \widetilde{\Xi}) := \inf_{\boldsymbol{\xi}\in\widetilde{\Xi}} \text{dist}(\boldsymbol{\xi}_j, \boldsymbol{\xi}) = \inf_{\boldsymbol{\xi}\in\widetilde{\Xi}} ||\boldsymbol{\xi}_j - \boldsymbol{\xi}||$. If $\alpha = 0$, then $\underline{\epsilon}_{N0} = \text{dist}(\boldsymbol{\xi}_{1:N}, \widetilde{\Xi})$.

Importantly, the minimum transportation budget to the power of p, i.e., $\underline{\epsilon}^p_{N\alpha}$, is the minimum value of $\widetilde{\rho}$ in (P) for this problem to be feasible. Furthermore, $\underline{\epsilon}_{N\alpha}$ is random, because it depends on the available data sample, but realizes before the decision $\mathbf{x}$ is to be made. It constitutes, therefore, input data to problem (P).

We note that, if the random vector $\mathbf{y}$ takes values in a set that is independent of the feature vector $\mathbf{z}$, i.e., for all $\mathbf{z}^* \in \Xi_{\mathbf{z}}$, $\{\mathbf{y} \in \Xi_{\mathbf{y}} : \boldsymbol{\xi} = (\mathbf{z}^*, \mathbf{y}) \in \Xi\} = \Xi_{\mathbf{y}}$, then $\text{dist}(\boldsymbol{\xi}_j, \widetilde{\Xi}) = \inf_{\boldsymbol{\xi} \in \widetilde{\Xi}} ||\boldsymbol{\xi}_j - \boldsymbol{\xi}|| = \inf_{\boldsymbol{\xi}=(\mathbf{z},\mathbf{y}) \in \widetilde{\Xi}} ||\mathbf{z}_j - \mathbf{z}||$.

Furthermore, in what follows, we assume that $\text{dist}(\boldsymbol{\xi}_j, \widetilde{\Xi})$ (interpreted as a random variable) conditional on $\xi_j \notin \widetilde{\Xi}$ has a continuous distribution function. This ensures that, in the case $\mathbb{Q}(\widetilde{\Xi}) = 0$, which we study in Sect. 3.3, there will be exactly K nearest data points to $\widetilde{\Xi}$ with probability one.

Next we present an interesting result, which deals with the inner supremum of problem (P) and adds more meaning to this problem by linking it to an alternative formulation more in the style of the Wasserstein data-driven DRO approach proposed in [35], where, however, no side information is taken into account. In fact, the distributionally robust approach to conditional stochastic optimization that is proposed in [38] is based on this alternative formulation (see Proposition A.4 in that work)[2].

Proposition 1 *Given $N \geq 1$, $\mathbb{Q}(\widetilde{\Xi}) = \alpha > 0$, and any positive value of $\widetilde{\rho}$, problem* (SP2) *is a relaxation of* (SP1)*, where* (SP1) *and* (SP2) *are given by*

$$\text{(SP1)} \quad \begin{cases} \sup_Q \ \mathbb{E}_Q\left[f(\mathbf{x}, \boldsymbol{\xi}) \mid \boldsymbol{\xi} \in \widetilde{\Xi}\right] \\ \ s.t. \ \ \mathscr{W}_p^p(Q, \widehat{\mathbb{Q}}_N) \leq \widetilde{\rho} \cdot \alpha \\ \qquad Q(\widetilde{\Xi}) = \alpha \end{cases}, \qquad \text{(SP2)} \quad \begin{cases} \sup_{Q_{\widetilde{\Xi}}} \ \mathbb{E}_{Q_{\widetilde{\Xi}}}\left[f(\mathbf{x}, \boldsymbol{\xi})\right] \\ \ s.t. \ \ \mathscr{W}_p^p(\mathscr{R}_{1-\alpha}(\widehat{\mathbb{Q}}_N), Q_{\widetilde{\Xi}}) \leq \widetilde{\rho} \\ \qquad Q_{\widetilde{\Xi}}(\widetilde{\Xi}) = 1 \end{cases}$$

and where by "relaxation" it is meant that any solution Q feasible in (SP1) *can be mapped into a solution $Q_{\widetilde{\Xi}}$ feasible in* (SP2) *with the same objective function value.*

Moreover, if $\widehat{\mathbb{Q}}_N(\widetilde{\Xi}) = 0$ or $\alpha = 1$, then (SP1) *and* (SP2) *are equivalent.*

Among other things, Proposition 1 reveals that parameter $\widetilde{\rho}$ in problem (SP2), and hence in problem (P), can be understood as a cost budget *per unit of transported mass*. Likewise, parameter α can be interpreted as the minimum amount of mass (in per unit) of the empirical distribution $\widehat{\mathbb{Q}}_N$ that must be transported to the support $\widetilde{\Xi}$. This interpretation of parameters $\widetilde{\rho}$ and α will be useful to follow the rationale behind the DRO solution approaches that we develop later on.

On the other hand, despite the connection between problems (SP1) and (SP2) that Proposition 1 unveils, the latter is qualitatively more amenable to further generalization and analysis. Examples of this are given by the relevant cases $\alpha = 0$, for which problem (SP1) is *ill-posed*, while problem (SP2) is not, and α unknown, for which the use of trimming sets in (SP2) allows for a more straightforward treatment. We will deal with both cases in Sects. 3.3 and 3.2 , respectively. Before that, we provide an implementable reformulation of the proposed DRO problem (P).

[2] Proposition 1 in this paper predates the release of preprint [38].

2.3 Towards a tractable reformulation of the partial mass transportation problem

In this section, we put the proposed DRO problem (P) in a form more suited to tackle its computational implementation and solution. For this purpose, we first need to introduce a technical result whereby we characterize the trimming sets of an empirical probability measure.

Lemma 2 *Consider the sample data* $\{\widehat{\boldsymbol{\xi}}_i\}_{i=1}^N$ *and their associated empirical measure* $\widehat{\mathbb{Q}}_N = \frac{1}{N}\sum_{i=1}^N \delta_{\widehat{\boldsymbol{\xi}}_i}$. *If* $\alpha > 0$, *the set of all* $(1-\alpha)$*-trimmings of* $\widehat{\mathbb{Q}}_N$ *is given by all probability distributions in the form* $\sum_{i=1}^N b_i \delta_{\widehat{\boldsymbol{\xi}}_i}$ *such that* $0 \leq b_i \leq \frac{1}{N\alpha}$, $\forall i = 1, \ldots, N$, *and* $\sum_{i=1}^N b_i = 1$. *Furthermore, if* $\alpha = 0$, *the set* $\mathscr{R}_{1-\alpha}(\widehat{\mathbb{Q}}_N)$ *of* $(1-\alpha)$*-trimmings of* $\widehat{\mathbb{Q}}_N$ *becomes* $\mathscr{R}_1(\widehat{\mathbb{Q}}_N) = \{\sum_{i=1}^N b_i \delta_{\widehat{\boldsymbol{\xi}}_i}$ *such that* $b_i \geq 0$, $\forall i = 1, \ldots, N$, *and* $\sum_{i=1}^N b_i = 1\}$.

Proof If $\alpha > 0$, the form of any $(1-\alpha)$-trimming of $\widehat{\mathbb{Q}}_N$ as $\sum_{i=1}^N b_i \delta_{\widehat{\boldsymbol{\xi}}_i}$, along with the condition $b_i \leq \frac{1}{N\alpha}$, follows directly from Definition 1 of a $(1-\alpha)$-trimming. Naturally, $b_i \geq 0$ and $\sum_{i=1}^N b_i = 1$ are then required because any $(1-\alpha)$-trimming is a probability distribution.

On the other hand, if $\alpha = 0$, the resulting trimming set $\mathscr{R}_1(\widehat{\mathbb{Q}}_N)$ is simply the family of all probability distributions supported on the data points $\{\widehat{\boldsymbol{\xi}}_i\}_{i=1}^N$. □

In short, Lemma 2 tells us that trimming a data sample of size N with level $1-\alpha$ involves reweighting the empirical distribution of such data by giving a new weight less than or equal to $\frac{1}{N\alpha}$ to each data point. Therefore, we can recast constraint $\mathscr{W}_p^p(\mathscr{R}_{1-\alpha}(\widehat{\mathbb{Q}}_N), Q_{\widetilde{\Xi}}) \leq \widetilde{\rho}$ in problem (P) as

$$\begin{aligned} \min_{b_i, \forall i \leq N} \ & \mathscr{W}_p\left(\sum_{i=1}^N b_i \delta_{\widehat{\boldsymbol{\xi}}_i}, Q_{\widetilde{\Xi}}\right) \leq \widetilde{\rho}^{1/p} \\ \text{s.t. } & 0 \leq b_i \leq \frac{1}{N\alpha}, \ \ \forall i \leq N \\ & \sum_{i=1}^N b_i = 1 \end{aligned}$$

We are now ready to introduce the main result of this section.

Theorem 1 (Reformulation based on strong duality) *For* $\alpha > 0$ *and any value of* $\widetilde{\rho} \geq \epsilon_{N\alpha}^p$, *subproblem* (SP2) *is equivalent to the following one:*

$$\begin{aligned} (\text{SP2}') \quad \inf_{\lambda \geq 0; \overline{\mu}_i, \forall i \leq N; \theta \in \mathbb{R}} \ & \lambda\widetilde{\rho} + \theta + \frac{1}{N\alpha}\sum_{i=1}^N \overline{\mu}_i \\ s.t. \ & \overline{\mu}_i + \theta \geq \sup_{(\mathbf{z},\mathbf{y}) \in \widetilde{\Xi}} \left(f(\mathbf{x}, (\mathbf{z}, \mathbf{y})) - \lambda \, \|(\mathbf{z}, \mathbf{y}) - (\widehat{\mathbf{z}}_i, \widehat{\mathbf{y}}_i)\|^p \right), \forall i \leq N \\ & \overline{\mu}_i \geq 0, \ \ \forall i \leq N \end{aligned}$$

Surely the most important takeaway message of Theorem 1 is that problem (P) is *as tractable as* the standard Wasserstein-metric-based DRO formulation proposed in [35] and [34]. In these two papers, conditions under which the inner supremum in (SP2′) can be recast in a more tractable form are provided. As an example, in Theorem EC.2 in the extended version of this paper [19], we provide a more refined reformulation of (SP2′), whereby the problems we solve in Sect. 4 can be directly handled.

In the following section, we show that problem (P) works, under certain conditions, as a statistically meaningful surrogate decision-making model for the target conditional stochastic program (1).

3 Finite sample guarantee and asymptotic consistency

Next we argue that the worst-case optimal expected cost provided by problem (P) for a fixed sample size N and a suitable choice of parameters $(\alpha, \widetilde{\rho})$ (dependent on N) leads to an upper confidence bound on the out-of-sample performance attained by the optimizers of (P) (*finite sample guarantee*) and that those optimizers almost surely converge to an optimizer of the true optimal expected cost as N grows to infinity (*asymptotic consistency*).

To be more precise, the *out-of-sample performance* of a given data-driven candidate solution $\widehat{\mathbf{x}}_N$ to problem (1) is defined as $\mathbb{E}_{\mathbb{Q}}[f(\widehat{\mathbf{x}}_N, \boldsymbol{\xi}) \mid \boldsymbol{\xi} \in \widetilde{\Xi}] = \mathbb{E}_{\mathbb{Q}_{\widetilde{\Xi}}}[f(\widehat{\mathbf{x}}_N, \boldsymbol{\xi})]$. We say that a data-driven method built to address problem (1) enjoys a *finite sample guarantee*, if it produces pairs $(\widehat{\mathbf{x}}_N, \widehat{J}_N)$ satisfying a relation in the form

$$\mathbb{Q}^N\Big[\mathbb{E}_{\mathbb{Q}}[f(\widehat{\mathbf{x}}_N, \boldsymbol{\xi}) \mid \boldsymbol{\xi} \in \widetilde{\Xi}] \leq \widehat{J}_N\Big] \geq 1 - \beta \tag{10}$$

and $\widehat{J}_N$ 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 data sample). The probability on the right-hand side of (10), i.e., $1 - \beta$, is known as the *reliability* of $(\widehat{\mathbf{x}}_N, \widehat{J}_N)$ and can be understood as a confidence level.

Our analysis relies on the lemma below, which immediately follows from setting $P_1 := \widehat{\mathbb{Q}}_N$, $Q := \mathbb{Q}_{\widetilde{\Xi}}$, $P_2 := \mathbb{Q}$ in Lemma 3.13 on probability trimmings in [1].

Lemma 3 *Assume that* $\mathbb{Q}_{\widetilde{\Xi}}, \mathbb{Q} \in \mathscr{P}_p(\mathbb{R}^d)$*, and take* $p \geq 1$*, then*

$$\mathscr{W}_p(\mathscr{R}_{1-\alpha}(\widehat{\mathbb{Q}}_N), \mathbb{Q}_{\widetilde{\Xi}}) \leq \mathscr{W}_p(\mathscr{R}_{1-\alpha}(\mathbb{Q}), \mathbb{Q}_{\widetilde{\Xi}}) + \frac{1}{\alpha^{1/p}} \mathscr{W}_p(\widehat{\mathbb{Q}}_N, \mathbb{Q}) \tag{11}$$

We notice that the term $\mathscr{W}_p(\mathscr{R}_{1-\alpha}(\mathbb{Q}), \mathbb{Q}_{\widetilde{\Xi}})$ in (11) is not random and depends exclusively on the true distributions $\mathbb{Q}_{\widetilde{\Xi}}, \mathbb{Q}$, and the trimming level α. It is, therefore, independent of the data sample (unlike the other two terms involved).

Inequality (11) reveals an interesting trade-off. On the one hand, the distance $\mathscr{W}_p(\mathscr{R}_{1-\alpha}(\mathbb{Q}), \mathbb{Q}_{\widetilde{\Xi}})$ diminishes as α decreases to zero, because the trimming set $\mathscr{R}_{1-\alpha}(\mathbb{Q})$ grows in size. On the other, the term $\frac{1}{\alpha^{1/p}} \mathscr{W}_p(\widehat{\mathbb{Q}}_N, \mathbb{Q})$ becomes larger as α approaches zero. As we will see later on, controlling this trade-off is key to endow-

ing problem (P) with performance guarantees. To this end, we will make use of the Proposition 2 below.

Assumption 1 Suppose that the true joint probability distribution $\mathbb{Q}$ is light-tailed, i.e., there exists a constant $a > p \geq 1$ such that $\mathbb{E}_{\mathbb{Q}}\left[\exp(\|\boldsymbol{\xi}\|^a)\right] < \infty$.

Proposition 2 (Concentration tail inequality) *Suppose that Assumption 1 holds. Then, there are constants $c, C > 0$ such that, for all $\epsilon > 0, \alpha > 0$, and $N \geq 1$, it holds*

$$\mathbb{Q}^N\left[\mathscr{W}_p\left(\mathscr{R}_{1-\alpha}(\widehat{\mathbb{Q}}_N), \mathbb{Q}_{\widetilde{\Xi}}\right) \geq \mathscr{W}_p(\mathscr{R}_{1-\alpha}(\mathbb{Q}), \mathbb{Q}_{\widetilde{\Xi}}) + \epsilon\right] \leq \beta_{p,\epsilon,\alpha}(N) \tag{12}$$

where

$$\begin{aligned}
&\beta_{p,\epsilon,\alpha}(N) \\
&\quad = \mathbb{I}_{\{\epsilon \leq 1/\alpha^{1/p}\}} C \begin{cases} \exp(-cN\,\alpha^2\,\epsilon^{2p}) & \text{if } p > d/2, \\ \exp(-cN(\alpha\epsilon^p/\log(2+1/\alpha\epsilon^p))^2) & \text{if } p = d/2, \\ \exp(-cN\,\alpha^{d/p}\,\epsilon^d) & \text{if } p \in [1, d/2),\ d > 2 \end{cases} \\
&\qquad + C\exp(-cN\,\alpha^{a/p}\,\epsilon^a)\mathbb{I}_{\{\epsilon > 1/\alpha^{1/p}\}}
\end{aligned} \tag{13}$$

with $d = d_{\mathbf{z}} + d_{\mathbf{y}}$.

