Distributionally Robust Optimization with Expected Constraints via Optimal Transport

Diego Fonseca and Mauricio Junca

Department of Mathematics, Universidad de los Andes, Bogotá, Colombia
{df.fonseca,mj.junca20}@uniandes.edu.co

Abstract. We consider a stochastic program with expected value constraints. We analyze this problem in a general context via Distributionally Robust Optimization (DRO) approach using 1 or 2-Wasserstein metrics where the ambiguity set depends on the decision. We show that this approach can be reformulated as a finite-dimensional optimization problem, and, in some cases, this can be convex. Additionally, we establish criteria to determine the feasibility of the problem in terms of the Wasserstein radius and the level of the constraint. Finally, we present numerical results in the context of inventory management and portfolio optimization. In the portfolio optimization context, we present the advantages that our approach has over some existing non-robust methods using real financial market data.

Keywords: Robust optimization · Expectation constraints· Wasserstein metric· Mean-variance model.

1 Introduction

In this work we consider stochastic programs with expected value constraints given by the following formulation:

$$J = \begin{cases} \min\limits_{x\in\mathbb{R}^m} & \Phi\left(x, F, \mathbb{P}\right) \\ \text{subject to } & \mathbb{E}_{\mathbb{P}}\left[G\left(x,\xi\right)\right] \geq \mu, \\ & x \in \mathcal{X}, \end{cases} \tag{1}$$

where F and G are functions such that $F, G : \mathbb{R}^m \times \mathbb{R}^n \to \mathbb{R}$, $\xi \in \mathbb{R}^n$ is a random vector with (unknown) probability distribution $\mathbb{P}$ supported in $\Xi \subseteq \mathbb{R}^n$, and $\mathcal{X} \subseteq \mathbb{R}^m$ is a set of constraints on the decision vectors. In addition, the objective function Φ is a risk function that depends on the performance function F.

When $\Phi\left(x, F, \mathbb{P}\right) := \mathbb{E}_{\mathbb{P}}\left[F(x,\xi)\right]$, this problem appears in various contexts such as finance, [16,9], operations research, [21], and, machine learning [25,22]. Most attempts to solve (1) use Sample Average Approximation (SAA) where samples of ξ are used to replace expected values by the sample means. Strategies based on the stochastic gradient descent methods are used in [1,15,37]. These strategies are sensitive to alterations in the quality of the sample, and the out-of-sample performance can be poor, specifically, the constraints tend not to

be satisfied out-of-sample when the sample size is small. Note that other risk functions, like Conditional Value-at-Risk, can be formulated as (1), see [26].

When $\Phi(x, F, \mathbb{P}) := \mathrm{Var}_{\mathbb{P}}[F(x, \xi)]$, (1) is known as the mean-variance model and this problem has been mostly explored in portfolio optimization and inventory management. The mean-variance model is formulated for the first time in portfolio optimization in [20]. In this case $m = n$ where m is the number of assets of the portfolio, ξ is a random vector of returns of each asset, x is a portfolio weights vector, and other constraints admissible for the investor are described by the set $\mathcal{X}$. Additionally, in this case, $F = G$ where $F(x, \xi) := \langle x, \xi \rangle$ where $\langle \cdot, \cdot \rangle$ is the Euclidean inner product in $\mathbb{R}^m$ so that $\langle x, \xi \rangle$ is the return of the portfolio x. The first attempts to solve this problem consider estimates of the vector of means and the covariance matrix of returns, but in [7] it is shown that the resulting portfolios do not perform well out of sample, and they are very sensitive to variations on the estimates. One of the first ideas to overcome this problem is to consider the vector of returns and the covariance matrix as variables, that is, the variables of the optimization problem will be the portfolio weights, the vector of means, and the covariance matrix. The choice of the feasible set for the vector and the matrix is crucial in this approach. Some works use sets based on a priori information about the returns or impose sets that are computationally tractable, see for example [11], [39], [23], [17], [18], and [36]. To impose unverifiable assumptions about the moments of the returns can also affect the out-of-sample performance of these methods. Most of the strategies that are used in portfolio optimization can be replicated in inventory management; however, the disadvantages of these strategies carry over to inventory management, so this motivates the search for other approaches. In the context of inventory management, mean-variance models are used in the newsvendor problem, where advances presented in [5] paved the way for this topic to become an attractive one. In this regard, there are research works such as [6], [27],and [38].

