Operations Research Letters 46 (2018) 262–267

Contents lists available at ScienceDirect

Operations Research Letters

journal homepage: www.elsevier.com/locate/orl

Data-driven risk-averse stochastic optimization with Wasserstein metric

Chaoyue Zhao [a], Yongpei Guan [b,*]

[a] School of Industrial Engineering and Management, Oklahoma State University, Stillwater, OK 74074, United States
[b] Department of Industrial and Systems Engineering, University of Florida, Gainesville, FL 32611, United States

ARTICLE INFO

Article history:
Received 25 June 2016
Received in revised form 29 January 2018
Accepted 29 January 2018
Available online 9 February 2018

Keywords:
Stochastic optimization
Data-driven decision making
Wasserstein metric

ABSTRACT

In this paper, we study a data-driven risk-averse stochastic optimization approach with Wasserstein Metric for the general distribution case. By using the Wasserstein Metric, we can successfully reformulate the risk-averse two-stage stochastic optimization problem with distributional ambiguity to a traditional two-stage robust optimization problem. In addition, we derive the worst-case distribution and perform convergence analysis to show that the risk aversion of the proposed formulation vanishes as the size of historical data grows to infinity.

1. Introduction

The traditional two-stage stochastic optimization problem can be described as follows (cf. [3] and [15]):

$$\text{(SP)}\ \min_{x\in X} c^\top x + \mathbb{E}_{\mathbb{P}}[\mathcal{Q}(x,\xi)],$$

where the first-stage decision variable x is in a compact set X and the second-stage problem is

$$\mathcal{Q}(x,\xi) = \min_{y\in Y}\{d(\xi)^\top y : (\xi)x + By \geq b(\xi)\}, \tag{1}$$

where $\mathcal{Q}(x,\xi)$ is assumed continuous on ξ (e.g., when $A(\xi)$ and $b(\xi)$ are continuous on ξ) and the uncertain random variable ξ is defined on a probability space $(\Omega, \mathcal{F}, \mathbb{P})$, in which Ω is a compact convex sample space for ξ, $\mathcal{F}$ is a σ-algebra of Ω, and $\mathbb{P}$ is the associated given true probability distribution. SP has broad applications. However, there are challenges. For instance, in practice, the true distribution of the random parameters is usually unknown and hard to predict accurately and the inaccurate estimation of the true distribution may lead to biased solutions and make the solutions sub-optimal. Meanwhile, there are usually a series of historical data available for the unknown true distribution. To incorporate distribution ambiguity and utilize the historical data, we consider a data-driven risk-averse stochastic optimization formulation:

$$\text{(DD-SP)}\ \min_{x\in X} c^\top x + \max_{\hat{\mathbb{P}}\in\mathcal{D}} \mathbb{E}_{\hat{\mathbb{P}}}[\mathcal{Q}(x,\xi)],$$

where $\hat{\mathbb{P}}$ represents an unknown distribution in the given confidence set $\mathcal{D}$. This formulation allows distribution ambiguity and introduces a confidence set $\mathcal{D}$ to ensure that the true distribution $\mathbb{P}$ is within this set with a certain confidence level based on non-parametric statistics. In general, if we let $d_{\mathrm{M}}(\mathbb{P}_0, \hat{\mathbb{P}})$ be any distance measure between a reference distribution $\mathbb{P}_0$ and an unknown distribution $\hat{\mathbb{P}}$ and use θ, as a function of the size of historical data, to represent the corresponding distance, then the confidence set $\mathcal{D}$ can be represented as follows:

$$\mathcal{D} = \{\hat{\mathbb{P}} : d_{\mathrm{M}}(\mathbb{P}_0, \hat{\mathbb{P}}) \leq \theta\}.$$

This approach is related to the distributional robustness study (see, e.g., [5]) in the literature. To construct the confidence set, in our approach, the empirical distribution based on the historical data is utilized as the reference distribution. Then, the confidence set $\mathcal{D}$ is constructed by utilizing Wasserstein metric to define the distance between the reference distribution and the unknown true distribution. The utilization of Wasserstein metric for stochastic optimization was previously studied in [13,11,12], among others. The advantage for this approach is that the convergence properties hold. That is, as the size of historical data increases, we can show that the confidence set $\mathcal{D}$ shrinks with the same confidence level guarantee, and accordingly the true distribution will be "closer" to the reference distribution. Although research progress on Wasserstein metric has been made for this "distribution-based" approach for the discrete distribution case (see, e.g., [13]), the study for the continuous distribution case is more challenging and is very limited. The recent related study on distance measure for the general ϕ-divergence [1,9,10] can be utilized for the continuous distribution case by defining the distance between the true density

* Corresponding author.
E-mail addresses: chaoyue.zhao@okstate.edu (C. Zhao), guan@ise.ufl.edu (Y. Guan).

https://doi.org/10.1016/j.orl.2018.01.011
0167-6377/

and the reference density, which however could not guarantee the convergence properties to the empirical distribution. In this paper, by studying the Wasserstein metric, which ensures convergence, and deriving the corresponding convergence rate, our proposed approach can fit well in a data-driven risk-averse two-stage stochastic optimization framework.

The main results of this paper are available in [17] and [18]. An independent work is described in [6]. Our contributions can be summarized as follows:

1. We apply the Wasserstein metric to construct the confidence set in a data-driven risk-averse two-stage stochastic optimization framework for general distributions (including both discrete and continuous distribution cases) to solve optimization under uncertainty problems. In particular, we can successfully reformulate the two-stage risk-averse stochastic optimization problem to an explicit two-stage robust optimization problem.
2. Our proposed approach provides the closed-form expression of the worst-case distribution whose parameters can be obtained by solving a traditional two-stage robust optimization model.
3. We show the convergence property of the proposed data-driven approach by proving that as the size of historical data increases, the risk-averse stochastic optimization problem converges to the risk-neutral one, i.e., the traditional two-stage stochastic optimization problem.