Proof Because of Lemma 3 we have

$$\begin{aligned}
&\mathbb{Q}^N\left(\mathscr{W}_p(\mathscr{R}_{1-\alpha}(\widehat{\mathbb{Q}}_N), \mathbb{Q}_{\widetilde{\Xi}}) - \mathscr{W}_p(\mathscr{R}_{1-\alpha}(\mathbb{Q}), \mathbb{Q}_{\widetilde{\Xi}}) \geq \epsilon\right) \\
&\quad \leq \mathbb{Q}^N\left(\mathscr{W}_p^p\left(\widehat{\mathbb{Q}}_N, \mathbb{Q}\right) \geq \alpha\epsilon^p\right)
\end{aligned}$$

where the right-hand side of this inequality is upper bounded by (13) according to [23, Theorem 2]. □

Assuming $p \neq d/2$, if we equate β to $\beta_{p,\epsilon,\alpha}(N)$ and solving for ϵ we get:

$$\epsilon_{N,p,\alpha}(\beta) := \begin{cases} \left(\frac{\log(C\beta^{-1})}{cN}\right)^{1/2p} \frac{1}{\alpha^{1/p}} & \text{if } N \geq \frac{\log(C\beta^{-1})}{c}, \quad p > d/2, \\ \left(\frac{\log(C\beta^{-1})}{cN}\right)^{1/d} \frac{1}{\alpha^{1/p}} & \text{if } N \geq \frac{\log(C\beta^{-1})}{c}, \quad p \in [1, d/2),\ d > 2 \\ \left(\frac{\log(C\beta^{-1})}{cN}\right)^{1/a} \frac{1}{\alpha^{1/p}} & \text{if } N < \frac{\log(C\beta^{-1})}{c} \end{cases} \tag{14}$$

In what follows, we distinguish three general setups that may appear in the real-life use of Conditional Stochastic Optimization, namely, the case $\mathbb{Q}(\widetilde{\Xi}) = \alpha > 0$ with α known, the case $\mathbb{Q}(\widetilde{\Xi}) = \alpha > 0$ with α unknown, and the case $\mathbb{Q} \ll \lambda^d$ with $\mathbb{Q}(\widetilde{\Xi}) = \alpha = 0$.

3.1 Case $\mathbb{Q}(\widetilde{\Xi}) = \alpha > 0$. Applications in data-driven decision making under contaminated samples

When $\mathbb{Q}(\widetilde{\Xi}) = \alpha > 0$ and *known*, we can solve the following DRO problem:

$$(\mathrm{P}_{(\alpha,\widetilde{\rho}_N)}) \quad \inf_{\mathbf{x}\in X} \sup_{Q_{\widetilde{\Xi}}} \mathbb{E}_{Q_{\widetilde{\Xi}}}\left[f(\mathbf{x},\boldsymbol{\xi})\right] \tag{15a}$$

$$\text{s.t. } \mathscr{W}_p^p(\mathscr{R}_{1-\alpha}(\widehat{\mathbb{Q}}_N), Q_{\widetilde{\Xi}}) \leq \widetilde{\rho}_N \tag{15b}$$

$$Q_{\widetilde{\Xi}}(\widetilde{\Xi}) = 1 \tag{15c}$$

As we show below, problem $\left(\mathrm{P}_{(\alpha,\widetilde{\rho}_N)}\right)$ enjoys a finite sample guarantee and produces solutions that are asymptotically consistent, i.e., that converge to the true solution (under complete information) given by problem (1). This is somewhat hinted at by the connection between problems (SP1) and (SP2) highlighted in Proposition 1.

Theorem 2 (Case $\alpha > 0$: Finite sample guarantee) *Suppose that the assumptions of Proposition 2 hold and take $p \neq d/2$. Given $N \geq 1$ and $\alpha > 0$, choose $\beta \in (0,1)$, and determine $\epsilon_{N,p,\alpha}(\beta)$ through (14). Then, for all $\widetilde{\rho}_N \geq \max(\epsilon^p_{N,p,\alpha}(\beta), \underline{\epsilon}^p_{N\alpha})$, where $\underline{\epsilon}^p_{N\alpha}$ is the minimum transportation budget as in Definition 2, the pair $(\widehat{\mathbf{x}}_N, \widehat{J}_N)$ that is solution to problem $\left(\mathrm{P}_{(\alpha,\widetilde{\rho}_N)}\right)$ enjoys the finite sample guarantee (10).*

Proof For problem $\left(\mathrm{P}_{(\alpha,\widetilde{\rho}_N)}\right)$ to be feasible, we must have $\widetilde{\rho}_N \geq \underline{\epsilon}^p_{N\alpha}$. Furthermore, $\mathscr{W}_p(\mathscr{R}_{1-\alpha}(\mathbb{Q}), \mathbb{Q}_{\widetilde{\Xi}}) = 0$ in (12) because of Lemma 1. Hence, Proposition 2 ensures that $\mathbb{Q}^N\left(\mathbb{Q}_{\widetilde{\Xi}} \in \widehat{\mathscr{U}}_N(\alpha, \widetilde{\rho}_N)\right) \geq 1-\beta$ for any $\widetilde{\rho}_N \geq \epsilon^p_{N,p,\alpha}(\beta)$. It follows then

$$\begin{aligned}\mathbb{E}_{\mathbb{Q}}[f(\widehat{\mathbf{x}}_N,\boldsymbol{\xi}) \mid \boldsymbol{\xi} \in \widetilde{\Xi}] &= \mathbb{E}_{\mathbb{Q}_{\widetilde{\Xi}}}[f(\widehat{\mathbf{x}}_N,\boldsymbol{\xi})] \\ &\leq \widehat{J}_N := \sup_{Q_{\widetilde{\Xi}}}\left\{\mathbb{E}_{Q_{\widetilde{\Xi}}}[f(\widehat{\mathbf{x}}_N,\boldsymbol{\xi})] \;:\; Q_{\widetilde{\Xi}} \in \widehat{\mathscr{U}}_N(\alpha,\widetilde{\rho}_N)\right\}\end{aligned}$$

with probability at least $1-\beta$. □

We point out that, in the case $\alpha > 0$, data points may fall into the set $\widetilde{\Xi}$. Logically, the contribution of these points to the minimum transportation budget $\underline{\epsilon}^p_{N\alpha}$ is null and their order (the way their tie is broken) is irrelevant to our purpose.

Now we show that the solutions of the distributionally robust optimization problem $\left(\mathrm{P}_{(\alpha,\widetilde{\rho}_N)}\right)$ converge to the solution of the target conditional stochastic program (1) as N increases, for a careful choice of the budget $\widetilde{\rho}_N$. This result is underpinned by the fact that, under that selection of $\widetilde{\rho}_N$, any distribution in $\widehat{\mathscr{U}}_N(\alpha,\widetilde{\rho}_N)$ converges to the true conditional distribution $\mathbb{Q}_{\widetilde{\Xi}}$. This is formally stated in the following lemma.

Lemma 4 (Case $\alpha > 0$: Convergence of conditional distributions) *Suppose that the assumptions of Proposition 2 hold. Choose a sequence $\beta_N \in (0,1)$, $N \in \mathbb{N}$, such that $\sum_{N=1}^{\infty} \beta_N < \infty$ and $\lim_{N\to\infty} \epsilon_{N,p,\alpha}(\beta_N) \to 0$. Then,*

$$\mathscr{W}_p(Q^N_{\widetilde{\Xi}}, \mathbb{Q}_{\widetilde{\Xi}}) \to 0 \;\; a.s.$$

for any sequence $Q^N_{\widetilde{\Xi}}$, $N \in \mathbb{N}$, *such that* $Q^N_{\widetilde{\Xi}} \in \widehat{\mathscr{U}}_N(\alpha, \widetilde{\rho}_N)$ *with* $\widetilde{\rho}_N = \max(\epsilon^p_{N,p,\alpha}(\beta_N), \underline{\epsilon}^p_{N\alpha})$.

Proof Take N large enough and let $\widehat{Q}_{N/\widetilde{\Xi}}$ be the conditional probability distribution of $\widehat{\mathbb{Q}}_N$ given $\boldsymbol{\xi} \in \Xi$. We have

$$\mathscr{W}_p(Q^N_{\widetilde{\Xi}}, \mathbb{Q}_{\widetilde{\Xi}}) \leq \mathscr{W}_p(Q^N_{\widetilde{\Xi}}, \widehat{Q}_{N/\widetilde{\Xi}}) + \mathscr{W}_p(\widehat{Q}_{N/\widetilde{\Xi}}, \mathbb{Q}_{\widetilde{\Xi}})$$

We show that the two terms on the right-hand side of the above inequality vanish with probability one as N grows to infinity. We start with $\mathscr{W}_p(\widehat{Q}_{N/\widetilde{\Xi}}, \mathbb{Q}_{\widetilde{\Xi}})$.

Let I denote the subset of observations $\widehat{\boldsymbol{\xi}}_i := (\widehat{\mathbf{z}}_i, \widehat{\mathbf{y}}_i)$ for $i = 1, \ldots, N$, such that $\widehat{\boldsymbol{\xi}}_i \in \widetilde{\Xi}$. It follows from the Strong Law of Large Numbers that $\widehat{\mathbb{Q}}_N(\widetilde{\Xi}) = \frac{|I|}{N} = \alpha_N \to \alpha$ almost surely. Besides, since the sequence $\beta_N, N \in \mathbb{N}$ is summable and $\lim_{N\to\infty} \epsilon_N(\beta_N) \to 0$, the Borel-Cantelli Lemma and Proposition 2 implies

$$\mathscr{W}_p\left(\mathscr{R}_{1-\alpha}(\widehat{\mathbb{Q}}_N), \mathbb{Q}_{\widetilde{\Xi}}\right) \to 0 \; a.s.$$

Then, from Lemma 1, we deduce that $\mathscr{W}_p(\widehat{Q}_{N/\widetilde{\Xi}}, \mathbb{Q}_{\widetilde{\Xi}}) \to 0$ with probability one.

We can deal with the term $\mathscr{W}_p(Q^N_{\widetilde{\Xi}}, \widehat{Q}_{N/\widetilde{\Xi}})$ in a similar fashion, except for the subtle difference that, in this case, we require $\widetilde{\rho}_N = \max(\epsilon^p_{N,p,\alpha}(\beta_N), \underline{\epsilon}^p_{N\alpha})$, so that, for *all* $N \in \mathbb{N}$, problem $P_{(\alpha,\widetilde{\rho}_N)}$ delivers a feasible $Q^N_{\widetilde{\Xi}}$ in the sequence. Hence, in order to prove that $\mathscr{W}_p(Q^N_{\widetilde{\Xi}}, \widehat{Q}_{N/\widetilde{\Xi}}) \to 0$ almost surely, we need to show that $\lim_{N\to\infty} \underline{\epsilon}_{N\alpha} = 0$ with probability one. This is something that can be directly deduced from the definition of $\underline{\epsilon}_{N\alpha}$, namely,

$$\underline{\epsilon}^p_{N\alpha} := \mathscr{W}^p_p(\mathscr{R}_{1-\alpha}(\widehat{\mathbb{Q}}_N), \mathscr{P}_p(\widetilde{\Xi})) = \min_{Q' \in \mathscr{P}_p(\widetilde{\Xi})} \mathscr{W}^p_p(\mathscr{R}_{1-\alpha}(\widehat{\mathbb{Q}}_N), Q') \tag{16}$$

$$\leq \mathscr{W}^p_p\left(\mathscr{R}_{1-\alpha}(\widehat{\mathbb{Q}}_N), \mathbb{Q}_{\widetilde{\Xi}}\right) \to 0 \; a.s. \tag{17}$$

□

Note that, by Eq. (9) in Definition 2, we have that $\underline{\epsilon}_{N\alpha} > 0$ if and only if

$$\lceil N\alpha \rceil > |I| \Leftrightarrow \frac{\lceil N\alpha \rceil}{N} > \frac{|I|}{N} = \alpha_N = \widehat{\mathbb{Q}}_N(\widetilde{\Xi}) \Leftrightarrow \alpha > \alpha_N$$

Once the convergence of $Q^N_{\widetilde{\Xi}}$ to the true conditional distribution $\mathbb{Q}_{\widetilde{\Xi}}$ in the p-Wasserstein metric has been established by the previous lemma, the following asymptotic consistency result, which is analogous to that of [35, Theorem 3.6], can also be derived.

Theorem 3 (Asymptotic consistency) *Consider that the conditions of Theorem 2 hold. Take a sequence* $\widetilde{\rho}_N$ *as in Lemma 4. Then, we have*

(i) If for any fixed value $\mathbf{x} \in X$, $f(\mathbf{x}, \boldsymbol{\xi})$ *is continuous in* $\boldsymbol{\xi}$ *and there is* $L \geq 0$ *such that* $|f(\mathbf{x}, \boldsymbol{\xi})| \leq L(1 + \|\boldsymbol{\xi}\|^p)$ *for all* $\mathbf{x} \in X$ *and* $\boldsymbol{\xi} \in \widetilde{\Xi}$, *then we have that* $\widehat{J}_N \to J^*$ *almost surely when* N *grows to infinity.*

(ii) If the assumptions in (i) are satisfied, $f(\mathbf{x}, \boldsymbol{\xi})$ is lower semicontinuous on X for any fixed $\boldsymbol{\xi} \in \widetilde{\Xi}$, and the feasible set X is closed, then we have that any accumulation point of the sequence $\{\widehat{\mathbf{x}}_N\}_N$ is almost surely an optimal solution of problem (1).

Proof We omit the proof, because it is essentially the same as the one in [35, Theorem 3.6], except that, since we are working with $p \geq 1$, we additionally require that $f(\mathbf{x}, \boldsymbol{\xi})$ be continuous in $\boldsymbol{\xi}$ so that we can make use of Theorem 7.12 from [45]. □

In the following remark, we show how problem $P_{(\alpha,\widetilde{\rho}_N)}$ can be used to make distributionally robust decisions in a context where the data available to the decision maker is contaminated.

Remark 1 (Data-driven decision-making under contaminated samples) Suppose that the dataset $\widehat{\boldsymbol{\xi}}_i := (\widehat{\mathbf{z}}_i, \widehat{\mathbf{y}}_i)$ for $i = 1, \ldots, N$ is composed of *correct* and *contaminated* samples. The decision maker only knows that a sample is correct with probability α and contaminated with probability $1 - \alpha$, but does not know which type each sample belongs to. Thus, the data have been generated from a mixture distribution given by $P = \alpha Q^* + (1 - \alpha)R$, where Q^* is the correct distribution and R a contamination.

In our context, this is equivalent to stating that $Q^* \in \mathscr{R}_{1-\alpha}(P)$, which, in turn, can be formulated as $\mathscr{W}_p(\mathscr{R}_{1-\alpha}(P), Q^*) = 0$. Since we only have limited information on P in the form of the empirical distribution $\widehat{P}_N$, we propose to solve problem $\mathrm{P}_{(\alpha,\widetilde{\rho}_N)}$, that is,

$$\inf_{\mathbf{x}\in X} \sup_{Q} \ \mathbb{E}_Q\left[f(\mathbf{x}, \boldsymbol{\xi})\right] \tag{18a}$$

$$\text{s.t. } \mathscr{W}_p^p(\mathscr{R}_{1-\alpha}(\widehat{P}_N), Q) \leq \widetilde{\rho}_N \tag{18b}$$

where we have assumed that the correct distribution Q^*, the contamination R and the data-generating distribution P are all supported on Ξ.