We propose a different data-driven approach to address (1) using Distributionally Robust Optimization (DRO). DRO was initially proposed for unconstrained stochastic problems such as

$$J^* := \min_{x \in \mathbb{X}} \mathbb{E}_{\mathbb{P}}[f(x, \xi)], \tag{2}$$

where $\mathbb{X}$ is a set of feasible solutions, ξ is a random vector of parameters with unknown probability distribution $\mathbb{P}$, and $f(x, \xi)$ is a cost function. In this sense, DRO approach for problem (2) is formulated as

$$J_{\mathcal{D}} := \min_{x \in \mathbb{X}} \sup_{\mathbb{Q} \in \mathcal{D}} \mathbb{E}_{\mathbb{Q}}[f(x, \xi)], \tag{3}$$

where $\mathcal{D}$ is a set of probability distributions, which is known as *ambiguity set*. Note that $J^* \leq J_{\mathcal{D}}$ if $\mathbb{P} \in \mathcal{D}$. Therefore, a natural DRO version of (1) is the following optimization problem

$$\begin{cases} \min\limits_{x\in\mathbb{R}^m} & \sup\limits_{\mathbb{Q}\in\mathcal{D}} \Phi\left(x, F, \mathbb{Q}\right) \\ \text{subject to} & \inf\limits_{\mathbb{Q}\in\mathcal{D}} \mathbb{E}_{\mathbb{Q}}\left[G\left(x,\xi\right)\right] \geq \mu, \\ & x \in \mathcal{X}. \end{cases} \tag{4}$$

The choice of the set $\mathcal{D}$ is decisive in the tractability of this problem, and there are several ways to define $\mathcal{D}$ in the literature. For example, in [14] and [30] it is defined as a set of distributions that are supported in a single point, while in [8], [24], [28], and [31], $\mathcal{D}$ is defined as the set of distributions that satisfy specific restrictions in their moments, or distributions belonging to a given family of parametric distributions. Another option is to endow the set of probability distributions with a notion of distance, so we define $\mathcal{D}$ as a ball respect to this distance. Usually, this ball is centered on an empirical distribution $\widehat{\mathbb{P}}_N$, given by a sample $\widehat{\xi}_1, \ldots, \widehat{\xi}_N$ of the random vector ξ, and the radius is chosen in such a way that the distribution $\mathbb{P}$ belongs to the ball with high probability, or such that the out-of-sample performance of the optimal solution is good. Again, the tractability of the resulting DRO depends on the notion of distance adopted. Some distances frequently used are Burg's entropy, see [35]; Kullback-Leibler divergence, see [13]; and Total Variation distance, which is adopted in [32]. In this work we use *Wasserstein distance*, that is, we define $\mathcal{D}$ as a ball respect to the Wasserstein metric with center at an empirical distribution and radius properly chosen. Note that if choose radius 0 in this approach we recover the SAA strategy.

Definition 1 (Wassertein distance). *The Wasserstein distance $W_p(\mu,\nu)$ between $\mu,\nu \in \mathcal{P}_p(\Xi)$ is defined by*

$$W_p(\mu,\nu) := \left(\inf_{\Pi\in\mathcal{P}(\Xi\times\Xi)} \left\{\int_{\Xi\times\Xi} d^p(\xi,\zeta)\Pi(d\xi,d\zeta) \;:\; \Pi(\cdot\times\Xi)=\mu(\cdot),\, \Pi(\Xi\times\cdot)=\nu(\cdot)\right\}\right)^{1/p}$$

where

$$\mathcal{P}_p(\Xi) := \left\{\mu\in\mathcal{P}(\Xi) \;:\; \int_{\Xi} d^p(\xi,\zeta_0)\mu(d\xi) < \infty \text{ for some } \zeta_0\in\Xi\right\}$$

and d is a metric in Ξ.

W_p defines a metric in $\mathcal{P}_p(\Xi)$ for $p\in[1,\infty)$, hence the ball with respect to some p-Wasserstein distance with radius $\varepsilon>0$ and center $\mu\in\mathcal{P}(\Xi)$ is given by

$$\mathcal{B}_\varepsilon\left(\mu\right) := \left\{\nu\in\mathcal{P}(\Xi) \mid W_p(\mu,\nu)\leq\varepsilon\right\}. \tag{5}$$

One of the first works in which this notion of distance is defined is in [33] although this notion of distance arises in different fields of science almost simultaneously, and, depending on the context, it is usually known by other names. In computer science, it is called Earth moving distance; in the field of physics, it is called the distance of Monge-Kantorovich-Rubinstein, and, in the context of optimization, some researchers called it the Optimal Transport distance.

There are theoretical reasons, many exposed in [34], and practical reasons that make this distance very appealing. One of the main advantages is its dual representation. In fact, this property allows to find a more tractable equivalent formulation of (3).

