# Data-Driven Risk-Averse Stochastic Optimization with Wasserstein Metric*

**Chaoyue Zhao† and Yongpei Guan‡**
†School of Industrial Engineering and Management
Oklahoma State University, Stillwater, OK 74074
‡Department of Industrial and Systems Engineering
University of Florida, Gainesville, FL 32611
Emails: chaoyue.zhao@okstate.edu; guan@ise.ufl.edu

**Abstract**

The traditional two-stage stochastic programming approach is to minimize the total expected cost with the assumption that the distribution of the random parameters is known. However, in most practices, the actual distribution of the random parameters is not known, and instead, only a series of historical data are available. Thus, the solution obtained from the traditional two-stage stochastic program can be biased and suboptimal for the true problem, if the estimated distribution of the random parameters is not accurate, which is usually true when only a limited amount of historical data are available. In this paper, we propose a data-driven risk-averse stochastic optimization approach. Based on the observed historical data, we construct the confidence set of the ambiguous distribution of the random parameters, and develop a risk-averse stochastic optimization framework to minimize the total expected cost under the worst-case distribution within the constructed confidence set. We introduce the Wasserstein metric to construct the confidence set and by using this metric, we can successfully reformulate the risk-averse two-stage stochastic program to its tractable counterpart. In addition, we derive the worst-case distribution and develop efficient algorithms to solve the reformulated problem. Moreover, we perform convergence analysis to show that the risk averseness of the proposed formulation vanishes as the amount of historical data grows to infinity, and accordingly, the corresponding optimal objective value converges to that of the traditional risk-neutral two-stage stochastic program. We further precisely derive the convergence rate, which indicates the value of data. Finally, the numerical experiments on risk-averse stochastic facility location and stochastic unit commitment problems verify the effectiveness of our proposed framework.

*Key words:* stochastic optimization; data-driven decision making; Wasserstein metric

*The basic idea and main results of this paper are also available in [29] and in the first author's dissertation [27].

# 1 Introduction

Solving optimization problems under uncertainty has always been challenging for most practitioners. In an uncertain environment, decision-makers commonly face the following two-stage optimization problem: $\min_{x\in X} c^\top x + \mathcal{Q}(x,\xi)$, where $x$ is the first-stage decision variable and accordingly $X \subseteq \mathbb{R}^n$ is a compact and convex set representing the feasible region for the first-stage decision variable. The random parameter $\xi \in \mathbb{R}^m$ represents the uncertainty and the corresponding second-stage cost

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

which is assumed continuous in $\xi$. To capture the uncertainty of $\xi$ and solve the problems effectively, two branches of optimization under uncertainty approaches, named robust and stochastic optimization approaches, are studied extensively recently.

For the robust optimization approach, the ambiguity of the random parameters is allowed. This approach defines an uncertainty set $\mathcal{U}$ for the random parameters, e.g., $\xi \in \mathcal{U}$, and achieves the objective of minimizing the total cost under the worst-case realization of $\xi$ in this predefined set $\mathcal{U}$ [2, 5], i.e., $\min_{x\in X} c^\top x + \max_{\xi\in\mathcal{U}} \mathcal{Q}(x,\xi)$. For the cases under the general settings, the robust optimization model (RP) is intractable. However, for certain special forms of the uncertainty set $\mathcal{U}$, such as ellipsoidal uncertainty set [3, 11, 13], polyhedral uncertainty set [3], and cardinality uncertainty set [5], the reformulations of robust optimization are possible to be obtained and thus the problem can be addressed efficiently. Considering the fact that the robust optimization approach utilizes limited information (e.g., lower/upper bound of the uncertain parameter plus a budget constraint) of the random parameters and considers the worst-case cost in the objective, practitioners are aware of its over-conservatism, although this approach has significant advantages in terms of providing reliable solutions for most practices.

Stochastic optimization is another effective approach to address uncertainty. The traditional two-stage stochastic optimization framework can be described as follows (cf. [7] and [21]):

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

where the uncertain random variable $\xi$ is defined on a probability space $(\Omega, \sigma(\Omega), \mathbb{P})$, in which $\Omega$

is the sample space for $\xi$, $\sigma(\Omega)$ is the $\sigma$-algebra of $\Omega$, and $\mathbb{P}$ is the associate probability measure. For this setting, the probability distribution $\mathbb{P}$ is given. To solve the problem, the sample average approximation method [21] is usually applied in which the random parameters are captured by a number of scenarios sampled from the true distribution. This approach becomes computationally heavy when the number of scenarios increases.

Considering the disadvantages of robust and stochastic optimization approaches, plus the fact that 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, the distributionally robust optimization approaches are proposed recently (see, [9], among others) and can be described as follows:

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

Instead of assuming that the distribution of the random parameters is known, this framework 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 statistical inference. The objective is to minimize the total cost under the worst-case distribution in the given confidence set. Moment-based approaches described below are initially introduced to build the set $\mathcal{D}$ [9]:

$$\mathcal{D}_1 = \{\hat{\mathbb{P}} \in \mathcal{M}_+ :\ (\mathbb{E}_{\hat{\mathbb{P}}}[\xi] - \mu_0)^\top \Sigma_0^{-1} (\mathbb{E}_{\hat{\mathbb{P}}}[\xi] - \mu_0) \le \gamma_1,\ \ \mathbb{E}_{\hat{\mathbb{P}}}[(\xi-\mu_0)(\xi-\mu_0)^\top] \preceq \gamma_2 \Sigma_0\}, \tag{2}$$

where $\mathcal{M}_+$ represents the set of all probability distributions, $\mu_0$ and $\Sigma_0 \succ 0$ represent the inferred mean and covariance matrix respectively, and $\gamma_1 > 0$ and $\gamma_2 > 1$ are two parameters obtained from the process of inference (e.g., through sample mean and covariance matrix). In (2), the first constraint ensures the mean of $\xi$ is in an ellipsoid with center $\mu_0$ and size $\gamma_1$, and the second constraint ensures the second moment matrix of $\xi$ is in a positive semidefinite cone. The moment-based approaches have advantages of transforming the original risk-averse stochastic optimization problem to a tractable conic or semidefinite program. The readers are referred to related pioneer works, including the special case without considering moment ambiguity, e.g., $\mathcal{D}_1^0 = \{\hat{\mathbb{P}} \in \mathcal{M}_+ : \mathbb{E}_{\hat{\mathbb{P}}}[\xi] = \mu_0, \mathbb{E}_{\hat{\mathbb{P}}}[\xi\xi^\top] = \mu_0\mu_0^\top + \Sigma_0\}$, in [12], [14], [8], [23], [30], and [31], among others. For these

moment-based approaches, the parameters of $\gamma_1$ and $\gamma_2$ can depend on the amount of historical data available. As the number of historical data increases, the values of $\gamma_1$ and $\gamma_2$ decrease and eventually $\mathcal{D}_1$ is constructed with fixed moments, i.e., $\mathcal{D}_1$ converges to $\mathcal{D}_1^0$. Note here that although the confidence set $\mathcal{D}_1$ can characterize the properties of the first and second moments of the random parameters, it cannot guarantee any convergence properties for the unknown distribution to the true distribution. Even for the case in which there are an infinite amount of historical data, it is only guaranteed that the first and second moments of the random parameters are estimated accurately. But there are still infinite distributions, including the true distribution, in $\mathcal{D}_1$ with the same first and second moments.

In this paper, instead of utilizing the moment-based approaches, we propose a "distribution-based" approach to construct the confidence set for the unknown true distribution. For our approach, we use the empirical distribution based on the historical data as the reference distribution. Then, the confidence set $\mathcal{D}$ is constructed by utilizing metrics to define the distance between the reference distribution and the true distribution. The advantage for this approach is that the convergence properties hold. That is, as the number 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. For instance, if we let $d_{\mathrm{M}}(\mathbb{P}_0, \hat{\mathbb{P}})$ be any metric between a reference distribution $\mathbb{P}_0$ and an unknown distribution $\hat{\mathbb{P}}$ and use $\theta$ to represent the corresponding distance, then the confidence set $\mathcal{D}$ can be represented as follows:

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

By learning from the historical data, we can build a reference distribution $\mathbb{P}_0$ of the uncertain parameter as well as a confidence set of the true distribution described above, and decisions are made in consideration of that the true distribution can vary within the confidence set. In particular, we use the Wasserstein metric to construct the confidence set for general distributions, including both discrete and continuous distributions. Although significant research progress has been made for the discrete distribution case (see, e.g., [18] and [16]), the study for the continuous distribution case is more challenging and is very limited. The recent study on the general $\phi$-divergence [1] can be utilized for the continuous distribution case by defining the distance between the true density and

the reference density, which however could not guarantee the convergence properties. In this paper, by studying the Wasserstein metric and deriving the corresponding convergence rate, our proposed approach can fit well in the data-driven risk-averse two-stage stochastic optimization framework. Our contributions can be summarized as follows:

1. We propose a data-driven risk-averse two-stage stochastic optimization framework to solve optimization under uncertainty problems by allowing distribution ambiguity. In particular, we introduce the Wasserstein metric to construct the confidence set and accordingly we can successfully reformulate the risk-averse stochastic program to its tractable counterpart.

2. Our proposed framework allows the true distribution to be continuous, and closed-form expressions of the worst-case distribution can be obtained by solving the proposed model.

3. We show the convergence property of the proposed data-driven framework by proving that as the number of historical data increases, the risk-averse problem converges to the risk-neutral one, i.e., the traditional stochastic optimization problem. We also provide a tighter convergence rate, which shows the value of data.

