Decomposition Methods for Solving Two-Stage Distributionally Robust Optimization Problems

Yannan Chen[1]
Department of Applied Mathematics, The Hong Kong Polytechnic University, Hong Kong
(yannan.chen@polyu.edu.hk)

Hailin Sun[2]
Jiangsu Key Laboratory for NSLSCS, School of Mathematical Sciences, Nanjing Normal University, Nanjing, 210023, China
Institute of Mathematics, Nanjing Normal University, Nanjing, 210023, China
(mathhlsun@163.com)

Huifu Xu
School of Mathematical Sciences, University of Southampton, SO17 1BJ, Southampton, UK
(H.Xu@soton.ac.uk)

December 11, 2018

Abstract. Decomposition methods have been well studied for solving two-stage and multi-stage stochastic programming problems, see [30, 33, 34]. In this paper, we propose an algorithmic framework based on the fundamental ideas of the methods for solving two-stage minimax distributionally robust optimization (DRO) problems where the underlying random variables take a finite number of distinct values. This is achieved by introducing nonanticipativity constraint for the first stage decision variables, rearranging the minimax problem through Lagrange decomposition and applying the well-known primal-dual hybrid gradient (PDHG) method to the new minimax problem. The algorithmic framework does not depend on specific structure of the ambiguity set. To extend the algorithm to the case that the underlying random variables are continuously distributed, we propose a discretization scheme and quantify the error arising from the discretization in terms of the optimal value and the optimal solutions when the ambiguity set is constructed through generalized prior moment conditions, the Kantorovich ball and ϕ-divergence centred at an empirical probability distribution. Some preliminary numerical tests show the proposed decomposition algorithm featured with parallel computing performs well.

Key Words. Distributionally robust optimization, decomposition method, moment conditions, Kantorovich ball, discrete approximation, parallel computing

[1]This author's work is supported in part by National Natural Science Foundation of China #11571178, #11771405.

[2]Corresponding author. This author's work is supported in part by National Natural Science Foundation of China #11871276, #11571056.

1 Introduction

We consider the following two-stage distributionally robust optimization problem:

$$\begin{array}{ll} \min\limits_{x\in\mathbb{R}^n} \sup\limits_{P\in\mathcal{P}} & \mathbb{E}_P[f_1(x,\xi)+v(x,\xi)] \\ \text{s.t.} & x\in X, \end{array} \tag{1.1}$$

where f_1 is a continuous function mapping from $\mathbb{R}^n\times\mathbb{R}^l$ to $\mathbb{R}$, x is a decision vector restricted to taking values from a convex and compact set $X\subset\mathbb{R}^n$, $\xi:\Omega\to\mathbb{R}^l$ is a random vector defined in the probability space $(\Omega,\mathcal{F},\mathbb{P})$ with support set Ξ, $v(x,\xi)$ is the optimal value function of the second stage problem

$$\begin{array}{ll} \min\limits_{y} & g(x,y,\xi) \\ \text{s.t.} & h(x,y,\xi)\le 0, \end{array} \tag{1.2}$$

where g and h are continuous functions mapping from $\mathbb{R}^n\times\mathbb{R}^m\times\mathbb{R}^l$ to $\mathbb{R}$ and $\mathbb{R}^s$ respectively, $\mathcal{P}$ is a set of probability distributions/measures on Ξ induced by ξ. In the case when $\mathcal{P}$ reduces to a singleton $P^*=\mathbb{P}\circ\xi^{-1}$, (1.1) reduces to an ordinary two-stage stochastic programming problem. We call P^* the true probability distribution. Our focus here is on the case that the true probability distribution is unknown, but it is possible to use empirical data, computer simulation or subjective judgement to construct a set $\mathcal{P}$ which constitutes or approximates P^*. In the literature of DRO, $\mathcal{P}$ is called ambiguity set reflecting ambiguity of the true probability distribution. DRO model is initiated by Scarf [37] and popularized by many others particularly over the past two decades for its wide applications in finance, economics and management sciences, we refer readers to [3, 6, 10, 12, 27, 35, 38, 44, 45, 46] and references therein for various ambiguity sets and DRO models.

For each fixed $(x,\xi)\in X\times\Xi$, let $Y(x,\xi)$ be the feasible set of (1.2). Under some moderate conditions, problem (1.1)-(1.2) can be equivalently written as

$$\begin{array}{ll} \min\limits_{x\in X, y(\cdot)\in\mathcal{Y}} \sup\limits_{P\in\mathcal{P}} & \mathbb{E}_P[f_1(x,\xi)+g(x,y(\xi),\xi)] \\ \text{s.t.} & x\in X, y(\xi)\in Y(x,\xi), \text{for P-a.e. } \xi\in\Xi, \end{array} \tag{1.3}$$

where $\mathcal{Y}$ is the space of measurable functions from Ξ to $\mathbb{R}^m$, see for instance [36, Chapter 1, section 2.4]. When the underlying functions in the second stage problem are piecewise linear or quadratic w.r.t. ξ and the ambiguity set takes some specific structure, the two-stage DRO can be reformulated as a tractable problem, see [2, 16, 19, 41]. It remains unclear whether similar reformulations can be obtained in general cases. Here we tackle the problem from a different angle, that is, by using the well-known decomposition methods.

Progressive hedging method (PHM) is one of the most well-known approaches as such. The fundamental idea of the decomposition method is to divide the solution process into two steps: 1. Find an admissible solution at each scenario. 2. Revise the admissible solution to an implementable solution in the nonanticipativity subspace via some proximal projection. One of the main advantages of the method is that step 1 can be implemented simultaneously through

parallel computations. We refer readers to [30, 31, 33, 34] for a comprehensive description and discussion of the method.

In this paper, we aim to extend the decomposition approach of PHM to the two-stage DRO problem (1.3). Since (1.3) is a minimax problem whereas PHM is proposed to solve minimization problems, we need to imbed PHM with some features which can handle the minimax structure. Primal-dual hybrid gradient (PDHG) method turns out to be the one which comes to our mind for the purpose. PDHG is proposed by Zhu and Chan [50] for solving saddle point problems such as

$$\min_{x\in X}\max_{y\in Y}\theta_1(x) - y^TAx - \theta_2(y), \tag{1.4}$$

where $A \in \mathbb{R}^{m\times n}$, $X \subset \mathbb{R}^n$ and $Y \subset \mathbb{R}^m$ are closed convex sets, $\theta_1 : \mathbb{R}^n \to \mathbb{R}$ and $\theta_2 : \mathbb{R}^m \to \mathbb{R}$ are convex functions. A number of variational image restoration problems with the total variation (TV) regularization can be reformulated as special cases of (1.4) and many papers focus on improving this approach and applying it to image restoration problems, to name a few, see [5, 7, 11, 18, 43, 47]. Recently, Liu et al. [21] apply the PDHG method to a one-stage DRO which provides a new iterative scheme for solving DRO. In this paper, we take a step further by combining the PDHG method with the decomposition feature of PHM to solve our two-stage DRO (1.3). The main contributions of this paper are summarized as follows.

- For the two-stage DRO problem (1.3) where the underlying random variables take a finite number of distinct values, we introduce nonanticipativity constraints for the first-stage decision vector, rearrange the minimax problem through Lagrange decomposition and apply the PDHG method to the new minimax problem. One of the main advantages of the alternative iterative algorithmic scheme is that it enables us to implement parallel computation at each iteration and this is indeed the main motivation of this paper.

- To extend the numerical approach to a two-stage DRO with continuous true probability distribution, we propose a discretization scheme and derive error bounds for the discretized problem in terms of the optimal values and the optimal solutions where the ambiguity sets are constructed through moment conditions, Kantorovich ball and ϕ-divergence centred at an empirical probability distribution.

- To examine the accuracy and efficiency of the algorithm, we test it with a distributionally robust multi-product newsvendor problem and a distributionally robust production planning problem with three different forms of the ambiguity set. Preliminary results obtained by parallel computing show that the new algorithm performs very well, in particular, by incorporating parallel computing, it may significantly increase the efficiency when the ambiguity set is constructed via moment conditions and ϕ-divergence.

The rest of the paper is organized as follows. Section 2 proposes the decomposition algorithm for solving two-stage DRO problem (1.3) when the support set of the random variables contains a finite number of points. Section 3 introduces the discrete approximation of two-stage DRO when the true probability distribution is continuous. Section 4 reports some preliminary numerical test results and shows the correctness and effectiveness of the algorithm.

2 A decomposition algorithmic scheme

In this section, we introduce a new decomposition algorithm for solving the two-stage DRO (1.3). To simplify the discussion, let us consider the case that for almost every fixed $\xi \in \Xi$ $g(x, y, \xi)$ is convex in (x, y) and each component of $h(x, y, \xi)$ is also convex in (x, y). This will effectively make the second stage problem convex and its optimal value function $v(x, \xi)$ is also convex in x for almost every ξ. By further assuming f_1 is convex in x for almost every ξ, we will have $\sup_{P\in\mathcal{P}} \mathbb{E}_P[f_1(x, \xi) + v(x, \xi)]$ being convex in x.

Throughout this section, we assume that ξ follows a discrete distribution with N scenarios and the true probability distribution is unknown, we will come back to the case when ξ is continuously distributed in the next section. Consequently we use $\mathcal{P}_N$ rather than $\mathcal{P}$ to denote the ambiguity set. This will also facilitate us to use $\mathcal{P}_N$ as a discrete approximation of $\mathcal{P}$ in Section 3 when the true unknown probability distribution is continuous.

For each $P \in \mathcal{P}_N$, we write $P(\xi_i) = p_i$ for $i = 1, \ldots, N$. Let $f(x, y, \xi) := f_1(x, \xi) + g(x, y(\xi), \xi)$. Consequently, we can rewrite (1.3) as

$$\text{(DRO-N)} \quad \min_{x,\boldsymbol{y}} \sup_{P\in\mathcal{P}_N} \sum_{i=1}^{N} p_i f(x, y_i, \xi_i) \quad \text{s.t.} \quad x \in X, y_i \in \mathcal{Y}(x, \xi_i),\ i = 1, \ldots, N, \tag{2.1}$$

where $\boldsymbol{y} := (y_1^T, \ldots, y_N^T)^T$. By exploiting the convexity of g and h in y, we can exchange the min and max operations w.r.t. p and y and consequently reformulate (2.1) as

$$\begin{aligned} \min_{\boldsymbol{x},\boldsymbol{y}} \max_{P\in\mathcal{P}_N} \quad & \sum_{i=1}^{N} p_i f(x_i, y_i, \xi_i) \\ \text{s.t.} \quad & x_i \in X, y_i \in \mathcal{Y}(x_i, \xi_i),\ i = 1, \ldots, N, \\ & x_i = x_{i+1}, \quad i = 1, \ldots, N. \end{aligned} \tag{2.2}$$

Here $x_{N+1} := x_1$ and we relax the first stage decision vector x to make it scenario dependent and then apply the nonanticipativity condition. Problem (2.2) is a minimax optimization problem and its solution is a saddle point of function $\sum_{i=1}^{N} p_i f(x_i, y_i, \xi_i)$ with variable P on one hand and (x, y) on the other over the space of $\mathcal{P}$ and $X \times \mathcal{Y}(X, \Xi)$. Through epigraphical reformulation, we can write (2.2) as

$$\begin{aligned} \min_{\boldsymbol{x},\boldsymbol{y},\boldsymbol{\sigma}} \max_{P\in\mathcal{P}_N} \quad & \sum_{i=1}^{N} p_i \sigma_i \\ \text{s.t.} \quad & f(x_i, y_i, \xi_i) \le \sigma_i,\ i = 1, \ldots, N, \\ & x_i \in X, y_i \in \mathcal{Y}(x_i, \xi_i),\ i = 1, \ldots, N, \\ & x_i = x_{i+1}, \quad i = 1, \ldots, N. \end{aligned} \tag{2.3}$$

To ease the exposition, let $v(\boldsymbol{x}, \boldsymbol{y}, \boldsymbol{\sigma}) := \max_{P\in\mathcal{P}_N} \sum_{i=1}^{N} p_i \sigma_i$ and

$$Z_N := \left\{ (\boldsymbol{x}, \boldsymbol{y}, \boldsymbol{\sigma}) : \begin{array}{l} f(x_i, y_i, \xi_i) \le \sigma_i, \\ x_i \in X, y_i \in \mathcal{Y}(x_i, \xi_i), \end{array} \quad i = 1, \ldots, N \right\}.$$

We can then rewrite problem (2.3) in a concise form

$$\begin{array}{ll} \min\limits_{\boldsymbol{x},\boldsymbol{y},\boldsymbol{\sigma}} & v(\boldsymbol{x},\boldsymbol{y},\boldsymbol{\sigma}) \\ \text{s.t.} & (\boldsymbol{x},\boldsymbol{y},\boldsymbol{\sigma}) \in Z_N, \\ & x_i = x_{i+1}, \;\; i = 1,\ldots,N. \end{array} \tag{2.4}$$

By introducing the partial Lagrange function of (2.4) w.r.t. the equality constraints

$$L(\boldsymbol{x},\boldsymbol{y},\boldsymbol{\sigma},\boldsymbol{w}) := v(\boldsymbol{x},\boldsymbol{y},\boldsymbol{\sigma}) + \frac{1}{2}\sum_{i=1}^{N} w_i^T(x_i - x_{i+1}),$$

we can reformulate (2.4) as

$$\min_{(\boldsymbol{x},\boldsymbol{y},\boldsymbol{\sigma})\in Z_N} \max_{\boldsymbol{w}} \; v(\boldsymbol{x},\boldsymbol{y},\boldsymbol{\sigma}) + \frac{1}{2}\sum_{i=1}^{N} w_i^T(x_i - x_{i+1}),$$

or

$$\min_{(\boldsymbol{x},\boldsymbol{y},\boldsymbol{\sigma})\in Z_N} \max_{P\in\mathcal{P}_N,\boldsymbol{w}} \; \sum_{i=1}^{N}\left(p_i\sigma_i + \frac{1}{2}w_i^T(x_i - x_{i+1})\right), \tag{2.5}$$

where $\boldsymbol{w} = (w_1^T,\ldots,w_N^T)^T$ and $w_i \in \mathbb{R}^n$ for $i = 1,\ldots,N$.

Problem (2.5) is a saddle-point problem and we can apply PDHG method to solve it. To this end, let

$$A := \begin{pmatrix} 0_{N\times nN} & 0_{N\times mN} & I_N \\ H_N \otimes I_n & 0_{nN\times mN} & 0_{nN\times N} \end{pmatrix},$$

where

$$H_N = \frac{1}{2}\begin{pmatrix} 1 & -1 & & \\ & \ddots & \ddots & \\ & & 1 & -1 \\ -1 & & & 1 \end{pmatrix} \in \mathbb{R}^{N\times N},$$

and $\otimes$ stands for the Kronecker product of matrices. Clearly, the matrix A is sparse and satisfies $\|A^TA\| = 1$. Consequently, the objective function of problem (2.5) can be written in a vector matrix product form

$$(P^T, \boldsymbol{w}^T)A\begin{pmatrix} \boldsymbol{x} \\ \boldsymbol{y} \\ \boldsymbol{\sigma} \end{pmatrix}$$

and we propose the following algorithmic procedures.

Algorithm 1 is a standard application of PDHG procedures to our problem (2.5), therefore its convergence is covered by existing convergence results for general PDHG, see i.e. [18]. The complexity of the algorithm is $O(1/\epsilon)$, see [17]. One of the main advantages of Algorithm 1 in

Algorithm 1 Decomposition method for solving two stage DRO.