Theorem 1. *Assume that* $\mathcal{D} := \mathcal{B}_\varepsilon\left(\widehat{\mathbb{P}}_N\right)$ *with respect to* W_p *and that* f *is upper semicontinuous. Then the problem (3) is equivalent to the optimization problem*

$$\begin{cases} \inf\limits_{x,\lambda,s} & \lambda\varepsilon^p + \dfrac{1}{N}\displaystyle\sum_{i=1}^{N} s_i \\ \textit{subject to} & \sup\limits_{\xi\in\Xi}\left(f(x,\xi) - \lambda d^p(\xi,\widehat{\xi}_i)\right) \leq s_i \ \forall i = 1,\ldots,N, \\ & \lambda \geq 0. \end{cases} \tag{6}$$

This theorem is formulated and proved in [3]. However, the reformulation (6) was also obtained in [12] and [19] under more restrictive assumptions. It is important to note that tractability of this reformulation does not depend on the form of the unknown true distribution $\mathbb{P}$.

Concretely, in this work we use DRO with Wasserstein metric to address stochastic programs with expected value constraints from a sample of the random vector. In that sense, our contributions are the following:

- We propose a data-driven robust formulation of (1) which prioritizes the expected value constraints without relying on regularization parameters.
- We show that our DRO approach of (1) can be reformulated as optimization problems with finite-dimensional variables. In portfolio optimization and inventory management contexts, for specific cases, we show that this problem is convex. We also establish criteria to determine the feasibility of the problems in terms of the radius of the ball and features of F and G.
- In inventory management, we focus on one version of the Newsvendor problem. Most of the works on this problem are parametric studies where they assume that the probability distribution of the random variable that influences demand is known. In contrast, our approach is non-parametric, so the performance of our strategy is compared to the SAA strategy.
- In portfolio optimization, we evaluate the performance of our approach and compare it with other traditional approaches. This evaluation is done on two types of data, the first type is synthetically generated return data, and the second is real market data. The results show the advantages of our proposal because the highest expected return and the highest Sharpe ratio are obtained compared to other benchmarks; in addition, a low turnover is obtained compared to the SAA strategy.

The organization of this paper is as follows. In Section 2, using Wasserstein distance, we describe our distributionally robust optimization model. In this section, we also derive tractable reformulations for the optimization problem and study its feasibility. Additionally, we establish a criterion to calibrate the size of

the ambiguity set. Simulation analysis of the proposed approaches are derived in Sections 3 and 4 for management inventory and portfolio optimization context respectively. Finally, conclusions are drawn in Section 5.

2 Problem formulation and main results

As stated before, in (1), the distribution $\mathbb{P}$ is unknown, and we assume we have access to realizations of the random vector ξ. That is, let $\widehat{\xi}_1, \ldots, \widehat{\xi}_N$ be a sample of ξ which allows to estimate $\mathbb{P}$ by means of the empirical distribution $\widehat{\mathbb{P}}_N := \frac{1}{N}\sum_{i=1}^{N} \delta_{\widehat{\xi}_i}$. Note that if we replace $\mathbb{P}$ with $\widehat{\mathbb{P}}_N$ in (1) we obtain the SAA strategy, with the drawbacks already mentioned. Another possible approach could be to consider $\mathcal{D} = \mathcal{B}_\varepsilon\left(\widehat{\mathbb{P}}_N\right)$ as ambiguity set in (4). However, this strategy has disadvantages. Note that if we use Theorem 1 to reformulate the objective function and the constraint in (4), for general functions F and G we will obtain a complicated and non-tractable semi-infinite optimization problem. This motivates our proposal, in which we seek to choose an ambiguity set that allows us to obtain a reformulation of (4) that is tractable and with good performance.

To present our approach we impose the following assumption.

Assumption 1 (Lipschitz). *We assume that F and G are Lipschitz functions respect to ξ. This is, for each x, there exists $\gamma_{x,F}, \gamma_{x,G} > 0$ such that $|F(x,\xi) - F(x,\zeta)| \leq \gamma_{x,F} \|\xi - \zeta\|$ and $|G(x,\xi) - G(x,\zeta)| \leq \gamma_{x,G} \|\xi - \zeta\|$ for all $\xi, \zeta \in \mathbb{R}^n$.*

