A Data-Driven Distributionally Robust Bound on the Expected Optimal Value of Uncertain Mixed 0-1 Linear Programming

Guanglin Xu* Samuel Burer†

August 24, 2017

Abstract

This paper studies the expected optimal value of a mixed 0-1 programming problem with uncertain objective coefficients following a joint distribution. We assume that the true distribution is not known exactly, but a set of independent samples can be observed. Using the Wasserstein metric, we construct an ambiguity set centered at the empirical distribution from the observed samples and containing the true distribution with a high statistical guarantee. The problem of interest is to investigate the bound on the expected optimal value over the Wasserstein ambiguity set. Under standard assumptions, we reformulate the problem into a copositive program, which naturally leads to a tractable semidefinite-based approximation. We compare our approach with a moment-based approach from the literature on three applications. Numerical results illustrate the effectiveness of our approach.

Keywords: Distributionally robust optimization; Wasserstein metric; copositive programming; semidefinite programming

1 Introduction

We consider the following uncertain mixed 0-1 linear programming problem:

$$v(\xi) := \max \left\{ (F\xi)^T x : \begin{array}{l} Ax = b,\ x \geq 0 \\ x_j \in \{0,1\}\ \forall\, j \in \mathcal{B} \end{array} \right\} \tag{1}$$

where $A \in \mathbb{R}^{m\times n}$, $F \in \mathbb{R}^{n\times k}$, and $b \in \mathbb{R}^m$ are the problem data, $x \in \mathbb{R}^n_+$ is the vector of decision variables, $\mathcal{B} \subseteq \{1,\ldots,n\}$ is an index set of binary variables, and the objective

*Department of Management Sciences, University of Iowa, Iowa City, IA, 52242-1994, USA. Email: guanglin-xu@uiowa.edu.

†Department of Management Sciences, University of Iowa, Iowa City, IA, 52242-1994, USA. Email: samuel-burer@uiowa.edu.

coefficients are linear in the random vector $\xi \in \mathbb{R}^k$ via F. Problem (1) entails two extreme classes of programs: if $\mathcal{B} = \emptyset$, then (1) represents the regular linear program with uncertain objective coefficients; if $\mathcal{B} = \{1, \ldots, n\}$, then (1) represents the regular binary program with uncertain coefficients. In general, problem (1) is NP-hard [48].

The optimal value $v(\xi)$ is a random variable as ξ is a random vector. We assume that ξ follows a multivariate distribution $\mathbb{P}$ supported on a nonempty set $\Xi \subseteq \mathbb{R}^k$, which is, in particular, defined as a slice of a closed, convex, full-dimensional cone $\widehat{\Xi} \subseteq \mathbb{R}_+ \times \mathbb{R}^{k-1}$:

$$\Xi := \left\{ \xi \in \widehat{\Xi} : e_1^T \xi = \xi_1 = 1 \right\},$$

where e_1 is the first standard basis vector in $\mathbb{R}^k$. In words, $\widehat{\Xi}$ is the homogenization of Ξ. We choose this homogenized version for notational convenience. Note that it, in fact, enables us to model affine effects of the uncertain parameters in (1).

The expected optimal value of (1), denoted by $v_{\mathbb{P}}$, is defined as

$$v_{\mathbb{P}} := \mathrm{E}_{\mathbb{P}}[v(\xi)] = \int_{\Xi} v(\xi)\, d\,\mathbb{P}(\xi).$$

The problem of computing $v_{\mathbb{P}}$ has been extensively studied in the literature. Hagstrom [26] showed that computing $v_{\mathbb{P}}$ for the longest path problem over a directed acyclic graph is $\#\mathcal{P}$-complete even if the arc lengths are each independently distributed and restricted to taking two possible values. Aldous [1] studied a linear assignment problem with random cost coefficients following either an independent uniform distribution on $[0, 1]$ or an exponential distribution with parameter 1 and proved that the asymptotic value of $v_{\mathbb{P}}$ approaches $\frac{\pi^2}{6}$ as the number of assignments goes to infinity. For additional studies, see [10, 18, 37].

In practice, it is difficult or impossible to know $\mathbb{P}$ completely, and computing $v_{\mathbb{P}}$ is thus not well defined in this situation. An alternative is to construct an ambiguity set, denoted by $\mathcal{D}$, that contains a family of distributions supported on Ξ and consistent with any known properties of $\mathbb{P}$. Ideally, the ambiguity set will possess some statistical guarantee, e.g., the probability that $\mathbb{P} \in \mathcal{D}$ will be at least $1 - \beta$, where β is the significance level. In analogy with $v_{\mathbb{P}}$, we define $v_{\mathbb{Q}}$ for any $\mathbb{Q} \in \mathcal{D}$. Then, we are interested in computing the maximum expected optimal value $v_{\mathbb{Q}}$ over the ambiguity set $\mathcal{D}$:

$$v_{\mathcal{D}}^+ := \sup_{\mathbb{Q} \in \mathcal{D}} v_{\mathbb{Q}}. \tag{2}$$

Note that, when the probability of $\mathbb{P} \in \mathcal{D}$ is at least $1 - \beta$, the probability of $v_{\mathbb{P}} \leq v_{\mathcal{D}}^+$ is at least $1 - \beta$.

There are three main issues (somehow conflicting) regarding the computation of $v_{\mathcal{D}}^+$. First, one would like an ambiguity set $\mathcal{D}$ with a high statistical guarantee to contain the true distribution $\mathbb{P}$. In this way, the computed $v_{\mathcal{D}}^+$ will be an upper bound on $v_{\mathbb{P}}$ with a high confidence level. (We will introduce several approaches in the following paragraph.) Second, one would like $v_{\mathcal{D}}^+$ to be tight in the sense that it is as close to $v_{\mathbb{P}}$ as possible. Generally, if $\mathcal{D}$ enforces more information about $\mathbb{P}$, then $v_{\mathcal{D}}^+$ will be closer to $v_{\mathbb{P}}$. Finally, the third concern is the complexity of the resulting optimization problem, i.e., whether the problem can be solved in polynomial time.

Bertsimas et al. [7, 8] constructed moment ambiguity sets using the first two marginal moments of each ξ_i. Denote the first and second of each uncertain parameter by μ_i and σ_i respectively. They computed $v_{\mathcal{D}}^+$ over all joint distributions sharing the same first two marginal moments and proved polynomial-time computability if the corresponding deterministic problem is solvable in polynomial time. However, the computed bound may not be tight with respect to $v_{\mathbb{P}}$ since the marginal-moment model does not capture the dependence of the random variables. In a closely related direction, Natarajan et al. [38] proposed an ambiguity set that was constructed from the known marginal distributions of each random variable ξ_i, and they computed $v_{\mathcal{D}}^+$ by solving a concave maximization problem. As an extension to the marginal moment-based approach, Natarajan et al. [40] proposed a cross-moment model that was based on an ambiguity set constructed using both marginal and cross moments. Compared to the marginal-moment approach, the cross-moment approach has tighter upper bounds as the model captures the dependence of the random variables. However, computing the bound requires solving a completely positive program, which itself can only be approximated in general. Thus, the authors proposed semidefinite programming (SDP) relaxations to approximate $v_{\mathcal{D}}^+$.

Moment-based ambiguity sets are also used prominently in a parallel vein of research, called distributionally robust optimization (DRO); see [9, 15, 16, 17, 20, 22, 24, 29, 39, 46, 49, 50]. The popularity of the moment-based approach is mainly due to the fact that it often leads to tractable optimization problems and relatively simple models. Its weakness, however, is that moment-based sets are not guaranteed to converge to the true distribution $\mathbb{P}$ when the sample size increases to infinity, even though the estimations of the first and second moments are themselves guaranteed to converge.

As an attractive alternative to moment-based ambiguity sets, distance-based ambiguity sets haven been proposed in recent years. This approach defines $\mathcal{D}$ as a ball in the space of probability distributions equipped with a distance measure, and the center of the ball is typically the empirical distribution derived from a series of independent realizations of the random vector ξ. The key ingredient of this approach is the distance function. Classical dis-

tance functions include the Kullback-Leibler divergence [31, 32], the ϕ-divergence [4, 19, 33], the Prohorov metric [22], empirical Burg-entropy divergence balls [34], and the Wasserstein metric [41, 47].

In this paper, we apply the Wasserstein metric to construct a data-driven ambiguity set $\mathcal{D}$ centered at the empirical distribution $\widehat{\mathbb{P}}_N$ derived from N independent observations of ξ. This approach has several benefits. The conservativeness of the ambiguity set can be controlled by tuning a single parameter, the radius of the Wasserstein ball; we will discuss this parameter in detail in Section 2. Also, under mild conditions on $\mathbb{P}$, the Wasserstein ambiguity provides a natural confidence set for $\mathbb{P}$. Specifically, the Wasserstein ball around the empirical distribution on N independent identical samples contains $\mathbb{P}$ with confidence $1-\beta$ if its radius exceeds an explicit threshold $\epsilon_N(\beta)$ that can be computed via a closed form equation [21, 23]. We then formulate $v_{\mathcal{D}}^+$ in (2) over the constructed Wasserstein ambiguity set. That is, we model the maximum value of $v_{\mathbb{Q}}$ over the ambiguity set $\mathcal{D}$ constructed by the Wasserstein metric. In Section 3, we reformulate problem (2) into a copositive problem under some standard assumptions. As the copositive reformulation is computationally intractable, we apply a standard approach based on semidefinite programming techniques to approximate $v_{\mathcal{D}}^+$ from above. In Section 4, we numerically verify our approach on three applications from the literature. In particular, we compare our approach with the moment-based approach proposed in [40]. We have several important observations from the experimental results. First, we find that the gaps between the bound from our semidefinite programs and the true expected optimal value becomes narrower as the sample size increases. However, the moment-based bound remains the same regardless of the increase in the sample size. Second, we observe that our bound converges to the true expected optimal value on the first two applications where the underlying deterministic problems are linear programs. Although our bound on the third application is not able to converge to the true expected optimal value, it is tighter than the moment-based bound after the sample size increases to a certain level. We conclude our research and discuss some future directions in Section 5.

We point out some similarities of our paper to a recent technical report by Hanasusanto and Kuhn [28]. In their report, they proposed a Wasserstein-metric ambiguity set for a two-stage DRO problem. In particular, they applied copositive programming techniques to reformulate the second-stage worst-case value function, which is essentially a max-min optimization problem, while we use copositive techniques to reformulate a max-max optimization problem; see (4). Furthermore, they directly used a hierarchy schema to approximate the copositive cones, while we derive natural SDP approximations based on the copositive reformulation. Note that their hierarchy of approximations lead to SDP approximations as well. Finally, they developed an approach to derive an empirical Wasserstein radius, which is in

spirit similar to our approach in this paper.