The decision maker can profit from the finite sample guarantee that the solution to problem (18a)–(18b) satisfies as per Theorem 2, with $\widetilde{\rho}_N \geq \epsilon^p_{N,p,\alpha}(\beta)$, $\beta \in (0, 1)$, since $\epsilon^p_{N\alpha} = 0$ in this case. Furthermore, if we choose a summable sequence of $\beta_N \in (0, 1)$, $N \in \mathbb{N}$, such that $\lim_{N\to\infty} \epsilon_N(\beta_N) = 0$, then we have that

$$P^\infty\left(\lim_{N\to\infty} \mathscr{W}_p\left(\mathscr{R}_{1-\alpha}(\widehat{P}_N), Q^*\right) = 0\right) = 1 \tag{19}$$

In plain words, for N large enough, the decision vector $\mathbf{x}$ is being optimized by way of problem (18a)–(18b) over the "smallest" ambiguity set that almost surely contains the correct distribution Q^* of the data (in the absence of any other information on Q^*). In fact, this means our DRO approach deals with contaminated samples in a way that is distinctly more convenient than that of [14,22]. Essentially, they suggest optimizing over a 1-Wasserstein ball centered at $\widehat{P}_N$ of radius $\widetilde{\rho}$, that is,

$$\inf_{\mathbf{x}\in X} \sup_{Q} \ \mathbb{E}_Q\left[f(\mathbf{x}, \boldsymbol{\xi})\right] \tag{20a}$$

$$\text{s.t. } \mathscr{W}_1(\widehat{P}_N, Q) \leq \widetilde{\rho} \tag{20b}$$

under the argument that for ρ sufficiently large, the Wasserstein ball contains the true distribution of the data Q^* with a certain confidence level. For instance, the author of [22] uses the triangle inequality and the convexity property of the Wasserstein distance to establish that $\mathscr{W}_1(\widehat{P}_N, Q^*) \leq \mathscr{W}_1(\widehat{P}_N, P) + (1-\alpha)\mathscr{W}_1(R, Q^*)$, so that the extra budget $(1-\alpha)\mathscr{W}_1(R, Q^*)$ would ensure that Q^* is within the Wasserstein ball with a given confidence level (a similar argument is made in [14]). In practice, though, this extra budget as such cannot be computed, because neither the correct distribution Q^* nor the contamination R are known to the decision maker. However, our approach naturally encodes it in the ambiguity set (18b). Indeed, for N large enough, result (19) tells us that the correct distribution Q^* belongs, almost surely, to the $(1-\alpha)$-trimming set of the empirical distribution $\widehat{P}_N$. It follows precisely from this and Proposition 4 in "Appendix A" that $\mathscr{W}_p(\widehat{P}_N, Q^*) \to \mathscr{W}_p(\alpha Q^* + (1-\alpha)R, Q^*) \leq \alpha\mathscr{W}_p(Q^*, Q^*) + (1-\alpha)\mathscr{W}_p(R, Q^*)$, i.e., $\mathscr{W}_p(\widehat{P}_N, Q^*) \leq (1-\alpha)\mathscr{W}_p(R, Q^*)$.

In short, our approach offers probabilistic guarantees in the finite-sample regime and, in the asymptotic one, naturally exploits all the information we have on Q^*, namely, $Q^* \in \mathscr{R}_{1-\alpha}(P)$, to robustify the decision $\mathbf{x}$ under contamination.

3.2 The case of unknown $\mathbb{Q}(\widetilde{\Xi}) = \alpha > 0$

In this section, we discuss how we can use the proposed DRO approach to deal with the case in which $\mathbb{Q}(\widetilde{\Xi}) = \alpha > 0$ is unknown. For this purpose, we first introduce a proposition that will allows us to design a distributionally robust strategy to tackle problem (1) by means of problem (P).

Proposition 3 *Suppose that $\mathbb{Q}(\widetilde{\Xi}) = \alpha > 0$. Take $0 < \alpha' < \alpha$ and any positive value of $\widetilde{\rho}$. Given $N \geq 1$, the following problem*

$$\begin{aligned} \text{(SP3)} \quad \sup_{Q_{\widetilde{\Xi}}} \ & \mathbb{E}_{Q_{\widetilde{\Xi}}}\left[f(\mathbf{x}, \boldsymbol{\xi})\right] \\ & \text{s.t. } \mathscr{W}_p^p(\mathscr{R}_{1-\alpha'}(\widehat{\mathbb{Q}}_N), Q_{\widetilde{\Xi}}) \leq \widetilde{\rho} \\ & Q_{\widetilde{\Xi}}(\widetilde{\Xi}) = 1 \end{aligned}$$

is either fully equivalent to (SP2), if $\frac{1}{N} \geq \alpha$ or a relaxation otherwise.

Proof The proof of the proposition is trivial and directly follows from the fact that $\mathscr{R}_{1-\alpha}(\widehat{\mathbb{Q}}_N) \subset \mathscr{R}_{1-\alpha'}(\widehat{\mathbb{Q}}_N)$, if $\alpha' \leq \alpha$, and that $\mathscr{R}_{1-\alpha}(\widehat{\mathbb{Q}}_N) = \mathscr{R}_{1-\alpha'}(\widehat{\mathbb{Q}}_N)$ if, besides, $\frac{1}{N\alpha} \geq 1$. □

Based on Proposition 3, we could use the following two-step *safe* strategy to handle the case of unknown $\mathbb{Q}(\widetilde{\Xi}) = \alpha > 0$:

1. First, solve the following uncertainty quantification problem (see [25,35] for further details),

$$\alpha_N := \inf_{Q \in \mathbb{B}_{\epsilon_N}(\widehat{\mathbb{Q}}_N)} Q(\boldsymbol{\xi} \in \widetilde{\Xi}) = 1 - \sup_{Q \in \mathbb{B}_{\epsilon_N}(\widehat{\mathbb{Q}}_N)} Q(\boldsymbol{\xi} \notin \widetilde{\Xi}) \tag{22}$$

where the radius ϵ_N of the Wasserstein ball has been chosen so that α_N represents the minimum probability that the joint true distribution $\mathbb{Q}$ of the data assigns to the event $\boldsymbol{\xi} \in \widetilde{\Xi}$ with confidence $1 - \beta_N$, $\beta_N \in (0, 1)$.

2. Next, solve problem $(P_{(\alpha_N, \widetilde{\rho}_N)})$, that is,

$$\inf_{\mathbf{x}\in X} \sup_{Q_{\widetilde{\Xi}}} \ \mathbb{E}_{Q_{\widetilde{\Xi}}}\left[f(\mathbf{x}, \boldsymbol{\xi})\right] \tag{23a}$$

$$\text{s.t. } \mathscr{W}_p^p(\mathscr{R}_{1-\alpha_N}(\widehat{\mathbb{Q}}_N), Q_{\widetilde{\Xi}}) \leq \widetilde{\rho}_N \tag{23b}$$

$$Q_{\widetilde{\Xi}}(\widetilde{\Xi}) = 1 \tag{23c}$$

with $\widetilde{\rho}_N \geq \epsilon_N^p(\beta_N)/\alpha_N$.

Now suppose that $\mathbb{Q} \in \mathbb{B}_{\epsilon_N(\beta_N)}(\widehat{\mathbb{Q}}_N)$ and therefore, $\alpha_N \leq \alpha$ (this is a random event that occurs with probability at least $1 - \beta_N$). According to Lemma 3, we have

$$\alpha^{1/p}\mathscr{W}_p\left(\mathscr{R}_{1-\alpha}(\widehat{\mathbb{Q}}_N), \mathbb{Q}_{\widetilde{\Xi}}\right) \leq \mathscr{W}_p\left(\widehat{\mathbb{Q}}_N, \mathbb{Q}\right) \leq \epsilon_N(\beta_N)$$

$$\mathscr{W}_p^p\left(\mathscr{R}_{1-\alpha_N}(\widehat{\mathbb{Q}}_N), \mathbb{Q}_{\widetilde{\Xi}}\right) \leq \mathscr{W}_p^p\left(\mathscr{R}_{1-\alpha}(\widehat{\mathbb{Q}}_N), \mathbb{Q}_{\widetilde{\Xi}}\right) \leq \frac{\epsilon_N^p(\beta_N)}{\alpha} \leq \frac{\epsilon_N^p(\beta_N)}{\alpha_N} = \widetilde{\rho}_N$$

Hence, $\mathbb{Q}_{\widetilde{\Xi}} \in \widehat{\mathscr{U}}_N(\alpha_N, \widetilde{\rho}_N)$ with probability at least $1 - \beta_N$. In other words, the two-step procedure here described does not degrade the reliability of the DRO solution. Furthermore, the minimum transportation budget $\underline{\epsilon}_{N\alpha_N}$ that makes problem $(P_{(\alpha_N, \widetilde{\rho}_N)})$ feasible is always zero here, if the event $\boldsymbol{\xi} \in \widetilde{\Xi}$ has been observed at least once. This is so because the uncertainty quantification problem of step 1 ensures that α_N is lower than or equal to the fraction of training data points falling in $\widetilde{\Xi}$. Moreover, when N grows to infinity, this uncertainty quantification problem reduces to computing such a fraction of points, which, by the Strong Law of Large Numbers converges to the real α, i.e., $\alpha_N \to \alpha$ with probability one. Therefore, in the asymptotic regime, this case resembles that of known $\alpha > 0$.

We notice, however, that, in practice, setting $\widetilde{\rho}_N \geq \epsilon_N^p(\beta_N)/\alpha_N$ may result in too large budgets $\widetilde{\rho}_N$, and thus, in overly conservative solutions, because, as ϵ_N is increased, α_N decreases to zero. For this reason, in the supplementary material, we provide an alternative data-driven procedure to address the case $\alpha > 0$, in which we simply set $\alpha_N = \widehat{\mathbb{Q}}_N(\widetilde{\Xi})$ in problem $(P_{(\alpha_N, \widetilde{\rho}_N)})$ and use the data to tune parameter $\widetilde{\rho}_N$.

3.3 The case $\mathbb{Q} \ll \boldsymbol{\lambda}^{\boldsymbol{d}}$ and $\mathbb{Q}(\widetilde{\boldsymbol{\Xi}}) = \boldsymbol{\alpha} = \mathbf{0}$

Suppose that the true joint distribution $\mathbb{Q}$ governing the random vector $\boldsymbol{\xi} := (\mathbf{z}, \mathbf{y})$ admits a density function with respect to the Lebesgue measure λ^d, with $d = d_{\mathbf{z}} + d_{\mathbf{y}}$. Without loss of generality, consider the event $\boldsymbol{\xi} \in \widetilde{\Xi}$, where $\widetilde{\Xi}$ is defined as $\widetilde{\Xi} = \{\boldsymbol{\xi} = (\mathbf{z}, \mathbf{y}) \in \Xi \ : \ \mathbf{z} = \mathbf{z}^*\}$. This means that $\mathbb{Q}(\widetilde{\Xi}) = \alpha = 0$.

Therefore, our focus in this case is on the particular variant of problem (1) given by

$$J^* := \inf_{\mathbf{x}\in X} \mathbb{E}_{\mathbb{Q}}\left[f(\mathbf{x},(\mathbf{z},\mathbf{y})) \mid \mathbf{z}=\mathbf{z}^*\right] \tag{24}$$

Problem (24) has become a central object of study in what has recently come to be known as *Prescriptive Stochastic Programming or Conditional Stochastic Optimization*, (see, e.g., [5,7–9,12,16,39,43], and references therein).

Devising a DRO approach to problem (24) using the standard Wasserstein ball $\mathscr{W}_p(\widehat{\mathbb{Q}}_N, Q) \leq \varepsilon$ is of no use here, because any point from the support of $\widehat{\mathbb{Q}}_N$ with an *arbitrarily small* mass can be transported to the set $\widetilde{\Xi}$ at an arbitrarily small cost in terms of $\mathscr{W}_p(\widehat{\mathbb{Q}}_N, Q)$. This way, one could always place this arbitrarily small particle at a point $(\mathbf{z}^*, \mathbf{y}') \in \arg\max_{(\mathbf{z},\mathbf{y})\in\widetilde{\Xi}} f(\mathbf{x},(\mathbf{z},\mathbf{y}))$. In contrast, problem (P), which is based on *partial* mass transportation, offers a richer framework to seek for a distributional robust solution to (24). To see this, consider again the inequality (11). If we could set $\alpha = 0$, the term $\mathscr{W}_p(\mathscr{R}_{1-\alpha}(\mathbb{Q}), \mathbb{Q}_{\widetilde{\Xi}})$ would vanish, because we could take random variables $\boldsymbol{\xi} \sim \mathbb{Q}_{\widetilde{\Xi}}$, $\boldsymbol{\xi}_m \sim \mathbb{Q}_m \in \mathscr{R}_1(\mathbb{Q})$, $m \in \mathbb{N}$, such that $\mathscr{W}_p(\mathbb{Q}_m, \mathbb{Q}_{\widetilde{\Xi}}) \to 0$. Unfortunately, fixing α to zero is not a real option due to the term $\frac{1}{\alpha^{1/p}}\mathscr{W}_p(\widehat{\mathbb{Q}}_N, \mathbb{Q})$ in the inequality. Therefore, what we propose instead is to solve a sequence of optimization problems in the form

$$\left(\mathrm{P}_{(\alpha_N,\widetilde{\rho}_N)}\right) \quad \inf_{\mathbf{x}\in X} \sup_{Q_{\widetilde{\Xi}}} \mathbb{E}_{Q_{\widetilde{\Xi}}}\left[f(\mathbf{x},\boldsymbol{\xi})\right] \tag{25a}$$

$$\text{s.t. } \mathscr{W}_p^p(\mathscr{R}_{1-\alpha_N}(\widehat{\mathbb{Q}}_N), Q_{\widetilde{\Xi}}) \leq \widetilde{\rho}_N \tag{25b}$$

$$Q_{\widetilde{\Xi}}(\widetilde{\Xi}) = 1 \tag{25c}$$

with both α_N and $\widetilde{\rho}_N$ tending to zero appropriately as N increases. Next we show that, under certain conditions, problem $\left(\mathrm{P}_{(\alpha_N,\widetilde{\rho}_N)}\right)$ enjoys a finite sample guarantee and is asymptotically consistent.

Assumption 2 (Condition (3.6) from [21]) Let $B(\mathbf{z}^*, r) := \{\mathbf{z} \in \Xi_{\mathbf{z}} : ||\mathbf{z} - \mathbf{z}^*|| \leq r\}$ denote the closed ball in $\mathbb{R}^{d_{\mathbf{z}}}$ with center $\mathbf{z}^*$ and radius r. The random vector $\boldsymbol{\xi} := (\mathbf{z}, \mathbf{y})$ has a joint density ϕ that verifies the following for some $r_0 > 0$.

1. It admits uniformly for $r \in [0, r_0]$ and $\mathbf{y} \in \mathbb{R}^{d_{\mathbf{y}}}$ the following expansion:

$$\phi(\mathbf{z}^* + r\,\mathbf{u}, \mathbf{y}) = \phi(\mathbf{z}^*, \mathbf{y})\left[1 + r\langle \mathbf{u}, \ell_1(\mathbf{y})\rangle + O(r^2\ell_2(\mathbf{y}))\right] \tag{26}$$

   where $\mathbf{u} \in \mathbb{R}^{d_{\mathbf{z}}}$ with $||\mathbf{u}|| = 1$, and where $\ell_1 : \mathbb{R}^{d_{\mathbf{y}}} \to \mathbb{R}^{d_{\mathbf{z}}}$ and $\ell_2 : \mathbb{R}^{d_{\mathbf{y}}} \to \mathbb{R}$ satisfy $\int(||\ell_1(\mathbf{y})||^2 + |\ell_2(\mathbf{y})|^2)\phi(\mathbf{z}^*, \mathbf{y})d\mathbf{y} < \infty$.
2. The marginal density of $\mathbf{z}$ is bounded away from zero in $B(\mathbf{z}^*, r_0)$.

Assumption 3 (Regularity and boundedness) We assume that

1. There exists $\widetilde{C} > 0$ and $r_0 > 0$ such that $\mathbb{P}(\|\mathbf{z}^* - \mathbf{z}\| \leq r) \geq \widetilde{C} r^{d_{\mathbf{z}}}$, for all $0 < r \leq r_0$.

2. The uncertainty $\mathbf{y}$ is bounded, that is, $\|\mathbf{y}\| \leq M$ a.s. for some constant $M > 0$.

We note that Assumption 3.1 is automatically implied by Assumption 2, but we explicitly state it here for ease of readability. Furthermore, under the boundedness condition established in Assumption 3.2, Assumption 2 is satisfied, for example, by a twice differentiable joint density $\phi(\mathbf{z}, \mathbf{y})$ with continuous and bounded partial derivatives in $B(\mathbf{z}^*, r) \times \Xi_{\mathbf{y}}$ and bounded away from zero in that set. These are standard regularity conditions in the technical literature on kernel density estimation and regression [39].