Our approach also uses an empirical distribution but this will depend on x. First, we establish the following notation: For $x \in \mathbb{R}^m$ we define $\zeta^{x,F} := F(x,\xi)$ and $\zeta^{x,G} := G(x,\xi)$, note that these are random variables. We called $\mathbb{P}^{x,F}$ and $\mathbb{P}^{x,G}$ to the probability distributions of $\zeta^{x,F}$ and $\zeta^{x,G}$ respectively. Because it depends on $\mathbb{P}$, $\mathbb{P}^{x,F}$ and $\mathbb{P}^{x,G}$ are also unknown. Additionally, we define $\widehat{\zeta}_i^{x,F} := F\left(x, \widehat{\xi}_i\right)$ and $\widehat{\zeta}_i^{x,G} := G\left(x, \widehat{\xi}_i\right)$, so $\widehat{\zeta}_1^{x,F}, \ldots, \widehat{\zeta}_N^{x,F}$ is a sample of $\zeta^{x,F}$, and $\widehat{\zeta}_1^{x,G}, \ldots, \widehat{\zeta}_N^{x,G}$ is a sample of $\zeta^{x,G}$. This allows us to define the empirical distributions of $\zeta^{x,F}$ and $\zeta^{x,G}$, which are given by $\widehat{\mathbb{P}}_N^{x,F} := \frac{1}{N}\sum_{i=1}^{N} \delta_{\widehat{\zeta}_i^{x,F}}$ and $\widehat{\mathbb{P}}_N^{x,G} := \frac{1}{N}\sum_{i=1}^{N} \delta_{\widehat{\zeta}_i^{x,G}}$. This dependence on x has its justification in the fact that the decision vector x has an influence on whether the constraint of (1) is satisfied or not. Specifically, the constraint $\mathbb{E}_{\mathbb{P}}[G(x,\xi)] \geq \mu$ must be satisfied by $G(x,\xi)$ which depends on x. Therefore, we consider the following optimization problem: For a given $\epsilon > 0$

$$\widehat{J}_N(\varepsilon) := \begin{cases} \min\limits_{x \in \mathbb{R}^m} & \sup\limits_{\mathbb{Q} \in \mathcal{B}_{\varepsilon\gamma_{x,F}}\left(\widehat{\mathbb{P}}_N^{x,F}\right)} \Phi(x, \zeta, \mathbb{Q}) \\ \text{subject to} & \inf\limits_{\mathbb{Q} \in \mathcal{B}_{\varepsilon\gamma_{x,G}}\left(\widehat{\mathbb{P}}_N^{x,G}\right)} \mathbb{E}_{\mathbb{Q}}[\zeta] \geq \mu, \\ & x \in \mathcal{X}. \end{cases} \tag{7}$$

where $\mathcal{B}_{\varepsilon\gamma_{x,F}}\left(\widehat{\mathbb{P}}_N^{x,F}\right)$ and $\mathcal{B}_{\varepsilon\gamma_{x,G}}\left(\widehat{\mathbb{P}}_N^{x,G}\right)$ are balls centered at $\widehat{\mathbb{P}}_N^{x,F}$ and $\widehat{\mathbb{P}}_N^{x,G}$ with radius $\varepsilon\gamma_{x,F}$ and $\varepsilon\gamma_{x,G}$ respectively. These balls are defined with respect to the p-Wassertein distance where p is 1 or 2. The type of p-Wasserstein distance that we use depend on Φ and the support of ξ. We see this in detail in the following subsections. However, the following lemma gives the reason why our ambiguity sets has radius $\varepsilon\gamma_{x,F}$ and $\varepsilon\gamma_{x,G}$, and also shows a relationship between $\mathcal{B}_{\varepsilon\gamma_{x,F}}\left(\widehat{\mathbb{P}}_N^{x,F}\right)$, $\mathcal{B}_{\varepsilon\gamma_{x,G}}\left(\widehat{\mathbb{P}}_N^{x,G}\right)$, and $\mathcal{B}_{\varepsilon}(\widehat{\mathbb{P}}_N)$.

Lemma 1. $W_p(\widehat{\mathbb{P}}_N^{x,F}, \mathbb{P}^{x,F}) \leq \gamma_{x,F} W_p\left(\widehat{\mathbb{P}}_N, \mathbb{P}\right)$ *for* $p \geq 1$.

Note that if $\varepsilon > 0$ is such that $\mathbb{P} \in \mathcal{B}_{\varepsilon}(\widehat{\mathbb{P}}_N)$, where the ball is taken with respect to the p-Wasserstein metric for $p = 1, 2$ and distributions supported in a subset of $\mathbb{R}^n$, then $\mathbb{P}^{x,F} \in \mathcal{B}_{\varepsilon\gamma_{x,F}}(\widehat{\mathbb{P}}_N^{x,F})$, where these balls are taken with respect to the p-Wasserstein metric for $p = 1, 2$ and distributions supported in a subset of $\mathbb{R}$. The proof of Lemma 1 is addressed in A.1. From now on, the optimal solutions of (7) will be denoted by $\widehat{x}_N(\varepsilon)$.