1.1 Notation, terminology, and basic techniques

We denote by $\mathbb{R}^n$ the n-dimensional Euclidean space and by $\mathbb{R}^n_+$ the nonnegative orthant in $\mathbb{R}^n$. For a scalar $p \geq 1$, the p-norm of $z \in \mathbb{R}^n$ is defined $\|z\|_p := (\sum_{i=1}^n |z_i|^p)^{1/p}$, e.g., $\|z\|_1 = \sum_{i=1}^n |z_i|$. We will drop the subscript for the 2-norm, i.e., $\|z\| := \|z\|_2$. For $v, w \in \mathbb{R}^n$, the inner product of v and w is denoted by $v^T w := \sum_{i=1}^n v_i w_i$. For the specific dimensions k and n of the problem in this paper, we denote by e_i the i-th standard basis vector in $\mathbb{R}^k$, and similarly, denote by f_j the j-th standard basis vector in $\mathbb{R}^n$. We will also define $g_1 := \binom{e_1}{0} \in \mathbb{R}^{k+n}$. We denote by δ_ξ the Dirac distribution concentrating unit mass at $\xi \in \mathbb{R}^k$. For any $N \in \mathbb{N}$, we define $[N] := \{1, \ldots, N\}$.

Let $\mathbb{R}^{m \times n}$ denote the space of real $m \times n$ matrices, and $A \bullet B := \text{trace}(A^T B)$ denote the trace of the inner product of two matrices $A, B \in \mathbb{R}^{m \times n}$. We denote by $\mathbb{S}^n$ the space of $n \times n$ symmetric matrices, and for $X \in \mathbb{S}^n$, $X \succeq 0$ represents that X is positive semidefinite. In addition, we denote by $\text{diag}(X)$ the vector containing the diagonal entries of X, and denote by $\text{Diag}(v)$ the diagonal matrix with vector v along its diagonal. $I \in \mathbb{S}^n$ denotes the identity matrix.

Finally, letting $\mathcal{K} \subseteq \mathbb{R}^n$ be a closed, convex cone, and $\mathcal{K}^*$ be its dual cone, we give a brief introduction to *copositive programming* with respect to the cone $\mathcal{K}$. The *copositive cone* with respect to $\mathcal{K}$ is defined as

$$\mathcal{COP}(\mathcal{K}) := \{M \in \mathbb{S}^n : x^T M x \geq 0 \ \forall \ x \in \mathcal{K}\},$$

and its dual cone, the *completely positive cone* with respect to $\mathcal{K}$, is given as

$$\mathcal{CP}(\mathcal{K}) := \{X \in \mathbb{S}^n : X = \textstyle\sum_i x^i (x^i)^T, \ x^i \in \mathcal{K}\},$$

where the summation over i is finite but its cardinality is unspecified. The term *copositive programming* refers to linear optimization over $\mathcal{COP}(\mathcal{K})$ or, via duality, linear optimization over $\mathcal{CP}(\mathcal{K})$. In fact, these problems are sometimes called *generalized copositive programming* or *set-semidefinite optimization* [14, 21] in contrast with the standard case $\mathcal{K} = \mathbb{R}^n_+$. In this paper, we work with generalized copositive programming, although we use the shorter phrase for convenience.

2 A Wasserstein-Based Ambiguity Set

In this section, we define the Wasserstein metric and discuss a standard method to construct a Wasserstein-based ambiguity set. Using this ambiguity set, we fully specify problem (2).

Denote by $\widehat{\Theta}_N := \{\hat{\xi}^1, \ldots, \hat{\xi}^N\}$ the set of N independent samples of ξ governed by $\mathbb{P}$. The uniform empirical distribution based on $\widehat{\Theta}_N$ is $\widehat{\mathbb{P}}_N := \frac{1}{N}\sum_{i=1}^N \delta_{\hat{\xi}^i}$ where δ_ζ is the Dirac distribution concentrating unit mass at $\zeta \in \mathbb{R}^k$.

Definition 1 (Definition 3 in [28]). *Let $\mathcal{M}^2(\Xi)$ be the set of all probability distributions $\mathbb{Q}$ that are supported on Ξ and that satisfy $\mathrm{E}_{\mathbb{Q}}[\|\xi - \xi'\|^2] = \int_\Xi \|\xi - \xi'\|^2 d\,\mathbb{Q}(\xi) < \infty$ where $\xi' \in \Xi$ is some reference point, e.g., $\xi' = \hat{\xi}^i$ for some $i \in [N]$.*

Definition 2 (Definition 3 in [28]). *The 2-Wasserstein distance between any $\mathbb{Q}, \mathbb{Q}' \in \mathcal{M}^2(\Xi)$ is*

$$W^2(\mathbb{Q}, \mathbb{Q}') := \inf \left\{ \left(\int_{\Xi^2} \|\xi - \xi'\|^2 \, \Pi(d\xi, d\xi') \right)^{1/2} : \begin{array}{l} \Pi \text{ is a joint distribution of } \xi \text{ and } \xi' \\ \text{with marginals } \mathbb{Q} \text{ and } \mathbb{Q}', \text{ respectively} \end{array} \right\}.$$

Remark 2.1. *The Wasserstein distance is essentially the minimum cost of redistributing mass from $\mathbb{Q}$ to $\mathbb{Q}'$. It is also called the "earth mover's distance" in the community of computer science; see [42]. In fact, the Wasserstein distance between two discrete distributions with a finite number of positive masses corresponds to a transportation planning problem in finite dimensions.*

Example 1 illustrates the Wasserstein distance between two discrete distributions.

Example 1. *Consider two discrete distributions: $\mathbb{Q} := \sum_{i=1}^M q_i \delta_{\xi_i}$ and $\mathbb{Q}' := \sum_{j=1}^{M'} q'_j \delta_{\xi'_j}$ where $q_i \geq 0$ $i = 1, \ldots, M$, $q'_j \geq 0$ $j = 1, \ldots, M'$, and $\sum_{i=1}^M q_i = \sum_{j=1}^{M'} q'_j = 1$. Define $c_{ij} = \|\xi_i - \xi'_j\|^2$ $\forall\, i = 1, \ldots, M$ $j = 1, \ldots, M'$. Then, the 2-Wasserstein distance between $\mathbb{Q}$ and $\mathbb{Q}'$ equals the square root of the optimal value of the following transportation planning problem:*

$$\begin{array}{ll} \min\limits_{\pi} & \sum_{i=1}^M \sum_{j=1}^{M'} c_{ij}\pi_{ij} \\ \text{s.t.} & \sum_{j=1}^{M'} \pi_{ij} = q_i \;\forall\, i = 1, \ldots, M \\ & \sum_{i=1}^{M} \pi_{ij} = q'_j \;\forall\, j = 1, \ldots, M' \\ & \pi_{ij} \geq 0 \;\forall\, i = 1, \ldots, M,\; j = 1, \ldots, M', \end{array} \tag{3}$$

where π is the joint distribution of ξ and ξ' with marginals of $\mathbb{Q}$ and $\mathbb{Q}'$ and π is the matrix variable in this optimization problem.

With this setting, our ambiguity set contains a family of distributions that are close to $\widehat{\mathbb{P}}_N$ with respect to the Wasserstein metric. In particular, we define our ambiguity set $\mathcal{D}$ as

a 2-Wasserstein ball of radius ϵ that is centered at the uniform empirical distribution $\widehat{\mathbb{P}}_N$:

$$\mathcal{D}(\widehat{\mathbb{P}}_N, \epsilon) := \left\{ \mathbb{Q} \in \mathcal{M}^2(\Xi) : W^2(\mathbb{Q}, \widehat{\mathbb{P}}_N) \leq \epsilon \right\}.$$

The reader is referred to [28] for the general case of $\mathcal{M}^r(\Xi)$ and $W^r(\mathbb{Q}, \mathbb{Q}')$ for any $r \geq 1$. We use the 2-Wasserstein distance in this paper for two reasons. First, the Euclidean distance is one of the most popular distances considered in the relevant literature; see [23, 28]. Second, we will find that problem (2) with an ambiguity set based on the 2-Wasserstein distance can be reformulated into a copositive program; see Section 3.

Then, we replace the generic ambiguity set $\mathcal{D}$ with the Wasserstein ball $\mathcal{D}(\widehat{\mathbb{P}}_N, \epsilon)$ in problem (2) to compute a data-driven upper bound:

$$\begin{array}{lll} v^+_{\mathcal{D}(\widehat{\mathbb{P}}_N, \epsilon)} = & \sup\limits_{\Pi,\ \mathbb{Q} \in \mathcal{M}^2(\Xi)} & \int_\Xi v(\xi)\ d\,\mathbb{Q}(\xi) \\ & \text{s. t.} & \int_{\Xi^2} \|\xi - \xi'\|^2\ \Pi(d\xi, d\xi') \leq \epsilon^2 \\ & & \Pi \text{ is a joint distribution of } \xi \text{ and } \xi' \\ & & \text{with marginals } \mathbb{Q} \text{ and } \widehat{\mathbb{P}}_N, \text{ respectively.} \end{array} \tag{4}$$

We next close this subsection by making some remarks. First, the Wasserstein ball radius in problem (4) controls the conservatism of the optimal value. A larger radius is more likely to contain the true distribution and thus a more likely valid upper bound on $v_{\mathbb{P}}$, but even if it is valid, it could be a weaker upper bound. Therefore, it is crucial to choose an appropriate radius for the Wasserstein ball. Second, the Kullback-Leibler divergence ball is also considered in recent research; see [31, 32]. However, in the case of our discrete empirical distribution, the Kullback-Leibler divergence ball is a singleton containing only the empirical distribution itself, with probability one. Third, the ambiguity sets constructed by goodness-of-fit tests in [5, 6] also possess statistical guarantees, however, they often lead to complicated and intractable optimization problems for the case of high-dimensional uncertain parameters.