Theorem 4 (Case $\alpha = 0$: Finite sample guarantee) *Suppose that Assumptions 2, 3 and those of Proposition 2 hold. Set* $\alpha_0 := \widetilde{C} r_0^{d_{\mathbf{z}}}$. *Given* $N \geq 1$, *choose* $\alpha_N \in (0, \alpha_0]$, $\beta \in (0, 1)$, *and determine* $\epsilon_{N,p,\alpha_N}(\beta)$ *through (14).*

Then, for all

$$\widetilde{\rho}_N \geq \max\left[\left(\epsilon_{N,p,\alpha_N}(\beta) + O\left(\alpha_N^{\min\{1,\, 2/p\}/d_{\mathbf{z}}}\right)\right)^p,\ \underline{\epsilon}^p_{N\alpha_N}\right] \tag{27}$$

we have that the pair $(\widehat{\mathbf{x}}_N,\ \widehat{J}_N)$ *delivered by problem* $\left(\mathrm{P}_{(\alpha_N, \widetilde{\rho}_N)}\right)$ *with parameters* $\widetilde{\rho}_N$ *and* α_N *enjoys the finite sample guarantee (10).*

Proof For problem $\left(\mathrm{P}_{(\alpha_N, \widetilde{\rho}_N)}\right)$ to be feasible, we need $\widetilde{\rho}_N \geq \underline{\epsilon}^p_{N\alpha_N}$.

The proof essentially relies on upper bounding the term $\mathscr{W}_p(\mathscr{R}_{1-\alpha}(\mathbb{Q}), \mathbb{Q}_{\widetilde{\Xi}})$ that appears in Equation (12) of Proposition 2. To that end, define $\alpha(r) = \widetilde{C} r^{d_{\mathbf{z}}}$, for all $0 < r \leq r_0$. Set $\alpha_0 := \alpha(r_0)$. Let $\mathbb{Q}_{B(\mathbf{z}^*, r) \times \Xi_{\mathbf{y}}}$ be the probability measure of $(\mathbf{z}, \mathbf{y})$ conditional on $(\mathbf{z}, \mathbf{y}) \in B(\mathbf{z}^*, r) \times \Xi_{\mathbf{y}}$ and let $\mathbb{Q}_{B(\mathbf{z}^*, r)}$ be its $\mathbf{y}$-marginal. Note that, by Assumption 3.1, $\mathbb{Q}_{B(\mathbf{z}^*, r) \times \Xi_{\mathbf{y}}} \in \mathscr{R}_{1-\alpha(r)}(\mathbb{Q})$ provided that $0 < r \leq r_0$.

Furthermore, according to Theorem 3.5.2 in [21], there exists a positive constant A such that

$$\mathrm{Hell}(\mathbb{Q}_{B(\mathbf{z}^*, r)}, \mathbb{Q}_{\widetilde{\Xi}}) \leq A r^2$$

uniformly for $0 < r < r_0$, where Hell stands for *Hellinger distance*.

From Equation (5.1) in [42] and Assumption 3.2 we know that

$$\mathscr{W}_p(\mathbb{Q}_{B(\mathbf{z}^*, r)}, \mathbb{Q}_{\widetilde{\Xi}}) \leq M^{\frac{p-1}{p}} \mathscr{W}_1(\mathbb{Q}_{B(\mathbf{z}^*, r)}, \mathbb{Q}_{\widetilde{\Xi}})^{1/p}$$

In turn, from [26] we have that $\mathscr{W}_1(\mathbb{Q}_{B(\mathbf{z}^*, r)}, \mathbb{Q}_{\widetilde{\Xi}}) \leq M \cdot \mathrm{Hell}(\mathbb{Q}_{B(\mathbf{z}^*, r)}, \mathbb{Q}_{\widetilde{\Xi}})$. Hence,

$$\begin{aligned} \mathscr{W}_p^p(\mathbb{Q}_{B(\mathbf{z}^*, r)}, \mathbb{Q}_{\widetilde{\Xi}}) &\leq M^p \mathrm{Hell}(\mathbb{Q}_{B(\mathbf{z}^*, r)}, \mathbb{Q}_{\widetilde{\Xi}}) \\ \mathscr{W}_p(\mathbb{Q}_{B(\mathbf{z}^*, r)}, \mathbb{Q}_{\widetilde{\Xi}}) &\leq M A^{1/p} r^{2/p},\ \ 0 < r \leq r_0 \end{aligned}$$

Thus,

$$\mathscr{W}_p(\mathbb{Q}_{B(\mathbf{z}^*, r) \times \Xi_{\mathbf{y}}}, \mathbb{Q}_{\widetilde{\Xi}}) \leq r + M A^{1/p} r^{2/p},\ \ 0 < r \leq r_0$$

Since $\mathbb{Q}_{B(\mathbf{z}^*,r)\times\Xi_\mathbf{y}} \in \mathscr{R}_{1-\alpha(r)}(\mathbb{Q})$ for all $0 < r \leq r_0$, it holds

$$\mathscr{W}_p(\mathscr{R}_{1-\alpha(r)}(\mathbb{Q}), \mathbb{Q}_{\widetilde{\Xi}}) \leq \mathscr{W}_p(\mathbb{Q}_{B(\mathbf{z}^*,r)\times\Xi_\mathbf{y}}, \mathbb{Q}_{\widetilde{\Xi}}) \leq r + MA^{1/p}r^{2/p}$$

which we can express in terms of α as

$$\begin{aligned}\mathscr{W}_p(\mathscr{R}_{1-\alpha}(\mathbb{Q}), \mathbb{Q}_{\widetilde{\Xi}}) &\leq \frac{\alpha^{1/d_\mathbf{z}}}{\widetilde{C}^{1/d_\mathbf{z}}} + A^{1/p}M\frac{\alpha^{2/(pd_\mathbf{z})}}{\widetilde{C}^{2/(pd_\mathbf{z})}}\\ \mathscr{W}_p(\mathscr{R}_{1-\alpha}(\mathbb{Q}), \mathbb{Q}_{\widetilde{\Xi}}) &= O\left(\alpha^{\min\{1,\,2/p\}/d_\mathbf{z}}\right)\end{aligned}$$

provided that $0 < \alpha \leq \alpha_0$. □

Remark 2 There are conditions on the smoothness of the true joint distribution $\mathbb{Q}$ around $\mathbf{z} = \mathbf{z}^*$, other than those stated in Assumptions 2 and 3, for which we can also upper bound the distance $\mathscr{W}_p(\mathscr{R}_{1-\alpha}(\mathbb{Q}), \mathbb{Q}_{\widetilde{\Xi}})$. We provide below two examples of these conditions, which have been invoked in [9,32,33], respectively, and neither of which requires the boundedness of the uncertainty $\mathbf{y}$.

Example 1 Suppose that the true data-generating model is given by $\mathbf{y} = f^*(\mathbf{z}) + \mathbf{e}$, where $f^*(\mathbf{z}) := \mathbb{E}[\mathbf{y} \mid \mathbf{z} = \mathbf{z}^*]$ is the regression function and $\mathbf{e}$ is a zero-mean random error. Furthermore, suppose that Assumption 3.1 holds and there exists a positive constant L such that $\|f^*(\mathbf{z}') - f^*(\mathbf{z})\| \leq L\|\mathbf{z}' - \mathbf{z}\|$, for all $0 \leq \|\mathbf{z}' - \mathbf{z}\| \leq r_0$.

Take $\alpha(r) = \widetilde{C}r^{d_\mathbf{z}}$, for all $0 < r \leq r_0$ and set $\alpha_0 := \alpha(r_0)$. With abuse of notation, we can write for any event within $B(\mathbf{z}^*, r) \times \Xi_\mathbf{y}$

$$\mathbb{Q}_{B(\mathbf{z}^*,r)\times\Xi_\mathbf{y}}(d\mathbf{z}, d\mathbf{y}) = \frac{1}{\mathbb{P}(B(\mathbf{z}^*, r))}\mathbb{Q}(d\mathbf{z}, d\mathbf{y}) = \frac{1}{\mathbb{Q}_\mathbf{z}(B(\mathbf{z}^*, r))}\mathbb{Q}_{\mathbf{z}=\mathbf{z}'}(d\mathbf{y})\mathbb{Q}_\mathbf{z}(d\mathbf{z}')$$

where $\mathbb{Q}_\mathbf{z}$ is the probability law of the feature vector $\mathbf{z}$ and $\mathbb{Q}_{\mathbf{z}=\mathbf{z}'}$ is the conditional measure of $\mathbb{Q}$ given that $\mathbf{z} = \mathbf{z}'$.

Since $\mathbb{Q}_{B(\mathbf{z}^*,r)\times\Xi_\mathbf{y}} \in \mathscr{R}_{1-\alpha(r)}(\mathbb{Q})$ for all $0 < r \leq r_0$, by the convexity of the Wasserstein distance, we have

$$\begin{aligned}&\mathscr{W}_p(\mathscr{R}_{1-\alpha}(\mathbb{Q}), \mathbb{Q}_{\widetilde{\Xi}}) \leq \mathscr{W}_p(\mathbb{Q}_{B(\mathbf{z}^*,r)\times\Xi_\mathbf{y}}, \mathbb{Q}_{\widetilde{\Xi}})\\ &\leq \int_{B(\mathbf{z}^*,r)} \left[\|\mathbf{z}' - \mathbf{z}^*\| + \mathscr{W}_p(\mathbb{Q}_{\mathbf{z}=\mathbf{z}'}, \mathbb{Q}_{\widetilde{\Xi}})\right] \frac{\mathbb{Q}_\mathbf{z}(d\mathbf{z}')}{\mathbb{Q}_\mathbf{z}(B(\mathbf{z}^*, r))}\\ &= \int_{B(\mathbf{z}^*,r)} \left[\|\mathbf{z}' - \mathbf{z}^*\| + \mathscr{W}_p(f^*(\mathbf{z}') + \mathbf{e},\, f^*(\mathbf{z}^*) + \mathbf{e})\right] \frac{\mathbb{Q}_\mathbf{z}(d\mathbf{z}')}{\mathbb{Q}_\mathbf{z}(B(\mathbf{z}^*, r))}\\ &\leq \int_{B(\mathbf{z}^*,r)} \left[\|\mathbf{z}' - \mathbf{z}^*\| + \|f^*(\mathbf{z}') - f^*(\mathbf{z}^*)\|\right] \frac{\mathbb{Q}_\mathbf{z}(d\mathbf{z}')}{\mathbb{Q}_\mathbf{z}(B(\mathbf{z}^*, r))}\\ &\leq (1 + L)\int_{B(\mathbf{z}^*,r)} \|\mathbf{z}' - \mathbf{z}^*\| \frac{\mathbb{Q}_\mathbf{z}(d\mathbf{z}')}{\mathbb{Q}_\mathbf{z}(B(\mathbf{z}^*, r))} = (1 + L)O(r) = O(\alpha^{1/d_\mathbf{z}})\end{aligned}$$

for all $0 < \alpha \leq \alpha_0$.

Example 2 Take $p = 1$. Suppose that there exists a positive constant L such that $\mathscr{W}_1(\mathbb{Q}_{\mathbf{z}=\mathbf{z}'}, \mathbb{Q}_{\mathbf{z}=\mathbf{z}^*}) \leq L\|\mathbf{z}' - \mathbf{z}^*\|$, for all $0 \leq \|\mathbf{z}' - \mathbf{z}\| \leq r_0$ and that Assumption 3.1 holds.

Following a line of reasoning that is parallel to that of the previous example, we also get

$$\mathscr{W}_1(\mathscr{R}_{1-\alpha}(\mathbb{Q}), \mathbb{Q}_{\widetilde{\Xi}}) = O(\alpha^{1/d_{\mathbf{z}}})$$

for all $0 < \alpha \leq \alpha_0$, with $\alpha_0 := \alpha(r_0)$.

Equation (27) and Examples 1 and 2 reveal that our finite sample guarantee is affected by the *curse of dimensionality*. Recently, powerful ideas to break this curse have been introduced in [24] under the standard Wasserstein-metric-based DRO scheme. In our setup, however, we also need distributional robustness against the (uncertain) error incurred when inferring conditional information from a sample of the true *joint* distribution. This implies increasing the robustness budget in our approach by an amount linked to the term $\mathscr{W}_p(\mathscr{R}_{1-\alpha}(\mathbb{Q}), \mathbb{Q}_{\widetilde{\Xi}})$. Consequently, we might need stronger assumptions on the data-generating model to break the dependence of this term with the dimension of the feature vector and thus extend the ideas in [24] to the realm of conditional stochastic optimization.

Now we state the conditions under which the sequence of problems $\left(\mathrm{P}_{(\alpha_N, \widetilde{\rho}_N)}\right)$, $N \to \infty$, is asymptotically consistent.

Lemma 5 (Convergence of conditional distributions) *Suppose that the support Ξ of the true joint distribution $\mathbb{Q}$ is compact and that Assumptions 2 and 3 .1 hold. Take $(\alpha_N, \widetilde{\rho}_N)$ such that $\alpha_N \to 0$, $\frac{N\alpha_N^2}{\log(N)} \to \infty$, and $\widetilde{\rho}_N \downarrow \epsilon^p_{N\alpha_N}$, where $\epsilon_{N\alpha_N}$ is the minimum transportation budget as in Definition 2. Then, we have that*

$$\mathscr{W}_p(Q^N_{\widetilde{\Xi}}, \mathbb{Q}_{\widetilde{\Xi}}) \to 0 \ a.s.$$

where $Q^N_{\widetilde{\Xi}}$ is any distribution from the ambiguity set $\widehat{\mathscr{U}}_N(\alpha_N, \widetilde{\rho}_N)$.

Proof First, we need to provide conditions under which $\mathscr{W}_p\left(\mathscr{R}_{1-\alpha}(\widehat{\mathbb{Q}}_N), \mathbb{Q}_{\widetilde{\Xi}}\right) \to 0$ a.s. Since Ξ is compact and $\mathscr{W}_{p-1}\left(\mathscr{R}_{1-\alpha}(\widehat{\mathbb{Q}}_N), \mathbb{Q}_{\widetilde{\Xi}}\right) \leq \mathscr{W}_p\left(\mathscr{R}_{1-\alpha}(\widehat{\mathbb{Q}}_N), \mathbb{Q}_{\widetilde{\Xi}}\right)$, we can take $p > d/2$ and α_N such that $\frac{N\alpha_N^2}{\log(N)} \to \infty$, so that the probabilities (12) becomes summable over N for any arbitrarily small ϵ. In this way, we can choose a sequence $\beta_N \in (0, 1)$, $N \in \mathbb{N}$, such that $\sum_{N=1}^{\infty} \beta_N < \infty$ and $\lim_{N\to\infty} \epsilon_{N,p,\alpha_N}(\beta_N) \to 0$. With this choice, we have

$$\begin{aligned}
&\mathbb{Q}^\infty\left[\lim_{N\to\infty} \mathscr{W}_p\left(\mathscr{R}_{1-\alpha_N}(\widehat{\mathbb{Q}}_N), \mathbb{Q}_{\widetilde{\Xi}}\right) - \mathscr{W}_p\left(\mathscr{R}_{1-\alpha_N}(\mathbb{Q}), \mathbb{Q}_{\widetilde{\Xi}}\right) = 0\right] \\
&\quad = \mathbb{Q}^\infty\left[\lim_{N\to\infty} \mathscr{W}_p\left(\mathscr{R}_{1-\alpha_N}(\widehat{\mathbb{Q}}_N), \mathbb{Q}_{\widetilde{\Xi}}\right) = 0\right] = 1
\end{aligned}$$

because $\mathscr{W}_p\left(\mathscr{R}_{1-\alpha_N}(\mathbb{Q}), \mathbb{Q}_{\widetilde{\Xi}}\right) = O\left(\alpha_N^{2/pd_{\mathbf{z}}}\right) \to 0$ for $\alpha_N \to 0$.

Since, $\mathbb{Q}_{\widetilde{\Xi}} \in \mathscr{R}_{1-\alpha_N}(\widehat{\mathbb{Q}}_N)$ a.s. in the limit and, by definition, $\mathbb{Q}_{\widetilde{\Xi}}(\widetilde{\Xi}) = 1$, we have that $\mathbb{Q}_{\widetilde{\Xi}} \in \widehat{\mathscr{U}}_N(\alpha_N, \widetilde{\rho}_N)$ for N sufficiently large, with both $\alpha_N, \widetilde{\rho}_N \to 0$.