Figure 1 shows an illustration of our approach for the case $F = G$ and $\Phi(x, F, \mathbb{P}) = \mathrm{Var}_{\mathbb{P}}[F(x, \xi)]$ in (7). In this figure, we have two decisions, x^1 and x^2, which induce two balls. The centers of these balls must belong to the blue shaded region. Additionally, these balls must be contained in the red shaded region that represents the semi-space induced by the constraint in (7). From the figure we can see that x^2 is a better decision than x^1 since the measures in the ball induced by x^2 have lower variance levels compared to the measures within the ball induced by x^1. This implies that the supreme of the variance in $\mathcal{B}_{\varepsilon\gamma_{x,G}}\left(\widehat{\mathbb{P}}_N^{x^2,G}\right)$ is less than the supreme of the variance in $\mathcal{B}_{\varepsilon\gamma_{x,G}}\left(\widehat{\mathbb{P}}_N^{x^1,G}\right)$.

2.1 Risk neutral case

In this case we assume that $\Phi(x, F, \mathbb{P}) := \mathbb{E}_{\mathbb{P}}[F(x, \xi)]$. The first task is to reformulate (7) as an optimization problem with finite-dimensional variables. This reformulation depends on the image of the support of ξ under the functions $F(x, \cdot)$ and $G(x, \cdot)$ for each $x \in \mathcal{X}$.

Theorem 2. *We have the following cases:*

1. 1*-Wasserstein distance. If* $F(x, \Xi) = [A_F(x), B_F(x)]$ *and* $G(x, \Xi) = [A_G(x), B_G(x)]$ *for each* $x \in \mathcal{X}$*, then the optimization problem (7) is equivalent to the following optimization problem*

$$\widehat{J}_N(\varepsilon) = \begin{cases} \underset{x\in\mathbb{R}^m}{\text{minimize}} \ \min\left\{\frac{1}{N}\sum_{i=1}^{N} F\left(x, \widehat{\xi}_i\right) + \varepsilon\gamma_{x,F}, B_F(x)\right\} \\ \textit{subject to} \max\left\{\dfrac{1}{N}\displaystyle\sum_{i=1}^{N} G\left(x, \widehat{\xi}_i\right) - \varepsilon\gamma_{x,G}, A_G(x)\right\} \geq \mu, \\ \qquad\qquad x \in \mathcal{X}. \end{cases} \tag{8}$$

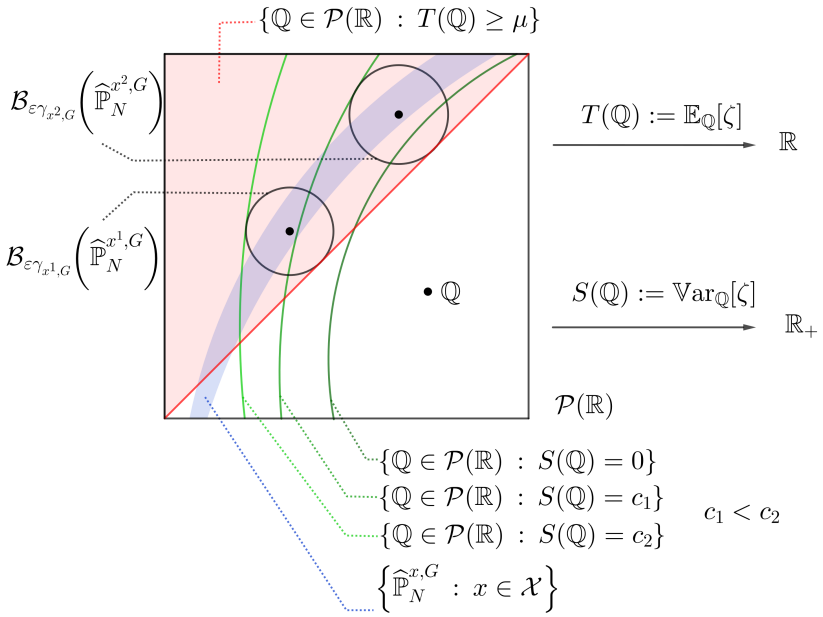


Fig. 1: Illustration of the optimization process proposed in (7) for the case $F = G$ and $\Phi\left(x, F, \mathbb{P}\right) = \mathrm{Var}_{\mathbb{P}}\left[F(x, \xi)\right]$.

2. 2-*Wasserstein distance. If If* $F(x, \Xi) = [A_F(x), \infty)$ *and* $G(x, \Xi) = (-\infty, B_G(x)]$ *for each* $x \in \mathcal{X}$*, then the optimization problem (7) is equivalent to the following optimization problem*

$$\widehat{J}_N(\varepsilon) = \begin{cases} \underset{x \in \mathbb{R}^m}{\text{minimize}} & \frac{1}{N}\sum_{i=1}^{N} F\left(x, \widehat{\xi}_i\right) + \varepsilon\gamma_{x,F} \\ \textit{subject to} & \frac{1}{N}\sum_{i=1}^{N} G\left(x, \widehat{\xi}_i\right) - \varepsilon\gamma_{x,G} \geq \mu, \\ & x \in \mathcal{X}. \end{cases} \tag{9}$$

In principle, problem (9) may be easier to solve than to solve (8). However, in certain cases, for example, in the context of inventory management (8) it is a tractable problem as we will see later.