4. The computational experiments on the data-driven risk-averse stochastic facility location and unit commitment problems numerically show the effectiveness of our proposed approach.

The remainder of this paper is organized as follows. In Section 2, we describe the data-driven risk-averse stochastic optimization framework, including the introduction of the Wasserstein metric and the reference distribution and confidence set construction based on the available historical data. In Section 3, we derive a tractable reformulation of the proposed data-driven risk-averse stochastic optimization problem, as well as the corresponding worst-case distribution. In Section 4, we show the convergence property by proving that as the number of historical data increases, the data-driven risk-averse stochastic optimization problem converges to the traditional risk-neutral one. In Section 5, we propose a solution approach to solve the reformulated problem corresponding to a given finite set of historical data. In Section 6, we perform numerical studies on stochastic facility location and unit commitment problems to verify the effectiveness of our solution framework. Finally, in Section 7, we conclude our study.

## 2 Data-Driven Risk-Averse Stochastic Optimization Framework

In this section, we develop 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 introduce the Wasserstein distribution 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. Finally, we derive the convergence rate and provide the formulation framework of the data-driven risk-averse two-stage stochastic optimization problem.

### 2.1 Wasserstein Metric

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

$$d_{\mathrm{W}}(\mathbb{P}, \hat{\mathbb{P}}) := \inf_{\pi}\{\mathbb{E}_{\pi}[\rho(X, Y)] : \mathbb{P} = \mathcal{L}(X), \hat{\mathbb{P}} = \mathcal{L}(Y)\}, \tag{3}$$

where $\rho(X, Y)$ is defined as the distance between random variables $X$ and $Y$, and the infimum is taken over all joint distributions $\pi$ with marginals $\mathbb{P}$ and $\hat{\mathbb{P}}$. As indicated in [25], the Wasserstein metric is indeed a metric since it satisfies the properties of metrics. That is, $d_{\mathrm{W}}(\mathbb{P}, \hat{\mathbb{P}}) = 0$ if and only if $\mathbb{P} = \hat{\mathbb{P}}$, $d_{\mathrm{W}}(\mathbb{P}, \hat{\mathbb{P}}) = d_{\mathrm{W}}(\hat{\mathbb{P}}, \mathbb{P})$ (symmetric property) and $d_{\mathrm{W}}(\mathbb{P}, \hat{\mathbb{P}}) \leq d_{\mathrm{W}}(\mathbb{P}, \mathcal{O}) + d_{\mathrm{W}}(\mathcal{O}, \hat{\mathbb{P}})$ for any probability distribution $\mathcal{O}$ (triangle equality). In addition, by the Kantorovich-Rubinstein theorem [15], the Wasserstein metric is equivalent to the Kantorovich metric, which is defined as

$$d_{\mathrm{K}}(\mathbb{P}, \hat{\mathbb{P}}) = \max_{h \in \mathcal{H}} \left| \int_{\Omega} h d\mathbb{P} - \int_{\Omega} h d\hat{\mathbb{P}} \right|,$$

where $\mathcal{H} = \{h : \|h\|_{\mathrm{L}} \leq 1\}$, and $\|h\|_{\mathrm{L}} := \sup\{(h(x) - h(y))/\rho(x, y) : x \neq y \text{ in } \Omega\}$. The Wasserstein metric is commonly applied in many areas. For example, many metrics known in statistics, measure theory, ergodic theory, functional analysis, etc., are special cases of the Wasserstein/Kantorovich metric [24]. The Wasserstein/Kantorovich metric also has many applications in transportation theory [19], and some applications in computer science like probabilistic concurrency, image retrieval, data mining, and bioinformatics, etc [10]. In this study, we use the Wasserstein metric to construct

the confidence set for the general probability distribution, including both discrete and continuous distributions.

### 2.2 Reference Distribution

After observing a series of historical data, we can have an estimation of the true probability distribution, which is called reference probability distribution. Intuitively, the more historical data we have, the more accurate estimation of the true probability distribution we can obtain. Many significant works have been made to obtain the reference distribution, with both parametric and nonparametric statistical estimation approaches. For instance, for the parametric approaches, the true probability distribution is usually assumed following a particular distribution function, e.g., normal distribution, and the parameters (e.g., mean and variance) are estimated by learning from the historical data. On the other hand, the nonparametric estimations (in which the true distribution is not assumed following any specific distribution), such as kernel density estimation, are also proved to be effective approaches to obtain the reference probability distribution (e.g., [20] and [17]). In this paper, we utilize a nonparametric estimation approach. More specifically, we use the empirical 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 $\xi_0^1, \xi_0^2, \cdots, \xi_0^N$, the empirical distribution is defined as

$$\mathbb{P}_0(x) = \frac{1}{N} \sum_{i=1}^{N} \delta_{\xi_0^i}(x),$$

where $\delta_{\xi_0^i}(x)$ is one if $x \geq \xi_0^i$ and zero elsewhere. Based on the strong law of large numbers, it can be proved that the reference distribution $\mathbb{P}_0$ pointwise converges to the true probability distribution $\mathbb{P}$ almost surely [22]. By Glivenko-Cantelli theorem, this result can be strengthened by proving the uniform convergence of $\mathbb{P}$ to $\mathbb{P}_0$ [26].

### 2.3 Confidence Set Construction

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 $\theta$ to represent the distance between the reference distribution and the true distribution, then the more historical data observed, the smaller the value of $\theta$ is, and the tighter the confidence set becomes. Therefore, the confidence set $\mathcal{D}$ can be represented as follows:

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

where the value of $\theta$ depends on the number of historical data. More specifically, according to the definition of the Wasserstein metric in (3), the confidence set $\mathcal{D}$ is:

$$\mathcal{D} = \left\{\hat{\mathbb{P}} \in \mathcal{M}_+ : \inf_{\pi}\{\mathbb{E}_\pi[\rho(Z, W)] : \mathbb{P}_0 = \mathcal{L}(Z), \hat{\mathbb{P}} = \mathcal{L}(W)\} \leq \theta\right\}. \tag{4}$$

We can further show that under the Wasserstein metric, the empirical distribution $\mathbb{P}_0$ converges to the true distribution $\mathbb{P}$ exponentially fast. As indicated in [28], the exact relationship between the number of historical data and the value of $\theta$ can be expressed in the following proposition:

**Proposition 1.** *For a general $m$-dimension supporting space $\Omega$, we have*

$$P(d_W(\mathbb{P}_0, \mathbb{P}) \leq \theta) \geq 1 - \exp\left(-\frac{\theta^2}{2B^2}N\right),$$

*where $N$ is the number of historical data and $B$ is the diameter of $\Omega$.*

Based on Proposition 1, if we set the confidence level of the confidence set $\mathcal{D}$ as $\beta$, then we can obtain the closed-form expression for $\theta$ as follows:

$$\theta = B\sqrt{\frac{2}{N}\log\left(\frac{1}{1-\beta}\right)}. \tag{5}$$

### 2.4 Data-driven Risk-Averse Stochastic Optimization Framework

As described in Section 1, for the data-driven risk-averse stochastic optimization approach, instead of knowing the true probability distribution $\mathbb{P}$, we assume $\mathbb{P}$ can vary within the confidence set $\mathcal{D}$ and consider $\mathbb{E}[\mathcal{Q}(x, \xi)]$ under the worst-case distribution in $\mathcal{D}$. Therefore, the data-driven risk-averse

two-stage stochastic optimization problem can be formulated as follows:

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

where $\mathcal{Q}(x,\xi) = \min_{y(\xi)\in Y}\{d^\top y(\xi) : A(\xi)x + By(\xi) \geq b(\xi)\}$ and $\hat{\mathbb{P}}$ is an arbitrary probability distribution in $\mathcal{D}$. In this case, the proposed approach is more conservative than the traditional stochastic optimization approach. That is, the proposed approach is risk-averse.

# 3 Tractable Reformulation and Worst-Case Distribution

In this section, we first derive the reformulation of the worst-case expectation $\max_{\hat{\mathbb{P}}\in\mathcal{D}} \mathbb{E}_{\hat{\mathbb{P}}}[\mathcal{Q}(x,\xi)]$, and then obtain the reformulation of problem (DD-SP) correspondingly. Following that, we derive the worst-case distribution corresponding to the Wasserstein metric.

Without loss of generality, we can define the supporting space $\Omega$ by assuming that the random parameter $\xi$ is between a lower bound $W^-$ and an upper bound $W^+$. That is,

$$\Omega := \left\{ \xi \in \mathbb{R}^m : \quad W^- \leq \xi \leq W^+ \right\}. \tag{6}$$

Now we can derive the reformulation of the worst-case expectation $\max_{\hat{\mathbb{P}}\in\mathcal{D}} \mathbb{E}_{\hat{\mathbb{P}}}[\mathcal{Q}(x,\xi)]$ in the following proposition.