2. Wasserstein metric and confidence set construction

In this section, we introduce the data-driven risk-averse two-stage stochastic optimization framework, for which instead of knowing the exact distribution of the random parameters, a series of historical data are observed. We first describe the Wasserstein metric. By using this metric, based on the observed historical data, we then build the reference distribution and the confidence set for the true probability distribution.

2.1. Wasserstein metric

The Wasserstein metric is defined as a distance function between two probability distributions on a given supporting space Ω. More specifically, given two probability distributions $\mathbb{P}_1$ and $\mathbb{P}_2$ on the supporting space Ω, the Wasserstein metric is defined as

$$d_{\mathrm{w}}(\mathbb{P}_1, \mathbb{P}_2) := \inf_{\pi}\{\mathbb{E}_{\pi}[\rho(X, Y)] : X \sim \mathbb{P}_1,\ Y \sim \mathbb{P}_2\}, \tag{2}$$

where $\rho(X, Y)$ is defined as the distance (continuous) between random variables X and Y, where X follows distribution $\mathbb{P}_1$ and Y follows distribution $\mathbb{P}_2$, and the infimum is taken over all joint distributions π with marginals $\mathbb{P}_1$ and $\mathbb{P}_2$. The Wasserstein metric is commonly used in many applications in transportation theory [14].

2.2. Confidence set construction

We use the empirical distribution as the reference distribution to estimate the true probability distribution. The empirical distribution function is a step function that jumps up by $1/N$ at each of the N independent and identically-distributed (i.i.d.) data points. That is, given N i.i.d. historical data samples $\xi^1, \xi^2, \ldots, \xi^N$, the empirical distribution is defined as

$$\mathbb{P}_{\mathrm{e}} = \frac{1}{N}\sum_{i=1}^{N}\delta_{\xi^i},$$

where δ_{ξ^i} is a Dirac measure (i.e., $\delta_{\xi^i}(A) = 1$ if $\xi^i \in A$ and $\delta_{\xi^i}(A) = 0$ otherwise for any set $A \in \sigma(\Omega)$). Based on the strong law of large numbers, it can be proved that the reference distribution $\mathbb{P}_{\mathrm{e}}$ pointwise converges to the true probability distribution $\mathbb{P}$ almost surely [16]. With the previously defined probability metric and reference probability distribution, we can now construct the confidence set for the true probability distribution $\mathbb{P}$. Intuitively, the more historical data observed, the "closer" the reference distribution is to the true distribution. If we use θ to represent the distance between the reference distribution and the true distribution, then the more historical data observed, the smaller the value of θ is, and the tighter the confidence set becomes. Therefore, the confidence set based on Wasserstein metric can be represented as follows:

$$\mathcal{D}_{\mathrm{w}} = \{\hat{\mathbb{P}} \in \mathcal{M}_+ : d_{\mathrm{w}}(\hat{\mathbb{P}}, \mathbb{P}_{\mathrm{e}}) \le \theta\},$$

where $\mathcal{M}_+$ represents the set of all probability measures on $(\Omega, \mathcal{F})$ and the value of θ depends on the size of historical data. More specifically, according to the definition of Wasserstein metric in (2), the confidence set can be written as

$$\mathcal{D}_{\mathrm{w}} = \left\{\hat{\mathbb{P}} \in \mathcal{M}_+ : \inf_{\pi}\{\mathbb{E}_{\pi}[\rho(Z, W)] : Z \sim \mathbb{P}_{\mathrm{e}}, W \sim \hat{\mathbb{P}}\} \le \theta\right\}. \tag{3}$$

We can further show that under the Wasserstein metric, the empirical distribution $\mathbb{P}_{\mathrm{e}}$ converges to the true distribution $\mathbb{P}$ exponentially fast. Several research works have discussed the convergence rates of the empirical distribution to the true distribution under Wasserstein metric [4,7]. For instance, as described in [7], if there exist an $\alpha > 1$ and a $\gamma > 0$ such that $\int_{\Omega} e^{\gamma|x|^{\alpha}}\mathbb{P}(dx) < \infty$, the following conclusion holds (We only list one setting here. The details on other settings are referred to Theorem 2 in [7]).

Proposition 1 (*[7]*). *For a general m-dimension (e.g., $m > 2$) supporting space Ω,*

$$P(d_{\mathrm{w}}(\mathbb{P}_e, \mathbb{P}) \le \theta) \ge 1 - C(exp(-cN\theta^m)1_{\{\theta\le 1\}} + exp(-cN\theta^{\alpha})1_{\{\theta>1\}}),$$

where N is the size of historical data, and C and c are positive constant numbers.

Based on Proposition 1, we can observe that if we set the confidence level of the confidence set $\mathcal{D}_{\mathrm{w}}$ as β, the corresponding θ can be approximated as follows:

$$\theta = \sqrt[m]{\log((1-\beta)/C)/(-cN)} \text{ if } \sqrt[m]{\log((1-\beta)/C)/(-cN)} \le 1 \text{ or } \theta = \max\{\sqrt[\alpha]{\log((1-\beta)/C)/(-cN)}, 1\}. \tag{4}$$

3. Reformulation to a two-stage robust optimization problem

In this section, we first derive the reformulation of the worst-case expectation $\max_{\hat{\mathbb{P}}\in\mathcal{D}_{\mathrm{w}}}\mathbb{E}_{\hat{\mathbb{P}}}[\mathcal{Q}(x, \xi)]$, and then obtain the reformulation of problem DD-SP correspondingly. In addition, we derive the worst-case distribution corresponding to the Wasserstein metric. Before describing our main result, we make the following assumption:

Assumption 1. DD-SP has relatively complete recourse and is bounded, i.e., $\sup_{\xi\in\Omega}|Q(x, \xi)| < \infty$ for each $x \in X$.