For its part, because $Q^N_{\widetilde{\Xi}} \in \widehat{\mathscr{U}}_N(\alpha_N, \widetilde{\rho}_N)$, this means that $\mathscr{W}_p\left(\mathscr{R}_{1-\alpha_N}(\widehat{\mathbb{Q}}_N), Q^N_{\widetilde{\Xi}}\right) \leq \widetilde{\rho}_N$. Take N large enough, set $\widetilde{\rho}_N$ arbitrarily close to $\underline{\epsilon}^p_{N\alpha_N}$ and notice that $\widehat{\mathscr{U}}_N(\alpha_N, \underline{\epsilon}^p_{N\alpha_N})$ boils down to one single probability measure, the one made up of the $N\alpha_N$ data points of $\widehat{\mathbb{Q}}_N$ that are the closest to $\widetilde{\Xi}$. In addition, we have $\underline{\epsilon}^p_{N\alpha_N} \to 0$ with probability one. To see this, take $K := \lceil N\alpha_N \rceil$ and note that

$$\underline{\epsilon}^p_{N\alpha_N} \leq \text{dist}(\widehat{\boldsymbol{\xi}}_{K:N}, \widetilde{\Xi}) \to \|\widehat{\mathbf{z}}_{K:N} - \mathbf{z}^*\| \to 0$$

almost surely provided that $\alpha_N \to 0$ (see [13, Lemmas 2.2 and 2.3]), where $\widehat{\mathbf{z}}_{K:N}$ is the $\mathbf{z}$-component of the K-*th* nearest neighbor to $\mathbf{z}^*$ after reordering the data sample $\{\widehat{\boldsymbol{\xi}}_i := (\widehat{\mathbf{z}}_i, \widehat{\mathbf{y}}_i)\}_{i=1}^N$ in terms of $\|\widehat{\mathbf{z}}_i - \mathbf{z}^*\|$ only.

Therefore, it must hold that $\mathscr{W}_p(Q^N_{\widetilde{\Xi}}, \mathbb{Q}_{\widetilde{\Xi}}) \to 0$ a.s. □

Remark 3 The compactness of the support set Ξ is assumed here just to simplify the proof. In fact, in the "Appendix EC.2" of the extended version of this paper [19], we use results from nearest neighbors to show that the convergence of conditional distributions can be attained under the less restrictive condition $\frac{N\alpha_N}{\log(N)} \to \infty$ even in some cases for which the uncertainty $\mathbf{y}$ and the feature vector $\mathbf{z}$ are unbounded. In addition, we also make use of those results to demonstrate that distributionally robust versions of some local nonparametric predictive methods, such as Nadaraya-Watson kernel regression and K-nearest neighbors, naturally emerge from our approach.

Remark 4 The convergence of conditional distributions allows us to establish an asymptotic consistency result analogous to that of Theorem 3, by simply replacing "Theorem 2", "$\widetilde{\rho}_N$" and "Lemma 4" with "Theorem 4", "$(\alpha_N, \widetilde{\rho}_N)$" and "Lemma 5", respectively.

Remark 5 Suppose that the event $\widetilde{\Xi}$ on which we condition problem (1) is given by $\widetilde{\Xi} := \{\boldsymbol{\xi} = (\mathbf{z}_1, \mathbf{z}_2, \mathbf{y}) \in \Xi \ : \ \mathbf{z}_1 = \mathbf{z}_1^*, \ \mathbf{z}_2 \in \mathscr{Z}_2\}$, with $\mathbb{Q}(\widetilde{\Xi}) = 0$ and $\mathbb{P}(\mathbf{z}_2 \in \mathscr{Z}_2) > 0$. Let $\mathbb{Q}_{\mathscr{Z}_2}$ be the probability measure of $(\mathbf{z}_1, \mathbf{y})$ conditional on $\mathbf{z}_2 \in \mathscr{Z}_2$. If we have that there is $\widetilde{C} > 0$ and $r_0 > 0$ such that $\mathbb{P}(\|\mathbf{z}_1^* - \mathbf{z}_1\| \leq r) \geq \widetilde{C} r^{d_{\mathbf{z}_1}}$, for all $0 < r \leq r_0$, and that $\mathbb{Q}_{\mathscr{Z}_2}$ satisfies the smoothness condition invoked in either Theorem 4, Example 1 or Example 2, then the analysis in this section extends to that type of event by setting $\alpha(r) = \widetilde{C} r^{d_{\mathbf{z}_1}} \cdot \mathbb{P}(\mathbf{z}_2 \in \mathscr{Z}_2)$ and noticing that $\mathbb{Q}_{B(\mathbf{z}_1^*, r) \times \mathscr{Z}_2 \times \Xi_{\mathbf{y}}} \in \mathscr{R}_{1-\alpha(r)}(\mathbb{Q})$, $0 < r \leq r_0$, where $\mathbb{Q}_{B(\mathbf{z}_1^*, r) \times \mathscr{Z}_2 \times \Xi_{\mathbf{y}}}$ is the probability measure of $(\mathbf{z}_1, \mathbf{z}_2, \mathbf{y})$ conditional on $(\mathbf{z}_1, \mathbf{z}_2, \mathbf{y}) \in B(\mathbf{z}_1^*, r) \times \mathscr{Z}_2 \times \Xi_{\mathbf{y}}$.

4 Numerical experiments

The following simulation experiments are designed to provide numerical evidence on the performance of the DRO framework with side information that we propose, with respect to other methods available in the technical literature. Here we only consider

the case $\alpha = 0$, while additional numerical experiments for the case $\alpha > 0$ can be found in the supplementary material.

To numerically illustrate the setting $\mathbb{Q}(\widetilde{\Xi}) = \alpha = 0$, we consider two well-known problems, namely, the (single-item) newsvendor problem and the portfolio allocation problem, both posed in the form $\inf_{\mathbf{x}\in X} \mathbb{E}_{\mathbb{Q}}\left[f(\mathbf{x}, \boldsymbol{\xi}) \mid \boldsymbol{\xi} \in \widetilde{\Xi}\right]$ to allow for side information. We compare four data-driven approaches to address the solution to these two problems: Our approach, i.e., problem $\mathrm{P}_{(\alpha_N, \widetilde{\rho}_N)}$ with $\alpha_N = K_N/N$, which we denote "DROTRIMM"; a Sample Average Approximation method based on a local predictive technique, in particular, the K_N nearest neighbors, which we refer to as "KNN" (see [7] for further details); this very same local predictive method followed by a standard Wasserstein-metric-based DRO approach to robustify it, as suggested in [9, Section 5], which we call "KNNDRO"; and the robustified KNN method (4), also proposed in [9], which we term "KNNROBUST." We clarify that KNNDRO uses the K nearest neighbors projected onto the set $\widetilde{\Xi}$ as the nominal "empirical" distribution that is used as the center of the Wasserstein ball in [35].

We also note that the newsvendor problem and the portfolio optimization problem are structurally different if seen from the lens of the standard Wasserstein-metric-based DRO approach. Indeed, the newsvendor problem features an objective function with a Lipschitz constant with respect to the uncertainty that is independent of the decision $\mathbf{x}$. Consequently, as per [35, Remark 6.7], KNNDRO renders the same minimizer for this problem as that of KNN whenever the support set $\widetilde{\Xi}$ is equal to the whole space. This is, in contrast, not true for the portfolio allocation problem, which has an objective function with a Lipschitz constant with regard to the uncertainty that depends on the decision $\mathbf{x}$.

In all the numerical experiments, we take the p-norm with $p = 1$ and, accordingly, we use the Wasserstein distance of order 1. Thus, all the optimization problems that we solve are linear programs. We consider a series of different values for the size N of the sample data. Unless stated otherwise in the text, for each N, we choose as the number of neighbors, K_N, the value $\lfloor N/\log(N+1)\rfloor$, where $\lfloor\cdot\rfloor$ stands for the floor function. Nevertheless, for the portfolio allocation problem, we also test the values $\lfloor N^{0.9}\rfloor$ and $\lfloor\sqrt{N}\rfloor$ to assess the impact of the number of neighbors on the out-of-sample performance of the four methods we compare.

We estimate $\mathbf{x}^* \in \arg\min_{\mathbf{x}\in X} \mathbb{E}_{\mathbb{Q}_{\widetilde{\Xi}}}\left[f(\mathbf{x}, \boldsymbol{\xi})\right]$ and $J^* = \mathbb{E}_{\mathbb{Q}_{\widetilde{\Xi}}}\left[f(\mathbf{x}^*, \boldsymbol{\xi})\right]$ using a discrete proxy of the true conditional distribution $\mathbb{Q}_{\widetilde{\Xi}}$. In the newsvendor problem, this proxy is made up of 1085 data points, resulting from applying the KNN method (with the logarithmic rule) to 10,000 samples from the true data-generating joint distribution. In the portfolio optimization problem, we have an explicit form of $\mathbb{Q}_{\widetilde{\Xi}}$, which we utilize to directly construct a 10,000-data-point approximation. To compare the four data-driven approaches under consideration, we use two performance metrics, specifically, the *out-of-sample performance* of the data-driven solution and its *out-of-sample disappointment*. The former is given by $J = \mathbb{E}_{\mathbb{Q}_{\widetilde{\Xi}}}\left[f(\widehat{\mathbf{x}}_N^m, \boldsymbol{\xi})\right]$, while the latter is calculated as $J - \widehat{J}_N^m$, where $m = \{\text{KNN, KNNROBUST, DROTRIMM, KNNDRO}\}$ and $\widehat{J}_N^m$ is the objective function value yielded by the data-driven optimization problem solved by method m. We note that a negative out-of-sample disappointment represents a favorable outcome.

Since $\mathbb{E}_{\mathbb{Q}_{\widetilde{\varepsilon}}}\left[f(\widehat{\mathbf{x}}_N^m, \boldsymbol{\xi})\right]$ and $\widehat{J}_N^m$ are functions of the sample data, we conduct a certain number of runs (400 for the newsvendor problem and 200 for the portfolio optimization problem) for every N, each run with an independent sample of size N. This way we can get (visual) estimates of the out-of-sample performance and disappointment for several values of the sample size N for different independent runs. These estimates are illustrated in the form of box plots in a series of figures, where the dotted black horizontal line corresponds to either the optimal solution $\mathbf{x}^*$ (only in the newsvendor problem) or to its associated optimal cost J^* with complete information.

As is customary in practice, we use a data-driven procedure to tune the robustness parameter of each method. In particular, for a desired value of reliability $1-\beta \in (0, 1)$ (in our numerical experiments, we set β to 0.15), and for each method j, where $j = \{\text{KNNROBUST, KNNDRO, DROTRIMM}\}$, we aim for the value of the robustness parameter for which the estimate of the objective value $\widehat{J}_N^j$ given by method j provides an upper $(1-\beta)$-confidence bound on the out-of-sample performance of its respective optimal solution (see Eq. (10)), while delivering the best out-of-sample performance. As the optimal robustness parameter is unknown and depends on the available data sample, we need to derive an estimator $param_N^{\beta,j}$ that is also a function of the training data. We construct $param_N^{\beta,j}$ and the corresponding reliability-driven solution as follows:

1. We generate *kboot* resamples (with replacement) of size N, each playing the role of a different training set. In our experiments we set $kboot = 50$. Moreover, we build a validation dataset determining the $K_{N_{val}}$-neighbors of the N_{val} data points of the original sample of size N that have not been used to form the training set.
2. For each resample $k = 1, \ldots, kboot$ and each candidate value for $param$, we compute a solution by method j with parameter $param$ on the k-th resample. The resulting optimal decision is denoted as $\widehat{x}_N^{j,k}(param)$ and its corresponding objective value as $\widehat{J}_N^{j,k}(param)$. Thereafter, we calculate 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 validation set.
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, we take as $param_N^{\beta,j}$ the one yielding the best out-of-sample performance averaged over the $kboot$ validation datasets.
4. Finally, we 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 $\widehat{J}_N^j := \widehat{J}_N^j(param_N^{\beta,j})$.

Recall that, in our approach, the robustness parameter $\widetilde{\rho}_N$ must be greater than or equal to the minimum transportation budget to the power of p, that is, $\underline{\varepsilon}_{N\alpha_N}^p$. Hence, if we decompose $\widetilde{\rho}_N$ as $\widetilde{\rho}_N = \underline{\varepsilon}_{N\alpha_N}^p + \Delta\widetilde{\rho}_N$, what one really needs to tune in DROTRIMM is the budget excess $\Delta\widetilde{\rho}_N$. Furthermore, for the same amount of budget $\Delta\widetilde{\rho}_N$, our approach will lead to more robust decisions $\mathbf{x}$ than KNNDRO, because the worst-case distribution in KNNDRO is also feasible in DROTRIMM. Consequently, in practice, the tuning of one of these methods could guide the tuning of the other.

Lastly, all the simulations have been run on a Linux-based server using up to 116 CPUs running in paralell, each clocking at 2.6 GHz with 4 GB of RAM. We have employed Gurobi 9.0 under Pyomo 5.2 to solve the associated linear programs.

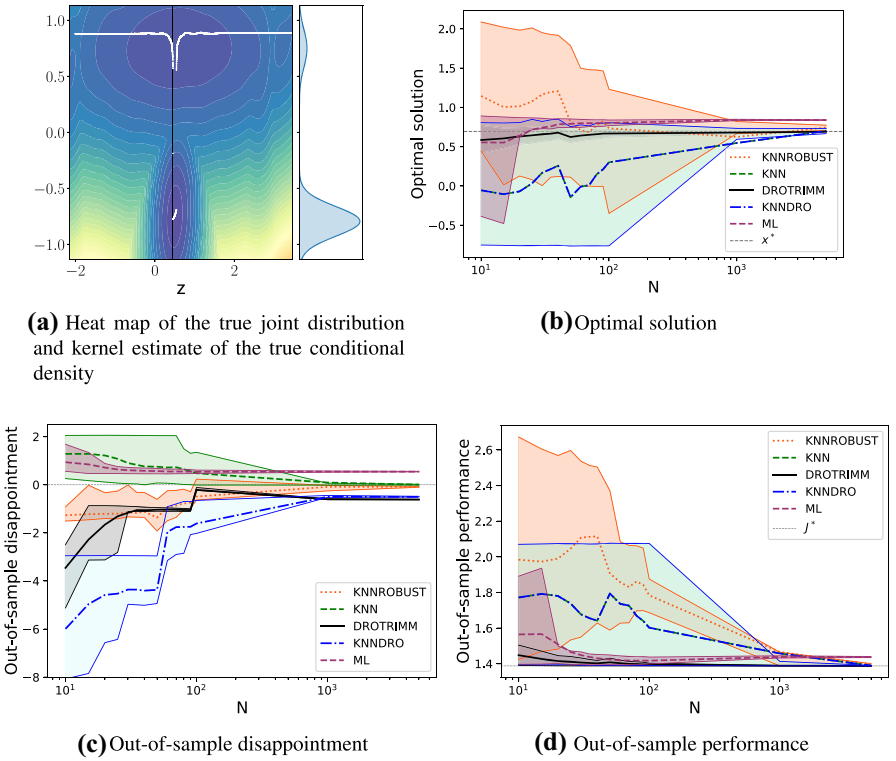
4.1 The single-item newsvendor problem

In this subsection, we deal with the popular single-item newsvendor problem, which has received a lot of attention lately (see, for example, [5,30] and references therein). It is known that the solution to the single-item newsvendor problem is equivalent to that of a quantile regression problem, where the goal is to estimate the quantile $b/(b+h)$ of the distribution of the uncertainty y, with h and b being the unit holding and backorder costs, respectively.

For the particular instance of this problem that we analyze next, we have considered $h = 1$ and $b = 10$. Furthemore, the true joint distribution of the data $\widehat{\boldsymbol{\xi}}_i := (\widehat{z}_i, \widehat{y}_i), i = 1, \ldots, N$ is assumed to follow a mixture (with equal weights) of two normal bivariate distributions with means $\mu_1 = [0.6, 0.75]^T$, $\mu_2 = [0.5, -0.75]^T$ and covariance matrices $\Sigma_1 = \begin{bmatrix} 0.5 & 0 \\ 0 & 0.01 \end{bmatrix}$, $\Sigma_2 = \begin{bmatrix} 0.0001 & 0 \\ 0 & 0.1 \end{bmatrix}$ respectively. Therefore, the support set of this distribution is the whole space $\mathbb{R}^{d_z+d_y}$, with $d_z = d_y = 1$. In addition, we consider as $\mathscr{Z}$ the singleton $\{z^* = 0.44\}$, with $\widetilde{\Xi}$ being the real line $\mathbb{R}$ as a result. Figure 1a shows a heat map of the true joint distribution, together with a kernel estimate of the probability density function of the random variable y conditional on z^*. Moreover, the white dotted curve in the figure corresponds to the optimal order quantity as a function of the feature z. Note that this curve is highly nonlinear around the context z^*. Also, the demand may be negative, which, in the context of the newsvendor problem, can be interpreted as items being returned to the stores due to, for example, some quality defect. The set of candidate values from which the robustness parameters in methods KNNROBUST, KNNDRO and DROTRIMM have been selected is the discrete set composed of the thirty linearly spaced numbers between 0 and 2. We also consider the machine learning algorithm proposed in [5], which was especially designed for the newsvendor problem with features. In this algorithm, a polynomial mapping between the optimal order quantity (i.e., the optimal quantile) and the covariates is presumed. The degree of the polynomial, up to the fourth degree, is tuned using the bootstrapping procedure described above. We denote this approach as ML from "Machine Learning".