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

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

*Proof.* As indicated in Section 2.2, if we have $N$ historical data samples $\xi^1, \xi^2, \cdots, \xi^N$, the reference distribution $\mathbb{P}_0$ can be defined as the empirical distribution, i.e., $\mathbb{P}_0 = \frac{1}{N}\sum_{i=1}^{N} \delta_{\xi^i}(x)$. In addition, the confidence set $\mathcal{D}$ is defined the same way as indicated in (4). Then, we can claim that, if $\hat{\mathbb{P}} \in \mathcal{D}$, then $\forall \epsilon \geq 0$, there exists a joint distribution $\pi$ such that $\mathbb{P}_0 = \mathcal{L}(Z), \hat{\mathbb{P}} = \mathcal{L}(W)$, and $\mathbb{E}_\pi[\rho(Z,W)] \leq \theta + \epsilon$. Based on the definition of $\mathbb{E}_\pi[\rho(Z,W)]$, we can obtain the following reformulation of $\mathbb{E}_\pi[\rho(Z,W)]$ (without loss of generality, we assume the cumulative distribution

function for $\mathbb{P}$ is absolute continuous):

$$\begin{aligned}\mathbb{E}_{\pi}[\rho(Z,W)] &= \int_{z\in\Omega}\int_{\xi\in\Omega} f(\xi,z)\rho(\xi,z)d\xi dz = \sum_{i=1}^{N}\int_{\xi\in\Omega} f_{W|Z}(\xi|\xi^i)\mathbb{P}_0(Z=\xi^i)\rho(\xi,\xi^i)d\xi \\ &= \frac{1}{N}\sum_{i=1}^{N}\int_{\xi\in\Omega} f_{W|Z}(\xi|\xi^i)\rho(\xi,\xi^i)d\xi, \end{aligned} \tag{7}$$

where $f(\xi,z)$ is the density function of $\pi$, and $f_{W|Z}(\xi|\xi^i)$ is the conditional density function when $Z=\xi^i$. The equation (7) holds since according to the definition of $\mathbb{P}_0$, $\mathbb{P}_0(Z=\xi^i)=1/N$ for $\forall i=1,\cdots N$. For notation brevity, we let $f^i(\xi)=f_{W|Z}(\xi|\xi^i)$ and $\rho^i(\xi)=\rho(\xi,\xi^i)$. Then the second-stage problem $\max_{\hat{\mathbb{P}}\in\mathcal{D}}\mathbb{E}_{\hat{\mathbb{P}}}[\mathcal{Q}(x,\xi)]$ of (DD-SP) can be reformulated as:

$$\begin{aligned} \max_{f_{\mathrm{W}}(\xi)\geq 0} \quad & \int_{\xi\in\Omega}\mathcal{Q}(x,\xi)f_{\mathrm{W}}(\xi)d\xi \\ s.t. \quad & f_{\mathrm{W}}(\xi)=\frac{1}{N}\sum_{i=1}^{N}f^i(\xi), && (8)\\ & \int_{\xi\in\Omega}f^i(\xi)d\xi=1, \quad \forall i, && (9)\\ & \frac{1}{N}\sum_{i=1}^{N}\int_{\xi\in\Omega}f^i(\xi)\rho^i(\xi)d\xi\leq\theta+\epsilon, && (10)\end{aligned}$$

where $f_{\mathrm{W}}(\xi)$ is the density function of $\hat{\mathbb{P}}$. Constraints (8) and (9) are based on the properties of conditional density function and constraint (10) follows the definition of $\mathcal{D}$ and equation (7). Note here that this reformulation holds for any $\epsilon\geq 0$. By substituting constraint (8) into the objective function, we can obtain its equivalent formulation as follows:

$$\begin{aligned} & \max_{f^i(\xi)\geq 0} \quad \frac{1}{N}\sum_{i=1}^{N}\int_{\xi\in\Omega}\mathcal{Q}(x,\xi)f^i(\xi)d\xi \\ (\mathrm{P}_1) \qquad & s.t. \quad \int_{\xi\in\Omega}f^i(\xi)d\xi=1, \quad \forall i, && (11)\\ & \frac{1}{N}\sum_{i=1}^{N}\int_{\xi\in\Omega}f^i(\xi)\rho^i(\xi)d\xi\leq\theta+\epsilon. && (12)\end{aligned}$$

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 $(\mathrm{P}_1)$ always has a feasible solution (e.g., the reference distribution $\mathbb{P}_0$ is obvious a feasible solution for $(\mathrm{P}_1)$). Meanwhile, since $\mathcal{Q}(x,\xi)$ is

assumed bounded above, ($P_1$) is bounded above. Furthermore, since the above problem is a convex programming, there is no duality gap. Then we can consider its Lagrangian dual problem that can be written as follows:

$$L(\lambda_i, \beta) = \max_{f^i(\xi)\geq 0} \frac{1}{N} \sum_{i=1}^{N} \int_{\xi\in\Omega} (\mathcal{Q}(x,\xi) - N\lambda_i - \beta\rho^i(\xi)) f^i(\xi) d\xi + \sum_{i=1}^{N} \lambda_i + (\theta+\epsilon)\beta,$$

where $\lambda_i$ and $\beta \geq 0$ are dual variables of constraints (11) and (12) respectively. The dual problem then is

$$\min_{\beta\geq 0, \lambda_i} L(\lambda_i, \beta).$$

Next, we argue that $\forall \xi \in \Omega$, $\mathcal{Q}(x,\xi) - N\lambda_i - \beta\rho^i(\xi) \leq 0$. If this argument does not hold, then there exists a $\xi_0$ such that $\mathcal{Q}(x,\xi_0) - N\lambda_i - \beta\rho^i(\xi_0) > 0$. It means there exists a strict positive constant $\sigma$, such that $\mathcal{Q}(x,\xi_0) - N\lambda_i - \beta\rho^i(\xi_0) > \sigma$. Based on the definition of $\mathcal{Q}(x,\xi)$, it is continuous with $\xi$. Also, the distance function $\rho^i(\xi)$ is continuous on $\xi$. Therefore, $\mathcal{Q}(x,\xi) - N\lambda_i - \beta\rho^i(\xi)$ is continuous with $\xi$. 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 $f^i(\xi)$ be 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 ($P_1$) bounded above. Hence, the argument $\mathcal{Q}(x,\xi) - N\lambda_i - \beta\rho^i(\xi) \leq 0$ for all $\xi \in \Omega$ holds. In this case,

$$\max_{f^i(\xi)\geq 0} \frac{1}{N} \sum_{i=1}^{N} \int_{\xi\in\Omega} (\mathcal{Q}(x,\xi) - N\lambda_i - \beta\rho^i(\xi)) f^i(\xi) d\xi + \sum_{i=1}^{N} \lambda_i + (\theta+\epsilon)\beta = \sum_{i=1}^{N} \lambda_i + (\theta+\epsilon)\beta, \quad (13)$$

with optimal solutions satisfying $(\mathcal{Q}(x,\xi) - N\lambda_i - \beta\rho^i(\xi)) f^i(\xi) = 0,\ i = 1, \cdots, N$ since $f^i(\xi) \geq 0$. Then, the dual formulation is reformulated as:

$$\begin{aligned} \min_{\beta\geq 0, \lambda_i} \quad & \sum_{i=1}^{N} \lambda_i + (\theta+\epsilon)\beta \\ s.t. \quad & \mathcal{Q}(x,\xi) - N\lambda_i - \beta\rho^i(\xi) \leq 0,\ \forall \xi \in \Omega, \forall i = 1, \cdots, N. \end{aligned} \quad (14)$$

From the above formulation, it is easy to observe that the optimal solution $\lambda_i$ should satisfy

$$\lambda_i = \frac{1}{N} \max_{\xi\in\Omega} \{\mathcal{Q}(x,\xi) - \beta\rho^i(\xi)\}, \quad (15)$$

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

$$\min_{\beta\geq 0}\left\{\frac{1}{N}\sum_{i=1}^{N}\max_{\xi\in\Omega}\{\mathcal{Q}(x,\xi)-\beta\rho^i(\xi)\}+(\theta+\epsilon)\beta\right\}. \tag{16}$$

Note here that reformulation (16) holds for $\forall\epsilon\geq 0$ and is continuous on $\epsilon$. Thus, reformulation (16) holds for $\epsilon=0$, which immediately leads to the following reformulation of $\max_{\hat{\mathbb{P}}\in\mathcal{D}} \mathbb{E}_{\hat{\mathbb{P}}}[\mathcal{Q}(x,\xi)]$:

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

□

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

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

we have the following proposition:

**Proposition 3.** *The function $g(\theta)$ is monotone increasing in $\theta$. In addition, $g(0)=\mathbb{E}_{\mathbb{P}_0}[\mathcal{Q}(x,\xi)]$ and $\lim_{\theta\to\infty} g(\theta)=\max_{\xi\in\Omega}\mathcal{Q}(x,\xi)$.*

*Proof.* The monotone property is obvious, since it can be easily observed that as $\theta$ decreases, the confidence set $\mathcal{D}$ shrinks and the worst-case expected value $\max_{\hat{\mathbb{P}}\in\mathcal{D}} \mathbb{E}_{\hat{\mathbb{P}}}[\mathcal{Q}(x,\xi)]$ decreases. Therefore, the reformulation $g(\theta)=\max_{\hat{\mathbb{P}}\in\mathcal{D}} \mathbb{E}_{\hat{\mathbb{P}}}[\mathcal{Q}(x,\xi)]$ is monotone in $\theta$.

When $\theta=0$, we have $d_{\mathrm{W}}(\mathbb{P}_0,\hat{\mathbb{P}})=0$. According to the properties of metrics, we have $\hat{\mathbb{P}}=\mathbb{P}_0$. That means the confidence set $\mathcal{D}$ is a singleton, only containing $\mathbb{P}_0$. Therefore, we have $g(0)=\mathbb{E}_{\mathbb{P}_0}[\mathcal{Q}(x,\xi)]$.

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 $\Omega$ is compact. Then we have

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

which indicates that $\lim_{\theta\to\infty} g(\theta)=\max_{\xi\in\Omega}\mathcal{Q}(x,\xi)$. 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 robust than the traditional stochastic optimization.