Proposition 2. *Assuming that there are N historical data samples $\xi^1, \xi^2, \ldots, \xi^N$ which are i.i.d. drawn from the true distribution $\mathbb{P}$, for any fixed first-stage decision x, we have*

$$\max_{\hat{\mathbb{P}}\in\mathcal{D}_{\mathrm{w}}}\mathbb{E}_{\hat{\mathbb{P}}}[\mathcal{Q}(x, \xi)] = \min_{\beta\ge 0}\left\{\frac{1}{N}\sum_{i=1}^{N}\max_{\xi\in\Omega}\{\mathcal{Q}(x, \xi) - \beta\rho(\xi, \xi^i)\} + \theta\beta\right\}. \tag{5}$$

Proof. As indicated in Section 2.2, if we have N historical data samples $\xi^1, \xi^2, \ldots, \xi^N$, the reference distribution $\mathbb{P}_e$ can be defined as the empirical distribution, i.e., $\mathbb{P}_e = \frac{1}{N}\sum_{i=1}^{N}\delta_{\xi^i}$. In addition, the confidence set $\mathcal{D}_w$ is defined the same way as indicated in (3). Then, we can claim that, if $\hat{\mathbb{P}} \in \mathcal{D}_w$, then $\forall \epsilon > 0$, there exists a joint distribution π such that Z follows distribution $\mathbb{P}_e$, W follows distribution $\hat{\mathbb{P}}$, and $\mathbb{E}_\pi[\rho(Z, W)] \le \theta+\epsilon$. Based on the definition of $\mathbb{E}_\pi[\rho(Z, W)]$, we can obtain the following reformulation of $\mathbb{E}_\pi[\rho(Z, W)]$:

$$\begin{aligned}\mathbb{E}_\pi[\rho(Z,W)] &= \int_{z\in\Omega}\int_{\xi\in\Omega}\rho(\xi,z)\pi(d\xi,dz)\\ &= \sum_{i=1}^{N}\int_{\xi\in\Omega}p_i^0\rho(\xi,\xi^i)\mathbb{P}_i(d\xi)\\ &= \frac{1}{N}\sum_{i=1}^{N}\int_{\xi\in\Omega}\rho(\xi,\xi^i)\mathbb{P}_i(d\xi), \qquad (6)\end{aligned}$$

where p_i^0 represents the probability for the event $Z = \xi^i$ and $\mathbb{P}_i$ is the conditional distribution of W when $Z = \xi^i$. Eq. (6) holds since according to the definition of $\mathbb{P}_e$, $p_i^0 = 1/N$ for each $i = 1, \ldots, N$. For notation brevity, we let $\rho^i(\xi) = \rho(\xi, \xi^i)$. Then the second-stage problem $\max_{\hat{\mathbb{P}}\in\mathcal{D}_w}\mathbb{E}_{\hat{\mathbb{P}}}[\mathcal{Q}(x,\xi)]$ of DD-SP can be reformulated as follows:

$$\begin{aligned}\max_{\hat{\mathbb{P}},\mathbb{P}_i\in\mathcal{M}_-}\ & \int_{\xi\in\Omega}\mathcal{Q}(x,\xi)\hat{\mathbb{P}}(d\xi)\\ s.t.\ \ & \hat{\mathbb{P}} = \frac{1}{N}\sum_{i=1}^{N}\mathbb{P}_i, \qquad (7)\\ & \int_{\xi\in\Omega}\mathbb{P}_i(d\xi) = 1,\ \ \forall i, \qquad (8)\\ & \frac{1}{N}\sum_{i=1}^{N}\int_{\xi\in\Omega}\rho^i(\xi)\mathbb{P}_i(d\xi) \le \theta+\epsilon, \qquad (9)\end{aligned}$$

where $\mathcal{M}_-$ represents the set of all (non-negative) measures on $(\Omega, \mathcal{F})$. Constraints (7) and (8) are based on the properties of conditional distribution and constraint (9) follows the definition of $\mathcal{D}_w$ and Eq. (6). Note here that this reformulation holds for any $\epsilon > 0$. By substituting constraint (7) into the objective function, we can obtain its equivalent formulation as follows:

$$\begin{aligned}\max_{\mathbb{P}_i\in\mathcal{M}_-}\ & \frac{1}{N}\sum_{i=1}^{N}\int_{\xi\in\Omega}\mathcal{Q}(x,\xi)\mathbb{P}_i(d\xi)\\ \text{(PEQ)}\quad s.t.\ \ & \int_{\xi\in\Omega}\mathbb{P}_i(d\xi) = 1,\ \ \forall i, \qquad (10)\\ & \frac{1}{N}\sum_{i=1}^{N}\int_{\xi\in\Omega}\rho^i(\xi)\mathbb{P}_i(d\xi) \le \theta+\epsilon. \qquad (11)\end{aligned}$$

Note here that N is a finite number and we can switch the integration and summation in the objective function. In addition, the above formulation PEQ has at least one feasible solution with $\mathbb{P}_i = \delta_{\xi^i}, i = 1, \ldots, N$, which is a relative interior point because $\theta, \epsilon > 0$ and constraint (11) is not binding. Thus, the Slater's condition holds. Meanwhile, since $\mathcal{Q}(x, \xi)$ is assumed bounded above following Assumption 1, PEQ is bounded above. Therefore, strong duality holds. Thus, we can consider its Lagrangian dual problem that can be written as follows:

$$\begin{aligned}L(\lambda_i,\beta) = \max_{\mathbb{P}_i\in\mathcal{M}_-}\ & \frac{1}{N}\sum_{i=1}^{N}\int_{\xi\in\Omega}(\mathcal{Q}(x,\xi)-N\lambda_i-\beta\rho^i(\xi))\mathbb{P}_i(d\xi)\\ & +\sum_{i=1}^{N}\lambda_i+(\theta+\epsilon)\beta,\end{aligned}$$

where λ_i and $\beta \ge 0$ are dual variables of constraints (10) and (11), respectively. The dual problem then is