2.1 An empirical Wasserstein radius

The papers [21, 23] present a theoretical radius $\epsilon_N(\beta)$ for datasets of size N, which guarantees a desired confidence level $1-\beta$ for $\mathbb{P} \in \mathcal{D}(\mathbb{P}_N, \epsilon_N(\beta))$ under the following standard assumption on $\mathbb{P}$:

Assumption 1 (Light-tailed distribution)**.** *There exists an exponent $a > 1$ such that*

$$\mathrm{E}_{\mathbb{P}}[\exp(\|\xi\|^a)] = \int_\Xi \exp(\|\xi\|^a) d\,\mathbb{P}(\xi) < \infty.$$

Note that $\epsilon_N(\beta)$ depends on N and β. However, $\epsilon_N(\beta)$ is known to be conservative in practice; see [21] for example. In other words, $\mathcal{D}(\mathbb{P}_N, \epsilon_N(\beta))$ might contain significantly more irrelevant distributions so that the computed $v^+_{\mathcal{D}(\mathbb{P}_N,\, \epsilon_N(\beta))}$ is significantly larger than $v_{\mathbb{P}}$. So, we propose a procedure to derive an empirical radius that provides a desired confidence level $1-\beta$ but is much smaller than $\epsilon_N(\beta)$. Our approach is based on the data set $\widehat{\Theta}_N$. In particular, we apply a procedure, similar to cross validation in spirit, that computes an empirical confidence level (between 0 and 1) for a given radius ϵ; see details in the next paragraphs. Our procedure guarantees that a larger radius leads to a higher confidence level. Therefore, by iteratively testing different ϵ, we can find a radius with a desired confidence level based on the data set $\widehat{\Theta}_N$. Although the derived $\epsilon(\widehat{\Theta}_N,\, \beta)$ depends on the data set $\widehat{\Theta}_N$, our experimental results in Section 4 indicate that it can be used for other datasets of the same sample size. We will show the numerical evidence in Section 4. Our approach is also similar in spirit to the one used in [23, 28].

Our procedure requires an oracle to compute (or approximate) $v^+_{\mathcal{D}(\mathbb{P}_N,\, \epsilon)}$. Later in Section 3, we will propose a specific approximation; see (21). Assume also that, in addition to the dataset $\widehat{\Theta}_N$, we predetermine a set $\mathcal{E}$ containing a large, yet finite, number of candidate radii ϵ. We randomly divide $\widehat{\Theta}_N$ into training and validation datasets K times. We enforce the same dataset size denoted by N_T on each of the K training datasets.

Next, for each $\epsilon \in \mathcal{E}$, we derive an empirical probability based on the following procedure: (i) we use each of the K training datasets to approximate $v^+_{\mathcal{D}(\widehat{\mathbb{P}}_{N_T},\, \epsilon)}$ with a value called $v_{\mathrm{WB}}(\epsilon)$ by calling the oracle, where $\mathbb{P}_{N_T}$ represents the empirical distribution from the training set; (ii) we then use the corresponding K validation datasets to simulate the expected optimal values denoted by v_{SB} [1]; and (iii) we finally compute the percentage of the K instances where $v_{\mathrm{WB}}(\epsilon) \geq v_{\mathrm{SB}}$. Let us call this empirical probability as the *empirical confidence level*. Thus, the empirical confidence level can roughly approximate the confidence level that the underlying distribution is contained in the Wasserstein-based ambiguity set with the radius ϵ. Note that the percentage computed is non-decreasing in ϵ and equal to 1 for some large ϵ_0. Therefore, the set containing all the empirical confidence levels is essentially an empirical cumulative distribution. Then, given a desired confidence level, we can choose a corresponding empirical radius $\epsilon \in \mathcal{E}$. The numerical results in Section 4 indicate that our choices of ϵ indeed return the desired confidence levels. We specify the above procedure in Algorithm 1.

[1] This process is to solve a linear program or integer program corresponding to each sample in the validation dataset and then to take the average of the optimal values.

Algorithm 1 Procedure to compute an empirical confidence level for any $\epsilon \in \mathcal{E}$

Inputs: A dataset $\widehat{\Theta}_N = \{\hat{\xi}^1, \ldots, \hat{\xi}^N\}$ and a radius $\epsilon \in \mathcal{E}$
Outputs: The empirical confidence level
for $k = 1, \ldots, K$ **do**
  Use the k^{th} training dataset to compute $v^k_{\text{WB}}(\epsilon)$
  Use the k^{th} validation dataset to simulate v^k_{SB}
end for
Calculate the empirical confidence level for ϵ as the percentage of the K instances where $v^k_{\text{WB}}(\epsilon) \geq v^k_{\text{SB}}$

3 Problem Reformulation and Tractable Bound

In this section, we propose a copositive programming reformulation for problem (4) under some mild assumptions. As copositive programs are computationally intractable, we then propose semidefinite-based relaxations for the purposes of computation.

Let us first define the feasible set for $x \in \mathbb{R}^n$ in (1) as follows:

$$\mathcal{X} := \left\{ x \in \mathbb{R}^n : \begin{array}{l} Ax = b,\ x \geq 0 \\ x_j \in \{0,1\}\ \forall\, j \in \mathcal{B} \end{array} \right\}.$$

We now introduce the following standard assumptions:

Assumption 2. *The set $\mathcal{X} \subseteq \mathbb{R}^n$ is nonempty and bounded.*

Assumption 3. $Ax = b,\ x \geq 0 \implies 0 \leq x_j \leq 1\ \forall\, j \in \mathcal{B}$.

Assumption 3 can be easily enforced. For example (see also [11], [40]), if $\mathcal{B} = \emptyset$, then the assumption is redundant; if problem (1) is derived from the network flow problems, for instance the longest path problem on a directed acyclic graph, then Assumption 3 is implied from the network flow constraints; if $\mathcal{B}$ is a nonempty set and the assumption is not implied by the constraints, we can add constraints $x_j + s_j = 1,\ s_j \geq 0\ \forall\, j \in \mathcal{B}$.

Assumption 4. *The support set $\Xi \subseteq \mathbb{R}^k$ is convex, closed, and computationally tractable.*

For example, Ξ could be represented using a polynomial number of linear, second-order-cone, and semidefinite inequalities. In particular, the set Ξ possesses a polynomial-time separation oracle [25].

Assumption 5. Ξ *is bounded.*

By Assumption 2, we know that $v(\xi)$ is finite and attainable for any $\xi \in \Xi$. Note that under Assumptions 2-5, $v^+_{\mathcal{D}}$ is finite and attainable and thus we can replace sup with max

in (4) under these conditions. Assumption 5 could be merged with Assumption 4, but it is stated separately to highlight its role in proving the exactness of the copositive programming reformulation below.

3.1 A copositive reformulation

We reformulate problem (4) via conic programming duality theory and probability theory. We introduce a useful result from the literature as follows.

Lemma 1. *$v^+_{\mathcal{D}(\widehat{\mathbb{P}}_N,\epsilon)}$ equals the optimal value of*

$$\begin{array}{ll} \sup & \frac{1}{N}\sum_{i=1}^N \int_\Xi v(\xi)\; d\,\mathbb{Q}_i(\xi) \\ \text{s.t.} & \frac{1}{N}\sum_{i=1}^N \int_\Xi \|\xi-\hat{\xi}_i\|^2\; d\,\mathbb{Q}_i(\xi) \le \epsilon^2 \\ & \mathbb{Q}_i \in \mathcal{M}^2(\Xi) \quad \forall\, i\in[N], \end{array} \tag{5}$$

where $\mathbb{Q}_i$ represents the distribution of ξ conditional on $\xi' = \hat{\xi}^i$ for all $i \in [N]$.

Proof. As $\mathbb{Q}_i$ represents the distribution of ξ conditional on $\xi' = \hat{\xi}^i$, the joint probability Π in problem (4) can be decomposed as $\Pi = \frac{1}{N}\sum\limits_{i\in[N]} \mathbb{Q}_i$ by the law of total probability. Thus, the optimal value of (5) coincides with $v^+_{\mathcal{D}(\widehat{\mathbb{P}}_N,\epsilon)}$, which completes the proof. □

We next provide a copositive programming reformulation for problem (5). As the first step, we use a standard duality argument to write the dual of (5) (see also [23]):

$$v^+_{\mathcal{D}(\widehat{\mathbb{P}}_N,\epsilon)} = \sup_{\mathbb{Q}_i\in\mathrm{M}^2(\Xi)} \inf_{\lambda\ge 0} \frac{1}{N}\sum_{i=1}^N \int_\Xi v(\xi)\; d\,\mathbb{Q}_i(\xi) + \lambda\left(\epsilon^2 - \frac{1}{N}\sum_{i=1}^N \int_\Xi \|\xi-\hat{\xi}_i\|^2\; d\,\mathbb{Q}_i(\xi)\right) \tag{6}$$

$$\le \inf_{\lambda\ge 0} \sup_{\mathbb{Q}_i\in\mathrm{M}^2(\Xi)} \lambda\,\epsilon^2 + \frac{1}{N}\sum_{i=1}^N \int_\Xi (v(\xi) - \lambda\|\xi-\hat{\xi}_i\|^2)\; d\,\mathbb{Q}_i(\xi) \tag{7}$$

$$= \inf_{\lambda\ge 0} \lambda\,\epsilon^2 + \frac{1}{N}\sum_{i=1}^N \sup_{\xi\in\Xi} (v(\xi) - \lambda\|\xi-\hat{\xi}_i\|^2), \tag{8}$$

where (7) follows from the max-min inequality, while equation (8) follows from the fact that $\mathcal{M}^2(\Xi)$ contains all the Dirac distributions supported on Ξ.

By Assumption 2, $v(\xi)$ is finite for all $\xi \in \Xi$. Then, the inequality in (7) becomes an equality for any $\epsilon > 0$ due to a straightforward generalization of a strong duality result for moment problems in Proposition 3.4 in [43]; see also Theorem 1 in [28] and Lemma 7 in [30].

By introducing auxiliary variables s_i, the minimization problem in (8) is equivalent to

$$\begin{array}{rll} v^+_{\mathcal{D}(\widehat{\mathbb{P}}_N,\epsilon)} = & \inf\limits_{\lambda,\, s_i} & \lambda\,\epsilon^2 + \frac{1}{N}\sum_{i=1}^N s_i \\ & \text{s.t.} & \sup\limits_{\xi\in\Xi} (v(\xi) - \lambda\|\xi - \hat{\xi}_i\|^2) \le s_i \quad \forall\, i \in [N] \\ & & \lambda \ge 0. \end{array} \tag{9}$$

For each $i \in [N]$, consider the following maximization problem corresponding to the left-hand side of the constraints in (9):

$$\begin{array}{rll} h^i(\lambda) := & \sup & (F\xi)^T x - \lambda(\xi^T\xi - 2\hat{\xi}_i^T\xi + \|\hat{\xi}_i\|^2) \\ & \text{s.t.} & Ax = b,\ x \ge 0 \\ & & x_j \in \{0,1\} \quad \forall\, j \in \mathcal{B} \\ & & e_1^T\xi = 1,\ \xi \in \hat{\Xi}, \end{array} \tag{10}$$

which is a mixed 0-1 bilinear program. Under Assumption 3, it holds also that the optimal value of (10) equals the optimal value of an associated copositive program [11, 12], which we now describe.