1: Choose parameters $\rho \in [1,2]$, $\tau, \eta > 0$ such that $\tau\eta < \frac{1}{\|A^T A\|}$, and initial points $(\boldsymbol{x}^0, \boldsymbol{y}^0, \boldsymbol{\sigma}^0) \in Z_N$, $\boldsymbol{w}^0$, $P^0 \in \mathcal{P}_N$. Set $k \leftarrow 0$.

2: **while** not converged **do**

3:
$$\begin{pmatrix} \bar{\boldsymbol{x}}^k \\ \bar{\boldsymbol{y}}^k \\ \bar{\boldsymbol{\sigma}}^k \end{pmatrix} := \begin{pmatrix} \boldsymbol{x}^k \\ \boldsymbol{y}^k \\ \boldsymbol{\sigma}^k \end{pmatrix} - \tau A^T \begin{pmatrix} P^k \\ \boldsymbol{w}^k \end{pmatrix}.$$

4: Solve a primal subproblem (Ps)
$$\begin{pmatrix} \widetilde{\boldsymbol{x}}^k \\ \widetilde{\boldsymbol{y}}^k \\ \widetilde{\boldsymbol{\sigma}}^k \end{pmatrix} = \arg\min_{(\boldsymbol{x},\boldsymbol{y},\boldsymbol{\sigma}) \in Z_N} \frac{1}{2\tau} \left\| \begin{pmatrix} \boldsymbol{x} \\ \boldsymbol{y} \\ \boldsymbol{\sigma} \end{pmatrix} - \begin{pmatrix} \bar{\boldsymbol{x}}^k \\ \bar{\boldsymbol{y}}^k \\ \bar{\boldsymbol{\sigma}}^k \end{pmatrix} \right\|_2^2.$$

5: Set
$$\begin{pmatrix} \bar{P}^k \\ \bar{\boldsymbol{w}}^k \end{pmatrix} := \begin{pmatrix} P^k \\ \boldsymbol{w}^k \end{pmatrix} + \eta A \left(2 \begin{pmatrix} \widetilde{\boldsymbol{x}}^k \\ \widetilde{\boldsymbol{y}}^k \\ \widetilde{\boldsymbol{\sigma}}^k \end{pmatrix} - \begin{pmatrix} \boldsymbol{x}^k \\ \boldsymbol{y}^k \\ \boldsymbol{\sigma}^k \end{pmatrix} \right).$$

6: Solve a dual subproblem (Ds)
$$\begin{pmatrix} \widetilde{P}^k \\ \widetilde{\boldsymbol{w}}^k \end{pmatrix} = \arg\min_{P \in \mathcal{P}_N, \boldsymbol{w}} \frac{1}{2\eta} \left\| \begin{pmatrix} P \\ \boldsymbol{w} \end{pmatrix} - \begin{pmatrix} \bar{P}^k \\ \bar{\boldsymbol{w}}^k \end{pmatrix} \right\|_2^2.$$

7: Let
$$\begin{pmatrix} \boldsymbol{x}^{k+1} \\ \boldsymbol{y}^{k+1} \\ \boldsymbol{\sigma}^{k+1} \\ P^{k+1} \\ \boldsymbol{w}^{k+1} \end{pmatrix} := (1-\rho) \begin{pmatrix} \boldsymbol{x}^k \\ \boldsymbol{y}^k \\ \boldsymbol{\sigma}^k \\ P^k \\ \boldsymbol{w}^k \end{pmatrix} + \rho \begin{pmatrix} \widetilde{\boldsymbol{x}}^k \\ \widetilde{\boldsymbol{y}}^k \\ \widetilde{\boldsymbol{\sigma}}^k \\ \widetilde{P}^k \\ \widetilde{\boldsymbol{w}}^k \end{pmatrix}.$$

8: $k \leftarrow k+1$.

9: **end while**

this context is that Step 4 allows us to solve the quadratic minimization problem per scenario simultaneously

$$\begin{aligned} \min_{x_i,y_i,\sigma_i} \quad & \frac{1}{2\tau}\|(x_i,y_i,\sigma_i)-(\bar{x}_i^k,\bar{y}_i^k,\bar{\sigma}_i^k)\|_2^2 \\ \text{s.t.} \quad & f(x_i,y_i,\xi_i)\leq \sigma_i, \\ & x_i\in X, y_i\in\mathcal{Y}(x_i,\xi_i), \end{aligned}$$

for $i=1,\ldots,N$ and hence effectively decompose the quadratic minimization problem into N-optimization problems with reduced size. Moreover, in Step 6, we also have decomposition structure $\widetilde{\boldsymbol{w}}^k=\bar{\boldsymbol{w}}^k$ and $\widetilde{P}^k=\arg\min_{P\in\mathcal{P}_N} \quad \|P-\bar{P}^k\|_2^2$.

3 Discrete approximation of the two-stage DRO

Algorithm 1 is derived under the condition that the true unknown probability distribution is discrete with a finite support set, and so is every probability distribution in the ambiguity set. In practice however the true probability distribution is often continuous particularly in finance. This requires us to discretize the random variable before Algorithm 1 is applied. The discretization in turn requires us to quantify the difference between the discretized DRO and the original DRO in terms of the optimal value and optimal solutions, which provides theoretical grounding for the usefulness of statistical estimators obtained from solving the discretized DRO. This section aims to address this. To this end, we need to recall some preliminary concepts, definitions and results on metrics of probability measures.

3.1 Preliminaries on metrics of probability measures

Let Ω be a sample space set and $\mathcal{F}$ be the associated sigma algebra. Let $\mathscr{P}(\Omega)$ be the set of all probability measures over the measurable space $(\Omega,\mathcal{F})$. We consider a vector-valued measurable function ξ mapping from Ω to $\Xi\subset\mathbb{R}^k$. Let $\mathscr{P}(\Xi)$ denoted the set of all probability measures over Ξ induced by ξ.

Let $\mathscr{G}$ be a family of real-valued bounded measurable functions defined on Ξ. For each pair of probability measures $P,Q\in\mathscr{P}(\Xi)$, let

$$\mathsf{dl}_{\mathscr{G}}(P,Q):=\sup_{g\in\mathscr{G}}\left|\mathbb{E}_P[g(\xi)]-\mathbb{E}_Q[g(\xi)]\right|.$$

$\mathsf{dl}_{\mathscr{G}}(P,Q)$ satisfies all properties of a metric except that $\mathsf{dl}_{\mathscr{G}}(P,Q)=0$ does not necessarily imply $P=Q$ unless the set $\mathscr{G}$ is sufficiently large. In the literature of stochastic programming, $\mathsf{dl}_{\mathscr{G}}(P,Q)$ is known as a pseudo-metric of ζ-structure or ζ-metric, see an excellent review paper by Römisch [32].

The pseudo-metric covers a wide range of metrics in probability theory including the total variation metric, Kantorovich metric, bounded Lipschitz metric and some other metrics; see

[9, 25, 28, 29, 32, 49] and references therein. Specifically, if

$$\mathscr{G} := \left\{ g : \Xi \to \mathbb{R} |\ g \text{ is } \mathscr{B} \text{ measurable, } \sup_{\xi \in \Xi} |g(\xi)| \leq 1 \right\},$$

then $\mathsf{dl}_{\mathscr{G}}(P, Q)$ reduces to the *total variation metric*, denoted by dl_{TV}. Note also that $\mathsf{dl}_{TV} = 2 \sup_{B \in \mathcal{B}} |P(B) - Q(B)|$[3]. If g is restricted further to Lipschitz continuous functions with modulus bounded by 1, i.e.,

$$\mathscr{G} = \left\{ g : \sup_{\xi \in \Xi} |g(\xi)| \leq 1,\ g \text{ is Lipschitz continuous with Lipschtiz constant } L_1(g) \leq 1 \right\}, \quad (3.1)$$

where $L_1(g) := \sup\{|g(u) - g(v)|/d(u, v) :\ u \neq v\}$, then the resulting metric is known as *bounded Lipschitz metric*, denoted by dl_{BL}. If the boundedness of g is lifted in (3.1), that is,

$$\mathscr{G} = \{ g : g \text{ is Lipschitz continuous and Lipschtiz modulus } L_1(g) \leq 1 \},$$

then we arrive at Kantorovich metric, denoted by dl_K.

It is well known that $\mathsf{dl}_{TV}(P, Q) \in [0, 2]$ and when Ξ is bounded $\mathsf{dl}_K(P, Q) \in [0, \mathrm{diam}(\Xi)]$, see [9]. Moreover, it follows by [48, Lemmas 1–4] and [9] that $\mathsf{dl}_K(P, Q) \leq \mathrm{diam}(\Xi) \mathsf{dl}_{TV}(P, Q)$. Note also that

$$\mathsf{dl}_K(P, Q) \leq \mathrm{diam}(\Xi) \mathsf{dl}_{TV}(P, Q). \quad (3.2)$$

Based on the ζ-metric, we can define the distance from a point to a set, deviation from one set to another and the Hausdorff distance between two sets in the space of probability measures $\mathscr{P}(\Xi)$. Specifically, for $\mathcal{C}, \mathcal{C}' \in \mathscr{P}(\Xi)$, let

$$\mathsf{dl}_{\mathscr{G}}(Q, \mathcal{C}) := \inf_{P \in \mathcal{C}} \mathsf{dl}_{\mathscr{G}}(Q, P), \quad \mathbb{D}(\mathcal{C}', \mathcal{C}; \mathsf{dl}_{\mathscr{G}}) := \sup_{Q \in \mathcal{C}'} \mathsf{dl}_{\mathscr{G}}(Q, \mathcal{C})$$

and

$$\mathbb{H}(\mathcal{C}', \mathcal{C}; \mathsf{dl}_{\mathscr{G}}) := \max \left\{ \mathbb{D}(\mathcal{C}', \mathcal{C}; \mathsf{dl}_{\mathscr{G}}),\ \mathbb{D}(\mathcal{C}, \mathcal{C}'; \mathsf{dl}_{\mathscr{G}}) \right\}.$$

Here $\mathbb{H}(\mathcal{C}', \mathcal{C}; \mathsf{dl}_{\mathscr{G}})$ is the Hausdorff distance between $\mathcal{C}'$ and $\mathcal{C}$ under the ζ-metric $\mathsf{dl}_{\mathcal{G}}$ in the space of $\mathscr{P}(\Xi)$.

Let $\{P_N\} \subset \mathscr{P}(\Xi)$ be a sequence of probability measure. Recall that $\{P_N\}$ is said to converge to $P \in \mathscr{P}(\Xi)$ *weakly* if $\lim_{N \to \infty} \mathbb{E}_{P_N}[h(\xi)] = \mathbb{E}_P[h(\xi)]$ for each bounded and continuous function $h : \Xi \to \mathbb{R}$. It is well known that the Kantorovich metric metrizes the weak convergence. In the case when Ξ is bounded and finite, the total variation metric also metrizes the weak convergence, see [9, Theorem 6]. For a set of probability measures $\mathcal{A} \subset \mathscr{P}(\Xi)$, $\mathcal{A}$ is said to be weakly compact if every sequence $\{A_N\} \subset \mathcal{A}$ contains a subsequence $\{A_{N'}\}$ and $A \in \mathcal{A}$ such that $A_{N'}$ converges to A weakly, see [1, Section 9].

[3] In some literatures, total variation metric is defined as $\mathsf{dl}_{TV} = \sup_{B \in \mathscr{B}} |P(B) - Q(B)|$, see [9].

3.2 Moment conditions

We start by looking into the case when the ambiguity set is constructed through moment conditions:

$$\mathcal{P} := \left\{ P \in \mathscr{P}(\Xi) : \begin{array}{ll} \mathbb{E}_P[\Psi_i(\xi)] = \mu_i, & \text{for } i = 1, \ldots, p \\ \mathbb{E}_P[\Psi_i(\xi)] \preceq \mu_i, & \text{for } i = p+1, \ldots, q \end{array} \right\}, \tag{3.3}$$

where $\Psi_i : \Xi \to \mathcal{S}^{n_i \times n_i}$ is a continuous mapping, $\mathcal{S}^{n_i \times n_i}$ denotes the space of $n_i \times n_i$ symmetric matrices, and $\mu_i \in \mathcal{S}^{n_i \times n_i}$ represents the mean value or an upper bound of the mean value of Ψ_i. In the case that $n_i = 1$, (3.8) reduces to classical generalized moment conditions. Here we write $A \preceq B$ when $A - B$ is negative semidefinite for any two matrices A, B. To ease the notation, we may write the functions of the moment system in vector-matrix form:

$$\mathcal{P} := \{ P \in \mathscr{P}(\Xi) : \mathbb{E}_P[\Psi_E(\xi)] = \mu_E, \mathbb{E}_P[\Psi_I(\xi)] \preceq \mu_I \},$$

where $\Psi_E(\xi) := (\Psi_1(\xi), \cdots, \Psi_p(\xi))$, $\Psi_I(\xi) := (\Psi_{p+1}(\xi), \cdots, \Psi_q(\xi))$, $\mu_E := (\mu_1, \cdots, \mu_p)$ and $\mu_I := (\mu_{p+1}, \cdots, \mu_q)$.

Let $\Xi_N := \{\xi_1, \cdots, \xi_N\} \subset \Xi$ be a discrete subset of Ξ and $\mathscr{P}(\Xi_N)$ the set of all probability measures defined on Ξ_N. Here $\xi_1, \cdots, \xi_N$ may be i.i.d. samples of ξ or selected in a deterministic manner. We leave this unspecified at this point. Let

$$\mathcal{P}_N := \{ P \in \mathscr{P}(\Xi_N) : \mathbb{E}_P[\Psi_E(\xi)] = \mu_E, \mathbb{E}_P[\Psi_I(\xi)] \preceq \mu_I \}. \tag{3.4}$$

Our idea is to replace $\mathcal{P}$ with $\mathcal{P}_N$ in the DRO model (1.3) so that Algorithm 1 is applicable to the (DRO-N) after the replacement. In what follows, we investigate the difference between $\mathcal{P}$ and $\mathcal{P}_N$ under some appropriate metric $\mathsf{dl}_{\mathscr{G}}$. Since $\mathcal{P}_N \subset \mathcal{P}$, the excess of $\mathcal{P}_N$ over $\mathcal{P}$, $\mathbb{D}(\mathcal{P}_N, \mathcal{P}; \mathsf{dl}_{\mathscr{G}})$ is zero. Our interest here is to quantify the difference of the two ambiguity sets under Hausdorff distance. To this end, we need to derive an intermediate technical result which characterizes an error bound for the moment system. Let $\mathcal{S}_+^{n \times n}$ denote the cone of $n \times n$ positive semidefinite matrices. We need the following assumption.