$$\min_{\beta\ge0,\lambda_i} L(\lambda_i,\beta).$$

Next, we argue that $\forall \xi \in \Omega$, $\mathcal{Q}(x,\xi) - N\lambda_i - \beta\rho^i(\xi) \le 0$. If this argument does not hold, then there exists a ξ_0 such that $\mathcal{Q}(x,\xi_0) - N\lambda_i - \beta\rho^i(\xi_0) > 0$. It means that there exists a strict positive constant σ, such that $\mathcal{Q}(x,\xi_0) - N\lambda_i - \beta\rho^i(\xi_0) > \sigma$. Based on the assumption that $\mathcal{Q}(x,\xi)$ is continuous on ξ and the fact that the distance function $\rho^i(\xi)$ is continuous on ξ, $\mathcal{Q}(x,\xi) - N\lambda_i - \beta\rho^i(\xi)$ is continuous on ξ. Thus, if $\mathcal{Q}(x,\xi_0) - N\lambda_i - \beta\rho^i(\xi_0) > \sigma$, there exists a small ball $B(\xi_0,\epsilon') \subseteq \Omega$ with a strictly positive measure, such that $\mathcal{Q}(x,\xi) - N\lambda_i - \beta\rho^i(\xi) > \sigma$ for $\forall \xi \in B(\xi_0,\epsilon')$. Accordingly, we can let $\mathbb{P}_i$ be continuous with its density function arbitrary large when $\xi \in B(\xi_0,\epsilon')$, then $L(\lambda_i,\beta)$ is unbounded, which leads to a contradiction to the strong duality corresponding to PEQ, which is bounded above. Hence, the argument

$$\mathcal{Q}(x,\xi) - N\lambda_i - \beta\rho^i(\xi) \le 0 \text{ for all } \xi \in \Omega \qquad (12)$$

holds. In this case,

$$\begin{aligned}\max_{\mathbb{P}_i\in\mathcal{M}_-}\ & \frac{1}{N}\sum_{i=1}^{N}\int_{\xi\in\Omega}(\mathcal{Q}(x,\xi)-N\lambda_i-\beta\rho^i(\xi))\mathbb{P}_i(d\xi)\\ & +\sum_{i=1}^{N}\lambda_i+(\theta+\epsilon)\beta = \sum_{i=1}^{N}\lambda_i+(\theta+\epsilon)\beta \qquad (13)\end{aligned}$$

with optimal solutions satisfying

$$\int_{\xi\in\Omega}(\mathcal{Q}(x,\xi)-N\lambda_i-\beta\rho^i(\xi))\mathbb{P}_i(d\xi) = 0,\ i = 1,\ldots,N. \qquad (14)$$

Note here that we can let $\mathbb{P}_i(d\xi) = 0$ when $\mathcal{Q}(x,\xi) - N\lambda_i - \beta\rho^i(\xi) < 0$ to make (14) hold since constraint (10) is relaxed. Then, the dual formulation can be reformulated as follows:

$$\begin{aligned}\min_{\beta\ge0,\lambda_i}\ & \sum_{i=1}^{N}\lambda_i+(\theta+\epsilon)\beta\\ s.t.\ \ & \mathcal{Q}(x,\xi)-N\lambda_i-\beta\rho^i(\xi)\le0,\ \forall\xi\in\Omega,\\ & \forall i=1,\ldots,N. \qquad (15)\end{aligned}$$

From the above formulation, it is easy to observe that the optimal solution λ_i should satisfy

$$\lambda_i = \frac{1}{N}\max_{\xi\in\Omega}\{\mathcal{Q}(x,\xi)-\beta\rho^i(\xi)\}, \qquad (16)$$

and therefore the worst-case expectation $\max_{\hat{\mathbb{P}}\in\mathcal{D}_w}\mathbb{E}_{\hat{\mathbb{P}}}[\mathcal{Q}(x,\xi)]$ is equivalent to

$$\min_{\beta\ge0}\left\{\frac{1}{N}\sum_{i=1}^{N}\max_{\xi\in\Omega}\{\mathcal{Q}(x,\xi)-\beta\rho^i(\xi)\}+(\theta+\epsilon)\beta\right\}. \qquad (17)$$

Note here that reformulation (17) holds for $\forall \epsilon > 0$ and is continuous on ϵ. Thus, reformulation (17) holds for $\epsilon = 0$, which immediately leads to the following reformulation of $\max_{\hat{\mathbb{P}}\in\mathcal{D}_w}\mathbb{E}_{\hat{\mathbb{P}}}[\mathcal{Q}(x,\xi)]$:

$$\begin{aligned}&\max_{\hat{\mathbb{P}}\in\mathcal{D}_w}\mathbb{E}_{\hat{\mathbb{P}}}[\mathcal{Q}(x,\xi)]\\ &= \min_{\beta\ge0}\left\{\frac{1}{N}\sum_{i=1}^{N}\max_{\xi\in\Omega}\{\mathcal{Q}(x,\xi)-\beta\rho^i(\xi)\}+\theta\beta\right\}. \quad \square\end{aligned}$$

Note here that the reformulation of $\max_{\hat{\mathbb{P}}\in\mathcal{D}_w}\mathbb{E}_{\hat{\mathbb{P}}}[\mathcal{Q}(x,\xi)]$ depends on θ. By defining

$$g(x,\theta) = \min_{\beta\ge0}\left\{\frac{1}{N}\sum_{i=1}^{N}\max_{\xi\in\Omega}\{\mathcal{Q}(x,\xi)-\beta\rho^i(\xi)\}+\theta\beta\right\},$$

we have the following proposition holds.

Proposition 3. *For a given first-stage decision x, the function $g(x,\theta)$ is monotonically increasing in θ. In addition, $g(x,0)=\mathbb{E}_{\mathbb{P}_e}[\mathcal{Q}(x,\xi)]$ and $\lim_{\theta\to\infty} g(x,\theta)=\max_{\xi\in\Omega}\mathcal{Q}(x,\xi)$.*

Proof. The monotone property is obvious, since it can be easily observed that as θ decreases, the confidence set $\mathcal{D}_w$ tends to shrink and the worst-case expected value $\max_{\hat{\mathbb{P}}\in\mathcal{D}_w}\mathbb{E}_{\hat{\mathbb{P}}}[\mathcal{Q}(x,\xi)]$ does not increase. Therefore, the reformulation $g(x,\theta)=\max_{\hat{\mathbb{P}}\in\mathcal{D}_w}\mathbb{E}_{\hat{\mathbb{P}}}[\mathcal{Q}(x,\xi)]$ is monotonically increasing in θ.