Note that some values of ε and μ can make problem (7) infeasible. The conditions to obtain feasibility are established in the following corollaries.

Corollary 1. *The feasibility of (8) is divided into the following cases:*

i) When there exists $x \in \mathcal{X}$ such that $A_G(x) \geq \mu$, the optimization problem (8) is feasible for all $\varepsilon > 0$, and for μ satisfying the following inequality

$$\mu < \widehat{\mu}_N^{\max}(\varepsilon) := \begin{cases} \sup\limits_{x \in \mathbb{R}^m} & \max\left\{ \dfrac{1}{N}\sum\limits_{i=1}^{N} G\left(x, \widehat{\xi}_i\right) - \varepsilon\gamma_{x,G}, A_G(x) \right\} \\ \textit{subject to} & A_G(x) \geq \mu \\ & x \in \mathcal{X}. \end{cases}. \tag{10}$$

ii) When $A_G(x) < \mu$ for all $x \in \mathcal{X}$, the optimization problem (8) is feasible if μ and ε satisfies the following inequalities

$$\varepsilon \leq \widehat{\varepsilon}_N^{\max}(\mu) := \begin{cases} \sup\limits_{x \in \mathbb{R}^m} & \dfrac{\frac{1}{N}\sum\limits_{i=1}^{N} G(x, \widehat{\xi}_i) - \mu}{\gamma_{x,G}} \\ \textit{subject to} & \frac{1}{N}\sum\limits_{i=1}^{N} G(x, \widehat{\xi}_i) \geq \mu, \\ & x \in \mathcal{X}. \end{cases}$$

and

$$\mu < \widehat{\mu}_N^{\max}(\varepsilon) := \begin{cases} \sup\limits_{x \in \mathbb{R}^m} & \max\left\{ \dfrac{1}{N}\sum\limits_{i=1}^{N} G\left(x, \widehat{\xi}_i\right) - \varepsilon\gamma_{x,G}, A_G(x) \right\} \\ \textit{subject to} & x \in \mathcal{X}. \end{cases}. \tag{11}$$

Note that in case *i)* of the previous corollary we can consider $\widehat{\varepsilon}_N^{\max}(\mu) = \infty$.

Corollary 2. *The optimization problem (9) is feasible if μ and ε satisfies the following inequalities*

$$\mu < \widehat{\mu}_N^{\max}(\varepsilon) := \begin{cases} \sup\limits_{x \in \mathbb{R}^m} & \frac{1}{N}\sum\limits_{i=1}^{N} G(x, \widehat{\xi}_i) - \varepsilon\gamma_{x,G} \\ \textit{subject to} & x \in \mathcal{X}. \end{cases}$$

and

$$\varepsilon \leq \widehat{\varepsilon}_N^{\max}(\mu) := \begin{cases} \sup\limits_{x \in \mathbb{R}^m} & \dfrac{\frac{1}{N}\sum\limits_{i=1}^{N} G(x, \widehat{\xi}_i) - \mu}{\gamma_{x,G}} \\ \textit{subject to} & \frac{1}{N}\sum\limits_{i=1}^{N} G(x, \widehat{\xi}_i) \geq \mu, \\ & x \in \mathcal{X}. \end{cases}$$

2.2 Variance case

In this case we assume that $\Phi(x, F, \mathbb{P}) := \mathrm{Var}_{\mathbb{P}}[F(x,\xi)]$. As in the previous case, the first task is to reformulate (7) but to make it computationally tractable we only consider 2-Wasserstein distance.

Theorem 3. *For 2-Wasserstein distance, if $F(x,\Xi) = [0,\infty)$ or $F(x,\Xi) = \mathbb{R}$, and $G(x,\Xi) = [0,\infty)$ or $G(x,\Xi) = \mathbb{R}$ for each $x \in \mathcal{X}$, then the optimization problem (7) is equivalent to the following optimization problem with finite-dimensional variables*

$$\widehat{J}_N(\varepsilon) = \begin{cases} \underset{x\in\mathbb{R}^m}{\text{minimize}} & \left(\sqrt{\frac{1}{N}\sum_{i=1}^{N} F\left(x,\widehat{\xi}_i\right)^2 - \frac{1}{N^2}\left(\sum_{i=1}^{N} F\left(x,\widehat{\xi}_i\right)\right)^2} + \varepsilon\gamma_{x,F}\right)^2 \\ \textit{subject to} & \frac{1}{N}\sum_{i=1}^{N} G\left(x,\widehat{\xi}_i\right) - \varepsilon\gamma_{x,G} \geq \mu, \\ & x \in \mathcal{X} \end{cases} \tag{12}$$

Note that Corollary 2 also applies in this case. One of the advantages of our approach is that the resulting optimization problem is finite dimensional optimization problem, and, for some functions F and G, convexity is obtained; for example, this happens in the context of portfolio optimization. The proofs of Theorem 2 and Theorem 3 can be consulted in A.4.