Figure 1b–d illustrate the box plots corresponding to the quantile estimators (i.e., the optimal solution of the problem), the out-of-sample disappointment and the out-of-sample performance delivered by each of the considered data-driven approaches for various sample sizes and runs, in that order. The shaded color areas have been obtained by joining the $15th$ and $85th$ percentiles of the box plots, while the associated bold colored lines link their means. The true optimal quantile (with complete information) and its out-of-sample performance are also depicted in Fig. 1c, b, respectively, using black dotted lines.

Interestingly, whereas the quantile estimators provided by DROTRIMM, KNNDRO and KNNROBUST all lead to negative out-of-sample disappointment in general,



(a) Heat map of the true joint distribution and kernel estimate of the true conditional density

(b) Optimal solution

(c) Out-of-sample disappointment

(d) Out-of-sample performance

Fig. 1 Newsvendor problem with features: true distributions, quantile estimate and performance metrics

KNNDRO and KNNROBUST exhibit substantially worse out-of-sample performance both in expectation and volatility. Recall that KNNDRO delivers the same solutions provided by KNN for this problem. Its behavior is, therefore, influenced by the bias introduced by the K-nearest neighbors estimation, which is particularly notorious for small-size samples in this case, given the shape of the true conditional density, see Fig. 1a. Actually, for some runs, the K-nearest neighbors, and hence KNNDRO, lead to negative quantile estimates, while the true one is positive and greater than 0.5. By construction, both KNNDRO and KNNROBUST are mainly affected by the estimation error of the conditional probability distribution incurred by the local predictive method. On the contrary, our approach DROTRIMM offers a natural protection against this error and a richer spectrum of data-driven solutions. Indeed, DROTRIMM is able to identify solutions that lead to a better out-of-sample performance with a negative out-of-sample disappointment.

Finally, both ML and DROTRIMM exhibit a notorious stable behavior against the randomness of the sample. The order quantity provided by the former, however, does not converge to the true optimal one, because the relationship between the true optimal order and the feature z is far from being polynomial. Note that ML is a *global* method that seeks to learn the optimal order quantity for *all* possible contexts by using a polynomial up to the fourth degree. However, the (true) optimal order curve (that is,

the white line in Fig. 1a) is highly nonlinear within a neighborhood of the context $z^* = 0.44$, but practically constant outside of it.

4.2 Portfolio optimization

We consider next an instance of the portfolio optimization problem that is based on that used in [8,12]. The instance corresponds to a single-stage portfolio optimization problem in which we wish to find an allocation of a fixed budget to six different assets. Thus, $\mathbf{x} \in \mathbb{R}^6_+$ denotes the decision variable vector, that is, the asset allocations, and their uncertain return is represented by $\mathbf{y} \in \mathbb{R}^6$. In practice, these uncertain returns may be influenced by a set of features. First, the decision maker observes auxiliary covariates and later, selects the portfolio. We consider three different covariates that can potentially impact the returns and that we denote as $\mathbf{z} = (z_1, z_2, z_3)$. The decision maker wishes to leverage this side information to improve his/her decision-making process in which the goal is to maximize the expected value of the return while minimizing the conditional value at risk (CVar) of the portfolio, that is, the risk that the loss $(-\langle \mathbf{x}, \mathbf{y}\rangle)^+ := \max(-\langle \mathbf{x}, \mathbf{y}\rangle, 0)$ is large. Using the reformulation of the CVar (see [12,41]) and introducing the auxiliary variable β', the decision maker aims to solve the following optimization problem given the value of the covariate $\mathbf{z}^*(= (1000, 0.01, 5)$ in the numerical experiments):

$$\min_{(\mathbf{x},\beta')\in X} \mathbb{E}\left[\beta' + \frac{1}{\delta}\left(-\langle \mathbf{x}, \mathbf{y}\rangle - \beta'\right)^+ - \lambda\langle \mathbf{x}, \mathbf{y}\rangle \;\middle|\; \mathbf{z} = \mathbf{z}^*\right] \tag{28}$$

where the feasible set of decision variables of the problem, that is, X is equal to $\{(\mathbf{x}, \beta') \in \mathbb{R}^6_+ \times \mathbb{R} : \sum_{j=1}^6 x_j = 1\}$. We set $\delta = 0.5$ and $\lambda = 0.1$ to simulate an investor with a moderate level of risk aversion. The parameter $\lambda \in \mathbb{R}_+$ serves to tradeoff between risk and return, and δ refers to the $(1-\delta)$-quantile of the loss distribution. We take the same marginal distributions for the covariates as in Section 5.2 of [12], i.e., $z_1 \rightsquigarrow \mathcal{N}(1000, 50)$, $z_2 \rightsquigarrow \mathcal{N}(0.02, 0.01)$ and $\log(z_3) \rightsquigarrow \mathcal{N}(0, 1)$. Furthermore, we follow their approach to construct the joint true distribution of the covariates and the asset returns. In particular, we take

$$\begin{aligned}\mathbf{y}/(\mathbf{z} = (z_1, z_2, z_3)) \rightsquigarrow \mathcal{N}_6(&\boldsymbol{\mu} + 0.1 \cdot (z_1 - 1000) \cdot \mathbf{v}_1 + 1000 \cdot z_2 \cdot \mathbf{v}_2 \\ &+10 \cdot \log(z_3 + 1) \cdot \mathbf{v}_3, \boldsymbol{\Sigma})\end{aligned}$$

with $\mathbf{v}_1 = (1, 1, 1, 1, 1, 1)^T$, $\mathbf{v}_2 = (4, 1, 1, 1, 1, 1)^T$, $\mathbf{v}_3 = (1, 1, 1, 1, 1, 1)^T$, and with $\boldsymbol{\mu}$, $\boldsymbol{\Sigma}^{1/2}$ given in [12,20].

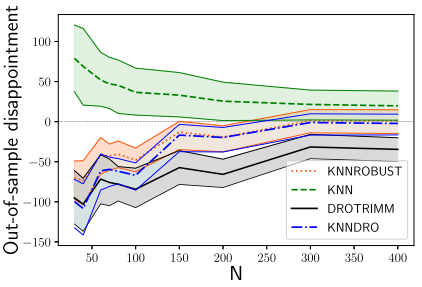
Note that, unlike in [12], not all the features affect equally all the asset returns. Moreover, feature z_3 is log-normal and therefore, Assumption 1 does not hold. However, as we show below, DROTRIMM performs satisfactorily, which reveals that the conditions we derive in this paper to guarantee that our approach performs well are sufficient, but not necessary. Indeed, the condition $\mathbb{Q}_{\widetilde{\Xi}} \in \widehat{\mathcal{U}}_N(\alpha_N, \widetilde{\rho}_N)$ is not required to ensure performance guarantees [24,34]. For all the methods, we have standardized the covariates $\mathbf{z}$ and the asset returns $\mathbf{y}$ using their means and variances. In all the simula-

tions, the robustness parameter each method uses (i.e., ε_N in KNNROBUST, the radius of the Wassertein ball, ρ_N, in KNNDRO, and the budget excess $\Delta\widetilde{\rho}_N$ in DROTRIMM) has been chosen from the discrete set $\{b \cdot 10^c \ : \ b \in \{0, \ldots, 9\}, \ c \in \{-2, -1, 0\}\}$, following the above data-driven procedure.

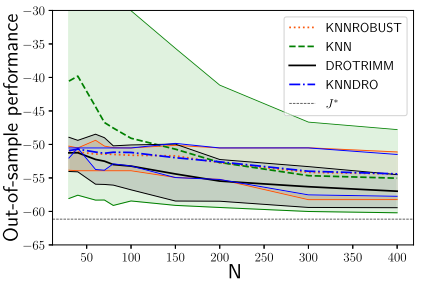
Similarly to the case of the single-item newsvendor problem, Fig. 2 shows, for various sample sizes and 200 runs, the box plots pertaining to the out-of-sample disappointment and performance associated with each of the considered data-driven approaches. Each of the three pairs of subplots at the top of the figure has been obtained with a different rule to determine the number K_N of nearest neighbors. Increasing this number seems to have a positive effect on the convergence speed of all the methods for this instance, although KNNROBUST (and KNNDRO to a lesser extent) has some trouble ensuring the desired reliability level, with the 85% line above 0 for the largest values of N we represent. In contrast, DROTRIMM manages to keep the disappointment negative. This is, in addition, accompanied by an important improvement of the the out-of-sample performance (in line with the criterion for selecting the best portfolio that we have established). In fact, DROTRIMM produces boxplots that appear to be shifted downward, i.e., in the direction of better objective function values. On the other hand, the KNN method substantially improves its performance by employing a larger number of neighbors. However, it is way too optimistic in any case.

The results shown in the pair of subplots at the bottom of Fig. 2 correspond to a number K_N of neighbors that has been tuned jointly with the robustness parameter and for each method independently. For this purpose, we have selected the best value of K_N for each approach from the discrete set $\{N^{0.1}, N^{0.2}, \ldots, N^{0.9}\}$ following the bootstrapping-based procedure previously described. The data-driven tuning of the number K_N of neighbors appears not to have a major effect on the performance of the different methods, especially in comparative terms. We do observe that the out-of-sample performance of KNNROBUST and KNNDRO is slightly improved on average. This improvement in cost performance is, however, accompanied by an increase in the number of sample sizes for which these methods do not satisfy the reliability requirement, particularly in the case of KNNROBUST and small sample sizes.

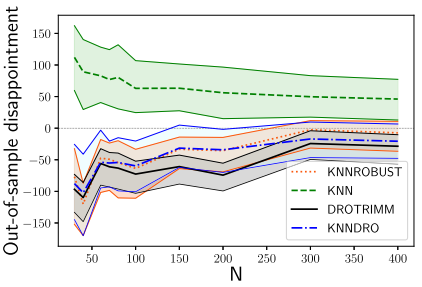
To facilitate the analysis of the results shown in Fig. 2, we also provide Fig. 3, which illustrates the (random) performance of the methods KNNROBUST, DROTRIMM and KNNDRO as a function of their respective robustness parameter, estimated over 200 independent runs. Again, the shaded areas cover the $15th$ and $85th$ percentiles, while the bold colored lines correspond to the average performance. The various plots are obtained for $N = 30$ and $N = 400$, with the number of neighbours given by the logarithmic rule. These plots are especially informative, because they are independent of the specific validation procedure used to tune the robustness parameters of the methods and thus, provide insight into the potential of each method to identify good solutions. Note that the out-of-sample performance of all the three methods stabilizes around the same value as their respective robustness parameters grow large enough. This phenomenon is analogous to that discussed in [35, Section 7.1]. However, the value we observe here does not correspond to the "equally weighted portfolio," because we have standardized the data on the asset returns. As a result, the "robust portfolio" that delivers this out-of-sample performance depends on and is solely driven by the standard deviations of the different assets. Very interestingly, DROTRIMM is able to



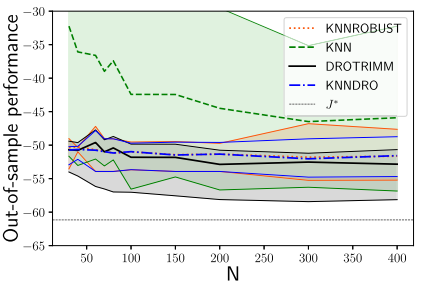
(a) $K_N = \lfloor N/(\log(N+1)) \rfloor$



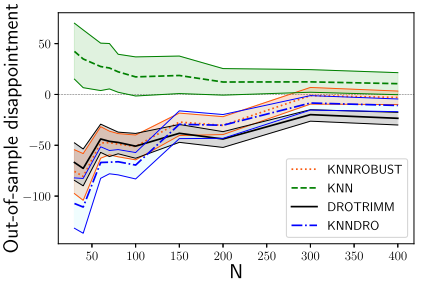
(b) $K_N = \lfloor N/(\log(N+1)) \rfloor$



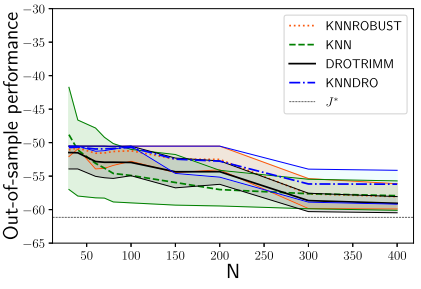
(c) $K_N = \lfloor N^{0.5} \rfloor$



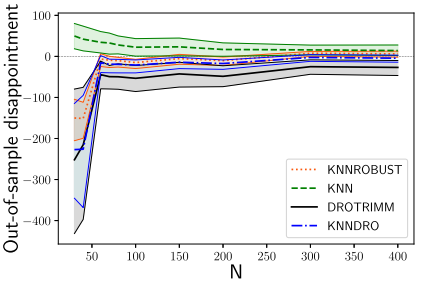
(d) $K_N = \lfloor N^{0.5} \rfloor$



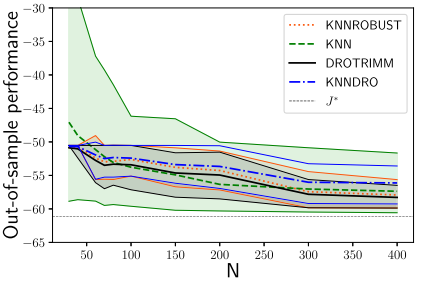
(e) $K_N = \lfloor N^{0.9} \rfloor$



(f) $K_N = \lfloor N^{0.9} \rfloor$



(g) Tuned K_N



(h) Tuned K_N

Fig. 2 Portfolio problem with features: performance metrics

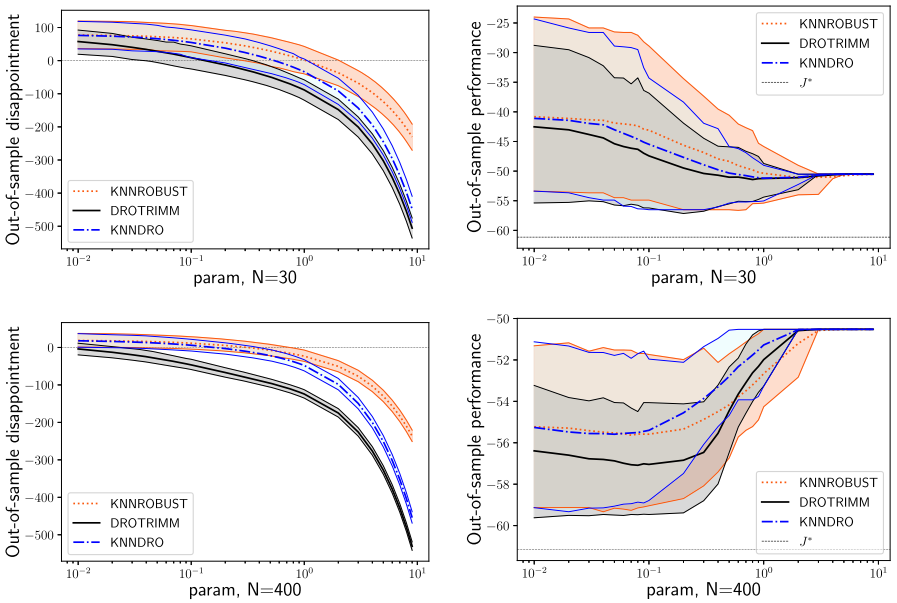


Fig. 3 Impact of the robustness parameter with 200 training samples, $K_N = \lfloor N/(\log(N+1)) \rfloor$ and $\delta = 0.5,\ \lambda = 0.1$

uncover portfolios whose out-of-sample performance features a better mean-variance trade-off, in general. Furthermore, it requires a smaller value of the robustness parameter to guarantee reliability. All this is more evident (and useful) for the case $N = 400$, as we explain next. When $N = 30$, all the considered methods need large values of their robustness parameter to ensure reliability, so they all tend to operate close to the "robust portfolio" we mentioned above. DROTRIMM can certainly afford lower values of $\Delta\widetilde{\rho}$ in an attempt to improve performance, but this proves not to be that profitable for such a small sample size, for which the robust portfolio performs very well. As N increases, the robust portfolio loses its appeal, since its performance gradually becomes comparatively worse. DROTRIMM is then able to identify portfolios that perform significantly better in expectation, while providing an estimate of their return such that the desired reliability is guaranteed. For their part, KNNDRO and KNNROBUST are also able to discover solutions with an actual average cost lower than that of the robust portfolio (albeit with a worse expectation and a higher variance than those given by DROTRIMM). However, they are more prone to overestimate their returns.