When $\theta=0$, we have $d_w(\mathbb{P}_e,\hat{\mathbb{P}})=0$. According to the properties of metrics, we have $\hat{\mathbb{P}}=\mathbb{P}_e$. That means the confidence set $\mathcal{D}_w$ is a singleton, only containing $\mathbb{P}_e$. Therefore, we have $g(x,0)=\mathbb{E}_{\mathbb{P}_e}[\mathcal{Q}(x,\xi)]$.

When $\theta\to\infty$, it is clear to observe that in the optimal solution we have $\beta=0$ since $\rho^i(\xi)<\infty$ following the assumption that Ω is compact. Then we have

$$\min_{\beta\geq 0}\left\{\frac{1}{N}\sum_{i=1}^{N}\max_{\xi\in\Omega}\{\mathcal{Q}(x,\xi)-\beta\rho^i(\xi)\}+\theta\beta\right\} =\frac{1}{N}\sum_{i=1}^{N}\max_{\xi\in\Omega}\mathcal{Q}(x,\xi)=\max_{\xi\in\Omega}\mathcal{Q}(x,\xi), \tag{18}$$

which indicates that $\lim_{\theta\to\infty} g(x,\theta)=\max_{\xi\in\Omega}\mathcal{Q}(x,\xi)$. Thus, the claim holds. □

From Proposition 3 we can observe that, the data-driven risk-averse stochastic optimization is less conservative than the traditional robust optimization and more conservative than the traditional stochastic optimization. In addition, by letting

$$\bar{\theta}=\operatorname*{argmin}_{\theta\geq 0}\left\{\min_{\beta\geq 0}\left\{\frac{1}{N}\sum_{i=1}^{N}\max_{\xi\in\Omega}\{\mathcal{Q}(x,\xi)-\beta\rho^i(\xi)\}+\theta\beta\right\} =\max_{\xi\in\Omega}\mathcal{Q}(x,\xi)\right\}, \tag{19}$$

following the monotone property described in Proposition 3, we have

$$\beta^*>0 \text{ when } \theta<\bar{\theta}. \tag{20}$$

Otherwise, if $\beta^*=0$, then $g(x,\theta)=\frac{1}{N}\sum_{i=1}^{N}\max_{\xi\in\Omega}\{\mathcal{Q}(x,\xi)-\beta\rho^i(\xi)\}+\theta\beta=\frac{1}{N}\sum_{i=1}^{N}\max_{\xi\in\Omega}\mathcal{Q}(x,\xi)=\max_{\xi\in\Omega}\mathcal{Q}(x,\xi)$ based on (18), which contradicts with $g(x,\theta)<g(x,\bar{\theta})$ based on (19) and the monotone property described in Proposition 3. Meanwhile, it can be observed that $\beta^*=0$ is a candidate solution when $\theta\geq\bar{\theta}$. Considering the fact that the data-driven risk-averse stochastic optimization is equivalent to (not less conservative than) the traditional robust optimization when $\theta\geq\bar{\theta}$, we analyze the worst-case distribution for the case $\theta<\bar{\theta}$ as follows.

Proposition 4. *When $Q(x,\xi)$ is concave with respect to ξ (e.g., $d(\xi)$ is affinely dependent on ξ and $A(\xi)$, $b(\xi)$ are deterministic), $\rho^i(\xi)$ is a strictly convex function of ξ (e.g., in the form of L_2-norm), and β^* is the unique solution for (5) as described in Proposition 2, there exists a worst-case distribution for $\max_{\hat{\mathbb{P}}\in\mathcal{D}_w}\mathbb{E}_{\hat{\mathbb{P}}}[\mathcal{Q}(x,\xi)]$ in the following form:*

$$\hat{\mathbb{P}}^*=\frac{1}{N}\sum_{i=1}^{N}\delta_{\xi_*^i}, \tag{21}$$

where ξ_*^i is the optimal solution of $\max_{\xi\in\Omega}\{\mathcal{Q}(x,\xi)-\beta^*\rho^i(\xi)\}$.

Proof. Based on (13), if $\hat{\mathbb{P}}$ is a worst-case distribution, then we must have

$$\frac{1}{N}\sum_{i=1}^{N}\int_{\xi\in\Omega}(\mathcal{Q}(x,\xi)-N\lambda_i^*-\beta^*\rho^i(\xi))\hat{\mathbb{P}}_i(d\xi)=0, \tag{22}$$

where $\hat{\mathbb{P}}=\frac{1}{N}\sum_{i=1}^{N}\mathbb{P}_i$ and (λ_i^*,β^*) is the optimal solution to the dual problem. Therefore, formulation (22) provides a necessary condition for any worst-case distribution.

As indicated in (12), we have $\mathcal{Q}(x,\xi)-N\lambda_i{}^*-\beta^*\rho^i(\xi)\leq 0$, $\forall\xi\in\Omega$, $\forall i$. Therefore,

$$(\mathcal{Q}(x,\xi)-N\lambda_i{}^*-\beta^*\rho^i(\xi))\mathbb{P}_i(d\xi)=0,\quad \forall\xi\in\Omega,\forall i.$$

Thus, for each i, since $\mathbb{P}_i(d\xi)\geq 0$, $\mathbb{P}_i(d\xi)$ can be non-zero only when $\mathcal{Q}(x,\xi)-N\lambda_i{}^*-\beta^*\rho^i(\xi)=0$. Since $\lambda_i^*=\frac{1}{N}\max_{\xi\in\Omega}\{\mathcal{Q}(x,\xi)-\beta^*\rho^i(\xi)\}$, $\mathcal{Q}(x,\xi)-N\lambda_i{}^*-\beta^*\rho^i(\xi)=0$ only when ξ is an optimal solution of $\max_{\xi\in\Omega}\{\mathcal{Q}(x,\xi)-\beta^*\rho^i(\xi)\}$. Since $\beta^*>0$ when $\theta<\bar{\theta}$ following (20), $\mathcal{Q}(x,\xi)-\beta^*\rho^i(\xi)$ is strictly concave with respect to ξ, the optimal solution to $\max_{\xi\in\Omega}\{\mathcal{Q}(x,\xi)-\beta^*\rho^i(\xi)\}$ is unique. Therefore, the following distribution (indicated as $\hat{\mathbb{P}}^i$) is the only distribution satisfying (22):