2.3 Choice of Wasserstein radius

To finish this section, we analyze how to choose ε. Our goal is to ensure that, out of sample, the constraint in (1) is satisfied. For example, in (7), we want ε such that if $\widehat{x}_N(\varepsilon)$ is a optimal solution, then $\mathbb{E}_{\mathbb{P}}[G(\widehat{x}_N(\varepsilon),\xi)] \geq \mu$ is satisfied with high probability. In that sense, the following lemma could provide a criterion to achieve this.

Lemma 2. *Let $\varepsilon > 0$ such that $\mathbb{P} \in \mathcal{B}_\varepsilon(\widehat{\mathbb{P}}_N)$. If ε is such that (7) is feasible, and $\widehat{x}_N(\varepsilon)$ is a optimal solution of (7), then $\mathbb{E}_{\mathbb{P}}[G(\widehat{x}_N(\varepsilon),\xi)] \geq \mu$.*

This lemma suggests that it is enough to find an ε such that $\mathbb{P} \in \mathcal{B}_\varepsilon(\widehat{\mathbb{P}}_N)$ and this is achieved with large values of ε, but we must bear in mind that ε could be limited by $\widehat{\varepsilon}_N^{\max}(\mu)$ in (7). Another problem is that there is no efficient method to determine from which ε the condition $\mathbb{P} \in \mathcal{B}_\varepsilon(\widehat{\mathbb{P}}_N)$ is satisfied. In that vein, our idea is not to concentrate on guaranteeing that condition, instead, we concentrate on satisfying $\mathbb{E}_{\mathbb{P}}[G(\widehat{x}_N(\varepsilon),\xi)] \geq \mu$ in (7). To do this we use a strategy based on a Bootstrap method assuming that ξ satisfies all the conditions for the Bootstrap-based technique to be valid (see [10]). Given $\varepsilon > 0$, the strategy is to estimate the probability that the constraint is satisfied, which we call the *confidence level of* ε. The following method estimates this probability.

- *Confidence level of ε for (7):* Given $\widehat{x}_N(\varepsilon)$, generate K bootstrap samples with repetition from $\widehat{\Xi}_N = \left\{\widehat{\xi}_1, \ldots, \widehat{\xi}_N\right\}$, all these K samples of size N. We denote these bootstrap samples as $\widehat{\Xi}^{\text{bt}}_{N,i} := \left\{\widehat{\xi}^{\text{bt},i}_1, \ldots, \widehat{\xi}^{\text{bt},i}_N\right\}$ for $i = 1, 2, \ldots, K$. We define $\widehat{\mathbb{P}}^{\text{bt}}_{N,i}$ as the empirical distribution generated by $\widehat{\Xi}^{\text{bt}}_{N,i}$. Next, we calculate the sample mean of $G\left(\widehat{x}_N(\varepsilon), \xi\right)$ induced by $\widehat{\mathbb{P}}^{\text{bt}}_{N,i}$, that is, $\mathbb{E}_{\widehat{\mathbb{P}}^{\text{bt}}_{N,i}}\left[G\left(\widehat{x}_N(\varepsilon), \xi\right)\right]$. Therefore, the estimate confidence level of ε is the percentage of times in which $\mathbb{E}_{\widehat{\mathbb{P}}^{\text{bt}}_{N,i}}\left[G\left(\widehat{x}_N(\varepsilon), \xi\right)\right] \geq \mu$.

If the estimate confidence level of ε is β, this means that the constraint is satisfied with a approximate probability $\beta/100$. Thus, the strategy is to find the smallest ε such that its confidence level β is acceptable, for example, $\beta = 75\%$. Since the confidence level decreases as ε gets smaller, so the maximum level of confidence that can be obtained is the one reached with $\varepsilon = \widehat{\varepsilon}^{\max}_N(\mu)$, we can start from this value whenever is finite. In Sections 3.1 and 4.1 we will show the performance of this procedure with two applications. The proof of Lemma 2 is relegated to A.1.