Define

$$z := \begin{pmatrix} \xi \\ x \end{pmatrix} \in \mathbb{R}^{k+n}, \quad E := \begin{pmatrix} -be_1^T & A \end{pmatrix} \in \mathbb{R}^{m\times(k+n)}, \tag{11}$$

$$H^i(\lambda) := \begin{pmatrix} -\lambda(I - \hat{\xi}_i e_1^T - e_1\hat{\xi}_i^T + \|\hat{\xi}_i\|^2 e_1 e_1^T) & \frac{1}{2}F^T \\ \frac{1}{2}F & 0 \end{pmatrix} \in \mathbb{S}^{k+n}, \tag{12}$$

and for any $j \in \mathcal{B}$, define

$$Q_j := \begin{pmatrix} 0 \\ f_j \end{pmatrix}\begin{pmatrix} 0 \\ f_j \end{pmatrix}^T - \frac{1}{2}\begin{pmatrix} 0 \\ f_j \end{pmatrix}\begin{pmatrix} e_1 \\ 0 \end{pmatrix}^T - \frac{1}{2}\begin{pmatrix} e_1 \\ 0 \end{pmatrix}\begin{pmatrix} 0 \\ f_j \end{pmatrix}^T \in \mathbb{S}^{k+n}. \tag{13}$$

where f_j denotes the j-th standard basis vector in $\mathbb{R}^n$.

Because both $\mathcal{X}$ and Ξ are bounded by Assumptions 2 and 5, there exists a scalar $r > 0$ such that the constraint $z^Tz = \xi^T\xi + x^Tx \le r$ is redundant for (10). Furthermore, it is well-known that we can use the following quadratic constraints to represent the binary variables in the description of $\mathcal{X}$:

$$x_j^2 - x_j = 0 \quad \Leftrightarrow \quad Q_j \bullet zz^T = 0.$$

After adding the redundant constraint and representing the binary variables, we homogenize

problem (10) as follows:

$$\begin{array}{ll} \max & H^i(\lambda) \bullet zz^T \\ \text{s.t.} & Ez = 0,\ g_1^T z = 1 \\ & I \bullet zz^T \le r \\ & Q_j \bullet zz^T = 0 \quad \forall\, j \in \mathcal{B} \\ & z \in \widehat{\Xi} \times \mathbb{R}^n_+, \end{array} \tag{14}$$

where $g_1 = \binom{e_1}{0} \in \mathbb{R}^{k+n}$ and e_1 denotes the standard basis vector in $\mathbb{R}^k$. The copositive representation is thus

$$\begin{array}{ll} \max & H^i(\lambda) \bullet Z \\ \text{s.t.} & \operatorname{diag}(EZE^T) = 0 \\ & g_1 g_1^T \bullet Z = 1 \\ & I \bullet Z \le r \\ & Q_j \bullet Z = 0 \quad \forall\, j \in \mathcal{B} \\ & Z \in \mathcal{CP}(\widehat{\Xi} \times \mathbb{R}^n_+). \end{array} \tag{15}$$

Letting $u^i \in \mathbb{R}^m$, $\rho^i \in \mathbb{R}_+$, $\alpha^i \in \mathbb{R}$, and $v^i \in \mathbb{R}^{|\mathcal{B}|}$ be the respective dual multipliers of $\operatorname{diag}(EZE^T) = 0$, $I \bullet Z \le r$, $g_1 g_1^T \bullet Z = 1$, and $Q_j \bullet Z = 0$, standard conic duality theory implies the dual of (15) is

$$\begin{array}{ll} \min\limits_{\alpha^i, \rho^i, u^i, v^i} & \alpha^i + r\rho^i \\ \text{s.t.} & \alpha^i g_1 g_1^T - H^i(\lambda) + E^T \operatorname{Diag}(u^i) E + \sum\limits_{j \in \mathcal{B}} v^i_j Q_j + \rho^i I \in \mathcal{COP}(\widehat{\Xi} \times \mathbb{R}^n_+) \\ & \rho^i \ge 0. \end{array} \tag{16}$$

Holding all other dual variables fixed, for $\rho^i > 0$ large, the matrix variable in (16) is strictly copositive—in fact, positive definite—which establishes that Slater's condition is satisfied, thus ensuring strong duality: the optimal value of (15) equals the optimal value of (16). Therefore, we can reformulate problem (9) as follows:

$$\begin{array}{rll} v^+_{\mathcal{D}(\widehat{\mathbb{P}}_N, \epsilon)} = & \min & \lambda\epsilon^2 + \frac{1}{N} \sum\limits_{i=1}^N (\alpha^i + r\rho^i) \\ & \text{s.t.} & \alpha^i g_1 g_1^T - H^i(\lambda) + E^T \operatorname{Diag}(u^i) E + \sum\limits_{j \in \mathcal{B}} v^i_j Q_j + \rho^i I \in \mathcal{COP}(\widehat{\Xi} \times \mathbb{R}^n_+)\ \forall\, i \in [N] \\ & & \rho^i \ge 0 \quad \forall\, i \in [N] \\ & & \lambda \ge 0. \end{array} \tag{17}$$

Note that if Assumption 5 fails, the constraint $I \bullet Z \le r$ should be excluded from (15) and thus the terms $r\rho^i$ and $\rho^i I$ in the objective function and the constraint, respectively,

should be excluded in (16) as well. As such, strong duality between (15) and (16) cannot be established in this case. However, the modified (16) still provides an upper bound on $h^i(\lambda)$. Accordingly, the modified problem (17) still provides an upper bound on $v^+_{\mathcal{D}(\widehat{\mathbb{P}}_N,\epsilon)}$.

3.2 A semidefinite-based relaxation

As problem (17) is difficult to solve in general, we propose a tractable approximation based on semidefinite programming techniques. In particular, we propose an inner approximation of $\mathcal{COP}(\widehat{\Xi}\times\mathbb{R}^n_+)$ in (17) so that the resulting problem has an optimal value that is an upper bound on $v^+_{\mathcal{D}}$. Now, define

$$\mathrm{IA}(\widehat{\Xi}\times\mathbb{R}^n_+) := \left\{ S+M \;:\; \begin{array}{c} S_{11}\in \mathrm{IA}(\widehat{\Xi}), \mathrm{Rows}(S_{21})\in \widehat{\Xi}^* \\ S_{22}\geq 0,\ M\succeq 0 \end{array} \right\},$$

where $\mathrm{IA}(\widehat{\Xi})$ is an inner approximation of $\mathcal{COP}(\widehat{\Xi})$, i.e., $\mathrm{IA}(\widehat{\Xi})\subseteq\mathcal{COP}(\widehat{\Xi})$. Immediately, we have a relationship between $\mathrm{IA}(\widehat{\Xi}\times\mathbb{R}^n_+)$ and $\mathcal{COP}(\widehat{\Xi}\times\mathbb{R}^n_+)$:

Lemma 2. $\mathrm{IA}(\widehat{\Xi}\times\mathbb{R}^n_+)\subseteq\mathcal{COP}(\widehat{\Xi}\times\mathbb{R}^n_+)$.

Proof. Let arbitrary $\binom{p}{q}\in\widehat{\Xi}\times\mathbb{R}^n_+$ be given. We need to show

$$\binom{p}{q}^T(S+M)\binom{p}{q} = \tbinom{p}{q}^T S\tbinom{p}{q} + \tbinom{p}{q}^T M\tbinom{p}{q} \geq 0.$$

$$\begin{aligned}
\binom{p}{q}^T(S+M)\binom{p}{q} &= \tbinom{p}{q}^T S\tbinom{p}{q} + \tbinom{p}{q}^T M\tbinom{p}{q} && (18)\\
&= p^T S_{11} p + 2q^T S_{21} p + q^T S_{22} q + \tbinom{p}{q}^T M\tbinom{p}{q} && (19)\\
&\geq 0 && (20)
\end{aligned}$$

The first term is nonnegative because $p\in\widehat{\Xi}$ and $S_{11}\in\mathrm{IA}(\widehat{\Xi})\subseteq\mathcal{COP}(\widehat{\Xi})$; the second term is nonnegative because $p\in\widehat{\Xi}, q\geq 0$, and $\mathrm{Rows}(S_{21})\in\widehat{\Xi}^*$; the third term is nonnegative because $q\geq 0$ and $S_{22}\geq 0$; the last term is nonnegative because $M\succeq 0$. □

When $\widehat{\Xi}=\{\xi\in\mathbb{R}^k : P\xi\geq 0\}$ is a polyhedral cone based on some matrix $P\in\mathbb{R}^{p\times k}$, a typical inner approximation $\mathrm{IA}(\widehat{\Xi})$ of $\mathcal{COP}(\widehat{\Xi})$ is given by

$$\mathrm{IA}(\widehat{\Xi}) := \{S_{11} = P^T Y P : Y\geq 0\},$$

where $Y \in \mathbb{S}^p$ is a symmetric matrix variable. This corresponds to the RLT approach of [2, 13, 44]. When $\widehat{\Xi} = \{\xi \in \mathbb{R}^k : \|(\xi_2, \ldots, \xi_k)^T\| \leq \xi_1\}$ is the second-order cone, it is known [45] that

$$\mathcal{COP}(\widehat{\Xi}) = \{S_{11} = \tau J + M_{11} : \tau \geq 0,\ M_{11} \succeq 0\},$$

where $J = \mathrm{Diag}(1, -1, \ldots, -1)$. Because of this simple structure, it often makes sense to take $\mathrm{IA}(\widehat{\Xi}) = \mathcal{COP}(\widehat{\Xi})$ in practice.

Now consider the following problem by replacing $\mathcal{COP}(\widehat{\Xi} \times \mathbb{R}^n_+)$ with $\mathrm{IA}(\widehat{\Xi} \times \mathbb{R}^n_+)$ in (17).

$$\begin{array}{rll} \bar{v}^+_{\mathcal{D}(\widehat{\mathbb{P}},\epsilon)} = & \min & \lambda\epsilon^2 + \frac{1}{N}\sum_{i=1}^{N}(\alpha^i + r\rho^i) \\ & \text{s.t.} & \alpha^i g_1 g_1^T - H^i(\lambda) + E^T \mathrm{Diag}(u^i) E + \sum_{j\in\mathcal{B}} v^i_j Q_j + \rho^i I \in \mathrm{IA}(\widehat{\Xi} \times \mathbb{R}^n_+)\ \forall\, i \in [N] \\ & & \rho^i \geq 0 \quad \forall\, i \in [N] \\ & & \lambda \geq 0. \end{array} \tag{21}$$

Obviously, we have the following result:

Theorem 1. $v^+_{\mathcal{D}(\widehat{\mathbb{P}}_N,\epsilon)} \leq \bar{v}^+_{\mathcal{D}(\widehat{\mathbb{P}},\epsilon)}$.