$$\hat{\mathbb{P}}^i=\delta_{\xi_*^i}, \tag{23}$$

where ξ_*^i is the optimal solution of $\max_{\xi\in\Omega}\{\mathcal{Q}(x,\xi)-\beta^*\rho^i(\xi)\}$. Since the ambiguity set $\mathcal{D}_w$ is compact and Ω is bounded, convergence under the Wasserstein metric implies weak convergence of the distributions in $\mathcal{D}_w$ as described in [8]. In addition, $Q(x,\xi)$ is assumed continuous on ξ and bounded following Assumption 1, based on the Helly–Bray Theorem [2], we can observe that the worst-case distribution always exists. Thus, (23) is the unique distribution that satisfies the necessary condition for any worst-case distribution. Accordingly, following (7), the following distribution serves as the worst-case distribution:

$$\hat{\mathbb{P}}^*=\frac{1}{N}\sum_{i=1}^{N}\hat{\mathbb{P}}^i=\frac{1}{N}\sum_{i=1}^{N}\delta_{\xi_*^i}.\quad\square$$

Based on Propositions 2 and 4, we can easily derive the following theorem.

Theorem 1. *The problem DD-SP under $\mathcal{D}_w$ is equivalent to the following two-stage robust optimization problem:*

$$\text{(RDD-SP)}\quad \min_{x\in X,\beta\geq 0}\ c^\top x+\theta\beta+\frac{1}{N}\sum_{i=1}^{N}\max_{\xi\in\Omega}\left\{\mathcal{Q}(x,\xi)-\beta\rho^i(\xi)\right\}. \tag{24}$$

Meanwhile, under the conditions described in Proposition 4, there exists a worst-case distribution in the following form:

$$\hat{\mathbb{P}}^*=\frac{1}{N}\sum_{i=1}^{N}\delta_{\xi_*^i},$$

where ξ_*^i is the optimal solution of $\max_{\xi\in\Omega}\{\mathcal{Q}(x,\xi)-\beta^*\rho^i(\xi)\}$ and β^* is the unique optimal solution for (5) as described in Proposition 2.

Remark 1. Note here that computational approaches can be derived to detect if β^* is the unique optimal solution for (5) as described in Proposition 2. For the existence and closed-form of the worst-case distribution for the general cases, one can refer to [6], which provides the detailed description.

Remark 2. In general, RDD-SP is a two-stage robust optimization problem. The solution approaches can be different based on the definition of the sample space Ω and the distance function ρ.

For instance, if Ω is convex and ρ is defined as L_1- or L_∞-norm, then RDD-SP can be further reformulated as a semi-infinite linear program and accordingly the Benders decomposition algorithm as described in [17] can be applied. If Ω is convex and ρ is defined as L_2-norm, which can be derived from inner products, then RDD-SP can be further reformulated as a finite convex program as described in [6].

4. Convergence analysis

In this section, we examine the convergence properties of DD-SP to SP as the size of historical data increases. We demonstrate that as the confidence set $\mathcal{D}_w$ shrinks with more historical data observed, the risk-averse problem DD-SP converges to the risk-neutral SP. We first assume that for each $x \in X$, there exists a constant number $L > 0$ such that $|\mathcal{Q}(x,\xi_1) - \mathcal{Q}(x,\xi_2)|/\rho(\xi_1,\xi_2) < L < \infty$, $\forall \xi_1, \xi_2 \in \Omega$ and $\xi_1 \neq \xi_2$. We now analyze the convergence property of the second-stage objective value, which can be shown as follows:

Proposition 5. *Corresponding to each predefined confidence level β and a given first-stage decision x, as the size of historical data $N \to \infty$, we have the distance value $\theta \to 0$ and the corresponding risk-averse second-stage objective value* $\lim_{N\to\infty} \max_{\hat{\mathbb{P}}\in\mathcal{D}_w} \mathbb{E}_{\hat{\mathbb{P}}}[\mathcal{Q}(x,\xi)] = \mathbb{E}_{\mathbb{P}}[\mathcal{Q}(x,\xi)]$.

Proof. First, following (4), it is obvious that $\theta \to 0$ as $N \to \infty$. Meanwhile, following Proposition 2, we have

$$\max_{\hat{\mathbb{P}}\in\mathcal{D}_w} \mathbb{E}_{\hat{\mathbb{P}}}[\mathcal{Q}(x,\xi)] = \min_{\beta\geq 0}\left\{\theta\beta + \frac{1}{N}\sum_{i=1}^{N}\max_{\xi\in\Omega}\{\mathcal{Q}(x,\xi) - \beta\rho^i(\xi)\}\right\}. \tag{25}$$

Therefore, in the following part, we only need to prove

$$\lim_{N\to\infty}\min_{\beta\geq 0}\left\{\theta\beta + \frac{1}{N}\sum_{i=1}^{N}\max_{\xi\in\Omega}\{\mathcal{Q}(x,\xi) - \beta\rho^i(\xi)\}\right\} \leq \mathbb{E}_{\mathbb{P}}[\mathcal{Q}(x,\xi)], \tag{26}$$

and

$$\lim_{N\to\infty}\max_{\hat{\mathbb{P}}\in\mathcal{D}_w}\mathbb{E}_{\hat{\mathbb{P}}}[\mathcal{Q}(x,\xi)] \geq \lim_{N\to\infty}\mathbb{E}_{\mathbb{P}_e}[\mathcal{Q}(x,\xi)] = \mathbb{E}_{\mathbb{P}}[\mathcal{Q}(x,\xi)]. \tag{27}$$