Now, we analyze the formulation of the worst-case distribution. We have the following proposition.

**Proposition 4.** *The worst-case distribution for* $\max_{\hat{\mathbb{P}}\in\mathcal{D}} \mathbb{E}_{\hat{\mathbb{P}}}[\mathcal{Q}(x,\xi)]$ *is*

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

*where* $\delta_{\xi^i}(\xi)$ *is one if* $\xi \geq \xi^i$ *and zero elsewhere, and* $\xi^i$ *is the optimal solution of* $\max_{\xi\in\Omega}\{\mathcal{Q}(x,\xi) - \beta\rho^i(\xi)\}$.

*Proof.* Based on (13), we have

$$\max_{f^i(\xi)\geq 0} \frac{1}{N}\sum_{i=1}^{N} \int_{\xi\in\Omega} (\mathcal{Q}(x,\xi) - N\lambda_i - \beta\rho^i(\xi)) f^i(\xi) d\xi = 0.$$

Therefore, any distribution (corresponding to $f^i(\xi)$) satisfying the above formulation is a worst-case distribution.

As indicated in (14), we have $\mathcal{Q}(x,\xi) - N\lambda_i - \beta\rho^i(\xi) \leq 0$, for $\forall\xi \in \Omega,\ \forall i$. Therefore,

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

Since $f^i(\xi) \geq 0$, $f^i(\xi)$ can be non-zero only when $\mathcal{Q}(x,\xi) - N\lambda_i - \beta\rho^i(\xi) = 0$, which means $\lambda_i = (\mathcal{Q}(x,\xi) - \beta\rho^i(\xi))/N$. Meanwhile, as indicated in (15), we have $\lambda_i = (\mathcal{Q}(x,\xi) - \beta\rho^i(\xi))/N$ only when $\xi$ is the optimal solution of $\max_{\xi\in\Omega}\{\mathcal{Q}(x,\xi) - \beta\rho^i(\xi)\}$. Therefore, there exists at least one distribution function (indicated as $\hat{\mathbb{P}}^i(\xi)$) of $f^i(\xi)$ described as follows:

$$\hat{\mathbb{P}}^i(\xi) = \delta_{\xi^i}(\xi),$$

where $\xi^i$ is an optimal solution of $\max_{\xi\in\Omega}\{\mathcal{Q}(x,\xi) - \beta\rho^i(\xi)\}$. Accordingly, following (8), there exists

at least one worst-case distribution $\hat{\mathbb{P}}^*$ that satisfies

$$\hat{\mathbb{P}}^*(\xi) = \frac{1}{N}\sum_{i=1}^{N}\hat{\mathbb{P}}^i(\xi) = \frac{1}{N}\sum_{i=1}^{N}\delta_{\xi^i}(\xi).$$

Therefore, we have Proposition 4 holds. □

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

**Theorem 1.** *The problem (DD-SP) is equivalent to the following two-stage robust optimization problem:*

$$\textit{(RDD-SP)} \qquad \min_{x\in X,\beta\geq 0} \quad c^\top x + \theta\beta + \frac{1}{N}\sum_{i=1}^{N}\max_{\xi\in\Omega}\left\{\mathcal{Q}(x,\xi) - \beta\rho^i(\xi)\right\}, \tag{17}$$

*with the worst-case distribution*

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

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

# 4 Convergence Analysis

In this section, we examine the convergence properties of the (DD-SP) to (SP) as the amount of historical data increases. We demonstrate that as the confidence set $\mathcal{D}$ shrinks with more historical data observed, the risk-averse problem (DD-SP) converges to the risk-neutral one (SP). We first analyze the convergence property of the second-stage objective value, which can be shown as follows:

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

*Proof.* First, following (5), it is obvious that $\theta \to 0$ as $N \to \infty$. Meanwhile, following Proposition 2, we have

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

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

$$\lim_{N\to\infty}\min_{\beta\geq 0}\left\{\theta\beta+\frac{1}{N}\sum_{i=1}^{N}\max_{\xi\in\Omega}\{\mathcal{Q}(x,\xi)-\beta\rho^i(\xi)\}\right\}\leq\mathbb{E}_{\mathbb{P}}[\mathcal{Q}(x,\xi)], \tag{18}$$

and

$$\lim_{\theta\to 0}\max_{\hat{\mathbb{P}}\in\mathcal{D}}\mathbb{E}_{\hat{\mathbb{P}}}[\mathcal{Q}(x,\xi)]\geq\mathbb{E}_{\mathbb{P}}[\mathcal{Q}(x,\xi)]. \tag{19}$$

Note here that (19) is obvious. We only need to prove (18).

Following our assumptions, we have that $\Omega$ is compact and $\mathcal{Q}(x,\xi)$ is continuous on $\xi$. Then, there exists a constant number $M>0$ such that for any given $x$,

$$-M\leq\mathcal{Q}(x,\xi)\leq M,\quad\forall\xi\in\Omega. \tag{20}$$

In addition, we have

$$0\leq\rho(\xi,z)\leq B,\quad\forall\xi,z\in\Omega, \tag{21}$$

where $B$ is the diameter of $\Omega$. Therefore, for any $\beta\geq 0$, based on (20) and (21), 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

$$\begin{aligned}
&\lim_{N\to\infty}\left\{\theta\beta+\frac{1}{N}\sum_{i=1}^{N}\max_{\xi\in\Omega}\{\mathcal{Q}(x,\xi)-\beta\rho(\xi,z^i)\}\right\}\\
&=\lim_{N\to\infty}\left\{\theta\beta+\mathbb{E}_{\mathbb{P}_0}\max_{\xi\in\Omega}\{\mathcal{Q}(x,\xi)-\beta\rho(\xi,z)\}\right\}\\
&=\lim_{N\to\infty}\theta\beta+\mathbb{E}_{\mathbb{P}}\max_{\xi\in\Omega}\{\mathcal{Q}(x,\xi)-\beta\rho(\xi,z)\}\\
&=\mathbb{E}_{\mathbb{P}}[\max_{\xi\in\Omega}\{\mathcal{Q}(x,\xi)-\beta\rho(\xi,z)\}],
\end{aligned} \tag{22}$$

where the first equality holds following the definition of $\mathbb{P}_0$, the second equality holds following the Helly-Bray Theorem [6], because $\mathbb{P}_0$ converges in probability to $\mathbb{P}$ as $N\to\infty$ (as indicated in

Proposition 1) and $\max_{\xi\in\Omega}\{\mathcal{Q}(x,\xi)-\beta\rho(\xi,z)\}$ is bounded for $\forall z\in\Omega$, and the third equality holds because $\theta\to 0$ as $N\to\infty$ following (5).

Now we show that for a given first-stage solution $x$ and any true distribution $\mathbb{P}$, we have the following claim holds:

$$\min_{\beta\geq 0}\ \mathbb{E}_{\mathbb{P}}[\max_{\xi\in\Omega}\{\mathcal{Q}(x,\xi)-\beta\rho(\xi,z)\}]=\mathbb{E}_{\mathbb{P}}[\mathcal{Q}(x,z)]. \tag{23}$$

To prove (23), we first denote $\xi^*(z)$ as the optimal solution to $\max_{\xi\in\Omega}\{\mathcal{Q}(x,\xi)-\beta\rho(\xi,z)\}$ corresponding to a given $z$. Accordingly, we can write

$$\mathbb{E}_{\mathbb{P}}[\max_{\xi\in\Omega}\{\mathcal{Q}(x,\xi)-\beta\rho(\xi,z)\}]=\mathbb{E}_{\mathbb{P}}[\mathcal{Q}(x,\xi^*(z))]-\beta\mathbb{E}_{\mathbb{P}}[\rho(\xi^*(z),z)].$$

Considering the assumption made in (20), we have

$$\mathbb{E}_{\mathbb{P}}[\mathcal{Q}(x,\xi^*(z))]-\beta\mathbb{E}_{\mathbb{P}}[\rho(\xi^*(z),z)\}]\leq M-\beta\mathbb{E}_{\mathbb{P}}[\rho(\xi^*(z),z)\}].$$

Now we can argue that $\mathbb{E}_{\mathbb{P}}[\rho(\xi^*(z),z)\}]=0$. If not, then $\mathbb{E}_{\mathbb{P}}[\rho(\xi^*(z),z)\}]>0$ since $\rho(\xi^*(z),z)$ is nonnegative. Accordingly, we have

$$\mathbb{E}_{\mathbb{P}}[\mathcal{Q}(x,\xi^*(z))]-\beta\mathbb{E}_{\mathbb{P}}[\rho(\xi^*(z),z)\}]\leq M-\beta\mathbb{E}_{\mathbb{P}}[\rho(\xi^*(z),z)\}]\to-\infty \text{ by letting } \beta\to\infty,$$

which contradicts the fact that $\min_{\beta\geq 0}\mathbb{E}_{\mathbb{P}}[\max_{\xi\in\Omega}\{\mathcal{Q}(x,\xi)-\beta\rho(\xi,z)\}]$ is bounded. Therefore, $\mathbb{E}_{\mathbb{P}}[\rho(\xi^*(z),z)\}]=0$. Furthermore, since for any $z$, $\rho(\xi^*(z),z)\geq 0$, we have $\rho(\xi^*(z),z)=0$ holds for any $z\in\Omega$. It means $\xi^*(z)=z$, $\forall z\in\Omega$. In this case,

$$\begin{aligned}
&\min_{\beta\geq 0}\mathbb{E}_{\mathbb{P}}[\max_{\xi\in\Omega}\{\mathcal{Q}(x,\xi)-\beta\rho(\xi,z)\}]\\
=\ &\min_{\beta\geq 0}\mathbb{E}_{\mathbb{P}}[\{\mathcal{Q}(x,\xi^*(z))-\beta\rho(\xi^*(z),z)\}]\\
=\ &\min_{\beta\geq 0}\mathbb{E}_{\mathbb{P}}[\{\mathcal{Q}(x,z)-\beta\rho(z,z)\}]=\mathbb{E}_{\mathbb{P}}[\mathcal{Q}(x,z)].
\end{aligned}$$