4 Numerical Experiments

In this section, we validate our proposed Wasserstein-ball approach (WB) on three applications. We will compare WB with the moment-based approach (MB) proposed in [40] where the exact values of the first two moments of the distributions are known. In practice, the moments of the distribution are often not known exactly. To this end, Delage and Ye [17] proposed a data-driven approach to handle this case. However, in this paper, we assume that the moments are known exactly for MB. Actually, this choice favors MB, but the goal of our experiments is to demonstrate that our approach provides a valid upper bound that gets closer to $v_{\mathbb{P}}$ as the size of the data set increases, while the MB provides an upper bound, which does not improve with the size of the data set.

All computations are conducted with Mosek version 8.0.0.28 beta [3] on an Intel Core i3 2.93 GHz Windows computer with 4GB of RAM and are implemented using the modeling language YALMIP [35] in MATLAB (R2014a) version 8.3.0.532. In order to demonstrate the effectiveness of WB, we also implement a Monte Carlo simulation-based approach (SB) which requires a sufficiently large number of randomly generated samples. For the project management problem in Section 4.2, a linear program is solved for each sample of the Monte Carlo simulation, while for the knapsack problem in Section 4.3, an integer program is solved for each sample. We employ CPLEX 12.4 to solve these linear programs and integer

programs.

4.1 Statistical sensitivity analysis of highest-order statistic

The problem of finding the maximum value from a set $\zeta = (\zeta_1, \ldots, \zeta_n)$ of n numbers can be formulated as the optimization problem:

$$\max\left\{\zeta^T x : e^T x = 1,\ x \geq 0\right\}. \tag{22}$$

For example, suppose $\zeta_1 = \max\{\zeta_1, \ldots, \zeta_n\}$, then the optimal solution to (22) is $x_1^* = 1, x_2^* = \cdots = x_n^* = 0$. For the statistical sensitivity analysis problem, we consider a random vector ζ following a joint distribution $\mathbb{P}$. In the situation where the true distribution is not known exactly, our focus is to investigate the upper bound on the expected maximum value over an ambiguity set containing distributions that possess partial shared information.

We consider an instance with $n = 3$ and the true distribution $\mathbb{P}$ of ζ is assumed to be jointly lognormal with first and second moments given by $\mu_{\log} \in \mathbb{R}^3$ and $\Sigma_{\log} \in \mathbb{S}^3$, respectively.

In our experiments, we use the following procedure to randomly generate $\mu_{\log}$ and $\Sigma_{\log}$. We first sample $\mu \in \mathbb{R}^3$ from a uniform distribution $[0, 2]^3$. Then, we randomly generate a matrix $\Sigma \in \mathbb{S}^3$ as follows: we set the vector of standard deviations to $\sigma = \frac{1}{4}e \in \mathbb{R}^3$, sample a random correlation matrix $C \in \mathbb{S}^3$ using the MATLAB command 'gallery('randcorr',3)', and set $\Sigma = \operatorname{diag}(\sigma) C \operatorname{diag}(\sigma) + \mu\mu^T$. We set μ and Σ as the first and second moments respectively of the corresponding normal distribution of $\mathbb{P}$. Then $\mu_{\log}$ and $\Sigma_{\log}$ can be computed based on the following formulae [27]:

$$\begin{aligned} (\mu_{\log})_i &= e^{\mu_i + 0.5\Sigma_{ii}}, \\ (\Sigma_{\log})_{ij} &= e^{\mu_i + \mu_j + 0.5(\Sigma_{ii} + \Sigma_{jj})}(e^{\Sigma_{ij}} - 1). \end{aligned} \tag{23}$$

We can cast this problem into our framework by setting $m = 1, k = n + 1$, $\xi = (1, \zeta_1, \ldots, \zeta_n)$, $F = (0, I)$, and $\mathcal{B} = \emptyset$. Obviously, Assumptions 2 and 4 are satisfied. Assumption 3 is vacuous. Although Assumption 5 does not hold, problem (21) can still provide a valid upper bound on the expected optimal value as discussed in Section 3.1.

4.1.1 The deviation of empirical Wasserstein radii

In this experiment, we consider a particular underlying distribution $\mathbb{P}$ that is generated by the procedure mentioned above. Also, we consider eight cases for the size of the dataset: $N \in \{10, 20, 40, 80, 160, 320, 640, 1280\}$. For each case, we randomly generate a dataset $\widehat{\Theta}_N$

containing N independent samples from $\mathbb{P}$ and use the procedure in Section 2 to determine a desired radius from a pre-specified set $\mathcal{E} = \{0.001, 0.005, 0.01, 0.05, 0.1, 0.2, 0.3, 0.4, 0.5, 0.6, 0.7, 0.8, 0.9, 1.0, 2.0\}$[2]. In particular, we set $K = 100$ in Algorithm 1. Figure 1 shows the trend of the reliabilities over different Wasserstein radii for $N \in \{20, 80, 320, 1280\}$. Clearly, smaller Wasserstein radii tend to have lower empirical confidence levels. Furthermore, as the sample size increases, the empirical confidence level increases as well for the same Wasserstein radius. The result of this experiment indicates that we can practically choose a Wasserstein radius with a desired statistical guarantee for each case of N. We remark that the derived radii can be used for datasets of the same sizes generated from different distributions of the same family (lognormal distributions in this application).

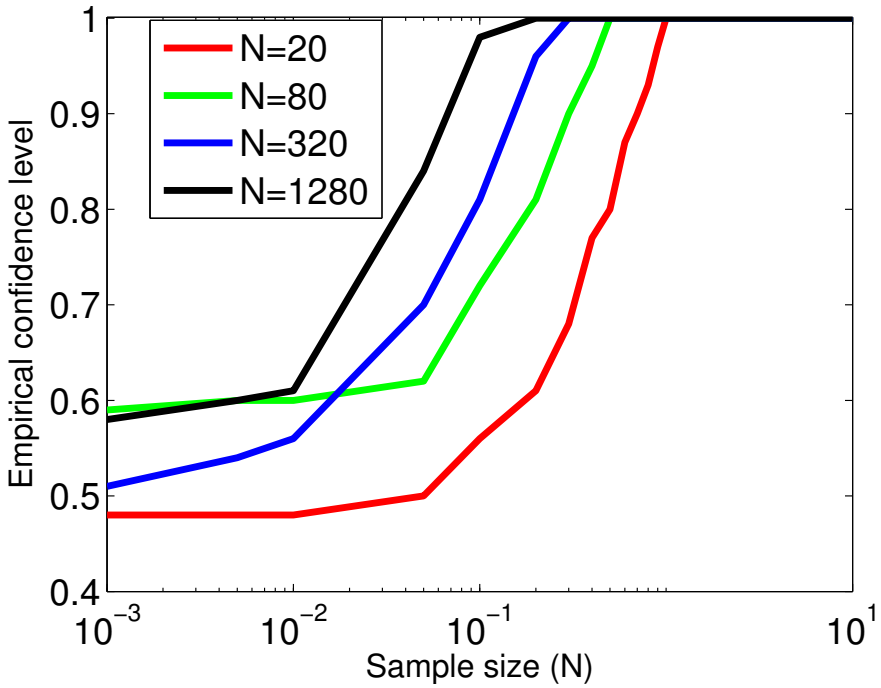

Figure 1: Empirical confidence levels of different Wasserstein radii for $N \in \{20, 80, 320, 1280\}$ respectively.

4.1.2 Instances with the same underlying distribution

Our next experiment is to focus on a particular joint lognormal distribution $\mathbb{P}$. We consider eight cases: $N \in \{10, 20, 40, 80, 160, 320, 640, 1280\}$. For each case, we test 100 trials and in each trial we randomly generate N independent samples from $\mathbb{P}$ and choose the Wasserstein

[2]From preliminary experiments, the largest element 2.0 in set $\mathcal{E}$ returned 1 as the empirical confidence level for all the experiments we conducted. Thus, we believe it is sufficient to have 2.0 as the largest element here.

radius with an empirical confidence level of 0.90. We compare our approach with MB where the first two moments are directly given as $\mu_{\log}$ and $\Sigma_{\log}$. We also randomly generate 100000 independent samples from $\mathbb{P}$ to simulate the true expected optimal value.

We demonstrate experimental results in Figure 2. Note that the solid black line represents the simulated value of the true expected optimal value, while the dashed black line represents the upper bound calculated by the moment-based approach. Furthermore, we solve an instance of (21) for each of the 100 trails in each case of N. We use the blue, red, and green lines to respectively represent the 80^{th} quantile, the median, and the 20^{th} quantile of the values from the 100 trials in each case. Figure 2 shows that our approach provides weaker bounds on the expected optimal value for smaller sample sizes. However, as the size of samples increases, our approach provides stronger bounds and the bounds get relatively close to the simulated value. In contrast, the value from MB remains the same regardless of sample size.

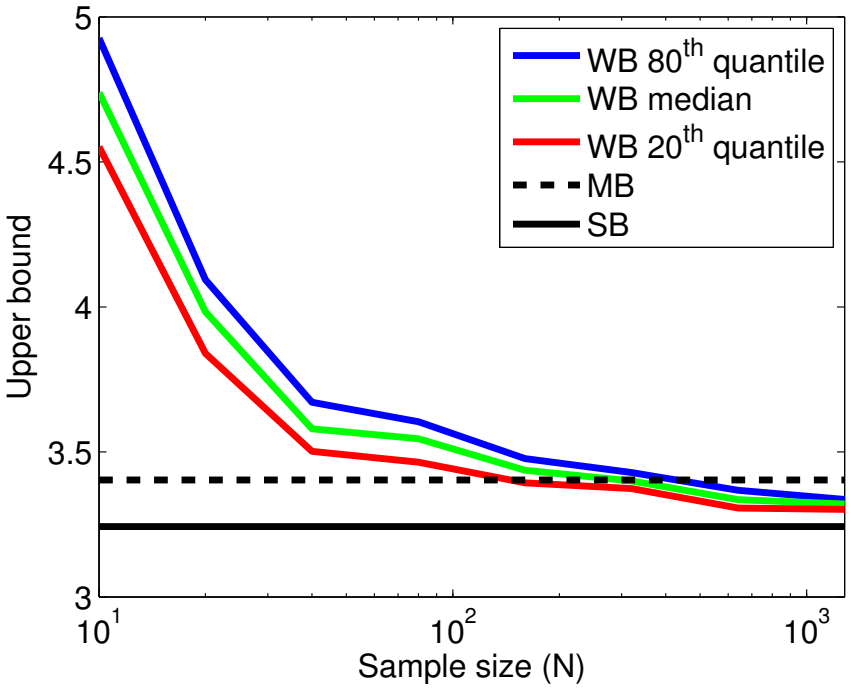

Figure 2: The comparison of WB and MB for the stochastic sensitivity analysis problem over different sample sizes for a particular randomly generated underlying distribution.