Assumption 3.1 (Slater type condition (STC)) *Let $\bar{\Xi}$ be a subset of Ξ and $\mathcal{B}$ the unit ball in the space of $\mathbb{R} \times \mathcal{S}_+^{n_1 \times n_1} \times \cdots \times \mathcal{S}_+^{n_q \times n_q}$, let $\mathscr{M}_+(\bar{\Xi})$ denote the space of positive measures generated by $\mathscr{P}(\bar{\Xi})$. There exists a positive number ϵ_0 such that*

$$\epsilon_0 \mathcal{B} \subset \text{int} \left[(\langle P, 1 \rangle - 1, \langle P, \Psi(\xi) \rangle - \mu) - \{0\} \times \{0_p\} \times \mathcal{K}_-^{q-p} : P \in \mathscr{M}_+(\bar{\Xi}) \right], \tag{3.5}$$

where $\Psi(\xi) := (\Psi_1(\xi), \cdots, \Psi_q(\xi))$ and $\langle P, \Psi(\xi) \rangle := \int_\Xi \Psi(\xi) P(d\xi)$ with the integration taken componentwise, $\{0_p\}$ denotes the Cartesian product of zero matrices in respective matrix spaces of $\mathcal{S}^{n_i \times n_i}$ corresponding to Φ_i for $i = 1, \ldots, p$, and $\mathcal{K}_-^{q-p} := \mathcal{S}_-^{n_{p+1} \times n_{p+1}} \times \cdots \times \mathcal{S}_-^{n_q \times n_q}$; see [39, condition (3.12)] for general moment problems.

Assumption 3.1 implies that

$$\epsilon_0 \mathcal{B} \subset \text{int} \left[(\langle P, 1 \rangle - 1, \langle P, \Psi(\xi) \rangle - \mu) - \{0\} \times \{0_p\} \times \mathcal{K}_-^{q-p} : P \in \mathscr{M}_+(\tilde{\Xi}) \right] \tag{3.6}$$

for any $\tilde{\Xi}$ with $\bar{\Xi} \subseteq \tilde{\Xi} \subseteq \Xi$. In particular, it implies the standard STC

$$\epsilon_0 \mathcal{B} \subset \text{int } \left[(\langle P, 1\rangle - 1, \langle P, \Psi(\xi)\rangle - \mu) - \{0\} \times \{0_p\} \times \mathcal{K}_-^{q-p} : P \in \mathscr{M}_+(\Xi) \right]. \tag{3.7}$$

The condition allows us to establish error bound for all moment problems of the following form

$$\tilde{\mathcal{P}} := \left\{ P \in \mathscr{P}(\tilde{\Xi}) : \mathbb{E}_P[\Psi_E(\xi)] = \mu_E, \mathbb{E}_P[\Psi_I(\xi)] \preceq \mu_I \right\} \tag{3.8}$$

so long as $\bar{\Xi} \subseteq \tilde{\Xi}$. Note that the condition depends on the choice of $\bar{\Xi}$. We may use a simple example to explain this.

Example 3.1 Let $\Xi := [-1, 1]$, $\Xi_1 := \{0, 0.5, 1\}$ and $\Xi_2 := \{-0.5, 0, 0.5\}$. Let

$$\mathcal{P}_1 := \{P \in \mathscr{P}(\Xi_1) : \mathbb{E}_P[\xi] = 0, \mathbb{E}_P[\xi_2] \leq 1\}$$

and

$$\mathcal{P}_2 := \{P \in \mathscr{P}(\Xi_2) : \mathbb{E}_P[\xi] = 0, \mathbb{E}_P[\xi_2] \leq 1\}.$$

It is easy to observe that the STC does not hold for the moment system in $\mathcal{P}_1$ in that for every $P \in \mathscr{M}_+(\Xi_1)$, $\langle P, \xi\rangle \geq 0$ and hence 0_3 does not lie in the interior of the set:

$$\left[\left(\langle P, 1\rangle - 1, \langle P, \Psi(\xi)\rangle - (0, 1)^T \right) - \{0\} \times \{0\} \times \mathbb{R}_- : P \in \mathscr{M}_+(\Xi_1) \right].$$

In contrast, the moment system in $\mathcal{P}_2$ satisfies the STC. To see this, consider any point $(a_1, a_2, a_3)^T \in \epsilon_0 \mathcal{B}$ with sufficiently small ϵ_0 and positive measures from $\mathscr{M}_+(\Xi_2)$ such that

$$P_1 := \begin{cases} b_1, & \text{for } \xi = -0.5, \\ b_2, & \text{for } \xi = 0.5, \\ b_3, & \text{for } \xi = 0, \end{cases} \qquad \text{where } b_1, b_2, b_3 \geq 0.$$

Then

$$\begin{aligned} &\left[\left(\langle P_1, 1\rangle - 1, \langle P_1, \Psi(\xi)\rangle - (0, 1)^T \right) - \{0\} \times \{0\} \times \mathbb{R}_- \right] \\ = \quad &\left(\sum_{i=1}^{3} b_i - 1, 0.5(b_2 - b_1), [0.25(b_1 + b_2) - 1, +\infty) \right) \end{aligned}$$

and the system $b_1 + b_2 + b_3 - 1 = a_1, 0.5(b_2 - b_1) = a_2, 0.25(b_1 + b_2) - 1 \leq a_3$ always has a solution which satisfies

$$\begin{cases} b_1 = b_2 - 2a_2 \leq \min\{2 + 2a_3 - a_2, 0.5 + a_2 + 0.5a_1\}, \\ b_2 \leq \min\{2 - a_2 + 3a_3, 0.5 + a_2 + 0.5a_1\}, \\ b_3 = 1 + 2a_2 + a_1 - 2b_2, \\ b_1, b_2, b_3 \geq 0, \end{cases}$$

e.g. by setting $b_1 = 0.5 + 0.5a_1 - a_2$, $b_2 = 0.5 + 0.5a_1 + a_2$ and $b_3 = 0$. Note that $b_1 \geq 0, b_2 \geq 0$ since ϵ_0 can be set sufficiently small. This implies that for any $(a_1, a_2, a_3)^T \in \epsilon_0 \mathcal{B}$, we can find a positive measure $P \in \mathscr{M}_+(\Xi_2)$ such that

$$(a_1, a_2, a_3)^T \in \left[\left(\langle P, 1\rangle - 1, \langle P, \Psi(\xi)\rangle - (0, 1)^T \right) - \{0\} \times \{0\} \times \mathbb{R}_- \right]$$

and then the origin $(0,0,0)$ lies in the range

$$\left[\left(\langle P,1\rangle - 1, \langle P, \Psi(\xi)\rangle - (0,1)^T\right) - \{0\}\times\{0\}\times \mathbb{R}_- : P \in \mathscr{M}_+(\Xi_2)\right],$$

and STC holds.

Lemma 3.1 (Hoffman's lemma for the moment problem) *Consider the ambiguity set defined as in (3.8). Suppose that Assumption 3.1 holds and $\mathcal{P}$ is weakly compact. For any $\tilde{\Xi}$ such that $\bar{\Xi} \subset \tilde{\Xi} \subset \Xi$. Then there exists a positive constant C_1 depending on Ψ (independent on $\tilde{\Xi}$) such that for any $Q \in \mathscr{P}(\tilde{\Xi})$,*

$$\mathsf{dl}_{TV}(Q, \tilde{\mathcal{P}}) \le C_1 \left(\|(\mathbb{E}_Q[\Psi_I(\xi)] - \mu_I)_+\| + \|\mathbb{E}_Q[\Psi_E(\xi)] - \mu_E\| \right), \tag{3.9}$$

where $C_1 \ge \frac{2}{\epsilon_0} + 1$ and dl_{TV} denotes the total variation metric.

The result is a modified version of [42, Lemma 2]. The key difference is that Lemma 3.1 provides a unified error bound for every ambiguity set $\mathcal{P}$ whose support set $\tilde{\Xi}$ contains $\bar{\Xi}$ and gives an explicit estimate for the parameter C_1. We attach a proof in the appendix.

Let $\Xi^N := \{\xi_1, \ldots, \xi_N\}$ be a discrete approximation of Ξ and

$$\beta_N := \max_{\xi\in\Xi} \min_{1\le i\le N} d(\xi, \xi_i). \tag{3.10}$$

Since $\Xi^N \subset \Xi$, it is easy to see that β_N is indeed the Hausdorff distance between Ξ and Ξ^N. Let $\{\Xi_1, \ldots, \Xi_N\}$ be a Voronoi tessellation of Ξ, that is,

$$\Xi_i \subseteq \left\{ y \in \Xi : \|y - \xi_i\| = \min_k \|y - \xi_k\| \right\} \quad \text{for } \ i = 1, \ldots, N \tag{3.11}$$

are pairwise disjoint subsets forming a partition of Ξ. For a fixed $P \in \mathscr{P}(\Xi)$, let $p_i = P(\Xi_i)$ for $i = 1, \ldots, N$ and define

$$P_N(\cdot) := \sum_{i=1}^N p_i \, \delta_{\xi_i}(\cdot). \tag{3.12}$$

We call P_N *VT-projection* of P on $\mathscr{P}(\Xi^N)$. The following result provides an upper bound for the discrete approximation of P by P_N.

Proposition 3.1 (cf. [26, Lemma 4.9]) *Let $P \in \mathscr{P}(\Xi)$ be fixed, P_N be defined as in* (3.12) *and β_N be defined in* (3.10)*. Then*

$$\mathsf{dl}_K(P, P_N) = \int \min_{1\le i\le N} d(\xi, \xi_i) dP = \sum_{i=1}^N \int_{\Xi_i} d(\xi, \xi_i) dP \le \beta_N, \tag{3.13}$$

where dl_K denotes the Kantorovich distance of two probability measures.

Lemma 3.2 (cf. [42, Lemma 3]) *Let $A, B \in S^{n\times n}$ and $A \succeq 0$. The following assertion hold.*

(i) $\mathrm{tr}(AB) \le \mathrm{tr}(AB_+)$, *where "*tr*" denotes the trace of a matrix, and for a symmetric matrix* $M \in \mathcal{S}^{n\times n}$ *with spectral decomposition* $Q\mathrm{diag}\{\iota_1,\ldots,\iota_n\}Q^T$,

$$M_+ := Q\mathrm{diag}\left\{\max\{\iota_1,0\},\ldots,\max\{\iota_n,0\}\right\}Q^T.$$

(ii) $\|(A+B)_+\|_F \le \|A_+\|_F + \|B_+\|_F$, *where* $\|\cdot\|_F$ *denotes the Frobenius norm.*

With Proposition 3.1, Lemma 3.1, Lemma 3.2 and the relationship between total variation metric and Kantorovich metric (see (3.2)), we are ready to present the main result in this section which quantifies the approximation of $\mathcal{P}$ by

$$\mathcal{P}_N := \left\{P \in \mathscr{P}(\Xi^N) : \mathbb{E}_P[\Psi_E(\xi)] = \mu_E, \mathbb{E}_P[\Psi_I(\xi)] \preceq \mu_I\right\} \tag{3.14}$$

under the Kantorovich metric.

Theorem 3.1 (Quantification of discrete approximation of the ambiguity set) *Suppose: (a) Assumption 3.1 holds with* $\bar{\Xi} \subset \Xi^N \subset \Xi$, *(b)* Ξ *is bounded and* β_N *tends to zero as* $N \to \infty$ *and (c) each component of* $\Psi(\cdot)$ *is Lipschitz continuous on* Ξ *with Lipschitz modulus* L. *Then for* N *sufficiently large*

$$\mathbb{H}(\mathcal{P}_N, \mathcal{P}; \mathsf{dl}_K) \le (1 + L\rho\mathrm{diam}(\Xi)C_1)\,\beta_N, \tag{3.15}$$

where $\mathrm{diam}(\Xi)$ *denotes the diameter of* Ξ, $\rho = \|\mathbf{E}\|$, *and* $\mathbf{E}$ *denotes a matrix of size* $\Psi(\xi)$ *with each component being* 1.

Proof. Since $\Xi_N \subset \Xi$, by (3.14), we have $\mathcal{P}_N \subset \mathcal{P}$ which implies $\mathbb{D}(\mathcal{P}_N, \mathcal{P}; \mathsf{dl}_K) = 0$. In what follows, we show (3.15) holds under $\mathbb{D}(\cdot,\cdot;\mathsf{dl}_K)$ which will in turn secure (3.15).

For any $\tilde{Q} \in \mathcal{P}$, by the Voronoi Tessellation (3.12), we can construct $\tilde{Q}_N \in \mathscr{P}(\Xi_N)$ such that $\mathsf{dl}_K(\tilde{Q}, \tilde{Q}_N) \le \beta_N$. If $\tilde{Q}_N \in \mathcal{P}_N$, then by (3.13), $\mathsf{dl}_K(\tilde{Q}, \mathcal{P}_N) \le \mathsf{dl}_K(\tilde{Q}, \tilde{Q}_N) \le \beta_N$. So we only need to consider the case that $\tilde{Q}_N \notin \mathcal{P}_N$ in order to complete the proof. By triangle inequality,

$$\mathsf{dl}_K(\tilde{Q}, \mathcal{P}_N) \le \mathsf{dl}_K(\tilde{Q}, \tilde{Q}_N) + \mathsf{dl}_K(\tilde{Q}_N, \mathcal{P}_N). \tag{3.16}$$

Moreover, by condition (a) and Lemma 3.1,

$$\begin{aligned}
\mathsf{dl}_{TV}(\tilde{Q}_N, \mathcal{P}_N) &\le C_1\left(\|(\mathbb{E}_{\tilde{Q}_N}[\Psi_I(\xi)] - \mu_I)_+\| + \|\mathbb{E}_{\tilde{Q}_N}[\Psi_E(\xi)] - \mu_E\|\right)\\
&\le C_1\left(\|(\mathbb{E}_{\tilde{Q}_N}[\Psi_I(\xi)] - \mathbb{E}_{\tilde{Q}}[\Psi_I(\xi)])_+\| + \|\mathbb{E}_{\tilde{Q}_N}[\Psi_E(\xi)] - \mathbb{E}_{\tilde{Q}}[\Psi_I(\xi)]\|\right)\\
&\le C_1\left(\|(\mathbb{E}_{\tilde{Q}_N}[\Psi_I(\xi)] - \mathbb{E}_{\tilde{Q}}[\Psi_I(\xi)])\| + \|\mathbb{E}_{\tilde{Q}_N}[\Psi_E(\xi)] - \mathbb{E}_{\tilde{Q}}[\Psi_I(\xi)]\|\right),
\end{aligned}$$

where the second inequality is due to the fact that $\tilde{Q} \in \mathcal{P}$ whereas the last inequality follows from the definition of $(A)_+$ when A is a vector or a matrix. Since Ξ is bounded under condition (b), it follows by (3.2) that the inequality above implies

$$\mathsf{dl}_K(\tilde{Q}_N, \mathcal{P}_N) \le \mathrm{diam}(\Xi)C_1\left(\|(\mathbb{E}_{\tilde{Q}_N}[\Psi_I(\xi)] - \mathbb{E}_{\tilde{Q}}[\Psi_I(\xi)])\| + \|\mathbb{E}_{\tilde{Q}_N}[\Psi_E(\xi)] - \mathbb{E}_{\tilde{Q}}[\Psi_I(\xi)]\|\right),$$

where $\mathrm{diam}(S)$ denotes the diameter of S. Moreover, under condition (c), every component Ψ_{ij} of Ψ is Lipschitz continuous on Ξ with modulus being bounded by L. By the definition of Kantorovich metric, we have

$$\|(\mathbb{E}_{\tilde{Q}_N}[\Psi_{ij}(\xi)] - \mathbb{E}_{\tilde{Q}}[\Psi_{ij}(\xi)])\|| \leq L\mathsf{dl}_K(\tilde{Q}, \tilde{Q}_N)$$

and hence

$$\|(\mathbb{E}_{\tilde{Q}_N}[\Psi_I(\xi)] - \mathbb{E}_{\tilde{Q}}[\Psi_I(\xi)])\| + \|\mathbb{E}_{\tilde{Q}_N}[\Psi_E(\xi)] - \mathbb{E}_{\tilde{Q}}[\Psi_I(\xi)]\| \leq L\rho\mathsf{dl}_K(\tilde{Q}, \tilde{Q}_N). \tag{3.17}$$

Combining (3.16)-(3.17) and (3.13), we arrive at

$$\mathsf{dl}_K(\tilde{Q}, \mathcal{P}_N) \leq (1 + L\rho\mathrm{diam}(\Xi)C_1)\beta_N,$$

which implies (3.15) holds under $\mathbb{D}(\cdot, \cdot; \mathsf{dl}_K)$. ■

The proof of Theorem 3.1 is similar to that of [20, Theorem 12] and [22, Theorem 3.2]. There are some subtle differences. (a) The Hoffman's Lemma in [20] (see [20, Theorem 12]) is established under Slater condition rather than Slater type condition where the former is not satisfied by moment systems with equality constraints. (b) [22, Theorem 3.2] focuses on the case that Ψ is a vector-valued function (each component is a real-valued function) and consequently it derives an inequality similar to (3.17) without any Slater or Slater type condition. However, in our case, we need Slater type condition as Ψ may contain some matrix-components and subsequently the discretized moment conditions constitute some semi-definite constraints. (c) Analogous to [22, Theorem 3.2], we derive the error bound (3.15) under the total variation metric and then use the relationship between the TV-metric and the Kantorovich metric (3.2) under the boundedness of Ξ, this allows the error bound to be applicable to a wider class of moment problems. Note that the boundedness condition is also needed in [20, Theorem 12] which means our result is stronger under the Kantorovich metric.