Thus, (23) holds. Therefore, we have

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

where the first equation holds following (22) and the second equation holds following (23). Therefore, the claim (18) is proved and accordingly the overall conclusion holds. □

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

**Theorem 2.** *Corresponding to each predefined confidence level $\beta$, as the amount of historical data increases to infinity, the optimal objective value of the data-driven risk-averse stochastic optimization problem 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 (5). 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 $\theta$ and $\psi(0)$ represents the optimal objective value of (SP). Meanwhile, for the convenience of analysis, corresponding to each given first-stage solution $x$ for (DD-SP) with the distance value $\theta$, 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 $\theta$ and the first-stage solution fixed as the optimal first-stage solution to (SP) (denoted as $x^*$). According to Proposition 5, for any arbitrary small positive number $\epsilon$, there exists a $\Delta_\theta > 0$ such that

$$
|V_\theta(x^*) - V_0(x^*)| \leq \epsilon, \quad \forall \theta \leq \Delta_\theta.
$$

Then, for any $\theta \leq \Delta_\theta$, by denoting the optimal solution to (DD-SP) as $x^*_\theta$, we have

$$
|\psi(\theta) - \psi(0)| = \psi(\theta) - \psi(0) = V_\theta(x^*_\theta) - V_0(x^*) \leq V_\theta(x^*) - V_0(x^*) \leq |V_\theta(x^*) - V_0(x^*)| \leq \epsilon,
$$

where the first inequality follows from the fact that $x^*_\theta$ is the optimal solution to (DD-SP) with the distance value $\theta$ and $x^*$ is a feasible solution to this same problem. Therefore, the claim is proved. □

# 5 Solution Approaches

In this section, we discuss a solution approach to solve the data-driven risk-averse stochastic optimization problem, i.e., formulation (RDD-SP) as shown in (17), when a finite set of historical data is given. We develop a Benders' decomposition algorithm to solve the problem. We first take the dual of the formulation for the second-stage cost (i.e., $\mathcal{Q}(x,\xi)$) (1) and combine it with the second-stage problem in (17) to obtain the following subproblem (denoted as SUB) corresponding to each sample $\xi^j$, $j = 1, \cdots, N$:

$$\begin{aligned} & \phi^j(x) = \quad \max_{\xi\in\Omega,\lambda} \ (b(\xi) - A(\xi)x)^\top \lambda - \beta\rho^j(\xi) \\ \text{(SUB)} \qquad & s.t. \qquad B^\top\lambda \le d, \\ & \qquad\qquad \lambda \ge 0, \end{aligned}$$

where $\lambda$ is the dual variable. Now letting $\psi^j$ represent the optimal objective value of the subproblem (SUB) with respect to sample $\xi^j$, we can obtain the following master problem:

$$\begin{aligned} \min_{x\in X,\beta\ge 0} \qquad & c^\top x + \theta\beta + \frac{1}{N}\sum_{j=1}^{N}\psi^j \\ s.t. \qquad & \text{Feasibility cuts,} \\ & \text{Optimality cuts.} \end{aligned}$$

The above problem can be solved by adding feasibility and optimality cuts iteratively. Note here that in the subproblem (SUB), we have a nonlinear term $(b(\xi) - A(\xi)x)^\top\lambda$. In the following part, we propose two separation approaches to address the nonlinear term.

### 5.1 Exact Separation Approach

There exist exact separation approaches for problem settings in which the supporting space $\Omega$ of the random parameters is defined in (6), the distance function $\rho(.,.)$ is defined as the $L^1$-norm, i.e, $\rho(\xi, \xi^j) = \sum_{i=1}^m |\xi_i - \xi_i^j|$ for each $j = 1, \cdots N$ (similar results hold for the $L^\infty$-norm case), and $A(\xi)$ and $b(\xi)$ are assumed affinely dependent on $\xi$, i.e., $A(\xi) = A_0 + \sum_{i=1}^m A_i \xi_i$, and $b(\xi) = b_0 + \sum_{i=1}^m b_i \xi_i$, where $A_0$ and $b_0$ are the deterministic parts of $A(\xi)$ and $b(\xi)$ and $\xi_i$ is the $i$th component of vector $\xi$. Note here that the right-hand-side uncertainty case (i.e., $Ax + By \geq \xi$) is a special case of this problem setting by letting $A_0 = A$, $A_i$ be a zero matrix, $b_0 = 0$, and $b_i = e_i$ (i.e., the unit vector with the $i$th component to be 1). Then, corresponding to each sample $\xi^j$, $j = 1, \cdots, N$, the objective function of (SUB) can be reformulated as

$$\phi^j(x) = \max_{\xi \in \Omega, \lambda} (b_0 - A_0 x)^\top \lambda + \sum_{i=1}^m (b_i - A_i x)^\top \lambda \xi_i - \beta \sum_{i=1}^m |\xi_i - \xi_i^j|.$$

Now we derive the optimality conditions for $\xi$. It can be observed that the optimal solution $\xi^*$ to (SUB) should satisfy the following proposition.

**Lemma 1.** *For the subproblem (SUB) corresponding to each historical sample $\xi^j$, there exists an optimal solution $(\xi^*, \lambda^*)$ such that $\xi_i^* = W_i^-$, $\xi_i^* = W_i^+$, or $\xi_i^* = \xi_i^j$ for each $i = 1, 2, \cdots, m$.*

*Proof.* For a fixed solution $\lambda^*$, obtaining an optimal solution $\xi$ to the problem (SUB) is equivalent to obtaining an optimal solution to $\max_{\xi \in \Omega} \sum_{i=1}^m (b_i - A_i x)^\top \lambda^* \xi_i - \beta \sum_{i=1}^m |\xi_i - \xi_i^j|$, so it can be easily observed that at least one optimal solution $\xi^*$ to the subproblem (SUB) satisfies $\xi_i^* = W_i^-$, $\xi_i^* = W_i^+$, or $\xi_i^* = \xi_i^j$, for each $i = 1, 2, \cdots, m$. □

We let binary variables $z_i^+$ and $z_i^-$ indicate the case in which $\xi_i^*$ achieves its upper bound and lower bound respectively, i.e., $z_i^+ = 1 \Leftrightarrow \xi_i^* = W_i^+$ and $z_i^- = 1 \Leftrightarrow \xi_i^* = W_i^-$. Then, based on Lemma 1, there exists an optimal solution $\xi^*$ of (SUB) satisfying the following constraints:

$$\xi_i = (W_i^+ - \xi_i^j) z_i^+ + (W_i^- - \xi_i^j) z_i^- + \xi_i^j, \quad \forall i = 1, \cdots, m, \tag{24}$$

$$z_i^+ + z_i^- \leq 1, \quad z_i^+, z_i^- \in \{0, 1\}, \quad \forall i = 1, \cdots, m, \tag{25}$$

which indicate that for each $i = 1, \cdots, m$, the $i$th component of optimal solution $\xi^*$ can achieve

its lower bound $W_i^-$ ($z_i^+ = 0, z_i^- = 1$), upper bound $W_i^+$ ($z_i^+ = 1, z_i^- = 0$), or the sample value $\xi_i^j$ ($z_i^+ = z_i^- = 0$). By letting $\sigma_i = (b_i - A_i x)^\top \lambda$, with constraints (24) and (25), we can linearize the bilinear term $\sum_{i=1}^m (b_i - A_i x)^\top \lambda \xi_i$ as follows:

$$\begin{aligned}
\textstyle\sum_{i=1}^m (b_i - A_i x)^\top \lambda \xi_i \quad &= \sum_{i=1}^m \xi_i \sigma_i = \sum_{i=1}^m \left( (W_i^+ - \xi_i^j) z_i^+ + (W_i^- - \xi_i^j) z_i^- + \xi_i^j \right) \sigma_i \\
&= \sum_{i=1}^m \left( (W_i^+ - \xi_i^j) z_i^+ \sigma_i + (W_i^- - \xi_i^j) z_i^- \sigma_i + \xi_i^j \sigma_i \right) \\
&= \sum_{i=1}^m \left( (W_i^+ - \xi_i^j) \sigma_i^+ + (W_i^- - \xi_i^j) \sigma_i^- + \xi_i^j \sigma_i \right) \qquad (26) \\
\text{s.t.} \quad & \sigma_i = (b_i - A_i x)^\top \lambda, \ \forall i = 1, 2, \cdots, m \qquad (27) \\
& \sigma_i^+ \le M z_i^+, \ \forall i = 1, 2, \cdots, m \qquad (28) \\
& \sigma_i^+ \le \sigma_i + M(1 - z_i^+), \ \forall i = 1, 2, \cdots, m \qquad (29) \\
& \sigma_i^- \ge -M z_i^-, \ \forall i = 1, 2, \cdots, m \qquad (30) \\
& \sigma_i^- \ge \sigma_i - M(1 - z_i^-), \ \forall i = 1, 2, \cdots, m \qquad (31) \\
& z_i^+ + z_i^- \le 1, \ \forall i = 1, 2, \cdots, m \qquad (32) \\
& z_i^+, z_i^- \in \{0, 1\}, \ \forall i = 1, 2, \cdots, m. \qquad (33)
\end{aligned}$$

That is, we can replace the bilinear term $\sum_{i=1}^m (b_i - A_i x)^\top \lambda \xi_i$ with (26) and add constraints (27)-(33) to the subproblem (SUB).