4.1.3 Instances with different underlying distributions

In this experiment, we consider eight cases $N \in \{10, 20, 40, 80, 160, 320, 640, 1280\}$. In each case, we randomly generate 100 trials. For each trial in each case we generate N samples

Case number	1	2	3	4	5	6	7	8
Empirical confidence level	1.00	1.00	1.00	1.00	0.97	0.98	0.97	0.98

Table 1: The percentage of the 100 trials where the optimal values from WB are greater or equal to the simulated values over the 8 cases for the stochastic sensitivity analysis problem.

from a random lognormal distribution whose first and second moments are generated by using the procedure at the beginning of this section. For each trial in each case, we solve an instance of (21) with a Wasserstein radius corresponding to an empirical confidence level of 0.90. We also simulate the true expected optimal values by randomly generating 100000 samples from the true distributions.

For each trial in each case, we denote the optimal value from (21) by $\bar{v}^+_{\text{WB}}$ and the simulated value by v_{SB}. Then, we calculate the relative gap between WB and SB as

$$\text{gap(WB)} := \frac{\bar{v}^+_{\text{WB}} - v_{\text{SB}}}{v_{\text{SB}}}.$$

We take the average of the relative gaps over the 100 trials for each case. Then, for each trial in each case, we solve MB with the first two moments computed by (23). Denote the optimal value from MB by $\bar{v}^+_{\text{MB}}$. Similarly, we calculate the relative gap between MB and SB as

$$\text{gap(MB)} := \frac{\bar{v}^+_{\text{MB}} - v_{\text{SB}}}{v_{\text{SB}}}.$$

We then take the average of the relative gaps over the 100 trials in each case. Figure 3 illustrates the average relative gaps from both WB and MB over the eight cases. Clearly, the upper bound from WB approaches the simulated value along with the increase of the size of samples, while the average relative gap between the bound from MB and the simulated value does not.

Table 1 shows the percentage of the 100 trials where the optimal values from WB are greater than or equal to the corresponding simulated optimal values in the eight cases. The result demonstrates that the derived empirical Wasserstein radii indeed provide desired statistical guarantees in practice.

4.2 Project management problem

In this application, we consider a project management problem, which can be formulated as a longest-path problem on a directed acyclic graph. The arcs denote activities and the nodes denote the completions of a set of activities. Arc lengths denote the time to complete the activities. Thus, the longest path from the starting node s to the ending node t gives

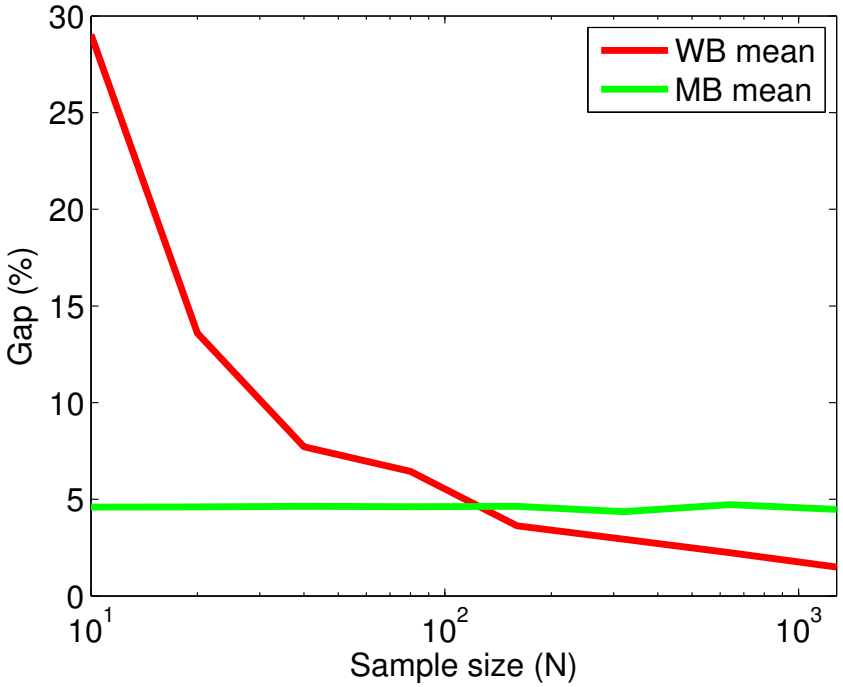

Figure 3: The average gaps from MB and WB for the stochastic sensitivity analysis problem over the eight cases: $N \in \{10, 20, 40, 80, 160, 320, 640, 1280\}$. The blue line represents the average relative gap between the optimal value from WB and the simulated value; the red line represents the average relative gap between the optimal value from MB and the simulated value.

the time needed to compete the whole project. Let ζ_{ij} be the length (time) of arc (activity) from node i to node j. The problem can be solved as a linear program due to the network flow structure as follows:

$$\begin{aligned} \max \quad & \sum_{(i,j)\in\mathcal{A}} \zeta_{ij} x_{ij} \\ \text{s. t.} \quad & \sum_{i:(i,j)\in\mathcal{A}} x_{ij} - \sum_{j:(i,j)\in\mathcal{A}} x_{ji} = \begin{cases} 1, & \text{if } i = s \\ 0, & \text{if } i \in \mathcal{N}, \text{ and } i \neq s, t \\ -1, & \text{if } i = t \end{cases} \\ & x_{ij} \geq 0, \ \forall\, (i,j) \in \mathcal{A}, \end{aligned} \tag{24}$$

where $\mathcal{A}$ denotes the set containing all the arcs, $\mathcal{N}$ denotes the set containing all nodes on the network, and x_{ij} denotes the number of units of flow sent from node i to node j through arc $(i,j) \in \mathcal{A}$. For the stochastic project management problem, the activity times are random. In such cases, due to the resource allocation and management constraints, the project manager would like to quantify the worst-case expected completion time of the project, which is corresponding to the worst-case longest path of the network.

We consider an instance with a network structure shown in Figure 4. This network con-

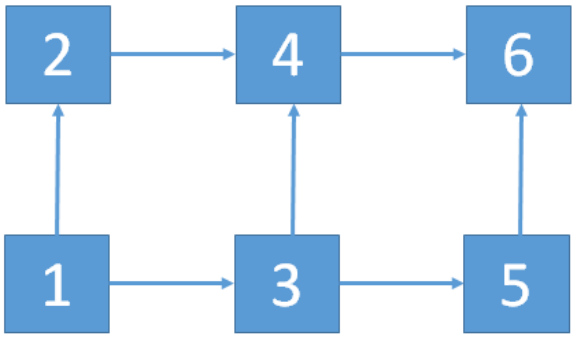

Figure 4: The structure of the project network where 1 and 6 are the starting and ending nodes respectively.

sists of 7 arcs and 6 nodes. There are 3 paths from the starting node to the ending node on the network. In the experiments of this example, we consider truncated joint normal distributions. We use the following procedure to generate a truncated joint normal distribution $\mathbb{P}$: denoting $|\mathcal{A}|$ by the cardinality of set $\mathcal{A}$, we generate $\zeta \geq 0$ from a jointly normal distribution with first and second moments given by $\mu \in \mathbb{R}^{|\mathcal{A}|}$ and $\Sigma \in \mathbb{S}^{|\mathcal{A}|}$, respectively. Specifically, we sample μ from a uniform distribution $[0, 5]^{|\mathcal{A}|}$ while the matrix Σ is generated randomly using the following procedure: we set the vector of standard deviations to $\sigma = e$, sample a random correlation matrix $C \in \mathbb{S}^{|\mathcal{A}|}$ using the MATLAB command 'gallery('randcorr',$|\mathcal{A}|$)', and set $\Sigma = \operatorname{diag}(\sigma) C \operatorname{diag}(\sigma) + \mu\mu^T$. Skipping the details, we can cast the network flow problem into our framework. It is straightforward to check that Assumptions 2, 4, and 5 are satisfied and Assumption 3 is vacuous.

4.2.1 Instances with the same underlying distribution

The first experiment of this example focuses on a particular underlying distribution $\mathbb{P}$. We consider seven cases: $N \in \{10, 20, 40, 80, 160, 320, 640\}$. For each case, we run 100 trials and in each trial we randomly generate a dataset $\widehat{\Theta}_N$ containing N independent samples from $\mathbb{P}$. We use the procedure in Section 2 to compute an empirical confidence level set for each case. Then, we use computed empirical confidence level sets to derive empirical Wasserstein radii for the following computations. For each trial in each case, we solve an instance of (21) with a Wasserstein radius corresponding to an empirical confidence level of 0.90. We compare WB with MB where the first two moments are approximated by using the sample mean and variance from 100000 samples. The computed moments are close to their theoretical counterparts as the sample size is considerably large. We also simulate the expected optimal value over the 100000 samples. Figure 5 shows that WB provides weaker

bounds on the expected optimal value for smaller sample sizes. However, as the size of samples increases, WB provides stronger bounds and the bounds get relatively close to the simulated value. In contrast, the bounds from MB remains the same regardless of the change of sample sizes.