Note here that (27) holds following the Helly–Bray Theorem [2] because $\mathbb{P}_e$ converges weakly to $\mathbb{P}$ as $N \to \infty$ under the Wasserstein Metric [8] and $\mathcal{Q}(x,\xi)$ is bounded and continuous in ξ. We only need to prove (26). Since

$$\lim_{N\to\infty}\min_{\beta\geq 0}\left\{\theta\beta + \frac{1}{N}\sum_{i=1}^{N}\max_{\xi\in\Omega}\{\mathcal{Q}(x,\xi) - \beta\rho^i(\xi)\}\right\} \leq \min_{\beta\geq 0}\limsup_{N\to\infty}\left\{\theta\beta + \frac{1}{N}\sum_{i=1}^{N}\max_{\xi\in\Omega}\{\mathcal{Q}(x,\xi) - \beta\rho^i(\xi)\}\right\}, \tag{28}$$

we only need to show

$$\min_{\beta\geq 0}\limsup_{N\to\infty}\left\{\theta\beta + \frac{1}{N}\sum_{i=1}^{N}\max_{\xi\in\Omega}\{\mathcal{Q}(x,\xi) - \beta\rho^i(\xi)\}\right\} \leq \mathbb{E}_{\mathbb{P}}[\mathcal{Q}(x,\xi)]. \tag{29}$$

Following Assumption 1, there exists a constant number $M > 0$ such that for any given x,

$$-M \leq \mathcal{Q}(x,\xi) \leq M, \quad \forall \xi \in \Omega. \tag{30}$$

In addition, we have

$$0 \leq \rho(\xi,z) \leq B, \quad \forall \xi, z \in \Omega, \tag{31}$$

where B is the diameter of Ω. Therefore, for any $\beta \geq 0$, based on (30) and (31), we have

$$\max_{\xi\in\Omega}\{\mathcal{Q}(x,\xi) - \beta\rho(\xi,z)\} \leq \max_{\xi\in\Omega}\{\mathcal{Q}(x,\xi)\} \leq M,$$

and

$$\max_{\xi\in\Omega}\{\mathcal{Q}(x,\xi) - \beta\rho(\xi,z)\} \geq \max_{\xi\in\Omega}\{\mathcal{Q}(x,\xi) - \beta B\} \geq -M - \beta B,$$

which means $\max_{\xi\in\Omega}\{\mathcal{Q}(x,\xi) - \beta\rho(\xi,z)\}$ is bounded for $\forall z \in \Omega$. Therefore, we have

$$\begin{aligned}
&\limsup_{N\to\infty}\left\{\theta\beta + \frac{1}{N}\sum_{i=1}^{N}\max_{\xi\in\Omega}\{\mathcal{Q}(x,\xi) - \beta\rho^i(\xi)\}\right\}\\
&= \limsup_{N\to\infty}\left\{\theta\beta + \mathbb{E}_{\mathbb{P}_e}\max_{\xi\in\Omega}\{\mathcal{Q}(x,\xi) - \beta\rho(\xi,z)\}\right\}\\
&= \lim_{N\to\infty}\theta\beta + \mathbb{E}_{\mathbb{P}}\max_{\xi\in\Omega}\{\mathcal{Q}(x,\xi) - \beta\rho(\xi,z)\}\\
&= \mathbb{E}_{\mathbb{P}}[\max_{\xi\in\Omega}\{\mathcal{Q}(x,\xi) - \beta\rho(\xi,z)\}],
\end{aligned} \tag{32}$$

where the first equality holds following the definition of $\mathbb{P}_e$, the second equality holds following the Helly–Bray Theorem [2], and the third equality holds because $\theta \to 0$ as $N \to \infty$ following (4).

Now we only need to show that for a given first-stage decision x and any true distribution $\mathbb{P}$, we have

$$\min_{\beta\geq 0}\mathbb{E}_{\mathbb{P}}[\max_{\xi\in\Omega}\{\mathcal{Q}(x,\xi) - \beta\rho(\xi,z)\}] \leq \mathbb{E}_{\mathbb{P}}[\mathcal{Q}(x,z)]. \tag{33}$$

Since $\min_{\beta\geq 0}\mathbb{E}_{\mathbb{P}}[\max_{\xi\in\Omega}\{\mathcal{Q}(x,\xi) - \beta\rho(\xi,z)\}] \leq \limsup_{\beta\to+\infty}\mathbb{E}_{\mathbb{P}}[\max_{\xi\in\Omega}\{\mathcal{Q}(x,\xi) - \beta\rho(\xi,z)\}]$, we only need to prove

$$\limsup_{\beta\to+\infty}\mathbb{E}_{\mathbb{P}}[\max_{\xi\in\Omega}\{\mathcal{Q}(x,\xi) - \beta\rho(\xi,z)\}] \leq \mathbb{E}_{\mathbb{P}}[\mathcal{Q}(x,z)]. \tag{34}$$

First of all, we notice that $\xi^*(z) = z$ in (34) makes it tight. Now we prove by contradiction that for $\beta > L$, the optimal solution for the maximization problem in (34) is $\xi^*(z) = z$. Otherwise, if $\xi^*(z) = \hat{z} \neq z$, $[\mathcal{Q}(x,z) - \beta\rho(z,z)] - [\mathcal{Q}(x,\hat{z}) - \beta\rho(\hat{z},z)] = \mathcal{Q}(x,z) - \mathcal{Q}(x,\hat{z}) + \beta\rho(\hat{z},z) > (\beta - L)\rho(\hat{z},z) > 0$, where the inequality follows the assumption $|\mathcal{Q}(x,\xi_1) - \mathcal{Q}(x,\xi_2)|/\rho(\xi_1,\xi_2) < L$. This is a contradiction. Thus,

$$\begin{aligned}
&\limsup_{\beta\to+\infty}\mathbb{E}_{\mathbb{P}}[\max_{\xi\in\Omega}\{\mathcal{Q}(x,\xi) - \beta\rho(\xi,z)\}]\\
&= \limsup_{\beta\to+\infty}\mathbb{E}_{\mathbb{P}}[\mathcal{Q}(x,z)] = \mathbb{E}_{\mathbb{P}}[\mathcal{Q}(x,z)].
\end{aligned}$$

The proof is completed. □

Now we prove that the objective value of DD-SP converges to that of SP as the size of historical data samples increases to infinity.