With the reformulation of the bilinear term $\sum_{i=1}^m (b_i - A_i x)^\top \lambda \xi_i$, we can now derive Benders' feasibility and optimality cuts corresponding to each sample $\xi^j, j = 1, \ldots, N$.

Corresponding to each particular sample $\xi^j$, the feasibility check problem can be described as follows:

$$\begin{aligned}
\varpi^j(x) = \max_{\hat{z}^+, \hat{z}^-, \hat{\sigma}^+, \hat{\sigma}^-, \hat{\sigma}, \hat{\lambda}} \quad & (b_0 - A_0 x)^\top \hat{\lambda} + (W^+ - \xi^j)^\top \hat{\sigma}^+ + (W^- - \xi^j)^\top \hat{\sigma}^- + (\xi^j)^\top \hat{\sigma} \\
& - \beta \rho((W^+ - \xi^j)^\top \hat{z}^+ + (W^- - \xi^j)^\top \hat{z}^- + \xi^j, \xi^j) \\
\text{s.t.} \quad & B^\top \hat{\lambda} \le d, \\
& \text{Constraints (27) to (32) with respect to } \hat{z}^+, \hat{z}^-, \hat{\sigma}^+, \hat{\sigma}^-, \hat{\sigma}, \text{ and } \hat{\lambda}, \\
& \hat{\lambda} \in [0, 1] \text{ and } \hat{z}^+, \hat{z}^- \in \{0, 1\}.
\end{aligned}$$

For a given first-stage solution $x_0$, if $\varpi^j(x_0) = 0$, then $x_0$ is a feasible solution to (SUB). Otherwise, if $\varpi^j(x_0) > 0$, we can generate a corresponding feasibility cut $\varpi^j(x) \leq 0$ to the master problem.

For the optimality cuts, after solving the master problem and obtaining the optimal solutions $x_0$ and $\psi_0^j$, we substitute $x_0$ into (SUB) and get $\phi^j(x_0)$. If $\phi^j(x_0) > \psi_0^j$, we can generate an optimality cut $\phi^j(x) \leq \psi^j$ to the master problem.

### 5.2 Bilinear Separation Approach

In this section, we discuss the bilinear heuristic separation approach to generate Benders' cuts. We employ the similar assumptions as the ones described in the exact separation approach, except allowing the distance function $\rho(.,.)$ to be more general. The feasibility check problem corresponding to each particular sample $\xi^j$ is shown as follows (denoted as FEA):

$$\begin{aligned} \theta^j(x) = \quad & \max_{\hat{\xi}\in\Omega,\hat{\lambda}} \ (b_0 - A_0x)^\top\hat{\lambda} + \sum_{i=1}^{m}(b_i - A_ix)^\top\hat{\lambda}\hat{\xi}_i - \beta\rho^j(\hat{\xi}) \qquad (34) \\ \text{(FEA)} \qquad \text{s.t.} \quad & B^\top\hat{\lambda} \leq d, \\ & \hat{\lambda} \in [0,1]. \end{aligned}$$

Similarly, we have the bilinear term $\sum_{i=1}^{m}(b_i - A_ix)^\top\hat{\lambda}\hat{\xi}_i$ in the objective function of problem (FEA). We use the alternative bilinear separation approach to solve (FEA). First, we initiate the value of $\hat{\xi}$ as one of the extreme points of $\Omega$. Next, with the fixed $\hat{\xi}$, we solve (FEA) to obtain the corresponding optimal objective value $\theta^j(x)$ (denoted as $\theta_1^j(x,\hat{\xi})$) and optimal solution $\hat{\lambda}^*$. Then, by fixing $\hat{\lambda} = \hat{\lambda}^*$ and maximizing the objective function (34) with respect to $\hat{\xi} \in \Omega$, we can obtain the corresponding optimal objective value (denoted as $\theta_2^j(x,\hat{\lambda})$) and optimal solution $\hat{\xi}^*$. If $\theta_2^j(x,\hat{\lambda}) > \theta_1^j(x,\hat{\xi})$, we let $\hat{\xi} = \hat{\xi}^*$ and process it iteratively. Otherwise, we check whether $\theta_1^j(x,\hat{\xi}) = 0$. If so, we can terminate the feasibility check (FEA); if not, add the feasibility cut $\theta_1^j(x,\hat{\xi}) \leq 0$ to the master problem.

We can also use the bilinear heuristic approach to generate the Benders' optimality cuts. Similarly, for each $j = 1, \cdots, N$, we initiate the value of $\xi$ as one extreme point of $\Omega$, and solve the corresponding subproblem (SUB) to obtain the optimal objective value (denoted as $\phi_1^j(x,\xi)$) and optimal solution $\lambda^*$. Then, by fixing $\lambda = \lambda^*$, we solve (SUB) and obtain the optimal objective

value (denoted as $\phi_2^j(x,\lambda)$) and the corresponding optimal solution $\xi^*$ of the following problem

$$\phi_2^j(x,\lambda) = \max_{\xi\in\Omega} \sum_{i=1}^{m} (b_i - A_i x)^\top \lambda\xi_i - \beta\rho^j(\xi) + (b_0 - A_0 x)^\top \lambda.$$

If $\phi_2^j(x,\lambda) > \phi_1^j(x,\xi)$, we let $\xi = \xi^*$ and process it iteratively until $\phi_2^j(x,\lambda) \leq \phi_1^j(x,\xi)$. Then we check whether $\phi_1^j(x,\lambda) > \psi^j$. If so, we generate the corresponding optimality cut $\phi_1^j(x,\lambda) \leq \psi^j$ to the master problem.

# 6 Numerical Studies

In this section, we conduct computational experiments to show the effectiveness of the proposed data-driven risk-averse stochastic optimization model with Wasserstein metric. We test the system performance through two instances: Risk-Averse Stochastic Facility Location Problem and Risk-Averse Stochastic Unit Commitment Problem. In our experiments, we use the $L^1$-norm as described in Section 5.1 to serve as the distance measure. In addition, we set the feasibility tolerance gap to be $10^{-6}$ and the optimality tolerance gap to be $10^{-4}$. The mixed-integer-programming tolerance gap for the master problem is set the same as the CPLEX default gap. We use C++ with CPLEX 12.1 to implement the proposed formulations and algorithms. All experiments are executed on a computer workstation with 4 Intel Cores and 8GB RAM.

## 6.1 Risk-Averse Stochastic Facility Location Problem

We consider a traditional facility location problem in which there are $M$ facility locations and $N$ demand sites. We let binary variable $y_i$ represent whether facility $i$ is open ($y_i = 1$) or not ($y_i = 0$) and continuous variable $x_{ij}$ represent the amount of products to be shipped from facility $i$ to demand site $j$. For the problem parameters, we assume that the fixed cost to open facility $i$ is $F_i$ and the capacity for facility $i$ is $C_i$. We let the transportation cost for shipping one unit product from facility $i$ to demand site $j$ be $T_{ij}$. The demand for each site is uncertain (denoted as $d_j(\xi)$). Based on this setting, the corresponding data-driven risk-averse two-stage stochastic facility

location problem (DD-FL) can be formulated as follows:

$$\begin{aligned}
\min_y \quad & \sum_{i=1}^{M} F_i y_i + \max_{\mathbb{P}\in\mathcal{D}} \mathbb{E}_{\mathbb{P}} \left[ \min_{x(\xi)} \sum_{i=1}^{M}\sum_{j=1}^{N} T_{ij} x_{ij}(\xi) \right] \\
\text{s.t.} \quad & \sum_{j=1}^{N} x_{ij} \le C_i y_i, i = 1, \cdots, M, \\
& \sum_{i=1}^{M} x_{ij} = d_j(\xi), j = 1, \cdots, N, \\
& y_i \in \{0,1\}, x_{ij} \ge 0, i = 1, \cdots, M, j = 1, \cdots, N.
\end{aligned}$$

For illustration purpose, in our experiment setting, we assume there are five facility locations and three demand sites. For each location $i$, the capacity is set as $15 + i$ and the fixed cost is set as $100 + i$. Meanwhile, the unit shipping cost from location $i$ to demand site $j$ is set as $5 + 0.008i$. Finally, the demand for each site is assumed within the interval $[100, 200]$.

We first study the value of data by comparing (DD-FL) with the traditional stochastic facility location model (S-FL). To generate the series of historical data, we take samples from a normal distribution for the demand, and compare the objective values with different number of samples. We set the confidence level to be 95% and test the sample sizes 10, 50, 100, 500 and 1000. We report the objective values and CPU times in seconds (in the columns labelled "T(S-FL)" and "T(DD-FL)") of each model in Table 1. In addition, we calculate the gap (in the column labelled "Gap") between two approaches and the value of data (in the column labelled "VoD") by using the following expressions:

$$\begin{aligned}
\text{Gap}(n) &= \text{Obj}(n) - \text{Obj}_0(n), \\
\text{VoD}(n) &= \text{Obj}(n) - \text{Obj}(n-1),
\end{aligned}$$

where $\text{Obj}(n)$ and $\text{Obj}_0(n)$ represent the optimal objective values of (DD-FL) and (S-FL) with the sample size $n$, respectively.