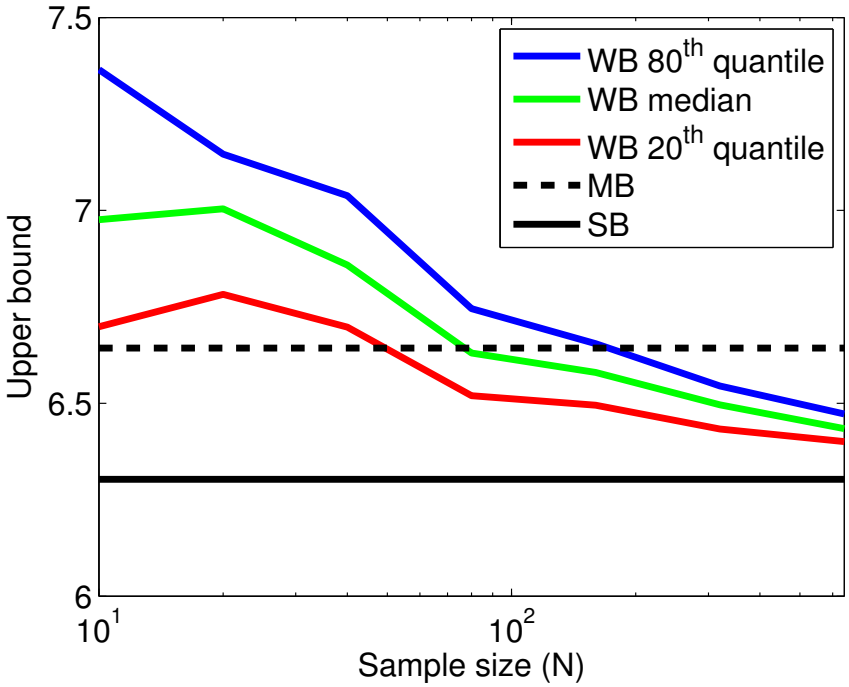

Figure 5: The comparison of WB and MB for the project management problem over different sample sizes for a particular randomly generated underlying distribution.

4.2.2 Instances with different underlying distributions

In this experiment, we consider seven cases: $N \in \{10, 20, 40, 80, 160, 320, 640\}$. For each case, we randomly generate 100 trials in which N independent samples are drawn from a randomly generated truncated joint normal distribution. Then, for each trial in each case, we solve and instance of (21) with a Wasserstein radius corresponding to a 0.90 empirical confidence level. We solve MB where the first two moments are approximated by computing the sample mean and variance of 100000 samples. We also simulate the expected optimal value over the 100000 samples for each trial in each case. We compute the relative gap between the WB and SB as well as the relative gap between MB and SB. Then, for each case, we take the average of the relative gaps from both WB and MB over the 100 trials. Figure 6 illustrates the average relative gaps over the seven cases. Clearly, the upper bound from WB approaches to the simulated value along with the increase in the size of samples,

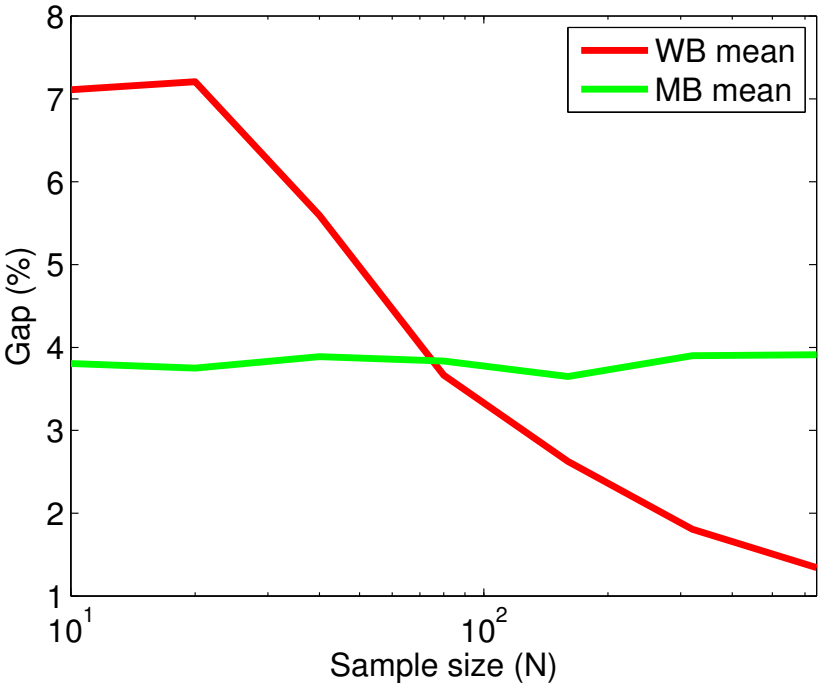

Figure 6: The average gaps from both MB and WB for the project management problem over the seven cases: $N \in \{10, 20, 40, 80, 160, 320, 640\}$.

Case number	1	2	3	4	5	6	7
Empirical confidence level	1.00	1.00	1.00	1.00	0.97	0.98	0.97

Table 2: the percentage of the 100 trials where the optimal values from WB are greater than or equal to the corresponding simulated optimal values over the 7 cases for the project management problem.

while the gap between the bound from MB and the simulated value remains relatively the same as the sample size increases. Table 2 shows the percentage of the 100 trials where the optimal values from WB are greater than or equal to the corresponding simulated optimal values over the seven cases.

4.3 Knapsack problem

A standard knapsack problem is defined as follows: given a set of items and each with a weight and a value, the problem is to determine the number of items to include in a knapsack such that the total weight is less than or equal to a given capacity limit and the total value is maximized; see the detail in [36]. Let w_i and ζ_i be the weight and value of item i $(i = 1, \ldots, n)$ respectively. Let W be the maximum weight capacity of the knapsack. Then, the knapsack

problem can be formulated as an integer program:

$$v(\zeta) := \max \left\{ \sum_{i=1}^{n} \zeta_i x_i : \sum_{i=1}^{n} w_i x_i \leq W,\ x_i \in \{0,1\} \right\},$$

where x_i represents the number of item i to include in the knapsack. Assume that the values of the items are random and follow an unknown joint distribution. Assume also that we can collect a dataset containing N samples with each corresponding to an observation of the n item values. In such cases, we would like to compute a data-driven distributionally robust upper bound on the expected maximum value of the knapsack. We can approximate the upper bound by solving problem (21).

We consider an instance with $n = 4, w = (5,\ 4,\ 6,\ 3)^T$, and $W = 10$. The true distribution $\mathbb{P}$ of ζ is assumed to be jointly lognormal with first and second moments given by $\mu_{\log} \in \mathbb{R}^4$ and $\Sigma_{\log} \in \mathbb{S}^4$, respectively. Similar to the procedure described in Section 4.1, we sample $\mu \in \mathbb{R}^4$ from a uniform distribution $[0,2]^4$. Then, we randomly generate a matrix $\Sigma \in \mathbb{S}^4$ as follows: we set the vector of standard deviations to $\sigma = \frac{1}{4}e \in \mathbb{R}^4$, sample a random correlation matrix $C \in \mathbb{S}^4$ using the MATLAB command 'gallery('randcorr',4)', and set $\Sigma = \operatorname{diag}(\sigma) C \operatorname{diag}(\sigma) + \mu\mu^T$. Then $\mu_{\log}$ and $\Sigma_{\log}$ can be computed based on (23). We can easily cast this problem into our framework. For simplicity, we skip the details. It is also straightforward to check that the conditions in Assumptions 2, 3, and 4 are satisfied. Although Assumption 5 is not satisfied, we still can solve (21) to obtain a valid upper bound on the expected optimal value of the knapsack problem.

4.3.1 Instances with the same underlying distribution

In the first experiment, we focus on a particular underlying distribution $\mathbb{P}$ and consider eight cases: $N \in \{10,\ 20,\ 40,\ 80,\ 160,\ 320,\ 640,\ 1280\}$. For each case, we run 100 trials and in each trial we randomly generate a dataset $\widehat{\Theta}_N$ containing N independent samples from $\mathbb{P}$. Similarly, we derive empirical Wasserstein radii for each case. Then, for each trial in each case, we solve an instance of (21) with an empirical Wasserstein radius corresponding to an empirical confidence level of 0.90. We compare our approach with MB where the first two moments are computed by (23). We simulate the expected optimal value over 100000 samples. Note that we solve an integer program for each sample in the simulation. Figure 7 shows that WB provides weaker bounds on the expected optimal value for smaller sample sizes. However, as the size of samples increases, WB provides stronger bounds and the bounds get relatively close to the simulated value. In contrast, the bounds from MB remains the same regardless of the change of sample sizes. We remark that the upper bound

computed by WB may not converge to the true expected optimal value as the sample size increases to infinity; see the trend shown in Figure 7. This is due to the fact that problem (21) is a relaxation of problem (17) and the fact the relaxation is not tight in this example.

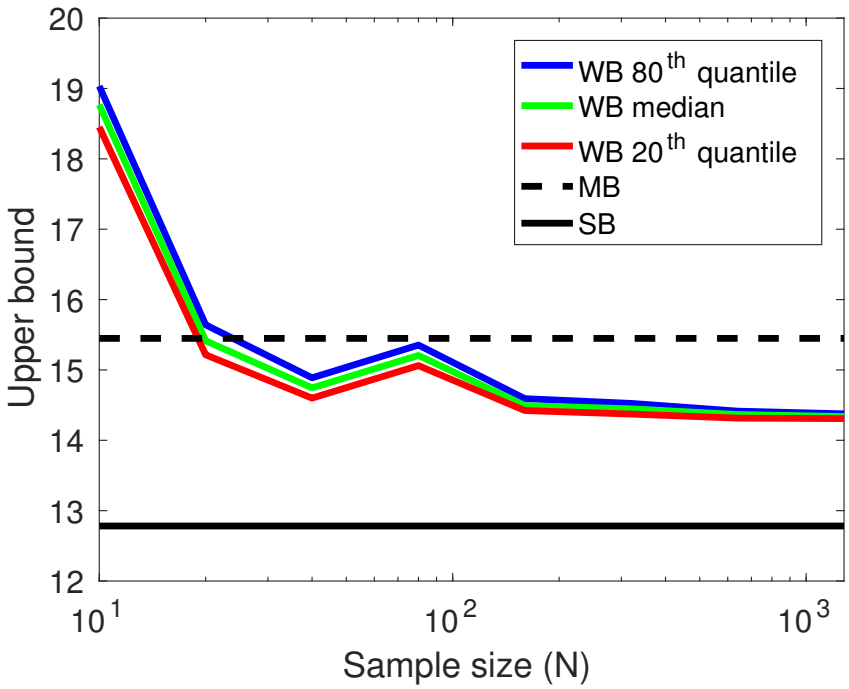

Figure 7: The comparison of WB and MB for the knapsack problem over different sample sizes for a particular randomly generated underlying distribution.

4.3.2 Instances with different underlying distributions

This experiment considers eight cases: $N \in \{10, 20, 40, 80, 160, 320, 640, 1280\}$. For each case, we randomly generate 100 trials with each trial is drawn from a randomly generated joint lognormal distribution. Then, for each trial in each case, we solve an instance of (21) with a Wasserstein radius corresponding to a 0.90 empirical confidence level. We simulate the expected optimal value over 100000 samples for each trial in each case. Next, we compute the relative gap between the values of WB and SB as well as the relative gap between the values of MB and SB. Then, for each case, we take the average of the relative gaps from both WB and MB over the 100 trials. Figure 8 illustrates the average relative gaps over the seven cases. Clearly, the WB gap becomes narrower as the sample size increases, while the MB gap remains relatively the same.