3.3 Kantorovich ball

We now turn to discuss the case when the information on the probability distribution is acquired through samples. In some practical data-driven problems such as mobile ad hoc networks and cyber security, sample size is often small. Under these circumstances, it might be more appropriate to construct an ambiguity set than a single empirical distribution to approximate the true probability distribution. Here we follow the approach of Esfahani and Kuhn [14], Gao and Kleywegt [8] and Pichler and Xu [28] to construct a Kantorovich ball centred at the empirical distribution with certain radius:

$$B(P_N, r; \mathsf{dl}_K) := \{P' \in \mathscr{P}(\Xi), \mathsf{dl}_K(P', P_N) \leq r\}, \tag{3.18}$$

where P_N is a discrete distribution constructed through samples. It can be an empirical distribution or some distribution constructed through optimal quantization method.

The main issue here is that $B(P_N, r; \mathsf{dl}_K)$ is a ball in the space of probability measures $\mathscr{P}(\Xi)$ which contains both discrete and continuous probability measures and hence when we substitute

it into the DRO model (1.3), we will not be able to apply Algorithm 1 to solve it. This motivates us to consider a Kantorovich ball in the space of discrete probability measures $\mathscr{P}(\Xi_N)$:

$$B_N(P_N, r; \mathsf{dl}_K) := \{P' \in \mathscr{P}(\Xi_N) : \mathsf{dl}_K(P', P_N) \leq r\}. \tag{3.19}$$

While $B_N(P_N, r; \mathsf{dl}_K)$ fulfills our requirement for Algorithm 1, it loses some continuous distributions which might affect the robustness of our DRO model (1.3). To address this concern, we establish a technical result which quantifies the difference between $B(P_N, r; \mathsf{dl}_K)$ and $B_N(P_N, r; \mathsf{dl}_K)$ under the Kantorovich metric.

Theorem 3.2 (Quantitative stability of discrete approximation of Kantorovich ball) *Let $B(P, r_1; \mathsf{dl}_K)$ and $B_N(P_N, r_2; \mathsf{dl}_K)$ be defined as in* (3.18) *and* (3.19) *respectively. Then*

$$\mathbb{H}(B(P, r_1; \mathsf{dl}_K), B_N(P_N, r_2; \mathsf{dl}_K); \mathsf{dl}_K) \leq \mathsf{dl}_K(P, P_N; \mathsf{dl}_K) + 2\beta_N + |r_1 - r_2|, \tag{3.20}$$

where β_N is defined as in (3.10).

Proof. By the triangle inequality of the Hausdorff distance and [28, Theorem 1],

$$\begin{aligned}
&\mathbb{H}(B(P, r_1; \mathsf{dl}_K), B_N(P_N, r_2; \mathsf{dl}_K); \mathsf{dl}_K) \\
&\leq \mathbb{H}(B(P, r_1; \mathsf{dl}_K), B(P_N, r_2; \mathsf{dl}_K); \mathsf{dl}_K) + \mathbb{H}(B(P_N, r_2; \mathsf{dl}_K), B_N(P_N, r_2; \mathsf{dl}_K); \mathsf{dl}_K) \\
&\leq \mathsf{dl}_K(P, P_N) + |r_1 - r_2| + \mathbb{H}(B(P_N, r_2; \mathsf{dl}_K), B_N(P_N, r_2; \mathsf{dl}_K); \mathsf{dl}_K).
\end{aligned}$$

In what follows, we show that

$$\mathbb{H}(B(P_N, r_2; \mathsf{dl}_K), B_N(P_N, r_2; \mathsf{dl}_K); \mathsf{dl}_K) \leq 2\beta_N.$$

Since $B_N(P_N, r_2; \mathsf{dl}_K) \subseteq B(P_N, r_2; \mathsf{dl}_K)$, it suffices to show

$$\mathbb{D}(B(P_N, r_2; \mathsf{dl}_K), B_N(P_N, r_2; \mathsf{dl}_K); \mathsf{dl}_K) \leq 2\beta_N. \tag{3.21}$$

Let $Q^* \in B(P_N, r_2; \mathsf{dl}_K)$ such that

$$\mathsf{dl}_K(Q^*, B_N(P_N, r_2; \mathsf{dl}_K)) = \mathbb{D}(B(P_N, r_2; \mathsf{dl}_K), B_N(P_N, r_2; \mathsf{dl}_K)).$$

Existence of Q^* is due to the continuity of $\mathsf{dl}_K(Q, B_N(P_N, r_2; \mathsf{dl}_K)$ in Q and weak compactness of $B(P_N, r_2; \mathsf{dl}_K)$. By the Voronoi Tessellation (3.12), we can construct $Q_N \in \mathscr{P}(\Xi_N)$ such that $\mathsf{dl}_K(Q^*, Q_N) \leq \beta_N$. If $Q_N \in B_N(P_N, r_2; \mathsf{dl}_K)$, then

$$\mathbb{D}(B(P, r_2; \mathsf{dl}_K), B_N(P_N, r_2; \mathsf{dl}_K); \mathsf{dl}_K) \leq \mathsf{dl}(Q^*, Q_N) \leq \beta_N$$

and hence (3.21). So we are left with the case when $Q_N \notin B_N(P_N, r_2; \mathsf{dl}_K)$. Let $t = \frac{r_2}{\mathsf{dl}_K(Q_N, P_N)}$ and $\tilde{Q}_N = tQ_N + (1-t)P_N$. Then $\tilde{Q}_N \in \mathscr{P}(\Xi_N)$ and

$$\mathsf{dl}_K(\tilde{Q}_N, P_N) = t\mathsf{dl}_K(Q_N, P_N) = r_2,$$

which implies $\tilde{Q}_N \in B_N(P_N, r_2; \mathsf{dl}_K)$. Moreover, since

$$\mathsf{dl}_K(Q_N, P_N) \leq \mathsf{dl}_K(Q_N, Q^*) + \mathsf{dl}_K(Q^*, P_N) \leq \beta_N + r_2,$$

we have

$$\mathsf{dl}_K(Q_N, \tilde{Q}_N) = (1-t)\mathsf{dl}_K(Q_N, P_N) = \mathsf{dl}_K(Q_N, P_N) - r_2 \leq \beta_N. \tag{3.22}$$

Thus

$$\mathbb{D}(Q^*, B_N(P_N, r_2; \mathsf{dl}_K); \mathsf{dl}_K) \leq \mathsf{dl}_K(Q^*, \tilde{Q}_N) \leq \mathsf{dl}_K(Q^*, Q_N) + \mathsf{dl}_K(Q_N, \tilde{Q}_N) \leq 2\beta_N. \tag{3.23}$$

This shows (3.21) as desired. ■

It might be interesting to ask whether the Q_N in the proof above will always be contained in $B_N(P_N, r_2; \mathsf{dl}_K)$. The following example shows that the answer is no. Moreover, the bound $2\beta_N$ at the right hand side of (3.20) is tight.

Example 3.2 *Let $\Xi = [-2, 2]$, $r = 1$, P_N and Q^* the Dirac probability measures supported at 0 and 1 respectively. It is easy to calculate $\mathsf{dl}_K(P_N, Q^*) = 1$ and hence $Q^* \in \mathcal{B}(P_N, 1; \mathsf{dl}_K)$. Let*

$$\Xi^N = \{-2, -1.7, -1.3, -0.7 + 2\epsilon, -0.2, 0, 0.2, 0.7 - 2\epsilon, 1.3, 1.7, 2\}$$

with a sufficiently small $\epsilon > 0$. By (3.10), $\beta_N = 0.3 + \epsilon$. Let Q_N be VT-projection of Q^* on $\mathscr{P}(\Xi^N)$. Then $\mathsf{dl}_K(Q_N, Q^*) \leq \beta_N$. Indeed, $Q_N(\xi = 1.3) = 1$ and $\mathsf{dl}_K(Q_N, Q^*) = 0.3$. Note also that $\mathsf{dl}_K(Q_N, P_N) = 1.3$ and hence $Q_N \notin B_N(P_N, 1; \mathsf{dl}_K)$. Moreover, by (3.22) and (3.23)

$$\begin{aligned} \mathbb{D}(Q^*, B_N(P_N, 1; \mathsf{dl}_K); \mathsf{dl}_K) &\leq \mathsf{dl}_K(Q^*, Q_N) + \mathsf{dl}_K(Q_N, \tilde{Q}_N) \\ &\leq \mathsf{dl}_K(Q^*, Q_N) + \mathsf{dl}_K(Q_N, P_N) - 1 \\ &= 0.6 \leq 2\beta_N = 0.6 + 2\epsilon. \end{aligned}$$

Since ϵ can be arbitrarily small, the inequality above implies that the bound $2\beta_N$ is tight.

The key difference between Theorem 3.2 and [28, Theorem 1] is that the latter considers the quantitative analysis of general ζ-ball with different centers and radii but the same support set, whereas Theorem 3.2 extends the result and considers the quantitative analysis of the ambiguity sets with not only different centers and radii but also different support sets in the case when the ambiguity sets are constructed by Kantorovich balls.

Theorem 3.2 extends [28, Theorem 1] by allowing one to quantify the difference between two Kantorovich balls supported over Ξ and Ξ_N respectively. The result may be further extended to balls with other metrics of ζ-structure.

3.4 ϕ-divergence

Alternative to Kantorovich ball, we may consider ϕ-divergence to construct the ambiguity set $\mathcal{P}$. ϕ-divergence is well studied [24, 23]. Here we focus on its definition over the space of discrete probability measures.

Let $p = (p_1, \ldots, p_N) \in \mathbb{R}_+^N$ and $q = (q_1, \ldots, q_N) \in \mathbb{R}_+^N$ be two probability vectors, that is, $\sum_{i=1}^N p_i = \sum_{i=1}^N q_i = 1$. The ϕ-divergence between p and q is defined as

$$I_\phi(p, q) := \sum_{i=1}^N q_i \phi(\frac{p_i}{q_i}),$$

where $\phi(t)$ is a convex function for $t \geq 0$, $\phi(1) = 0$, $0\phi(a/0) := a \lim_{t\to\infty} \phi(t)/t$ for $a > 0$ and $0\phi(0/0) := 0$. There are several common ϕ-divergence which are used in the paper as follows:

(a) Kullback-Leibler: $I_{\phi_{KL}}(p, q) = \sum_i p_i \log(\frac{p_i}{q_i})$ with $\phi_{KL}(t) = t \log t - t + 1$;

(b) Burg entropy: $I_{\phi_B}(p, q) = \sum_i q_i \log(\frac{q_i}{p_i})$ with $\phi_B(t) = -\log t + t - 1$;

(c) J-divergence: $I_{\phi_J} = \sum_i (p_i - q_i) \log(\frac{p_i}{q_i})$ with $\phi_J(t) = (t - 1) \log t$;

(d) χ^2-distance: $I_{\phi_{\chi^2}}(p, q) = \sum_i \frac{(p_i - q_i)^2}{p_i}$ with $\phi_{\chi^2}(t) = \frac{1}{t}(t - 1)^2$;

(e) Modified χ^2-distance: $I_{\phi_{m\chi^2}}(p, q) = \sum_i \frac{(p_i - q_i)^2}{q_i}$ with $\phi_{m\chi^2}(t) = (t - 1)^2$;

(f) Helinger distance: $I_{\phi_H}(p, q) = \sum_i (\sqrt{p_i} - \sqrt{q_i})^2$ with $\phi_H(t) = (\sqrt{t} - 1)^2$;

(g) Variation distance: $I_{\phi_V}(p, q) = \sum_i |p_i - q_i|$ with $\phi_V(t) = |t - 1|$.

Let $P^* \in \mathscr{P}(\Xi)$ be the true distribution of random variable ξ and $\{\zeta_1, \cdots, \zeta_V\} \subset \Xi$ denote a set of V-distinct points in the support set of ξ, let Ξ_i denote the Voronoi tessellation of Ξ as in (3.11) centered at ζ_i for $i = 1, \ldots, V$. Let $\{\xi_1, \cdots, \xi_N\}$ be an i.i.d sample of ξ where $N \gg V$ and N_i denotes the number of samples falling into area Ξ_i. Define empirical distribution $P_V(\cdot) := \sum_{i=1}^V \frac{N_i}{N} \delta_{\zeta_i}(\cdot)$ and ambiguity set

$$\mathcal{P}_N^V := \left\{ \sum_{i=1}^V p_i \delta_{\zeta_i}(\cdot) : I_\phi(p, p_V) \leq r_V, \sum_{i=1}^V p_i = 1, p_i \geq 0, \forall i = 1, \ldots, V \right\}, \tag{3.24}$$

where $p_V = (\frac{N_1}{N}, \cdots, \frac{N_V}{N})^T$ and I_ϕ is defined above. From the definition, we can see that the nominal distribution p_V is defined through reallocation of empirical probabilities to a specified set of points with size V via Voronoi tessellation. This is to avoid potentially large samples in the second stage DRO, that is, V is often much smaller than N. Convergence of $\mathcal{P}_V^N$ to P^* is established by Guo and Xu [13].

3.5 Discretization of the two-stage DRO

Having discussed discretization of the ambiguity sets in the preceding section, we now move on to investigate the impact of the discretization on the two-stage DRO problem (1.3) with $\mathcal{P}$ being replaced by $\mathcal{P}_N$:

$$\begin{array}{ll} \min\limits_{x, y(\cdot)} \sup\limits_{P \in \mathcal{P}_N} & \mathbb{E}_P[f_1(x, \xi) + g(x, y(\xi), \xi)] \\ \text{s.t.} & x \in X, y(\xi_i) \in \mathcal{Y}(x, \xi_i), i = 1, \ldots, N. \end{array} \tag{3.25}$$

Note that problem (3.25) can also be written as

$$\begin{array}{ll} \min\limits_{x, \mathbf{y}} \sup\limits_{P \in \mathcal{P}_N} & \sum_{i=1}^N p_i f_1(x, \xi_i) + g(x, \mathbf{y}_i, \xi_i) \\ \text{s.t.} & x \in X, \mathbf{y}_i \in \mathcal{Y}(x, \xi_i), i = 1, \ldots, N, \end{array} \tag{3.26}$$

where $P = (p_1, \cdots, p_N)$, and $\mathbf{y} = (\mathbf{y}_1, \cdots, \mathbf{y}_N)^T$.