Theorem 2. *Corresponding to each predefined confidence level β, as the size of historical data increases to infinity, the optimal objective value of the data-driven risk-averse stochastic optimization problem uniformly converges to that of the traditional two-stage risk-neutral stochastic optimization problem.*

Proof. First, notice that $N \to \infty$ is equivalent to $\theta \to 0$ following (4). We only need to prove $\lim_{\theta\to 0}\psi(\theta) = \psi(0)$, where $\psi(\theta)$ represents the optimal objective value of DD-SP with the distance value θ and $\psi(0)$ represents the optimal objective value of SP. Meanwhile, for the convenience of analysis, corresponding to each given first-stage decision x for DD-SP with the distance value θ, we denote $V_\theta(x)$ as its corresponding objective value and $V_0(x)$ as the objective value of SP.

Now consider the DD-SP problem with the distance value θ. Denote the first-stage decision of SP as x^* and the first-stage decision of DD-SP as x^*_θ. According to Proposition 5, $V_\theta(x)$ converges pointwise to $V_0(x)$. In addition, we notice that X is compact, $V_\theta(x)$ and $V_0(x)$ are continuous in x (this can be proved following the fact that the worst-case probability distribution exists under our setting and by applying Lebesgue dominated convergence theorem based on the continuity of $\mathcal{Q}(x,\xi)$ in x as indicated in (1), $|\mathcal{Q}(x,\xi)| < \infty$ as shown in Assumption 1, and the compact sample space Ω), and $V_\theta(x)$ is monotonic in θ based on Proposition 3. Then, following Dini's theorem, we have $V_\theta(x)$ converges to $V_0(x)$ uniformly. Therefore, for any arbitrary small positive number ϵ, there exists a $\Delta_\epsilon > 0$ such that $\forall \theta \le \Delta_\epsilon$:

$$|V_\theta(x^*) - V_0(x^*)| \le \epsilon, \quad |V_\theta(x^*_\theta) - V_0(x^*_\theta)| \le \epsilon.$$

Then, for any $\theta \le \Delta_\epsilon$, we have

$$\begin{aligned}\psi(\theta) - \psi(0) &= V_\theta(x^*_\theta) - V_0(x^*) \le V_\theta(x^*) - V_0(x^*) \\ &\le |V_\theta(x^*) - V_0(x^*)| \le \epsilon,\end{aligned}$$

where the first inequality follows from the fact that x^*_θ is the optimal solution to DD-SP with the distance value θ and x^* is a feasible solution to this same problem. Similarly, we have

$$\begin{aligned}\psi(0) - \psi(\theta) &= V_0(x^*) - V_\theta(x^*_\theta) \le V_0(x^*_\theta) - V_\theta(x^*_\theta) \\ &\le |V_0(x^*_\theta) - V_\theta(x^*_\theta)| \le \epsilon.\end{aligned}$$

Therefore, $|\psi(\theta) - \psi(0)| \le \epsilon$, which proves the claim. □

Acknowledgments

The authors thank the editors and two anonymous referees for their sincere suggestions on improving the quality of this paper. This research was partially supported by NSF under grant ECCS1609794.

References

[1] A. Ben-Tal, D. Den Hertog, A. De Waegenaere, B. Melenberg, G. Rennen, Robust solutions of optimization problems affected by uncertain probabilities, Manage. Sci. 59 (2) (2013) 341–357.
[2] P. Billingsley, Convergence of Probability Measures, John Wiley & Sons, 2013.
[3] J. Birge, F. Louveaux, Introduction to Stochastic Programming, Springer, 1997.
[4] F. Bolley, A. Guillin, C. Villani, Quantitative concentration inequalities for empirical measures on non-compact spaces, Probab. Theory Related Fields 137 (3–4) (2007) 541–593.
[5] E. Delage, Y. Ye, Distributionally robust optimization under moment uncertainty with application to data-driven problems, Oper. Res. 58 (3) (2010) 595–612.
[6] P.M. Esfahani, D. Kuhn, Data-driven distributionally robust optimization using the Wasserstein metric: Performance guarantees and tractable reformulations, 2015. Available at Optimization-Online: www.optimization-online.org/DB_FILE/2015/05/4899.pdf.
[7] N. Fournier, A. Guillin, On the rate of convergence in Wasserstein distance of the empirical measure, Probab. Theory Related Fields 162 (3–4) (2015) 707–738.
[8] A.L. Gibbs, F.E. Su, On choosing and bounding probability metrics, Internat. Statist. Rev. 70 (3) (2002) 419–435.
[9] R. Jiang, Y. Guan, Data-driven chance constrained stochastic program, Math. Program. 158 (2016) 291–327.
[10] D. Love, G. Bayraksan, Phi-Divergence Constrained Ambiguous Stochastic Programs. Technical report, University of Arizona, 2012.
[11] G.C. Pflug, A. Pichler, Multistage Stochastic Optimization, Springer, 2014.
[12] G.C. Pflug, A. Pichler, D. Wozabal, The 1/N investment strategy is optimal under high model ambiguity, J. Banking Finance 36 (2) (2012) 410–417.
[13] G.C. Pflug, D. Wozabal, Ambiguity in portfolio selection, Quant. Finance 7 (4) (2007) 435–442.
[14] S.T. Rachev, Mass Transportation Problems, Vol. 2, Springer, 1998.
[15] A. Shapiro, D. Dentcheva, A. Ruszczyński, Lectures on Stochastic Programming: Modeling and Theory, Vol. 9, SIAM, 2009.
[16] A.W. Van der Vaart, Asymptotic Statistics, Vol. 3, Cambridge University Press, 2000.
[17] C. Zhao, Data-Driven Risk-Averse Stochastic Program and Renewable Energy Integration (Ph.D. Dissertation), University of Florida, 2014.
[18] C. Zhao, Y. Guan, Data-driven risk-averse stochastic optimization with Wasserstein metric, 2015. Available at Optimization-Online: http://www.optimization-online.org/DB_HTML/2015/05/4902.html.