Finally, we study the behavior of the different methods under other contexts. For this, we consider several values of N, one random data sample for each N, and 200 different contexts $\mathbf{z}^*$ sampled from the marginal distributions of the features. The performance metrics (i.e., the out-of-sample disappointment and performance) are plotted in Fig. 4a, 4b, respectively, under an optimal selection of the robustness parameters (that is, for each method we use the value of the robustness parameter that, while ensuring a negative disappointment, delivers the best out-of-sample performance). We observe

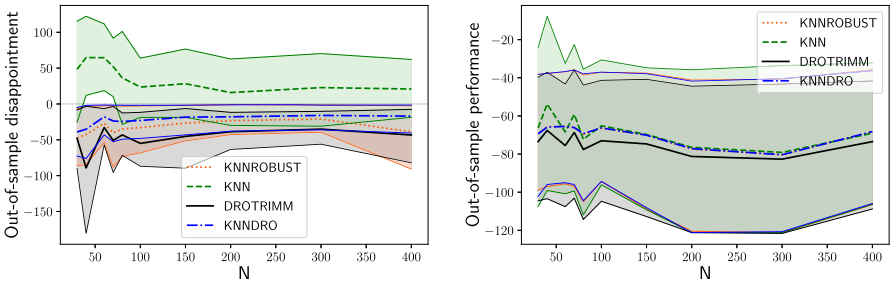


Fig. 4 Portfolio problem with features: varying context under an optimal selection of the robustness parameters, $K_N = \lfloor N/(\log(N+1)) \rfloor$ and $\delta = 0.5,\ \lambda = 0.1$

that DROTRIMM systematically performs better, with an actual cost averaged over the 200 contexts that is lower irrespective of the sample size.

5 Conclusions

In this paper, we have exploited the connection between probability trimmings and partial mass transportation to provide an easy, but powerful and novel way to extend the standard Wasserstein-metric-based DRO to the case of *conditional* stochastic programs. Our approach produces decisions that are distributionally robust against the uncertainty in the whole process of inferring the conditional probability measure of the random parameters from a finite sample coming from the true joint data-generating distribution. Through a series of numerical experiments built on the single-item newsvendor problem and a portfolio allocation problem, we have demonstrated that our method attains notably better out-of-sample performance than some existing alternatives. We have supported these empirical findings with theoretical analysis, showing that our approach enjoys attractive performance guarantees.

Supplementary Information The online version contains supplementary material available at https://doi.org/10.1007/s10107-021-01724-0.

Author Contributions AE-P: Conceptualization, Methodology, Investigation, Formal analysis, Software, Writing - original draft. JMM: Conceptualization, Methodology, Investigation, Formal analysis, Software, Writing - original draft, Supervision, Funding acquisition.

Funding Open Access funding provided by Universidad de Málaga / CBUA thanks to the CRUE-CSIC agreement with Springer Nature. This research has received funding from the European Research Council (ERC) under the European Union's Horizon 2020 research and innovation programme (Grant agreement No. 755705). This work was also supported in part by the Spanish Ministry of Science and Innovation (AEI/10.13039/501100011033) through project PID2020-115460GB-I00 and in part by the Junta de Andalucía through the research project P20_00153. Finally, the authors thankfully acknowledge the computer resources, technical expertise, and assistance provided by the SCBI (Supercomputing and Bioinformatics) center of the University of Málaga.

A Proofs of theoretical results

This appendix compiles the proofs of some of the theoretical derivations that appear in the paper. The following technical results are needed to develop these proofs.

Definition 3 (Contamination of a distribution) Given two probabilities P, Q on $\mathbb{R}^d$, we say that P is a $(1-\alpha)$-contaminated version of Q, if $P = \alpha Q + (1-\alpha)R$, where R is some probability. A $(1-\alpha)$-contamination neighbourhood of Q is the set of all $(1-\alpha)$-contaminated versions of Q and will be denoted as $\mathscr{F}_{1-\alpha}(Q)$.

Proposition 4 (Section 2.2. from [2] and p.18 in [1]) *Let P, Q be probabilities on $\mathbb{R}^d$ and $\alpha \in (0, 1]$, then*

$$Q \in \mathscr{R}_{1-\alpha}(P) \iff P = \alpha Q + (1-\alpha)R \iff P \in \mathscr{F}_{1-\alpha}(Q) \tag{29}$$

for some probability R. Moreover, if D is a probability metric such that $\mathscr{R}_{1-\alpha}(P)$ is closed for D over an appropriate set of probability distributions, then (29) is equivalent to $D(Q, \mathscr{R}_{1-\alpha}(P)) = 0$.

Remark 6 As a particular case, if we consider $D = \mathscr{W}_p$ over the set of probability distributions with finite p-th moment, $\mathscr{P}_p$, we have that, if $P, Q \in \mathscr{P}_p$, then $Q \in \mathscr{R}_{1-\alpha}(P)$ if and only if $\mathscr{W}_p(Q, \mathscr{R}_{1-\alpha}(P)) = 0$.

Corollary 1 (Corollary 3.12 from [1]) *Given two probabilities $P, Q \in \mathscr{P}_p(\mathbb{R}^d)$ and $\alpha \in (0, 1)$, there exists $P_{1-\alpha} \in \mathscr{F}_{1-\alpha}(Q)$ such that $P_{1-\alpha} = \alpha Q + (1-\alpha)R_{1-\alpha}$ for some $R_{1-\alpha} \in \mathscr{R}_\alpha(P)$ and $\mathscr{W}_p(P, P_{1-\alpha}) = \min_{R \in \mathscr{F}_{1-\alpha}(Q)} \mathscr{W}_p(P, R)$.*

Proposition 5 (Proposition 3.14 from [1]) *Take $P, Q \in \mathscr{P}_p(\mathbb{R}^d)$. If $\alpha \in (0, 1)$, then*

$$\mathscr{W}_p^p\left(P, \mathscr{F}_{1-\alpha}(Q)\right) = \alpha \mathscr{W}_p^p\left(\mathscr{R}_{1-\alpha}(P), Q\right)$$

Moreover, if $\widehat{P}_{1-\alpha} \in \mathscr{R}_{1-\alpha}(P)$ is such that $\mathscr{W}_p(\widehat{P}_{1-\alpha}, Q) = \mathscr{W}_p\left(\mathscr{R}_{1-\alpha}(P), Q\right)$, then if we construct the probability measure $\widetilde{P}_{1-\alpha} = \frac{1}{1-\alpha}\left(P - \alpha \widehat{P}_{1-\alpha}\right)$, we have that $P_{1-\alpha} := \alpha Q + (1-\alpha)\widetilde{P}_{1-\alpha} \in \mathscr{F}_{1-\alpha}(Q)$ and $\mathscr{W}_p\left(P, P_{1-\alpha}\right) = \mathscr{W}_p\left(P, \mathscr{F}_{1-\alpha}(Q)\right)$.

A.1 Proof of Lemma 1

We will prove the lemma by contradiction. Suppose there are two different probability distributions $Q_{\widetilde{\Xi}}$ and $Q'_{\widetilde{\Xi}}$ such that

$$D\left(\mathscr{R}_{1-\alpha}(Q), Q_{\widetilde{\Xi}}\right) = D(\mathscr{R}_{1-\alpha}(Q), Q'_{\widetilde{\Xi}}) = 0$$

and $Q_{\widetilde{\Xi}}(\widetilde{\Xi}) = Q'_{\widetilde{\Xi}}(\widetilde{\Xi}) = 1$.

Because $D\left(\mathscr{R}_{1-\alpha}(Q), Q_{\widetilde{\Xi}}\right) = D(\mathscr{R}_{1-\alpha}(Q), Q'_{\widetilde{\Xi}}) = 0$, we know by Proposition 4 above that $Q_{\widetilde{\Xi}}, Q'_{\widetilde{\Xi}} \in \mathscr{R}_{1-\alpha}(Q)$. Therefore, again applying Proposition 4, we have

$$Q = \alpha Q_{\widetilde{\Xi}} + (1-\alpha)R$$
$$Q = \alpha Q'_{\widetilde{\Xi}} + (1-\alpha)R'$$

for some probabilities R and R' with $R(\widetilde{\Xi}) = R'(\widetilde{\Xi}) = 0$.

Since, by hypothesis, $Q_{\widetilde{\Xi}}$ and $Q'_{\widetilde{\Xi}}$ are different, there must exist an event $A \subset \widetilde{\Xi}$ such that $Q_{\widetilde{\Xi}}(A) \neq Q'_{\widetilde{\Xi}}(A)$. We take that event and compute $Q(A)$ as follows:

$$Q(A) = \alpha Q_{\widetilde{\Xi}}(A) + (1-\alpha)R(A) = \alpha Q'_{\widetilde{\Xi}}(A) + (1-\alpha)R'(A),$$

which renders a contradiction given that $R(A) = R'(A) = 0$. □

A.2 Proof of Proposition 1

We begin by proving the first claim of Proposition 1.

We show that every feasible solution of (SP1) can be mapped into a feasible solution of (SP2) with the same objective function value. To this end, take Q as a feasible solution of (SP1) and let $Q_{\widetilde{\Xi}}$ be the Q-conditional probability measure given $\boldsymbol{\xi} \in \widetilde{\Xi}$. Take $\widehat{\mathbb{Q}}_N$ and $Q_{\widetilde{\Xi}}$ as the two probabilities in Corollary 1 with $\alpha \in (0,1)$. There exists $Q_{1-\alpha} \in \mathscr{F}_{1-\alpha}(Q_{\widetilde{\Xi}})$ such that $Q_{1-\alpha} = \alpha Q_{\widetilde{\Xi}} + (1-\alpha)\widetilde{Q}_{1-\alpha}$, with $\widetilde{Q}_{1-\alpha} \in \mathscr{R}_\alpha(\widehat{\mathbb{Q}}_N)$ and $\mathscr{W}_p(\widehat{\mathbb{Q}}_N, Q_{1-\alpha}) = \mathscr{W}_p(\widehat{\mathbb{Q}}_N, \mathscr{F}_{1-\alpha}(Q_{\widetilde{\Xi}}))$. Furthermore, it automatically follows from Proposition 5 that $\mathscr{W}_p^p(\widehat{\mathbb{Q}}_N, \mathscr{F}_{1-\alpha}(Q_{\widetilde{\Xi}})) = \alpha \mathscr{W}_p^p(\mathscr{R}_{1-\alpha}(\widehat{\mathbb{Q}}_N), Q_{\widetilde{\Xi}})$.

Since $Q \in \mathscr{F}_{1-\alpha}(Q_{\widetilde{\Xi}})$, we deduce that $\mathscr{W}_p^p(\widehat{\mathbb{Q}}_N, \mathscr{F}_{1-\alpha}(Q_{\widetilde{\Xi}})) \leq \mathscr{W}_p^p(\widehat{\mathbb{Q}}_N, Q) \leq \widetilde{\rho} \cdot \alpha$. Hence, it holds that $\mathscr{W}_p^p(\mathscr{R}_{1-\alpha}(\widehat{\mathbb{Q}}_N), Q_{\widetilde{\Xi}}) \leq \widetilde{\rho}$. In other words, $Q_{\widetilde{\Xi}}$ is feasible in (SP2). Besides, since $Q_{\widetilde{\Xi}}$ is the Q-conditional probability measure given $\boldsymbol{\xi} \in \widetilde{\Xi}$, we have that $\mathbb{E}_Q\left[f(\mathbf{x}, \boldsymbol{\xi}) \mid \boldsymbol{\xi} \in \widetilde{\Xi}\right] = \frac{1}{\alpha}\mathbb{E}_Q\left[f(\mathbf{x}, \boldsymbol{\xi})\mathbb{I}_{\widetilde{\Xi}}(\boldsymbol{\xi})\right] = \mathbb{E}_{Q_{\widetilde{\Xi}}}\left[f(\mathbf{x}, \boldsymbol{\xi})\right]$ a.s.

Next we prove the second claim of the proposition. To this end, first we show that, if $\widehat{\mathbb{Q}}_N(\widetilde{\Xi}) = 0$, then every feasible solution of (SP2) can also be mapped into a feasible solution of (SP1) with the same objective function value. To this end, take $Q_{\widetilde{\Xi}}$ feasible in (SP2) and consider $\widehat{Q}_{1-\alpha} \in \mathscr{R}_{1-\alpha}(\widehat{\mathbb{Q}}_N)$ such that $\mathscr{W}_p(\widehat{Q}_{1-\alpha}, Q_{\widetilde{\Xi}}) = \mathscr{W}_p(\mathscr{R}_{1-\alpha}(\widehat{\mathbb{Q}}_N), Q_{\widetilde{\Xi}})$. Fix $\widetilde{Q}_{1-\alpha} = \frac{1}{1-\alpha}(\widehat{\mathbb{Q}}_N - \alpha\widehat{Q}_{1-\alpha})$. By Proposition 5, we have

$$Q_{1-\alpha} = \alpha Q_{\widetilde{\Xi}} + (1-\alpha)\widetilde{Q}_{1-\alpha} = \alpha Q_{\widetilde{\Xi}} + \widehat{\mathbb{Q}}_N - \alpha\widehat{Q}_{1-\alpha} \in \mathscr{F}_{1-\alpha}(Q_{\widetilde{\Xi}})$$

Hence, $Q_{1-\alpha}(\widetilde{\Xi}) = \alpha$, because $\widehat{\mathbb{Q}}_N(\widetilde{\Xi})$ gives zero measure to $\widetilde{\Xi}$ and so does any of its $(1-\alpha)$-trimmings. Besides, we have that

$$\mathscr{W}_p^p(\widehat{\mathbb{Q}}_N, Q_{1-\alpha}) = \mathscr{W}_p^p(\widehat{\mathbb{Q}}_N, \mathscr{F}_{1-\alpha}(Q_{\widetilde{\Xi}})) = \alpha\mathscr{W}_p^p(\mathscr{R}_{1-\alpha}(\widehat{\mathbb{Q}}_N), Q_{\widetilde{\Xi}}) \leq \alpha\widetilde{\rho}.$$

Therefore, $Q_{1-\alpha}$ is feasible in (SP1) and $Q_{\widetilde{\Xi}}$ is the $Q_{1-\alpha}$-conditional probability measure given $\boldsymbol{\xi} \in \widetilde{\Xi}$.

Finally, if $\alpha = 1$, then $\mathscr{R}_{1-\alpha}(\widehat{\mathbb{Q}}_N) = \widehat{\mathbb{Q}}_N$, $\mathbb{E}_Q\left[f(\mathbf{x},\boldsymbol{\xi}) \mid \boldsymbol{\xi} \in \widetilde{\Xi}\right] = \mathbb{E}_Q\left[f(\mathbf{x},\boldsymbol{\xi})\right]$ and the mapping is direct, namely, $Q = Q_{\widetilde{\Xi}}$. □

A.3 Proof of Theorem 1

Thanks to Lemma 2, the subproblem (SP2) can be written equivalently as follows:

$$\begin{aligned}(\text{SP2}) \quad & \sup_{Q_{\widetilde{\Xi}};\ \mathbf{b}\in\Delta(\alpha_N)} \ \mathbb{E}_{Q_{\widetilde{\Xi}}}\left[f(\mathbf{x},\boldsymbol{\xi})\right] \\ & \text{s.t. } Q_{\widetilde{\Xi}}(\widetilde{\Xi}) = 1 \\ & \quad \mathscr{W}_p\left(\sum_{i=1}^N b_i \delta_{\widehat{\boldsymbol{\xi}}_i},\, Q_{\widetilde{\Xi}}\right) \le \widetilde{\rho}^{1/p}\end{aligned}$$

where $\Delta(\alpha_N)$ stands for the set of constraints $\{0 \le b_i \le \frac{1}{N\alpha_N},\ \forall i \le N,\ \sum_{i=1}^N b_i = 1\}$.