We denote

$$y^N(\xi) := \sum_{i=1}^{N} \mathbf{1}_{\Xi_i^N}(\xi)\mathbf{y}_i.$$

Let ϑ^* and ϑ^N denote respectively the optimal value of (1.3) and (3.26), S_x^* and S_x^N denote the set of corresponding x-component of optimal solutions and $\mathcal{Y} : X \times \Xi \to \mathbb{R}^m$ be the feasible set-valued mapping of the second stage problem.

Proposition 3.2 (cf. [28, Proposition 4]) *Assume: (i) $\mathcal{Y}(x, \xi)$ is pseudo-Lipschitzian at every pair of $(y_0, (x_0, \xi_0)) \in \mathcal{Y}(x_0, \xi_0) \times \{(x_0, \xi_0)\}$, (ii) there exist positive constants L_c and β such that*

$$|g(x, y, \xi) - g(x_0, y_0, \xi_0)| \le L_c[d(x, x_0) + d(\xi, \xi_0)^\beta + d(y, y_0)]$$

for $(x, \xi) \in X \times \Xi$ and $y \in \mathcal{N}(y_0)$, where $\mathcal{N}(y_0)$ denotes a neighborhood of y_0. Then there exists a positive constant L such that

$$|v(x, \xi) - v(x_0, \xi_0)| \le L[d(x, x_0) + d(\xi, \xi_0) + d(\xi, \xi_0)^\beta].$$

The following theorem summarizes the convergence of the optimal values and optimal solutions when the ambiguity set is constructed via moment conditions, Kantorovich ball or ϕ-divergence.

Theorem 3.3 *Let $\mathcal{H} := \{f(x, \cdot) + v(x, \cdot) : x \in X\}$. Assume (i) for N sufficiently large,*

$$\sup_{h\in\mathcal{H}} |\sup_{P\in\mathcal{P}_N} \mathbb{E}_P[h(\xi)]| < \infty \text{ and } \sup_{h\in\mathcal{H}} |\sup_{P\in\mathcal{P}} \mathbb{E}_P[h(\xi)]| < \infty,$$

(ii) $f(\cdot, \xi)$ is Lipschitz continuous with Lipschitz modules κ_f and conditions in Proposition 3.2 hold. Then the following assertions hold:

(i) If the ambiguity set is defined by moment conditions and the conditions in Theorem 3.1 hold, then

$$|\vartheta^* - \vartheta^N| \le L(1 + L\rho\mathrm{diam}(\Xi)C_1)\beta_N,$$

and $\mathbb{D}(S_x^N, S_x^*) \to 0$ as $N \to \infty$; moreover, if $\sup_{P\in\mathcal{P}} \mathbb{E}_P[f(x, \cdot) + v(x, \cdot)]$ satisfies the following growth condition at the optimal solution set S_x^*, that is,

$$\sup_{P\in\mathcal{P}} \mathbb{E}_P[f(x, \xi) + v(x, \xi)] \ge \vartheta^* + ud(x, S_x^*)^\nu, \ \ \forall x \in X, \tag{3.27}$$

where u, ν are some positive constants, then

$$\mathbb{D}(S_x^N, S_x^*) \le \left(\frac{3L}{u}(1 + L\rho\mathrm{diam}(\Xi)C_1)\beta_N\right)^{\frac{1}{\nu}}.$$

(ii) If the ambiguity set is defined by Kantorovich ball $\mathcal{B}(P, r; \mathsf{dl}_K)$ and $B_N(P_N, r_N; \mathsf{dl}_K)$ is its discrete approximation, where P_N is constructed by (3.12) and $|r - r_N| \to 0$ as $N \to \infty$, then

$$|\vartheta^* - \vartheta^N| \le L(3\beta_N + |r - r_N|),$$

and $\mathbb{D}(S_x^N, S_x^*) \to 0$ as $N \to \infty$. If, in addition, $\sup_{P\in\mathcal{P}} \mathbb{E}_P[f(x,\cdot) + v(x,\cdot)]$ satisfies the growth condition (3.27), then

$$\mathbb{D}(S_x^N, S_x^*) \le \left(\frac{3L}{u}(3\beta_N + |r - r_N|)\right)^{\frac{1}{\nu}}.$$

(iii) Let $\mathcal{P}_N^V$ be defined as in (3.24). Let σ be a positive number such that $\sigma V < 1$. If ϕ is chosen from one of the functions listed in (a)-(g) of Section 3.4 and $\frac{V}{\sqrt{N}} \to 0$ and $r_V \to 0$ as $V \to \infty$ and $N \to \infty$. Then with probability at least $1 - V\sigma$,

$$|\vartheta^* - \vartheta^V| \le L\left(\beta_V + \frac{\operatorname{diam}(\Xi)}{2}\max\{2\sqrt{r}, r\} + \frac{\operatorname{diam}(\Xi)}{2}\Delta(V, N, \sigma)\right),$$

where $\Delta(V, N, \sigma) := \min\left(\frac{V}{\sqrt{N}}(2 + \sqrt{2\ln\frac{1}{\sigma}}), 4 + \frac{1}{\sqrt{N}}(2 + \sqrt{2\ln\frac{1}{\sigma}})\right)$ and $\mathbb{D}(S_x^V, S_x^*) \to 0$ as $V \to \infty$, $N \to \infty$. If, in addition, $\sup_{P\in\mathcal{P}} \mathbb{E}_P[f(x,\cdot)+v(x,\cdot)]$ satisfies the growth condition (3.27), then

$$\mathbb{D}(S_x^V, S_x^*) \le \left(\frac{3L}{u}\left(\beta_V + \frac{\operatorname{diam}(\Xi)}{2}\max\{2\sqrt{r}, r\} + \frac{\operatorname{diam}(\Xi)}{2}\Delta(V, N, \sigma)\right)\right)^{\frac{1}{\nu}}.$$

Theorem 3.3 follows from [28, Theorem 3], Theorem 3.1, Theorem 3.2 and [13, Proposition 3.1].

4 Numerical tests

To examine the efficiency of Algorithm 1, we carry out some numerical tests on a one-stage multi-product newsvendor problem and a multi-product production planning problem. In this section, we report some of the test results.

To facilitate reading, we rewrite the two-stage DRO problem (1.3) as

$$\begin{array}{ll} \min\limits_{\boldsymbol{x},\boldsymbol{y},\boldsymbol{\sigma}} \max\limits_{P\in\mathcal{P}_N,\boldsymbol{w}} & \sum\limits_{i=1}^{N}\left(p_i\sigma_i + \frac{1}{2}w_i^T(x_i - x_{i+1})\right) \\ \text{s.t.} & f(x_i, y_i, \xi_i) \le \sigma_i,\ i = 1, \ldots, N, \\ & x_i \in X, y_i \in \mathcal{Y}(x_i, \xi_i),\ i = 1, \ldots, N. \end{array} \tag{4.1}$$

We consider the cases that $\mathcal{P}_N$ is constructed by moment conditions, Kantorovich ball or modified χ^2-distance in N-dimensional space of probability measures $\mathscr{P}(\mathbb{R}^N)$.

4.1 Multi-product Newsvendor problem

Consider a single period multi-product newsvendor problem where a retailer sells a seasonal product. At the beginning of a selling season, the retailer has to make a decision on its order quantity $x \in \mathbb{R}^n$ (where each component represents a product) before market demand is observed. The market demand is characterized by a random vector $\xi \in \mathbb{R}^n_+$ with each component representing the demand of a product. For the test of Algorithm 1, we assume that ξ follows a discrete distribution with support set $\Xi_N = \{\xi_1, \cdots, \xi_N\}$ although in reality it is more likely to be continuously distributed. The assumption may be viewed as a result of discretiztion as we discussed in the previous section. We also assume that the support Ξ_N is bounded below by 0 and bounded above by $\bar{d} \in \mathbb{R}^n_+$ componentwise.

In this setup, the true probability distribution of ξ is unknown, but it is possible to construct an ambiguity set using some partial moment information

$$\mathcal{P}_N := \left\{ P \in \mathbb{R}^N_+ : \begin{array}{l} \sum_{i=1}^N p_i = 1, \\ \sum_{i=1}^N p_i \xi_i = \mu, \\ \sum_{i=1}^N p_i (\xi_i - \mu)(\xi_i - \mu)^T \preceq \Sigma \end{array} \right\}, \tag{4.2}$$

where $\mu \in \mathbb{R}^n$ is the mean value of ξ, $\Sigma \in \mathbb{R}^{n \times n}$ is an upper bound of its covariance matrix.

Note that the newsvendor problem is one-stage by nature. Therefore we can develop an one-stage DRO formulation

$$\min_{x \in \{x \in \mathbb{R}^n_+, c^T x \leq W\}} \sup_{P \in \mathcal{P}_N} \mathbb{E}_P[f(x, \xi)], \tag{4.3}$$

where $\mathcal{P}_N$ is defined in (4.2),

$$f(x, \xi) = (\mathbf{c} - \mathbf{v})^T x + (\mathbf{s} - \mathbf{v})^T \max(-x, -\xi),$$

where $\mathbf{s} \in \mathbb{R}^n_+$ denotes the unit selling price in the market, $\mathbf{c} \in \mathbb{R}^n_+$ denotes the unit purchase cost, and $\mathbf{v} \in \mathbb{R}^n_+$ denotes the net salvage value for each unit of leftover, all of which are constant with $\mathbf{v}_j < \mathbf{c}_j < \mathbf{s}_j, j = 1, \ldots, n$. We assume total cost $\mathbf{c}^T x$ is bounded above by $W \in \mathbb{R}_+$.

We can also artificially represent the one-stage decision making problem (4.3) as a two-stage DRO

$$\begin{array}{rll} \min\limits_{x \in \mathbb{R}^n_+, y(\cdot)} \sup\limits_{P \in \mathcal{P}_N} & \mathbb{E}_P[(\mathbf{c} - \mathbf{v})^T x + (\mathbf{s} - \mathbf{v})^T y(\xi)], \\ \text{s.t.} & \mathbf{c}^T x \leq W, \\ & -y(\xi) - x \leq 0, \\ & -y(\xi) - \xi \leq 0, \end{array} \tag{4.4}$$

and cast the latter as

$$\begin{aligned}\min_{\boldsymbol{x},\boldsymbol{y},\boldsymbol{\sigma}} \max_{P\in\mathcal{P}_N,\boldsymbol{w}} \quad & \sum_{i=1}^{N}\left(p_i\sigma_i + \frac{1}{2}w_i^T(x_i - x_{i+1})\right) \\ \text{s.t.} \quad & (\mathbf{c}-\mathbf{v})^T x_i + (\mathbf{s}-\mathbf{v})^T y_i \le \sigma_i,\ i=1,\ldots,N, \\ & -y_i - x_i \le 0,\ \ i=1,\ldots,N, \\ & -y_i - \xi_i \le 0,\ \ i=1,\ldots,N, \\ & x_i \ge 0, \mathbf{c}^T x_i \le W,\ i=1,\ldots,N.\end{aligned} \tag{4.5}$$

We use Algorithm 1 to solve problem (4.5) with $n = 10$, $\mathbf{s} = (8,\ldots,8)^T$, $\mathbf{c} = (4,\ldots,4)^T$, and $\mathbf{v} = (2,\ldots,2)^T$. To construct the ambiguity set $\mathcal{P}_N$, we first generate the support set Ξ_N of ξ with $N = 100$. Specifically, for $j = 1,\ldots,n$, let ζ_j be a random variable such that $\zeta_j/10$ follows a beta-distribution with parameters $\alpha_j = \beta_j = \frac{1}{2}(\frac{25}{j} - 1)$. For each j, let $\zeta_j/10$ generate 100 samples and multiply them by 10. We then treat the 100 points as the support set of the j-th component of ξ. Let ξ_{ji} be the i-th sample. Then $\xi_i = (\xi_{1i},\cdots,\xi_{ni})^T$ for $i = 1,\ldots,100$. This effectively defines Ξ_{100}. Note that since the support set of ζ_j is $[0, 10]$, $\mathbb{E}[\zeta_j] = 5$ and $\mathrm{var}[\zeta_j] = j$, then we set $\mathbb{E}[\xi_j] = 5$ and $\mathrm{var}[\xi_j] = j$, where ξ_j denotes the j-th component of ξ. This gives $\mu = (5,\ldots,5)^T$. The covariance of ξ is unknown, but we set its upper bound Σ with diagonal elements being $1,\ldots,n$ and the off-diagonal elements 0.1. The total capital W is set 160. We also apply a cutting plane method in [45] to solve (4.3) using the same set of data.

We have carried out 30 independent tests each of which generates a support set Ξ_{100} and solves the one-stage and two-stage distributionally robust models. The numerical test results are displayed in Figure 1. The horizontal axis represents 10 products and the vertical axis represents the purchase quantities of each of the products. Each strip represents the $\frac{1}{4}$ to $\frac{3}{4}$ quantile of the 30 computed quantities of a product. The performance of these two models are almost the same in the sense that they both show a tendency of decrease of purchase quantity x as the variance of ξ_j increases. Similar results are observed when the sample size is increased from 100 to 1000, see Figure 2. In short, we claim that the two-stage DRO model performs as good as the one-stage DRO model, which shows the correctness of the two-stage DRO model and the associated algorithm.

We set the selling prices, the unit purchase costs and the unit net salvage values of the ten produces identical so that we may concentrate on examining the performance of algorithm with different demand on each of the products. Figures 1-2 show that the newsvendor tends to purchase products with lower variance of demand uncertainty which is consistent with practice. Note also that this is one-stage decision making problem, we artificially make it a two-stage DRO only for the purpose of examining whether our algorithm delivers the right results.

4.2 Two-stage distributionally robust production planning

Consider a two-stage production planning problem under uncertain environment. A firm plans to produce n products and sell them to customers and retailers. Market demand for the products

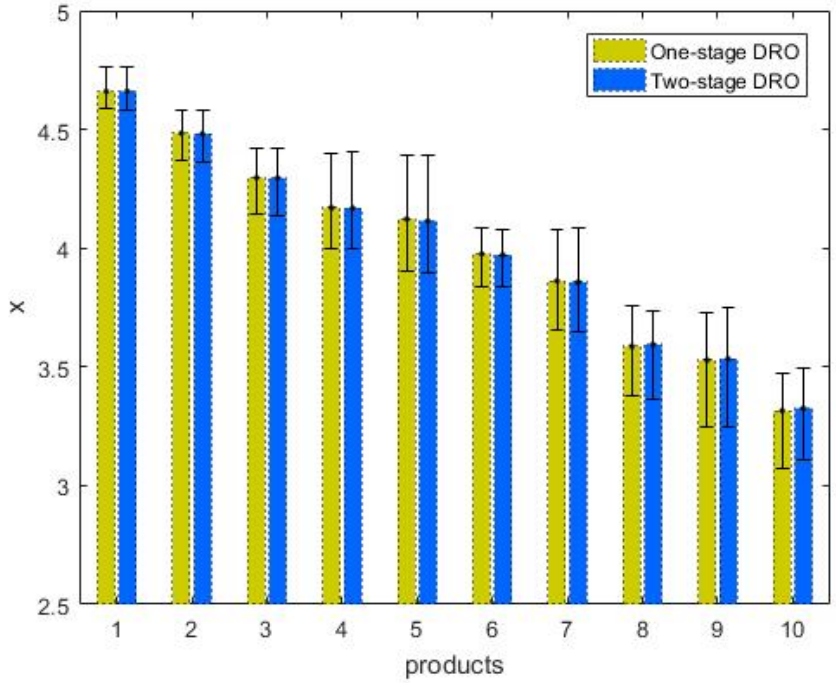

Figure 1: The newsvendor problem with $n = 10$ products and $N = 100$ samples.