Table 3 shows the percentage of the 100 trials where the optimal values from WB are greater than or equal to the corresponding simulated optimal values over the eight cases.

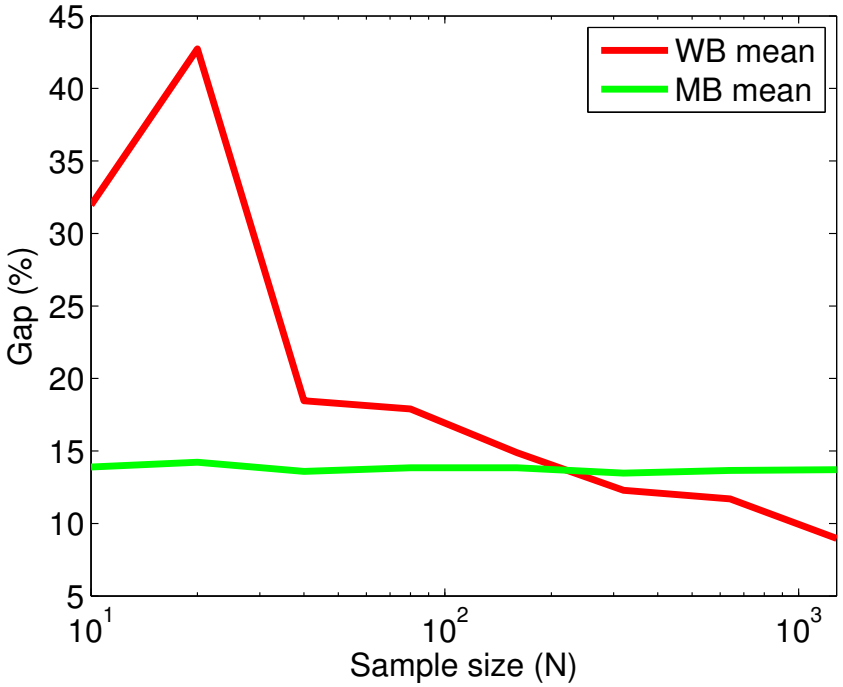

Figure 8: Illustration of the average gaps from both MB and WB in the case of $N \in \{10, 20, 40, 80, 160, 320, 640\}$. The blue line represents the average gap between the optimal values from WB and the simulated values; the red line represents the average gap between the optimal values from MB and the simulated values.

Case number	1	2	3	4	5	6	7	8
Empirical confidence level	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00

Table 3: The percentage of the 100 trials where the optimal values from WB are greater than or equal to the corresponding simulated optimal values over the 8 cases for the knapsack problem.

5 Concluding Remarks

In this paper, we have studied the expected optimal value of a mixed 0-1 programming problem with uncertain objective coefficients following a joint distribution whose information is not known exactly but a set of independent samples can be collected. Using the samples, we have constructed a Wasserstein-based ambiguity set that contains the true distribution with a desired confidence level. We proposed an approach to compute the upper bound on the expected optimal value. Then under mild assumption, the problem was reformulated to a copositive program, which leads to a semidefinite-based relaxation. We have validated the effectiveness of our approach over three applications.

Acknowledgements

We would like to thank Kurt Anstreicher, Qihang Lin, Luis Zuluaga, and Tianbao Yang for many useful suggestions, which helped us improve the results of the paper.

References

[1] D. J. Aldous. The ζ (2) limit in the random assignment problem. *Random Structures & Algorithms*, 18(4):381–418, 2001.

[2] K. M. Anstreicher. Semidefinite programming versus the reformulation-linearization technique for nonconvex quadratically constrained quadratic programming. *Journal of Global Optimization*, 43(2-3):471–484, 2009.

[3] M. ApS. *The MOSEK optimization toolbox for MATLAB manual. Version 8.0.*, 2016.

[4] A. Ben-Tal, 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.

[5] D. Bertsimas, V. Gupta, and N. Kallus. Data-driven robust optimization. *arXiv preprint arXiv:1401.0212*, 2013.

[6] D. Bertsimas, V. Gupta, and N. Kallus. Robust saa. *arXiv preprint arXiv:1408.4445*, 2014.

[7] D. Bertsimas, K. Natarajan, and C.-P. Teo. Probabilistic combinatorial optimization: Moments, semidefinite programming, and asymptotic bounds. *SIAM Journal on Optimization*, 15(1):185–209, 2004.

[8] D. Bertsimas, K. Natarajan, and C.-P. Teo. Persistence in discrete optimization under data uncertainty. *Mathematical programming*, 108(2):251–274, 2006.

[9] D. Bertsimas and I. Popescu. Optimal inequalities in probability theory: A convex optimization approach. *SIAM Journal on Optimization*, 15(3):780–804, 2005.

[10] R. Bowman. Efficient estimation of arc criticalities in stochastic activity networks. *Management Science*, 41(1):58–67, 1995.

[11] S. Burer. On the copositive representation of binary and continuous nonconvex quadratic programs. *Mathematical Programming*, 120(2):479–495, 2009.

[12] S. Burer. Copositive programming. In *Handbook on semidefinite, conic and polynomial optimization*, pages 201–218. Springer, 2012.

[13] S. Burer. A gentle, geometric introduction to copositive optimization. *Mathematical Programming*, 151(1):89–116, 2015.

[14] S. Burer and H. Dong. Representing quadratically constrained quadratic programs as generalized copositive programs. *Operations Research Letters*, 40:203–206, 2012.

[15] 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.

[16] Z. Chen, M. Sim, and H. Xu. Distributionally robust optimization with infinitely constrained ambiguity sets. Working Paper, 2016.

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

[18] B. Dodin. Bounding the project completion time distribution in pert networks. *Operations Research*, 33(4):862–881, 1985.

[19] J. Duchi, P. Glynn, and H. Namkoong. Statistics of robust optimization: a generalized empirical likelihood approach. *arXiv preprint arXiv:1610.03425*, 2016.

[20] J. Dupačová. The minimax approach to stochastic programming and an illustrative application. *Stochastics: An International Journal of Probability and Stochastic Processes*, 20(1):73–88, 1987.

[21] G. Eichfelder and J. Jahn. Set-semidefinite optimization. *Journal of Convex Analysis*, 15:767–801, 2008.

[22] E. Erdoğan and G. Iyengar. Ambiguous chance constrained problems and robust optimization. *Mathematical Programming*, 107(1-2):37–61, 2006.

[23] P. M. Esfahani and D. Kuhn. Data-driven distributionally robust optimization using the wasserstein metric: Performance guarantees and tractable reformulations. *arXiv preprint arXiv:1505.05116*, 2015.

[24] L. E. 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.

[25] M. Grötschel, L. Lovász, and A. Schrijver. The ellipsoid method and its consequences in combinatorial optimization. *Combinatorica*, 1(2):169–197, 1981.

[26] J. N. Hagstrom. Computational complexity of pert problems. *Networks*, 18(2):139–147, 1988.

[27] L. J. Halliwell. The lognormal random multivariate. In *Casualty Actuarial Society E-Forum, Spring 2015*.

[28] G. A. Hanasusanto and D. Kuhn. Conic programming reformulations of two-stage distributionally robust linear programs over wasserstein balls. *arXiv preprint arXiv:1609.07505*, 2016.

[29] G. A. Hanasusanto, D. Kuhn, S. W. Wallace, and S. Zymler. Distributionally robust multi-item newsvendor problems with multimodal demand distributions. *Mathematical Programming*, 152(1):1–32, 2015.

[30] G. A. Hanasusanto, V. Roitch, D. Kuhn, and W. Wiesemann. Ambiguous joint chance constraints under mean and dispersion information. *Operations Research*, 2017.

[31] Z. Hu and L. J. Hong. Kullback-leibler divergence constrained distributionally robust optimization. *Available at Optimization Online*, 2013.

[32] R. Jiang and Y. Guan. Data-driven chance constrained stochastic program. *Mathematical Programming*, pages 1–37, 2015.

[33] D. Klabjan, D. Simchi-Levi, and M. Song. Robust stochastic lot-sizing by means of histograms. *Production and Operations Management*, 22(3):691–710, 2013.

[34] H. Lam. Recovering best statistical guarantees via the empirical divergence-based distributionally robust optimization. *arXiv preprint arXiv:1605.09349*, 2016.

[35] J. Lofberg. Yalmip: A toolbox for modeling and optimization in matlab. In *Computer Aided Control Systems Design, 2004 IEEE International Symposium on*, pages 284–289. IEEE, 2004.

[36] S. Martello and P. Toth. *Knapsack Problems: Algorithms and Computer Implementations*. John Wiley & Sons, Inc., New York, NY, USA, 1990.

[37] R. H. Möhring. Scheduling under uncertainty: Bounding the makespan distribution. In *Computational Discrete Mathematics*, pages 79–97. Springer, 2001.

[38] K. Natarajan, M. Song, and C.-P. Teo. Persistency model and its applications in choice modeling. *Management Science*, 55(3):453–469, 2009.

[39] K. Natarajan and C. P. Teo. On reduced semidefinite programs for second order moment bounds with applications. *Mathematical Programming*, 161(1):487–518, 2017.

[40] K. Natarajan, C. P. Teo, and Z. Zheng. Mixed 0-1 linear programs under objective uncertainty: A completely positive representation. *Operations research*, 59(3):713–728, 2011.

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

[42] Y. Rubner, C. Tomasi, and L. J. Guibas. The earth mover's distance as a metric for image retrieval. *International journal of computer vision*, 40(2):99–121, 2000.

[43] A. Shapiro. On duality theory of conic linear problems. In *Semi-infinite programming*, pages 135–165. Springer, 2001.

[44] H. D. Sherali and W. P. Adams. *A reformulation-linearization technique for solving discrete and continuous nonconvex problems*, volume 31. Springer Science & Business Media, 2013.

[45] J. F. Sturm and S. Zhang. On cones of nonnegative quadratic functions. *Mathematics of Operations Research*, 28(2):246–267, 2003.

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

[47] W. Wiesemann, D. Kuhn, and M. Sim. Distributionally robust convex optimization. *Operations Research*, 62(6):1358–1376, 2014.

[48] L. A. Wolsey. *Integer programming*, volume 42. Wiley New York, 1998.

[49] 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.

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