From Table 1 and Figure 1, we can observe that, the more historical data we have, the smaller value the $\theta$ is, and the smaller total cost incurs. This is because as the number of samples increases, the confidence set shrinks and the conservatism of the problem decreases. Therefore the worst-case

| # of Samples | $\theta$ | Obj (DD-FL) | Obj (S-FL) | Gap | VoD | T (S-FL) | T (DD-FL) |
|---|---|---|---|---|---|---|---|
| 10 | 5.47 | 1789.35 | 1761.94 | 27.40 | NA | 0 | 0 |
| 50 | 2.45 | 1746.03 | 1733.78 | 12.25 | 1.08 | 1 | 1 |
| 100 | 1.73 | 1733.84 | 1725.17 | 8.67 | 0.24 | 3 | 3 |
| 500 | 0.77 | 1719.02 | 1715.15 | 3.87 | 0.04 | 57 | 18 |
| 1000 | 0.55 | 1682.80 | 1681.30 | 1.50 | 0.07 | 279 | 50 |

Table 1: Comparison between S-FL and DD-FL Approaches

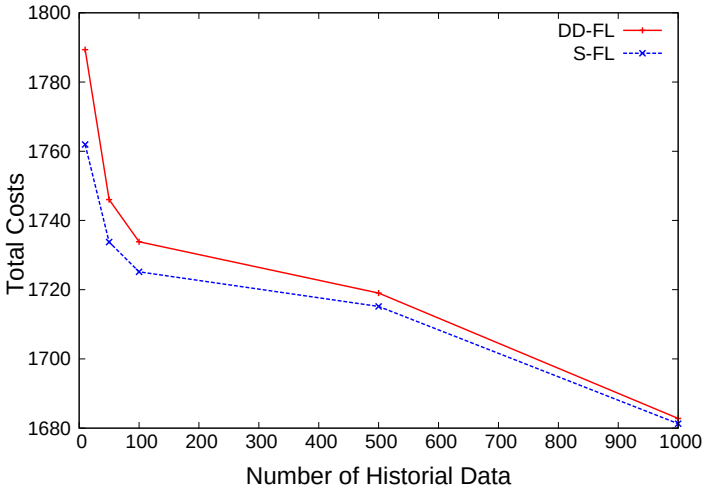


Figure 1: Effects of Historical Data

cost is reduced. Also, as the number of historical data samples increases, the CPU time increases, which is due to the fact that we have more subproblems to solve as the number of historical data samples increases. In addition, as the number of historical data samples increases, the "Gap" and "VoD" value decrease. When the number of data is larger than 100, there is little value to obtain with additional data, which means it is not necessary to collect a huge amount of historical data samples in order to obtain a robust and cost-effective solution.

Next, we study the effect of confidence level as described in Table 2 and Figure 2. We set the number of historical data samples to be 100, and consider the confidence levels to be 0.99, 0.9, 0.8, 0.7, and 0.6, respectively. We report the objective values and CPU times in Table 2.

From the results in Table 2 and Figure 2, we can observe that, as the confidence level increases, the total cost increases. This is because as the confidence level increases, we have a larger confidence set, and accordingly the problem becomes more conservative.

| C.L. | $\theta$ | Obj | CPU (sec) |
|---|---|---|---|
| 0.6 | 0.96 | 1709.55 | 4 |
| 0.7 | 1.10 | 1710.26 | 3 |
| 0.8 | 1.27 | 1711.11 | 3 |
| 0.9 | 1.16 | 1712.36 | 4 |
| 0.99 | 2.15 | 1715.51 | 3 |

Table 2: Effects of Confidence Level

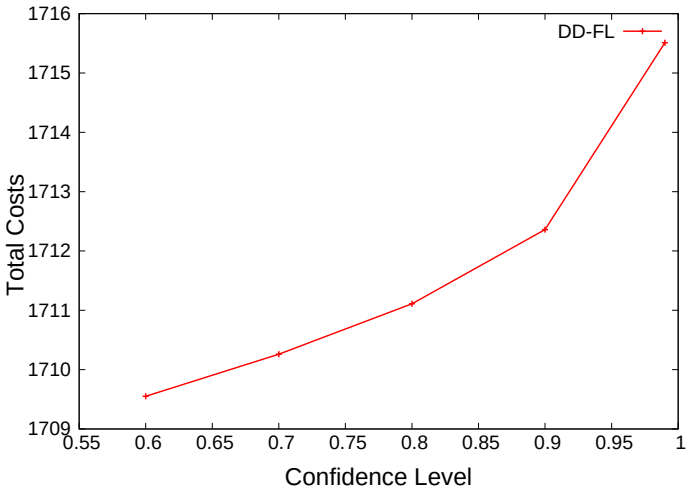


Figure 2: Effects of Confidence Level

### 6.2 Risk-Averse Stochastic Unit Commitment Problem

Now we apply our proposed modeling and solution framework to solve the risk-averse stochastic unit commitment problem. We first introduce the detailed formulation of the problem. For a $T$-period network-constrained unit commitment problem, we let $\mathcal{B}$, $\mathcal{G}$, and $\mathcal{E}$ represent the sets of buses, generators, and transmission lines, respectively. The parameters of each generator $i$ include start-up cost ($\bar{S}_i$), shut-down cost ($\hat{S}_i$), minimum-up time ($\bar{H}_i$), minimum-down time ($\hat{H}_i$), minimum ($L_i$) and maximum ($U_i$) generating capacity, and ramp-up ($\bar{R}_i$) and ramp-down ($\hat{R}_i$) rate limits. For each transmission line $(i,j) \in \mathcal{E}$, we let $C_{ij}$ represent the transmission capacity and $K_{ij}^n$ represent the line flow distribution factor due to the net injection at bus $n$. In addition, for each bus $b$, we let $D_t^b$ represent the demand and $W_t^b(\xi)$ represent the uncertain renewable generation in time $t$. For decision variables, we let the first-stage decision variables include the unit on/off status $y_{it}$, start-up status $u_{it}$, and shut-down status $v_{it}$, of generator $i$ at time $t$. In addition, we let the

second-stage decision variables include the economic dispatch amount $x_{it}$ and the auxiliary variable $F_i(.)$ to help represent the fuel cost function. Accordingly, the data-driven risk-averse stochastic unit commitment problem (DD-SUC) can be described as follows:

$$\min \quad \sum_{t\in\mathcal{T}}\sum_{i\in\mathcal{G}}(\bar{S}_i u_{it} + \hat{S}_i v_{it}) + \max_{\hat{\mathbb{P}}\in\mathcal{D}} \mathbb{E}_{\hat{\mathbb{P}}}[Q(y,u,v,\xi)] \tag{35}$$

$$\text{s.t.} \quad -y_{i(t-1)} + y_{it} - y_{ik} \leq 0, \qquad \forall t\in\mathcal{T}, \forall i\in\mathcal{G}, 1 \leq k-(t-1) \leq \bar{H}_i, \tag{36}$$

$$y_{i(t-1)} - y_{it} + y_{ik} \leq 1, \qquad \forall t\in\mathcal{T}, \forall i\in\mathcal{G}, 1 \leq k-(t-1) \leq \hat{H}_i, \tag{37}$$

$$-y_{i(t-1)} + y_{it} - u_{it} \leq 0, \qquad \forall t\in\mathcal{T}, \forall i\in\mathcal{G}, \tag{38}$$

$$y_{i(t-1)} - y_{it} - v_{it} \leq 0, \qquad \forall t\in\mathcal{T}, \forall i\in\mathcal{G}, \tag{39}$$

$$y_{it}, u_{it}, v_{it} \in \{0,1\}, \qquad \forall t\in\mathcal{T}, \forall i\in\mathcal{G}, \tag{40}$$

where $Q(y,u,v,\xi)$ is equal to

$$\min \quad \sum_{t\in\mathcal{T}}\sum_{i\in\mathcal{G}} F_i(x_{it}(\xi)) \tag{41}$$

$$s.t. \quad L_i y_{it} \leq x_{it}(\xi) \leq U_i y_{it}, \quad \forall t\in\mathcal{T}, \forall i\in\mathcal{G}, \tag{42}$$

$$x_{it}(\xi) - x_{i(t-1)}(\xi) \leq (2 - y_{i(t-1)} - y_{it})L_i + (1 + y_{i(t-1)} - y_{it})\bar{R}_i, \quad \forall t\in\mathcal{T}, \forall i\in\mathcal{G}, \tag{43}$$

$$x_{i(t-1)}(\xi) - x_{it}(\xi) \leq (2 - y_{i(t-1)} - y_{it})L_i + (1 - y_{i(t-1)} + y_{it})\hat{R}_i, \quad \forall t\in\mathcal{T}, \forall i\in\mathcal{G}, \tag{44}$$

$$\sum_{i\in\mathcal{G}} x_{it}(\xi) + \sum_{b\in\mathcal{B}} W_t^b(\xi) = \sum_{b\in\mathcal{B}} D_t^b, \quad \forall t\in\mathcal{T}, \tag{45}$$

$$-C_{ij} \leq \sum_{b\in\mathcal{B}} K_{ij}^b \left( \sum_{i\in\mathcal{G}_b} x_{it}(\xi) + W_t^b(\xi) - D_t^b \right) \leq C_{ij}, \quad \forall t\in\mathcal{T}, (i,j)\in\mathcal{E}. \tag{46}$$

In the above formulation, the objective function (35) is composed of the unit commitment cost and the expected fuel cost under the worst-case distribution. Constraints (36) and (37) describe minimum up-time and minimum down-time for each unit, respectively. Constraints (38) and (39) indicate the start-up and shut-down status of each generator, and the corresponding costs are incurred in the objective. Constraints (42) represent the lower and upper generation amount limits of each generator. Constraints (43) and (44) indicate the ramping up and ramping down restrictions. Constraints (45) ensure that the energy supply (thermal generation and renewable generation) and the load are balanced. Finally, constraints (46) represent the transmission capacity restrictions.