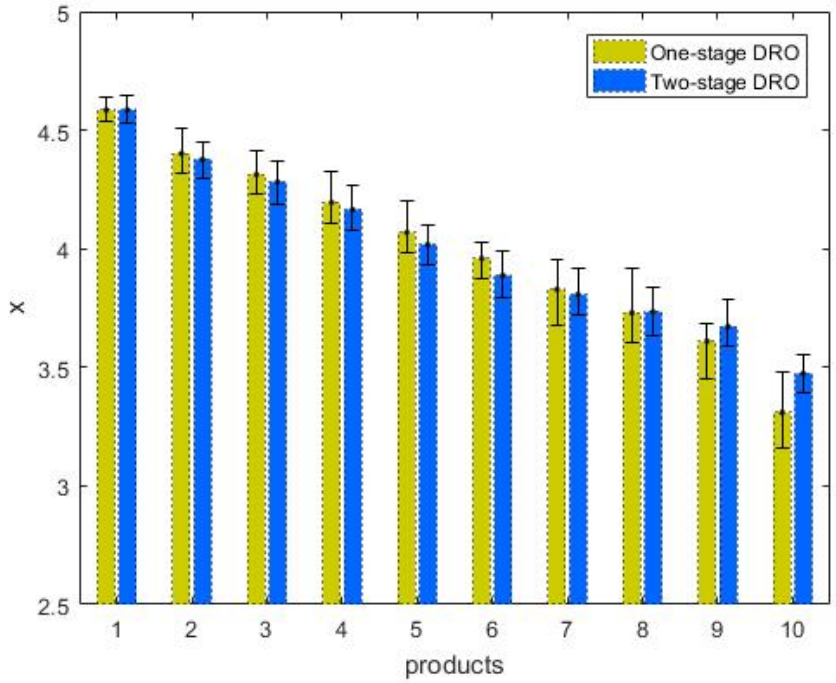

Figure 2: The newsvendor problem with $n = 10$ products and $N = 1000$ samples.

is described by inverse demand functions, that is, if the total supply of product t to the market is y_t, then the price of the product will be

$$\varrho_t(y_t, \xi) = \xi_{t1} - \sqrt{y_t + \xi_{t2}},$$

where ξ_{t1} and ξ_{t2} are uncertain parameters. While the quantity of production of each product is determined after observation of the uncertainties, materials to be used for producing each of the products need to be pre-ordered from manufactures before realization of the uncertainty. Let $\mathbf{a}_s^t$, $s = 1, \ldots, m$ denote the quantity of type s material required to produce one unit of product t and $\mathbf{c}_s$, $s = 1, \ldots, m$ denote the unit cost of material s. Let x_s denote the total quantity to be ordered for material s. Then the total order cost of all materials will be $\sum_{s=1}^m \mathbf{c}_s x_s$.

Since the decision of production will be made after realization of uncertainties in market demand, the firm may buy more materials from local wholesale market at higher prices provided that it is profitable to do so. Let $r_s(\xi_{s3}) = \xi_{s3}$, $s = 1, \ldots, m$ denote the unit cost of material s where ξ_{s3} is an uncertain parameter underlying the uncertainty of the price in the wholesale market. Assuming that the firm aims to maximize the overall expected profit, we can develop a two-stage stochastic programming model for this problem

$$\begin{array}{rl} \min\limits_{x \in X, y(\xi) \geq 0} & \mathbf{c}^T x + \mathbb{E}_P[-\varrho(y(\xi), \xi)^T y(\xi) + r(\xi)^T z(\xi)] \\ \text{s.t.} & My(\xi) \leq x + z(\xi), \end{array}$$

where we write $\mathbf{c}$ for $(\mathbf{c}_1, \cdots, \mathbf{c}_m)$, $x = (x_1, \cdots, x_m)$ for the first stage decision vector, M is an $m \times n$ matrix with the (s,t)-th element $\mathbf{a}_s^t$, $y(\xi)$ is the vector of n products to be produced at the second stage, $z(\xi)$ is the vector of quantities of m materials to be purchased from a wholesale market and $r(\xi) = (r_1(\xi), \cdots, r_m(\xi))$ is the vector of the associated unit price/cost. To simplify the notation here, we write ξ for any of the random variables in this problem although the ξ in $r(\xi)$ may have different components from those in $y(\xi)$ and $z(\xi)$. By applying the robust argument as we discussed in the previous sections on the probability distribution of ξ, we may write down the DRO model for this problem

$$\begin{array}{rl} \min\limits_{x \in X, y(\xi) \geq 0} \sup\limits_{P \in \mathcal{P}_N} & \mathbf{c}^T x + \mathbb{E}_P[-\varrho(y(\xi), \xi)^T y(\xi) + r(\xi)^T z(\xi)] \\ \text{s.t.} & My(\xi) \leq x + z(\xi). \end{array}$$

To fit into our proposed numerical framework directly, we assume that ξ follows a finite discrete distribution which may result from a discretization approach as we discussed in the previous section in the case when ξ is continuously distributed.

We propose three ways to construct the ambiguity set $\mathcal{P}_N$.

1. Moment condition:

$$\mathcal{P}_N^1 := \left\{ P \in \mathbb{R}_+^N : \begin{array}{l} \sum_{i=1}^N p_i = 1, \\ \sum_{i=1}^N p_i \xi_i^{tk} = \mu_{tk},\ t = 1, \ldots, n, k = 1, 2, \\ \sum_{i=1}^N p_i \xi_i^{s3} = \nu_s,\ s = 1, \ldots, m, \\ \sum_{i=1}^N p_i (\xi_i - \mu)(\xi_i - \mu)^T \preceq \Sigma \end{array} \right\}. \tag{4.6}$$

2. Kantorovich ball:

$$\mathcal{P}_N^2 := \{P \in \mathscr{P}(\Xi_N) : \mathsf{dl}_K(P, P_N) \le d\}, \tag{4.7}$$

where $P_N = (p_1^N, \cdots, p_N^N)$ is a nominal distribution constructed through empirical data. Following a discussion in [26, Section 2.7], we can easily rewrite (4.7) as

$$\mathcal{P}_N^2 = \left\{ p \in \mathbb{R}_+^N : \begin{array}{l} \pi \in \mathbb{R}_+^{N\times N}, \\ \sum_{j=1}^N \pi_{ij} = p_i^N, \\ \sum_{i=1}^N \pi_{ij} = p_j, \sum_{j=1}^N p_j = 1, \\ \sum_{i=1}^N \sum_{j=1}^N \pi_{ij}\|\xi_i - \xi_j\| \le d \end{array} \right\}. \tag{4.8}$$

3. Modified χ^2-distance

Consider the random variable ξ with support set Ξ and true distribution P^*. Let P_N be an empirical distribution with support points $\{\zeta_1, \cdots, \zeta_N\} \subset \Xi$

$$P_N(\cdot) := \sum_{i=1}^N \frac{1}{N} \mathbb{1}_{\zeta_i}(\cdot),$$

where

$$\mathbb{1}_{\zeta_i}(\zeta) = \begin{cases} 1, & \zeta = \zeta_i, \\ 0, & \text{otherwise.} \end{cases}$$

Moreover, let $I_{m\chi^2}(P, Q) = \sum_i \frac{(p_i - q_i)^2}{q_i}$ be the modified χ^2-distance. Define the ambiguity set

$$\mathcal{P}_N^3 := \left\{ p \in \mathbb{R}_+^N : \sum_i \left(p_i - \frac{1}{N} \right)^2 \le \frac{r_N}{N}, \sum_{i=1}^N p_i = 1 \right\}. \tag{4.9}$$

$\mathcal{P}_N^3$ may be considered as an approximation of the true distribution P^*, see subsection 3.4, Theorem 3.3 (iii) and [13].

We have undertaken three sets of numerical experiments corresponding to the three ambiguity sets. The first two are carried out on a ThinkPad T450 notebook with two Intel Core i7 2.40GHz CPUs and 8GB RAM. The aim is to examine the performance of the algorithm with different types of the ambiguity sets when the sample size is moderate. The third set of numerical experiments is carried out in a cluster computer system with 20 Intel Xeon 2.40GHz CPUs and 64GB RAM. Our aim is to examine the effectiveness of the algorithm when parallel structure is exploited and the sample size is large.

Setup of the tests. We set $m = 3$ and $n = 2$, $X = \mathbb{R}_+^3$, $c = (\frac{2}{3}, \frac{1}{2}, \frac{1}{3})$, $a^1 = (a_{11}, a_{21}, a_{31}) = (3, 2, 1)$, and $a^2 = (a_{12}, a_{22}, a_{32}) = (1, 2, 3)$. This means the unit costs of the two products are $\frac{10}{3}$ and $\frac{8}{3}$.

Setup of the support Ξ_N. For the ambiguity sets $\mathcal{P}_N^1$ and $\mathcal{P}_N^2$, $\{\xi_i^{12}\}_{i=1}^N$ and $\{\xi_i^{22}\}_{i=1}^N$ are generated from uniform distribution over $[0, 1]$, and $\{(\xi_i^{11} - 3)/5\}_{i=1}^N$ are generated via beta

distribution with parameters $\alpha_1 = \beta_1 = \frac{21}{8}$. We also generate $\{\xi_i^{s3}\}_{i=1}^N$, $s = 1, 2, 3$, via uniform distribution over $[1.5c_i, 2.5c_i]$. In this experiment, we enlarge the support set Ξ_N of the ambiguity set $\mathcal{P}_N$ w.r.t. ξ_{21} by an increment of 20. We start with $N = 20$ and generate $\{(\xi_{21}^k - 1)/10\}_{k=1}^{20}$ via beta distribution with parameters $\alpha_2 = \beta_2 = \frac{1}{2}(\frac{25}{v} - 1)$ where $v = 22$. We then test $N = 40$ and generate $\{(\xi_{21}^k - 1)/10\}_{k=1}^{20}$ and $\{(\xi_{21}^k - 1)/10\}_{k=21}^{40}$ via beta distribution with parameters $\alpha_2 = \beta_2 = \frac{1}{2}(\frac{25}{v} - 1)$ and $v = 22$ and 18 respectively. Finally, we test $N = 120$ and generate $\{(\xi_{21}^k - 1)/10\}_{k=1}^{20}, \cdots, \{(\xi_{21}^k - 1)/10\}_{k=101}^{120}$ via beta distribution with parameters $\alpha_2 = \beta_2 = \frac{1}{2}(\frac{25}{v} - 1)$ and $v = 22, 18, 14, 10, 6, 2$ respectively.

In the numerical tests with $\mathcal{P}_N^3$, the setting of the support set is almost same as above, the only difference is the sample size N is 10 times of the above.

Setting of the parameters in the ambiguity sets. The parameters in (4.6) are set with $\mu_{11} = 5.5$, $\mu_{12} = 0.5$, $\mu_{21} = 6$, $\mu_{22} = 0.5$, $\nu_s = 2c_s$, $s = 1, \ldots, m$, and

$$\Sigma = \begin{bmatrix} 2 & \frac{1}{20} & \frac{1}{20} & \frac{1}{20} & \frac{1}{20} & \frac{1}{20} & \frac{1}{20} \\ \frac{1}{20} & \frac{1}{6} & \frac{1}{20} & \frac{1}{20} & \frac{1}{20} & \frac{1}{20} & \frac{1}{20} \\ \frac{1}{20} & \frac{1}{20} & 20 & \frac{1}{20} & \frac{1}{20} & \frac{1}{20} & \frac{1}{20} \\ \frac{1}{20} & \frac{1}{20} & \frac{1}{20} & \frac{1}{6} & \frac{1}{20} & \frac{1}{20} & \frac{1}{20} \\ \frac{1}{20} & \frac{1}{20} & \frac{1}{20} & \frac{1}{20} & 2c_1^2 & \frac{1}{20} & \frac{1}{20} \\ \frac{1}{20} & \frac{1}{20} & \frac{1}{20} & \frac{1}{20} & \frac{1}{20} & 2c_2^2 & \frac{1}{20} \\ \frac{1}{20} & \frac{1}{20} & \frac{1}{20} & \frac{1}{20} & \frac{1}{20} & \frac{1}{20} & 2c_3^2 \end{bmatrix}.$$

The parameters in (4.8) are set $p_i^N = \frac{1}{N}$, $\xi_i = (\xi_i^{11}, \xi_i^{12}, \xi_i^{21}, \xi_i^{22}, \xi_i^{31}, \xi_i^{32}, \xi_i^{33})$, and $d = 1/2$. In addition, we set $r_N = 0.1$ in (4.9).

Numerical results. When the ambiguity set is constructed through moment conditions and the Kantorovich ball, the production planning results are reported in Tables 1 and 2 respectively. We first solve these problems with general serial computation. The CPU times are listed in columns 12–14, where the "Ps time" counts the CPU time for the primal $(\boldsymbol{x}, \boldsymbol{y}, \boldsymbol{\sigma})$-subproblem, the column "Ds time" counts the CPU time for the dual $(P, \boldsymbol{w})$-subproblem, and the column "Total" is the total CPU time of Algorithm 1 which is almost add-up of the two. Since the primal $(\boldsymbol{x}, \boldsymbol{y}, \boldsymbol{\sigma})$-subproblem is decomposable and can be solved in parallel, we implement it in MATLAB using the "parfor" command with two computing cores. The corresponding CPU times are shown in columns 15–17.

The performance of the algorithm is similar in the first two sets of experiments in terms of the optimal values and optimal solutions. By the construction of the support set Ξ_N of the ambiguity set $\mathcal{P}_N^1$ ($\mathcal{P}_N^2$), we note that as N increases, ξ has a wider range and this may be interpreted as increase of uncertainty. Subsequently, the ambiguity set becomes larger and hence the profit (the absolute value of the optimal value of the DRO problem) goes down as a more and more conservative decision is taken.

However, the performance of the algorithm is very different with the two ambiguity sets in terms of CPU times. In the moment case, the "Ps time" is much larger than "Ds time", we can see clear benefit by exploiting parallel computing. Indeed, by running the parallel computing with the two computer cores, we can effectively reduce almost 30% of the overall CPU time.