Problem (SP2) can be, in turn, reformulated as

$$\left\{\begin{array}{ll} \sup\limits_{Q_{\widetilde{\Xi}};\ \Pi;\ \mathbf{b}\in\Delta(\alpha_N)} & \displaystyle\int_{\widetilde{\Xi}} f(\mathbf{x},(\mathbf{z},\mathbf{y}))\, Q_{\widetilde{\Xi}}(d\mathbf{z},d\mathbf{y}) \\ \text{s.t.} & \\ & \displaystyle\int_{\widetilde{\Xi}} Q_{\widetilde{\Xi}}(d\mathbf{z},d\mathbf{y}) = 1 \\ & \displaystyle\left(\int_{\widetilde{\Xi}\times\Xi} \left\|(\mathbf{z},\mathbf{y}) - (\mathbf{z},\mathbf{y})'\right\|^p \Pi(d(\mathbf{z},\mathbf{y}), d(\mathbf{z},\mathbf{y})')\right)^{1/p} \le \widetilde{\rho}^{1/p} \\ & \left\{\begin{array}{l}\Pi \text{ is a joint distribution of } (\mathbf{z},\mathbf{y}) \text{ and } (\mathbf{z},\mathbf{y})' \\ \text{with marginals } Q_{\widetilde{\Xi}} \text{ and } \sum_{i=1}^N b_i \delta_{\widehat{\boldsymbol{\xi}}_i}, \text{ respectively}\end{array}\right. \end{array}\right. \tag{30}$$

$$= \left\{\begin{array}{ll} \sup\limits_{Q^i_{\widetilde{\Xi}};\ \mathbf{b}\in\Delta(\alpha_N)} & \displaystyle\sum_{i=1}^N b_i \int_{\widetilde{\Xi}} f(\mathbf{x},(\mathbf{z},\mathbf{y}))\, Q^i_{\widetilde{\Xi}}(d\mathbf{z},d\mathbf{y}) \\ \text{s.t.} & \displaystyle\int_{\widetilde{\Xi}} Q^i_{\widetilde{\Xi}}(d\mathbf{z},d\mathbf{y}) = 1,\ \forall i \le N \\ & \displaystyle\sum_{i=1}^N b_i \int_{\widetilde{\Xi}} \|(\mathbf{z},\mathbf{y}) - (\widehat{\mathbf{z}}_i,\widehat{\mathbf{y}}_i)\|^p\, Q^i_{\widetilde{\Xi}}(d\mathbf{z},d\mathbf{y}) \le \widetilde{\rho} \end{array}\right. \tag{31}$$

where reformulation (31) follows from the fact that the marginal distribution of $(\mathbf{z},\mathbf{y})'$ is the discrete distribution supported on points $(\widehat{\mathbf{z}}_i,\widehat{\mathbf{y}}_i)$, with probability masses b_i, $i = 1,\ldots,N$. Thus, Π is completely determined by the conditional distributions $Q^i_{\widetilde{\Xi}}$ of $(\mathbf{z},\mathbf{y})$ given $(\mathbf{z},\mathbf{y})' = (\widehat{\mathbf{z}}_i,\widehat{\mathbf{y}}_i)$, $i = 1,\ldots,N$, that is,

$$\Pi(d(\mathbf{z},\mathbf{y}), d(\mathbf{z},\mathbf{y})') = \sum_{i=1}^N b_i \delta_{(\widehat{\mathbf{z}}_i,\widehat{\mathbf{y}}_i)}(d(\mathbf{z},\mathbf{y})')\, Q^i_{\widetilde{\Xi}}(d(\mathbf{z},\mathbf{y}))$$

Now we split up the supremum into two:

$$\sup_{\mathbf{b}\in\Delta(\alpha_N)} \sup_{Q^i_{\widetilde{\Xi}},\forall i\le N} \sum_{i=1}^{N} b_i \int_{\widetilde{\Xi}} f(\mathbf{x},(\mathbf{z},\mathbf{y}))\, Q^i_{\widetilde{\Xi}}(d\mathbf{z},d\mathbf{y}) \tag{32a}$$

$$\text{s.t} \int_{\widetilde{\Xi}} Q^i_{\widetilde{\Xi}}(d\mathbf{z},d\mathbf{y}) = 1,\ \ \forall i \le N \tag{32b}$$

$$\sum_{i=1}^{N} b_i \int_{\widetilde{\Xi}} \|(\mathbf{z},\mathbf{y}) - (\widehat{\mathbf{z}}_i,\widehat{\mathbf{y}}_i)\|^p\, Q^i_{\widetilde{\Xi}}(d\mathbf{z},d\mathbf{y}) \le \widetilde{\rho} \tag{32c}$$

If we set λ as the dual variable of constraint (32c), then using standard duality arguments, we can equivalently rewrite the inner supremun as

$$\sup_{\mathbf{b}\in\Delta(\alpha_N)} \inf_{\lambda\ge 0} \sup_{Q^i_{\widetilde{\Xi}},\forall i\le N} \lambda\widetilde{\rho} + \sum_{i=1}^{N} b_i \int_{\widetilde{\Xi}} \left(f(\mathbf{x},(\mathbf{z},\mathbf{y})) - \lambda\,\|(\mathbf{z},\mathbf{y}) - (\widehat{\mathbf{z}}_i,\widehat{\mathbf{y}}_i)\|^p\right) Q^i_{\widetilde{\Xi}}(d\mathbf{z},d\mathbf{y}) \tag{33}$$

$$\text{s.t} \int_{\widetilde{\Xi}} Q^i_{\widetilde{\Xi}}(d\mathbf{z},d\mathbf{y}) = 1,\ \ \forall i \le N \tag{34}$$

$$= \sup_{\mathbf{b}\in\Delta(\alpha_N)} \inf_{\lambda\ge 0} \lambda\widetilde{\rho} + \sum_{i=1}^{N} b_i \sup_{(\mathbf{z},\mathbf{y})\in\widetilde{\Xi}} \left(f(\mathbf{x},(\mathbf{z},\mathbf{y})) - \lambda\,\|(\mathbf{z},\mathbf{y}) - (\widehat{\mathbf{z}}_i,\widehat{\mathbf{y}}_i)\|^p\right) \tag{35}$$

$$= \inf_{\lambda\ge 0} \sup_{\mathbf{b}\in\Delta(\alpha_N)} \lambda\widetilde{\rho} + \sum_{i=1}^{N} b_i \sup_{(\mathbf{z},\mathbf{y})\in\widetilde{\Xi}} \left(f(\mathbf{x},(\mathbf{z},\mathbf{y})) - \lambda\,\|(\mathbf{z},\mathbf{y}) - (\widehat{\mathbf{z}}_i,\widehat{\mathbf{y}}_i)\|^p\right) \tag{36}$$

$$= \inf_{\lambda\ge 0;\overline{\mu}_i,\forall i\le N;\theta\in\mathbb{R}} \lambda\widetilde{\rho} + \theta + \frac{1}{N\alpha}\sum_{i=1}^{N}\overline{\mu}_i \tag{37}$$

$$\text{s.t.}\ \overline{\mu}_i + \theta \ge \sup_{(\mathbf{z},\mathbf{y})\in\widetilde{\Xi}} \left(f(\mathbf{x},(\mathbf{z},\mathbf{y})) - \lambda\,\|(\mathbf{z},\mathbf{y}) - (\widehat{\mathbf{z}}_i,\widehat{\mathbf{y}}_i)\|^p\right),\ \ \forall i \le N \tag{38}$$

$$\overline{\mu}_i \ge 0,\ \ \forall i \le N \tag{39}$$

where we have swapped the supremum and the infimum in (35) by appealing to Sion's min-max theorem [44], given that the objective function in (35) is linear in the $b_i, i = 1,\ldots,N$, over a compact convex set, and a positively weighted sum of convex functions in λ. □

Remark 7 (Limiting case $\alpha = 0$) If $\alpha = 0$, $\mathscr{R}_1(\widehat{\mathbb{Q}}_N) = \{\sum_{i=1}^{N} b_i \delta_{\widehat{\boldsymbol{\xi}}_i}$ such that $b_i \ge 0$, $\forall i = 1,\ldots,N$, and $\sum_{i=1}^{N} b_i = 1\}$. Therefore, dual variables $\overline{\mu}_i, \forall i \le N$, do not appear in (37)–(39) in this case. Similarly, if $\frac{1}{N\alpha} \ge 1$, the constraints $b_i \le \frac{1}{N\alpha}, \forall i \le N$, become redundant and hence we can set $\overline{\mu}_i = 0, \forall i \le N$.

References

1. Agulló Antolín, M.: Trimming methods for model validation and supervised classification in the presence of contamination. Ph.D. thesis (2018). http://uvadoc.uva.es/handle/10324/31682
2. Álvarez-Esteban, del Barrio, E., Cuesta-Albertos, J.A., Matrán, C.: Similarity of samples and trimming. Bernoulli **18**(2), 606–634 (2012)
3. Álvarez-Esteban, P.C., del Barrio, E., Cuesta-Albertos, J.A., Matrán, C.: Uniqueness and approximate computation of optimal incomplete transportation plans. Ann. Inst. Henri Poincare-Probab. Stat. **47**(2), 358–375 (2011)
4. Balghiti, O.E., Elmachtoub, A.N., Grigas, P., Tewari, A.: Generalization bounds in the predict-then-optimize framework (2019). arXiv:1905.11488
5. Ban, G.Y., Rudin, C.: The big data newsvendor: Practical insights from machine learning. Oper. Res. **67**(1), 90–108 (2019). https://doi.org/10.1287/opre.2018.1757
6. del Barrio, E., Matrán, C.: Rates of convergence for partial mass problems. Probab. Theory Relat. Field **155**(3–4), 521–542 (2013)
7. Bertsimas, D., Kallus, N.: From predictive to prescriptive analytics. Manag. Sci. **66**(3), 1025–1044 (2020)
8. Bertsimas, D., McCord, C.: Optimization over continuous and multi-dimensional decisions with observational data (2018). arXiv:1807.04183
9. Bertsimas, D., McCord, C., Sturt, B.: Dynamic optimization with side information (2019). arXiv:1907.07307
10. Bertsimas, D., Shtern, S., Sturt, B.: A data-driven approach for multi-stage linear optimization (2018). http://www.optimization-online.org/DB_HTML/2018/11/6907.html
11. Bertsimas, D., Shtern, S., Sturt, B.: Technical note-Two-stage sample robust optimization. Oper. Res. (2021). https://doi.org/10.1287/opre.2020.2096
12. Bertsimas, D., Van Parys, B.: Bootstrap robust prescriptive analytics (2017). arXiv:1711.09974
13. Biau, G., Devroye, L.: Lectures on the Nearest Neighbor Method. Springer Series in the Data Sciences. Springer (2015)
14. Chen, R.: Distributionally robust learning under the Wasserstein metric. Ph.D. thesis (2019). https://open.bu.edu/handle/2144/38236
15. Chen, Z., Sim, M., Xiong, P.: Robust stochastic optimization made easy with rsome. Manag. Sci. **66**(8), 3329–3339 (2020)
16. Diao, S., Sen, S.: Distribution-free algorithms for learning enabled predictive stochastic programming (2020). http://www.optimization-online.org/DB_HTML/2020/03/7661.html
17. Donti, P., Amos, B., Kolter, J.Z.: Task-based end-to-end model learning in stochastic optimization. Adv. Neural Inf. Process Syst. **1**, 5484–5494 (2017)
18. Elmachtoub, A.N., Grigas, P.: Smart "predict, then optimize." Manage. Sci. (2021). https://doi.org/10.1287/mnsc.2020.3922
19. Esteban-Pérez, A., Morales, J.M.: Distributionally robust stochastic programs with side information based on trimmings—Extended version (2020). arXiv:2009.10592
20. Esteban-Pérez, A., Morales, J.M.: Distributionally robust stochastic programs with side information based on trimmings - Codes and Data. GitHub repository (2021). https://github.com/groupoasys/DRO_CONDITIONAL_TRIMMINGS
21. Falk, M., Hüsler, J., Reiss, R.D.: Laws of small numbers: extremes and rare events. Springer (2010)
22. Farokhi, F.: Why does regularization help with mitigating poisoning attacks? Neural Process. Lett. (2021). https://doi.org/10.1007/s11063-021-10539-1
23. Fournier, N., Guillin, A.: On the rate of convergence in wasserstein distance of the empirical measure. Probab. Theory Relat. Field **162**(3), 707–738 (2015)
24. Gao, R.: Finite-sample guarantees for Wasserstein distributionally robust optimization: Breaking the curse of dimensionality (2020). arXiv:2009.04382
25. Gao, R., Kleywegt, A.J.: Distributionally Robust Stochastic Optimization with Wasserstein Distance (2016). arXiv:1604.02199
26. Gibbs, A.L., Su, F.E.: On choosing and bounding probability metrics. Int. Stat. Rev. **70**(3), 419–435 (2002)
27. Gray, R.M.: Probability, Random Processes, and Ergodic Properties. Springer (2009). https://doi.org/10.1007/978-1-4419-1090-5

28. Hanasusanto, G.A., Kuhn, D.: Robust data-driven dynamic programming. Adv. Neural Inf. Process. Syst. **26**, 827–835 (2013)
29. Hannah, L., Powell, W., Blei, D.: Nonparametric density estimation for stochastic optimization with an observable state variable. Adv. Neural Inf. Process Syst. **23**, 820–828 (2010)
30. Huber, J., Müller, S., Fleischmann, M., Stuckenschmidt, H.: A data-driven newsvendor problem: from data to decision. Eur. J. Oper. Res. **278**(3), 904–915 (2019)
31. Kannan, R., Bayraksan, G., Luedtke, J.: Heteroscedasticity-aware residuals-based contextual stochastic optimization. arXiv preprint arXiv:2101.03139 (2021)
32. Kannan, R., Bayraksan, G., Luedtke, J.R.: Data-driven sample average approximation with covariate information. Optimization. Online. http://www.optimization-online.org/DB_HTML/2020/07/7932.html (2020)
33. Kannan, R., Bayraksan, G., Luedtke, J.R.: Residuals-based distributionally robust optimization with covariate information (2020). arXiv:2012.01088
34. Kuhn, D., Esfahani, P.M., Nguyen, V.A., Shafieezadeh-Abadeh, S.: Wasserstein distributionally robust optimization: Theory and applications in machine learning. In: Operations Research & Management Science in the Age of Analytics, pp. 130–166. INFORMS (2019). https://doi.org/10.1287/educ.2019.0198
35. Mohajerin Esfahani, P., Kuhn, D.: Data-driven distributionally robust optimization using the Wasserstein metric: performance guarantees and tractable reformulations. Math. Program. **171**(1–2), 115–166 (2018)
36. Muñoz, M.A., Pineda, S., Morales, J.M.: A bilevel framework for decision-making under uncertainty with contextual information (2020). arXiv:2008.01500
37. Nguyen, V.A., Zhang, F., Blanchet, J., Delage, E., Ye, Y.: Distributionally robust local non-parametric conditional estimation (2020). arXiv:2010.05373
38. Nguyen, V.A., Zhang, F., Blanchet, J., Delage, E., Ye, Y.: Robustifying conditional portfolio decisions via optimal transport (2021). arXiv:2103.16451
39. Pang Ho, C., Hanasusanto, G.A.: On data-driven prescriptive analytics with side information: A regularized Nadaraya–Watson approach (2019). http://www.optimization-online.org/DB_HTML/2019/01/7043.html
40. Rahimian, H., Mehrotra, S.: Distributionally robust optimization: a review (2019). arXiv:1908.05659
41. Rockafellar, R.T., Uryasev, S.: Optimization of conditional value-at-risk. J. Risk **2**, 21–41 (2000)
42. Santambrogio, F.: Optimal transport for applied mathematicians (2015). https://doi.org/10.1007/978-3-319-20828-2
43. Sen, S., Deng, Y.: Learning enabled optimization: Towards a fusion of statistical learning and stochastic programming (2018). http://www.optimization-online.org/DB_HTML/2017/03/5904.html
44. Sion, M.: On general minimax theorems. Pac. J. Math. **1103040253** (1958)
45. Villani, C.: Topics in Optimal Transportation, Graduate Studies in Mathematics, vol. 58. American Mathematical Society (2003)