We test the performance of our proposed approach on a modified IEEE 118-bus system, based on the one given online at *http://motor.ece.iit.edu/data*. The system contains 118 buses, 33 generators, and 186 transmission lines. The operational time interval is set as 24 hours. In our experiment, we compare the performance of our approach with the traditional two-stage stochastic optimization approach. We introduce a penalty cost at \$5000/MWh [4] for any power imbalance or transmission capacity/ramp-rate limit violation. Besides, we assume the uncertain renewable energy generation follows a multivariate normal distribution to generate samples to create the series of historical data. In addition, we set the confidence level as 99%. In the implementation of comparing our data-driven risk-averse stochastic optimization approach (DD-SUC) with the traditional two-stage stochastic optimization approach (SO), we obtain the total costs corresponding to each approach through the following steps: 1) Obtain the unit commitment decisions by using the (DD-SUC) and (SO) approaches, respectively; 2) Fix the obtained unit commitment decisions and solve the second-stage problem repeatedly for 50 randomly generated samples, following the true distribution, to obtain the total costs for the (DD-SUC) and (SO) approaches, respectively.

We report the results in Table 3. The first column represents the number of samples. The third column computes the value of $\theta$ based on Proposition 1. The numbers of start-ups are given in the fourth column, and unit commitment costs are provided in the fifth column. Finally, the sixth column reports the total costs.

| # of Samples | Model | $\theta$ | # of Start-ups | UC.C.(\$) | T.C.(\$) |
|---|---|---|---|---|---|
| 1 | SO | 26.63 | 21 | 5440 | 602828 |
| | DD-SUC | 26.63 | 29 | 5905 | 604425 |
| 10 | SO | 8.42 | 21 | 5440 | 602823 |
| | DD-SUC | 8.42 | 29 | 5865 | 544465 |
| 50 | SO | 3.77 | 22 | 5495 | 602884 |
| | DD-SUC | 3.77 | 28 | 5805 | 544394 |
| 100 | SO | 2.66 | 23 | 5555 | 602960 |
| | DD-SUC | 2.66 | 28 | 5805 | 543703 |

Table 3: Comparison between (DD-SUC) and (SO) Approaches

From Table 3, we have the following two key observations:

(1) On one hand, as the number of historical data samples increases, the number of start-ups obtained by the proposed (DD-SUC) approach, the unit commitment cost, and the total cost decrease. This is because when more historical data samples are observed, the confidence

set shrinks, which leads to less conservative solutions. On the other hand, as the number of historical data samples increases, the number of start-ups obtained by the (SO) approach increases, and the unit commitment cost increases, since more historical data lead to more robust solutions for the (SO) approach.

(2) As compared to the (SO) approach, the (DD-SUC) approach brings more generators online to provide sufficient generation capacity to maintain the system reliability. As a result, the (DD-SUC) approach has a larger unit commitment cost than the (SO) approach. This result verifies that the proposed (DD-SUC) approach is more robust than the (SO) approach. In addition, since the (DD-SUC) approach is more robust than the (SO) approach due to committing more generators, this leads to smaller penalty costs for certain generated scenarios, which eventually makes the total expected costs reduced.

Therefore, in general, our (DD-SUC) approach provides a more reliable while cost-effective solution as compared to the (SO) approach.

# 7 Summary

In this paper, we proposed one of the first studies on data-driven stochastic optimization with Wasserstein metric. This approach can fit well in the data-driven environment. Based on a given set of historical data, which corresponds to an empirical distribution, we can construct a confidence set for the unknown true probability distribution using the Wasserstein metric through statistical nonparametric estimation, and accordingly develop a data-driven risk-averse two-stage stochastic optimization framework. The derived formulation can be reformulated into a tractable two-stage robust optimization problem. Moreover, we obtained the corresponding worst-case distribution and demonstrated the convergence of our risk-averse model to the traditional risk-neutral one as the number of historical data samples increases to infinity. Furthermore, we obtained a stronger result in terms of deriving the convergence rate and providing the closed-form expression for the distance value (in constructing the confidence set) as a function of the size of available historical data, which shows the value of data. Finally, we applied our solution framework to solve the stochastic facility location and stochastic unit commitment problems respectively, and the experiment results verified the effectiveness of our proposed approach.

## References


[1] A. Ben-Ta, D. Den Hertog, A. De Waegenaere, B. Melenberg, and G. Rennen. Robust solutions of optimization problems affected by uncertain probabilities. *Management Science*, 59(2):341–357, 2013.

[2] A. Ben-Tal and A. Nemirovski. Robust convex optimization. *Mathematics of Operations Research*, 23(4):769–805, 1998.

[3] A. Ben-Tal and A. Nemirovski. Robust solutions of uncertain linear programs. *Operations Research Letters*, 25(1):1–13, 1999.

[4] D. Bertsimas, E. Litvinov, X. A. Sun, J. Zhao, and T. Zheng. Adaptive robust optimization for the security constrained unit commitment problem. *IEEE Transactions on Power Systems*, to be published.

[5] D. Bertsimas and M. Sim. The price of robustness. *Operations Research*, 52(1):35–53, 2004.

[6] P. Billingsley. *Convergence of Probability Measures*. John Wiley & Sons, 2013.

[7] J. Birge and F. Louveaux. *Introduction to Stochastic Programming*. Springer, 1997.

[8] G. C. Calafiore and L. El Ghaoui. On distributionally robust chance-constrained linear programs. *Journal of Optimization Theory and Applications*, 130(1):1–22, 2006.

[9] E. Delage and Y. Ye. Distributionally robust optimization under moment uncertainty with application to data-driven problems. *Operations Research*, 58(3):595–612, 2010.

[10] Y. Deng and W. Du. The Kantorovich metric in computer science: A brief survey. *Electronic Notes in Theoretical Computer Science*, 253(3):73–82, 2009.

[11] L. El Ghaoui and H. Lebret. Robust solutions to least-squares problems with uncertain data. *SIAM Journal on Matrix Analysis and Applications*, 18(4):1035–1064, 1997.

[12] L. El Ghaoui, M. Oks, and F. Oustry. Worst-case value-at-risk and robust portfolio optimization: A conic programming approach. *Operations Research*, 51(4):543–556, 2003.

[13] L. El Ghaoui, F. Oustry, and H. Lebret. Robust solutions to uncertain semidefinite programs. *SIAM Journal on Optimization*, 9(1):33–52, 1998.

[14] E. Erdoğan and G. Iyengar. Ambiguous chance constrained problems and robust optimization. *Mathematical Programming*, 107(1):37–61, 2006.

[15] L. V. Kantorovich and G. S. Rubinshtein. On a space of totally additive functions. *Vestn Lening. Univ.*, 13(7):52–59, 1958.

[16] S. Mehrotra and H. Zhang. Models and algorithms for distributionally robust least squares problems. *Mathematical Programming*, in press, 2013.

[17] E. Parzen. On estimation of a probability density function and mode. *The Annals of Mathematical Statistics*, pages 1065–1076, 1962.

[18] G. Pflug and D. Wozabal. Ambiguity in portfolio selection. *Quantitative Finance*, 7(4):435–442, 2007.

[19] S. T. Rachev. *Mass Transportation Problems*, volume 2. Springer, 1998.

[20] M. Rosenblatt. Remarks on some nonparametric estimates of a density function. *The Annals of Mathematical Statistics*, 27(3):832–837, 1956.

[21] A. Shapiro, D. Dentcheva, and A. Ruszczyński. *Lectures on Stochastic Programming: Modeling and Theory*, volume 9. SIAM, 2009.

[22] A. W. Van der Vaart. *Asymptotic Statistics*, volume 3. Cambridge University Press, 2000.

[23] L. Vandenberghe, S. Boyd, and K. Comanor. Generalized Chebyshev bounds via semidefinite programming. *SIAM Review*, 49(1):52–64, 2007.

[24] A. M. Vershik. Kantorovich metric: Initial history and little-known applications. *Journal of Mathematical Sciences*, 133(4):1410–1417, 2006.

[25] C. Villani. *Topics in Optimal Transportation*. Number 58. American Mathematical Society, 2003.

[26] J. Wolfowitz. Generalization of the theorem of glivenko-cantelli. *The Annals of Mathematical Statistics*, pages 131–138, 1954.

[27] C. Zhao. *Data-Driven Risk-Averse Stochastic Program and Renewable Energy Integration*. Ph.D. Dissertation, University of Florida, August 5, 2014.

[28] C. Zhao and Y. Guan. Data-driven risk-averse two-stage stochastic program with $\zeta$-structure probability metrics. *Submitted for Publication*, 2015.

[29] C. Zhao and Y. Guan. Data-driven risk-averse two-stage stochastic program. *Technical Report, University of Florida*, March 5, 2014.

[30] S. Zymler, D. Kuhn, and B. Rustem. Distributionally robust joint chance constraints with second-order moment information. *Mathematical Programming*, 137(1-2):167–198, 2013.

[31] S. Zymler, D. Kuhn, and B. Rustem. Worst-case value at risk of nonlinear portfolios. *Management Science*, 59(1):172–188, 2013.