N	Opt. Val.	x_1	x_2	x_3	$\mathbb{E}_P[z_1]$	$\mathbb{E}_P[z_2]$	$\mathbb{E}_P[z_3]$	$\mathbb{E}_P[y_1]$	$\mathbb{E}_P[y_2]$
20	−16.4316	16.0020	12.4493	9.4716	0.2280	8.4508	16.8289	2.8200	7.3256
40	−11.6923	12.6698	21.5979	30.5247	0.0015	0.3530	0.7099	2.1618	6.1862
60	−7.8850	9.9919	15.1094	17.1114	1.1464	0.6157	2.4897	2.0292	5.0449
80	−6.3867	10.2763	12.1432	14.0228	0.0000	0.8828	1.8019	1.8608	4.6535
100	−6.1205	9.9868	12.7495	15.5146	0.0344	0.0006	0.0030	1.8221	4.5530
120	−6.1154	9.9520	12.6831	15.4130	0.0002	0.0006	0.0021	1.8055	4.5357

N	Serial computing			Parallel computing (2 cores)		
	Ps time	Ds time	Total	Ps time	Ds time	Total
20	140.0	41.4	182.5	143.2	43.5	188.2
40	513.6	93.8	607.6	399.3	91.4	490.9
60	694.3	107.7	802.2	493.2	105.4	598.8
80	776.4	105.6	882.1	530.5	103.4	634.1
100	1189.0	153.9	1343.2	799.1	152.6	951.9
120	1591.9	208.6	1800.7	1062.0	206.9	1269.1

Table 1: A firm produces two products using three materials under the moment condition.

N	Opt. Val.	x_1	x_2	x_3	$\mathbb{E}_P[z_1]$	$\mathbb{E}_P[z_2]$	$\mathbb{E}_P[z_3]$	$\mathbb{E}_P[y_1]$	$\mathbb{E}_P[y_2]$
20	−12.8261	13.0131	18.7939	33.8260	0.1533	3.1917	1.8831	2.3431	6.0713
40	−11.1879	13.2210	20.2886	27.3558	0.1619	1.5068	2.9351	2.6173	5.5019
60	−9.6323	11.9528	17.7935	23.7946	0.2660	1.3727	2.7003	2.3234	5.0974
80	−8.3201	11.1095	16.3659	21.6232	0.3143	1.0678	2.1920	2.1719	4.7728
100	−7.2256	10.4216	14.3997	18.3767	0.2728	1.1353	2.3775	2.0385	4.4678
120	−6.3794	9.6893	12.7195	15.7503	0.2855	1.1167	2.2970	1.9089	4.1739

N	Serial computing			Parallel computing (2 cores)		
	Ps time	Ds time	Total	Ps time	Ds time	Total
20	256.2	41.9	299.3	242.5	42.8	286.4
40	976.3	560.5	1537.2	696.9	483.6	1180.8
60	928.4	808.3	1736.9	680.1	791.2	1471.5
80	2053.2	2429.2	4482.7	1437.9	2472.5	3910.9
100	2231.3	4059.6	6291.3	1478.7	3989.7	5468.8
120	1626.8	4515.6	6142.5	1057.0	4457.9	5515.1

Table 2: A firm produces two products using three materials under the Kantorovich ball.

In contrast, when the ambiguity is constructed via Kantorovich ball, we can see that the "Ds time" constitutes the main CPU time, consequently, running parallel computing to reduce "Ps time" does not significantly contribute to the reduction of the overall CPU time.

The numerical results of the ambiguity set constructed through modified χ^2-distance are reported in Table 3. The obtained approximate optimal values and optimal solutions are similar to those in the previous two tests. However, the impact of parallel computing on the CPU time is much more significant because the CPU time is dominated by the "Ps time" in the serial computing. We employ 20 computing cores in parallel computation to speed up the process. From Table 3, we can see that the parallel computing effectively reduces the "Ps time" to $\frac{1}{10}$.

N	Opt. Val.	x_1	x_2	x_3	$\mathbb{E}_P[z_1]$	$\mathbb{E}_P[z_2]$	$\mathbb{E}_P[z_3]$	$\mathbb{E}_P[y_1]$	$\mathbb{E}_P[y_2]$
200	−12.2039	14.0245	9.3495	4.6748	0.0587	6.0099	12.1365	3.1131	4.5661
400	−10.4623	12.6333	8.4222	4.2112	0.0684	5.2466	10.6385	2.8241	4.0085
600	−9.1444	11.7322	8.1821	4.7089	0.0882	4.5108	9.1072	2.6727	3.6157
800	−8.1243	10.9453	8.2498	5.5699	0.1002	3.7006	7.4957	2.5284	3.3050
1000	−7.0371	9.9456	8.2813	6.6652	0.1325	2.8679	5.8102	2.2928	3.0626
1200	−6.1809	9.2225	8.5590	7.9592	0.1672	2.0893	4.2205	2.1266	2.9103

N	Serial computing			Parallel computing (20 cores)		
	Ps time	Ds time	Total	Ps time	Ds time	Total
200	517.8	16.1	534.8	82.1	16.1	99.2
400	1637.7	31.9	1669.7	179.1	31.4	210.7
600	3266.9	56.0	3323.2	310.3	55.0	365.9
800	5067.8	77.5	5145.9	443.8	76.0	520.4
1000	6709.2	109.1	6819.0	576.0	109.0	685.8
1200	8305.1	134.1	8440.0	742.7	137.5	881.1

Table 3: A firm produces two products using three materials under the ambiguity set constructed through modified χ^2-distance.

Acknowledgement. We would like to thank Shabbir Ahmed for initiating the research in a private discussion with the 3rd author during the 14th international conference on stochastic programming in Búzios and his further encouragement during the preparation of the paper.

References

[1] Athreya K. B., Lahiri. S. N.: Measure theory and probability theory. Springer Science & Business Media, New York (2006)

[2] Bertsimas, D., Doan, X. V., Natarajan, K., Teo, C. P.: Models for minimax stochastic linear optimization problems with risk aversion, Math. Oper. Res. **35**, 580–602 (2010)

[3] Bertsimas, D., Parys, B. V.: Bootstrap robust prescriptive analytics, arXiv preprint arXiv:1711.09974 (2017)

[4] Billingsley, P.: Convergence of Probability Measures. John Wiley, New York (1968)

[5] Chambolle, A., Pock, T.: A first-order primal-dual algorithm for convex problems with applications to imaging, J. Math. Imaging. Vis. **40**, 120–145 (2011).

[6] Delage, E., Ye, Y.: Distributionally robust optimization under moment uncertainty with application to data-driven problems, Oper. Res. **58**, 592–612 (2010)

[7] Esser, E., Zhang, X., Chan, T.: A general framework for a class of first order primal-dual algorithms for Convex Optimization in Imaging Science, SIAM J. Imaging Sci. **3**, 1015–1046 (2010)

[8] Gao, R., Kleywegt, A.: Distributionally robust stochastic optimizationwithwasserstein distance, arXiv preprint arXiv:1604.02199 (2016)

[9] Gibbs, A. L., Su, F. E.: On choosing and bounding probability metrics, Int. Stat. Rev. **70**, 419–435 (2002)

[10] Goh. J., Sim, M.: Distributionally robust optimization and its tractable approximations, Oper. Res. **58**, 902–917 (2010)

[11] Goldstein, T., Li, M., Yuan, X., Esser, E., Baraniuk, R.: Adaptive primal-dual hybrid gradient methods for saddle-point problems, arXiv preprint arXiv:1305.0546 (2013)

[12] Guo, S., Xu, H., Zhang, L.: Convergence analysis for mathematical programs with distributionally robust chance constraint, SIAM J. Optim. **27**, 784–816 (2017)

[13] Guo, S., Xu, H.: Distributionally robust shortfall risk optimization model and its approximation, Math. Program. (2018) https://link.springer.com/article/10.1007%2Fs10107-018-1307-z

[14] Mohajerin Esfahani, P., Kuhn, D.: Data-driven distributionally robust optimization using the Kantorovich metric: performance guarantees and tractable reformulations, Math. Program. **171**, 115-166 (2018)

[15] Fan, K., Minimax theorems. Izv. Nats. Akad. Nauk Armen. Mekh, **39**, 42-47 (1953)

[16] Hanasusanto, G. A., Kuhn, D.: Conic programming reformulations of two-stage distributionally robust linear programs over Kantorovich balls, Oper. Res. **66**, 849–869 (2018)

[17] He, B., Ma, F., Yuan, X., An algorithm framework of generalized primal-dual hybrid gradient methods for saddle point problems, J. Math. Imaging. Vis. **58**, 279–293 (2017)

[18] He, B., Yuan, X., Convergence analysis of primal-dual algorithms for a saddle-point problem: From contraction perspective, SIAM J. Imaging Sci. **5**, 119–149 (2012)

[19] Jiang, R., Guan, Y.: Risk-averse two-stage stochastic program with distributional ambiguity, Oper. Res. **66**, 1390–1405 (2018)

[20] Liu, Y., Pichler A., Xu, H., Discrete approximation and quantification in distributionally robust optimization, Math. Oper. Res. (2018) https://doi.org/10.1287/moor.2017.0911

[21] Liu, Y., Yuan, X., Zeng, S., Zhang, J.: Primal-dual hybrid gradient method for distributionally robust optimization problems, Oper. Res. Lett. **45**, 625–630 (2017)

[22] Liu, Y., Yuan, X., Zhang, J.: Quantitative stability analysis of stochastic programs with distributionally robust second order dominance constraints, manuscript, (2017)

[23] Love, D., Bayrakcan, G.: Phi-divergence constrained ambiguous stochastic programs for data-driven optimization, available on Optimization Online (2016) http://www.optimization-online.org/DB_HTML/2016/03/5350.html

[24] Pardo, L.: Statistical Inference Based on Divergence Measures, Chapman and Hall/CRC, Boca Raton, FL (2005)

[25] Pflug, G. Ch., Pichler, A.: Approximations for probability distributions and stochastic optimization problems. In Bertocchi, M., Consigli, G., Dempster, M. A. H., editors, Stochastic Optimization Methods in Finance and Energy, vol. 163 of International Series in Operations Research & Management Science, Springer, New York (2011)

[26] Pflug, G. Ch., Pichler, A.: Multistage Stochastic Optimization. Springer Series in Operations Research and Financial Engineering. Springer, New York (2014)

[27] Pflug, G. Ch., Wozabal., D.: Ambiguity in portfolio selection, Quant. Financ., **7**, 435–442 (2007)

[28] Pichler, A., Xu, H.: Quantitative stability analysis for minimax distributionally robust risk optimization, To appear in Math. Program. (2018)

[29] Rachev., S. T.: Probability Metrics and the Stability of Stochastic Models. John Wiley & Sons, West Sussex, England (1991)

[30] Rockafellar, R., Wets, R. J-B.: Scenarios and policy aggregation in optimization under uncertainty, Math. Oper. Res. **16**, 119–147 (1991)

[31] Rockafellar, R., Sun, J.: Solving monotone stochastic variational inequalities and complementarity problems by progressive hedging, Math. Program. (2018) https://link.springer.com/article/10.1007/s10107-018-1251-y

[32] Römisch, W.: Stability of stochastic programming problems. Rusczyński, A., Shapiro, A., eds. *Stochastic Programming*, Handbook in OR & MS, Vol. 10, North-Holland Publishing Company, Amsterdam (2003)

[33] Ruszczyński, A.: Decomposition method. In Rusczyński, A., Shapiro, A., eds. *Stochastic Programming*, Handbook in OR & MS, Vol. 10, North-Holland Publishing Company, Amsterdam (2003)

[34] Ruszczyński, A.: Decomposition methods in stochastic programming, Math. Program. **79**, 333-353 (1997)

[35] Rahimian, H., Bayraksan, G., Homem-de-Mello, T.: Identifying effective scenarios in distributionally robust stochastic programs with total variation distance, Math. Program. (2018) https://link.springer.com/content/pdf/10.1007/s10107-017-1224-6

[36] Rusczyński, A., Shapiro, A.: *Stochastic Programming*, Handbook in OR & MS, Vol. 10, North-Holland Publishing Company, Amsterdam (2003)

[37] H. Scarf, A min-max solution of an inventory problem. K. S. Arrow, S. Karlin, H. E. Scarf. Studies in the Mathematical Theory of Inventory and Production, Stanford University Press, 201-209 (1958)

[38] Shapiro, A., Ahmed, S.: On a class of minimax stochastic programs, SIAM J. Optim., **14**, 1237–1249 (2004)

[39] Shapiro, A.: On duality theory of conic linear problems. In Goberna, M. A. López, M. A., eds., Semi-Infinite Programming. Nonconvex Optimization and Its Applications, vol 57. Springer, Boston, MA (2001)

[40] Shapiro, A., Dentcheva, D., Ruszczyński, A.: Lectures on Stochastic Programming: Modeling and Theory. SIAM, Philadelphia (2009)

[41] Sun, J., Liao, L. Z., Rodrigues, B.: Quadratic two-stage stochastic optimization with coherent measures of risk, Math. Program. **168**, 599–613 (2018)

[42] Sun, H., Xu, H.: Convergence analysis for distributionally robust optimization and equilibrium problems, Math. Oper. Res. **41**, 377-401 (2016)

[43] Weiss, P., Blanc-Feraud, L., Aubert, G.: Efficient schemes for total variation minimization under constraints in image processing, SIAM J. Sci. Comput. **31**, 2047–2080 (2009)

[44] Wiesemann, W., Kuhn, D. Sim, M.: Distributionally robust convex optimization, Oper. Res., **62**, 1358—376 (2014)

[45] Xu, H., Liu, Y., Sun., H.: Distributionally robust optimization with matrix moment constraints: Lagrange duality and cutting plane methods, Math. Program. **169**, 489–529 (2017)

[46] Zhang, Y., Jiang, R., Shen, S.: Ambiguous chance-constrained binary programs under mean-covariance information, SIAM J. Optim. **28**, 2922–2944 (2018)

[47] Zhang, X., Burger, M., Osher, S.: A unified primalšCdual algorithm framework based on Bregman iteration, J. Sci. Comput. **46**, 20–46 (2010)

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

[49] Zolotarev, V. M.: Probability metrics. Teoriya Veroyatnostei i ee Primeneniya, **28**, 264–287 (1983)

[50] Zhu, M., Chan, T. F.: An efficient primal dual hybrid gradient algorithm for total variation image restoration. CAM Report 08-34, UCLA, Los Angeles, CA (2008)

Appendix

Proof of Lemma 3.1. The inequality holds trivially if $Q \in \tilde{\mathcal{P}}$. So we only consider the case when if $Q \notin \tilde{\mathcal{P}}$. We proceed the proof in three steps.

Step 1. By the definition of the total variation norm (see [1, 4]), $||P|| = \sup_{\|\phi\|^* \leq 1} \langle P, \phi \rangle$, where $\|\cdot\|^*$ is the dual of norm $\|\cdot\|$. Moreover, by the definition of the total variation metric

$$\begin{aligned}
\mathsf{dl}_{TV}(Q, \tilde{\mathcal{P}}) &= \inf_{P \in \tilde{\mathcal{P}}} \mathsf{dl}_{TV}(Q, P) \\
&= \inf_{P \in \mathrm{cl}\{P \in \mathscr{P}(\tilde{\Xi}) : \mathbb{E}_P[\Psi_E] = \mu_E, \mathbb{E}_P[\Psi_I] \preceq \mu_I\}} \sup_{\|\phi\|^* \leq 1} \langle Q - P, \phi \rangle \\
&= \sup_{\|\phi\|^* \leq 1} \min_{P \in \mathrm{cl}\{P \in \mathscr{P}(\tilde{\Xi}) : \mathbb{E}_P[\Psi_E] = \mu_E, \mathbb{E}_P[\Psi_I] \preceq \mu_I\}} \langle Q - P, \phi \rangle \\
&= \sup_{\|\phi\|^* \leq 1} \inf_{P \in \{P \in \mathscr{P}(\tilde{\Xi}) : \mathbb{E}_P[\Psi_E] = \mu_E, \mathbb{E}_P[\Psi_I] \preceq \mu_I\}} \langle Q - P, \phi \rangle,
\end{aligned}$$

where $\mathrm{cl}(\cdot)$ denotes closure of a set under topology of weak convergence, $\langle P, \phi \rangle := \int_{\tilde{\Xi}} \phi(\xi) P(d\xi)$ and the exchange is justified by [15, Theorem 2] under our assumption that $\tilde{\mathcal{P}}$ is weakly compact. Note that we write $\langle P, \phi \rangle$ for $\mathbb{E}_P[\phi]$ in that later on we will relax P from a probability measure to a positive measure, and we will be able to see P clearly as a variable in the moment system and the moment system is linear in P. It is easy to observe that $\mathsf{dl}_{TV}(P, \mathcal{P}) \leq 2$. Moreover, under the Slater type condition (3.5), it follows by [39, Proposition 3.4] that

$$\begin{aligned}
&\inf_{P \in \{P \in \mathscr{P}(\tilde{\Xi}) : \mathbb{E}_P[\Psi_E] = \mu_E, \mathbb{E}_P[\Psi_I] \preceq \mu_I\}} \langle Q - P, \phi \rangle \\
&= \sup_{\lambda \in \Lambda, \lambda_0} \inf_{P \in \mathscr{M}_+(\tilde{\Xi})} \langle Q - P, \phi \rangle + \lambda \bullet (\langle P, \Psi \rangle - \mu) + \lambda_0 (\langle P, 1 \rangle - 1) \qquad (4.10) \\
&= \sup_{\lambda \in \Lambda, \lambda_0} \inf_{P \in \mathscr{M}_+(\tilde{\Xi})} \langle Q - P, \phi - \lambda \bullet \Psi - \lambda_0 \rangle + \langle Q, \lambda \bullet \Psi + \lambda_0 \rangle - \lambda \bullet \mu - \lambda_0,
\end{aligned}$$

where $\mathscr{M}_+(\tilde{\Xi})$ denotes the set of all positive measures defined on $\tilde{\Xi}$, $\Lambda := \{(\lambda_1, \cdots, \lambda_q) : \lambda_i \succeq 0, \text{ for } i = p+1, \ldots, q\}$, and $A \bullet B$ denotes the Frobenius product of the two matrices A and B. If there exists some ξ_i such that $\phi(\xi_i) - \lambda \bullet \Psi(\xi_i) - \lambda_0 > 0$, then the value of (4.10) is $-\infty$ because we can choose $P = \alpha \delta_{\xi_i}(\cdot)$, where $\delta_{\xi_i}(\cdot)$ denotes the Dirac probability measure at ξ_i, and drive α to $+\infty$. Thus we are left to consider that case with

$$\phi(\xi) - \lambda \bullet \Psi(\xi) - \lambda_0 \leq 0, \xi \in \tilde{\Xi}.$$

Consequently we can rewrite (4.10) as

$$\begin{aligned}
&\inf_{P \in \{P \in \mathscr{P}(\tilde{\Xi})_+ : \mathbb{E}_P[\Psi_E] = \mu_E, \mathbb{E}_P[\Psi_I] \preceq \mu_I\}} \langle Q - P, \phi \rangle \\
&= \sup_{\lambda \in \Lambda, \lambda_0} \inf_{P \in \mathscr{M}_+(\tilde{\Xi}), \xi \in \Xi_N} \langle Q - P, \phi - \lambda \bullet \Psi - \lambda_0 \rangle + \langle Q, \lambda \bullet \Psi + \lambda_0 \rangle - \lambda \bullet \mu - \lambda_0 \\
&= \sup_{\lambda \in \Lambda, \lambda_0} \langle Q, \phi - \lambda \bullet \Psi - \lambda_0 \rangle + \langle Q, \lambda \bullet \Psi + \lambda_0 \rangle - \lambda \bullet \mu - \lambda_0 \\
&= \sup_{\lambda \in \Lambda, \lambda_0} \langle Q, \phi \rangle - \lambda \bullet \mu - \lambda_0.
\end{aligned}$$

The second inequality is due to the fact that the optimum is attained at $P = 0$. Summarizing the discussions above, we arrive at

$$\begin{aligned}
\mathsf{dl}_{TV}(Q, \tilde{\mathcal{P}}) &= \left\{ \begin{array}{ll} \sup\limits_{\|\phi\|^* \le 1} \sup\limits_{\lambda \in \Lambda, \lambda_0} & \langle Q, \phi \rangle - \lambda \bullet \mu - \lambda_0 \\ \text{s.t.} & \phi(\xi) - \lambda \bullet \Psi(\xi) - \lambda_0 \le 0, \ \ \xi \in \tilde{\Xi} \end{array} \right. \\
&= \left\{ \begin{array}{ll} \sup\limits_{\lambda \in \Lambda, \lambda_0} \sup\limits_{\|\phi\|^* \le 1} & \langle Q, \phi \rangle - \lambda \bullet \mu - \lambda_0 \\ \text{s.t.} & \phi(\xi) - \lambda \bullet \Psi(\xi) - \lambda_0 \le 0, \ \ \xi \in \tilde{\Xi} \end{array} \right. \\
&= \left\{ \begin{array}{ll} \sup\limits_{\lambda \in \Lambda, \lambda_0} & \langle Q, \min\{\lambda \bullet \Psi(\xi) + \lambda_0, 1\} \rangle - \lambda \bullet \mu - \lambda_0 \\ \text{s.t.} & -1 - \lambda \bullet \Psi(\xi) - \lambda_0 \le 0, \ \ \text{a.e. } \xi \in \tilde{\Xi}. \end{array} \right.
\end{aligned} \tag{4.11}$$

The first equality is obtained by swapping the two "sup" operations because their optimal values are bounded. To see how the second equality holds, we may compare the optimal value of the second and third programs above. Let's denote the second one by (P) and the third one by (P'). Observe that (P') is transferred from (P) by replacing $\phi(\xi)$ in the objective by the largest possible value $\min\{\lambda \bullet \Psi(\xi) + \lambda_0, 1\}$ and in the constraint the smallest possible value -1 making the feasible set largest. This means the optimal value of (P) is less or equal to that of (P'). On the other hand, for any optimal solution (λ^*, λ_0^*) of (P'), let $\phi^* := \min\{\lambda^* \bullet \Psi(\xi) + \lambda_0^*, 1\}$. Then $(\lambda^*, \lambda_0^*, \phi^*)$ is a feasible solution of (P) which implies in turn the optimal value of (P') is less or equal to that of (P). Note also that the optimal value is $\mathsf{dl}_{TV}(Q, \tilde{\mathcal{P}}) \in [0, 2]$.

Step 2. We show that the optimization problem at the right hand side of (4.11) has a bounded optimal solution. Let

$$\mathcal{F} := \{(\lambda, \lambda_0) \in \Lambda \times \mathbb{R} : -1 - \lambda \bullet \Psi(\xi) - \lambda_0 \le 0, \ \ \forall \xi \in \tilde{\Xi}\}$$

denote the feasible set of (4.11) and

$$\mathcal{C} := \{(\lambda, \lambda_0) \in \Lambda \times \mathbb{R} : \|\lambda\| + |\lambda_0| = 1, -\lambda \bullet \Psi(\xi) - \lambda_0 \le 0, \ \ \forall \xi \in \tilde{\Xi}\}. \tag{4.12}$$

Then

$$\mathcal{F} = \{(0_q, -1) + t\mathcal{C} : t \in [0_q, +\infty)\}.$$

To see this, let $(\lambda, \lambda_0) \in \mathcal{F}$ and $t = \|\lambda\| + |\lambda_0 + 1|$. If $t = 0$, then $(0_q, -1) \in \{(0_q, -1) + t\mathcal{C} : t \in [0, +\infty)\}$. Consider the case that $t \neq 0$. Then it is easy to verify that $(\lambda, \lambda_0 + 1)/t \in \mathcal{C}$ and hence $(\lambda, \lambda_0) \in (0_q, -1) + t\mathcal{C}$. This shows $\mathcal{F} \subset \{(0_q, -1) + t\mathcal{C} : t \in [0, +\infty)\}$. The converse inclusion is obvious.

Note that $\mathcal{C} \neq \emptyset$. Otherwise we would have

$$\begin{aligned}
&\left\{ \begin{array}{ll} \sup\limits_{\lambda \in \Lambda, \lambda_0} & \langle Q, \min\{\lambda \bullet \Psi + \lambda_0, 1\} \rangle - \lambda \bullet \mu - \lambda_0 \\ \text{s.t.} & -1 - \lambda \bullet \Psi(\xi) - \lambda_0 \le 0, \ \ \text{a.e. } \xi \in \tilde{\Xi} \end{array} \right. \\
&\le \sup_{(\lambda, \lambda_0) = (0, -1)} \langle Q, \lambda \bullet \Psi \rangle - \lambda \bullet \mu = 0,
\end{aligned}$$

which implies $P \in \tilde{\mathcal{P}}$, a contradiction. In what follows, we consider the case when $\mathcal{C} \neq \emptyset$. From (4.12) we immediately have

$$-\lambda \bullet \langle P, \Psi \rangle - \lambda_0 \langle P, 1 \rangle \le 0, \forall (\lambda, \lambda_0) \in \mathcal{C}, P \in \mathscr{M}_+(\tilde{\Xi}). \tag{4.13}$$

On the other hand, by Assumption 3.1 and (3.6),

$$(0, 0_q) \in \text{int } [(\langle P, 1\rangle - 1, \langle P, \Psi(\xi)\rangle - \mu, -\{0\} \times \{0_p\} \times \mathcal{K}_-^{q-p} : P \in \mathscr{M}_+(\tilde{\Xi})]$$

and there exists a closed neighborhood of $(0, 0_q)$ with radius ϵ_0, denoted by $\epsilon_0\mathcal{B}$, such that

$$\epsilon_0\mathcal{B} \subset \text{int } [(\langle P, 1\rangle - 1, \langle P, \Psi(\xi)\rangle - \mu) - \{0\} \times \{0_p\} \times \mathcal{K}_-^{q-p} : P \in \mathscr{M}_+(\tilde{\Xi})].$$

Let $(\tilde{\lambda}, \tilde{\lambda}_0) \in \mathcal{C}$. Then there exists $\tilde{w} \in \epsilon_0\mathcal{B}$ (depending on $(\tilde{\lambda}, \tilde{\lambda}_0)$) such that $\langle \tilde{w}, (\tilde{\lambda}, \tilde{\lambda}_0)\rangle \leq -\epsilon_0$. In other words, there exist $\tilde{P} \in \mathscr{M}_+(\tilde{\Xi})$ and $\eta \in \mathcal{K}_-^{q-p}$ such that

$$\tilde{\lambda}_0(\langle \tilde{P}, 1\rangle - 1) + \tilde{\lambda} \bullet (\langle \tilde{P}, \Psi(\xi)\rangle - \mu) - \eta \bullet \tilde{\lambda}_I \leq -\epsilon_0. \tag{4.14}$$

Since $-\eta \bullet \tilde{\lambda}_I \geq 0$, we deduce from (4.13) and (4.14) that $-\tilde{\lambda}_0 - \tilde{\lambda} \bullet \mu \leq -\epsilon_0$. The inequality holds for every $(\tilde{\lambda}, \tilde{\lambda}_0) \in \mathcal{C}$. Note that for any $(\lambda, \lambda_0) \in \mathcal{F}$, we may write it in the form

$$(\lambda, \lambda_0) = (\hat{\lambda}, \hat{\lambda}_0) + t(\tilde{\lambda}, \tilde{\lambda}_0),$$

where $(\hat{\lambda}, \hat{\lambda}_0) = (0_q, -1)$, $(\tilde{\lambda}, \tilde{\lambda}_0) \in \mathcal{C}$ and $t \geq 0$. Observe that $\langle Q, \min\{\lambda \bullet \Psi + \lambda_0, 1\}\rangle \in [-1, 1]$ and

$$-\lambda \bullet \mu - \lambda_0 = -\mu \bullet \hat{\lambda} - \hat{\lambda}_0 - t(\mu \bullet \tilde{\lambda} + \tilde{\lambda}_0) \leq -\mu \bullet \hat{\lambda} - \hat{\lambda}_0 - t\epsilon_0 = 1 - t\epsilon_0.$$

Note that the optimal value of the optimization problem at the right hand side of (4.11) is positive, which implies for any optimal solution $(\lambda^*, \lambda_0^*) = (0_q, -1) + (\bar{\lambda}^*, \bar{\lambda}_0^*)$ with $(\bar{\lambda}^*, \bar{\lambda}_0^*) \in t^*\mathcal{C}$ we must have $1 - \mu \bullet 0_q - (-1) - t^*\epsilon_0 = 2 - t^*\epsilon_0 > 0$ or equivalently $t^* < \frac{2}{\epsilon_0}$. Let $t_1 = \frac{2}{\epsilon_0}$ and $\mathcal{F}_1 := \mathcal{F}_0 + \{t\mathcal{C} : t_1 \geq t \geq 0\}$. Based on the discussions above, we conclude that the optimization problem at the right hand side of (4.11) has and optimal solution in $\mathcal{F}_1$.

Step 3. Let $C_1 := \max_{(\lambda,\lambda_0)\in\mathcal{F}_1} \|\lambda\|$. Then

$$\begin{aligned}
d_{TV}(Q, \mathcal{P}) &= \sup_{(\lambda,\lambda_0)\in\mathcal{F}_1} \langle Q, \min\{\lambda \bullet \Psi(\xi(\omega)) + \lambda_0, 1\}\rangle - \lambda \bullet \mu - \lambda_0 \\
&\leq \sup_{(\lambda,\lambda_0)\in\mathcal{F}_1} \langle Q, \lambda \bullet \Psi(\xi(\omega)) + \lambda_0\rangle - \lambda \bullet \mu - \lambda_0 \\
&= \sup_{(\lambda,\lambda_0)\in\mathcal{F}_1} \lambda \bullet (\mathbb{E}_Q[\Psi(\xi)] - \mu) \\
&\leq \sup_{(\lambda,\lambda_0)\in\mathcal{F}_1} \sum_{i=1}^{p} \lambda_i(\mathbb{E}_Q[\Psi_i(\xi)] - \mu_i) + \sum_{i=p+1}^{q} \lambda_i(\mathbb{E}_Q[\Psi_i(\xi)] - \mu_i) \\
&\leq \sup_{(\lambda,\lambda_0)\in\mathcal{F}_1} \sum_{i=1}^{p} |\lambda_i||\mathbb{E}_Q[\Psi_i(\xi)] - \mu_i| + \sum_{i=p+1}^{q} \lambda_i(\mathbb{E}_Q[\Psi_i(\xi)] - \mu_i)_+ \\
&\leq C_1(\|(\mathbb{E}_Q[\Psi_E(\xi)] - \mu_E)\| + \|(\mathbb{E}_Q[\Psi_I(\xi)] - \mu_I)_+\|),
\end{aligned}$$

where $C_1 = \frac{2}{\epsilon_0} + 1$. The inequality also holds for the case when $\mathcal{C} = \emptyset$ because $\mathcal{F}_0 \subset \mathcal{F}_1$. Note that the above result holds for all $\widetilde{\mathcal{P}}$ when $\bar{\Xi} \subset \tilde{\Xi} \subset \Xi$, which include both continuous and discrete support set case, the proof is complete. ■