3 Newsvendor problem

In this section, we consider the following newsvendor problem

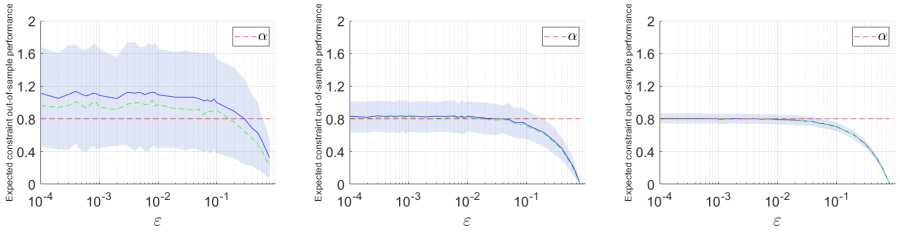
$$J = \begin{cases} \max\limits_{x} & \mathbb{E}_{\mathbb{P}}\left[p \min\left\{\xi, x\right\} - cx\right] \\ \text{subject to} & \mathrm{E}_{\mathbb{P}}\left[(\xi - x)^+\right] \leq \alpha, \\ & x \geq 0. \end{cases}$$

where $c > 0$ is the manufacturing cost, $p \geq c$ is the sell price, and x is the inventory level set for the good. In addition, ξ is the demand of the good where the probability distribution $\mathbb{P}$ of ξ is unknown, but with support $\Xi = [0, \infty)$. In this problem the constraint excludes inventory levels that underestimate the demand. Note that this problem is equivalent to

$$-\begin{cases} \min\limits_{x} & \mathbb{E}_{\mathbb{P}}\left[cx - p \min\left\{\xi, x\right\}\right] \\ \text{subject to} & \mathrm{E}_{\mathbb{P}}\left[-(\xi - x)^+\right] \geq -\alpha, \\ & x \geq 0, \end{cases}$$

and this is a particular case of (1). Indeed, in (1), we consider $m = n = 1$, F and G defined as $F(x, \xi) := cx - p \min\left\{\xi, x\right\}$ and $G(x, \xi) := -(\xi - x)^+ = \min\left\{x - \xi, 0\right\}$ respectively, and $\mu = -\alpha$. Note that F and G are Lipschitz functions with Lipschitz constants $\gamma_{x,F} = p$ and $\gamma_{x,G} = 1$, respectively. Moreover, in this case, $A_F(x) = (c - p)x$, $B_F(x) = cx$, $A_G(x) = -\infty$, and $B_G(x) = 0$. Therefore, given $\widehat{\xi}_1, \ldots, \widehat{\xi}_N$ sample of ξ, we solve the following problem

$$\widehat{J}_N(\varepsilon) = -\begin{cases} \min\limits_{x} & \sup\limits_{\mathbb{Q} \in \mathcal{B}_{\varepsilon p}\left(\widehat{\mathbb{P}}^{x,F}_N\right)} \mathbb{E}_{\mathbb{Q}}\left[cx - p \min\left\{\xi, x\right\}\right] \\ \text{subject to} & \inf\limits_{\mathbb{Q} \in \mathcal{B}_{\varepsilon}\left(\widehat{\mathbb{P}}^{x,G}_N\right)} \mathrm{E}_{\mathbb{Q}}\left[-(\xi - x)^+\right] \geq -\alpha, \\ & x \geq 0. \end{cases} \tag{13}$$



$N = 30$. $N = 300$. $N = 3000$.

Fig. 2: Out-of-sample performance of expected constraint $\mathbb{E}_{\mathbb{P}}\left[(\xi - \widehat{x}_N(\varepsilon))^+\right]$ as a function of the Wasserstein radius ε and estimated on the basis of 1000 simulations. The blue solid line is the mean, the green dashed line is the median, and the blue shaded area is the tube between the 20% and 80% quantile of data generated by 1000 simulations. In this case, $\alpha = 0.8$.

From Theorem 2-1 it follows the next proposition.

Proposition 1. *The problem (13) is equivalent to the optimization program*

$$\begin{cases} \max\limits_{x\in\mathbb{R}, z\in\mathbb{R}^N} & \frac{p}{N}\sum\limits_{i=1}^{N} z_i - p\varepsilon - cx \\ \textit{subject to} & -\frac{1}{N}\sum\limits_{i=1}^{N} z_i + \frac{1}{N}\sum\limits_{i=1}^{N}\widehat{\xi}_i + \varepsilon \leq \alpha, \\ & \frac{1}{N}\sum\limits_{i=1}^{N} z_i \geq \varepsilon, \\ & z_i \leq \widehat{\xi}_i, \qquad \forall i = 1, 2, \ldots, N, \\ & z_i \leq x, \qquad \forall i = 1, 2, \ldots, N, \\ & x \geq 0. \end{cases} \tag{14}$$

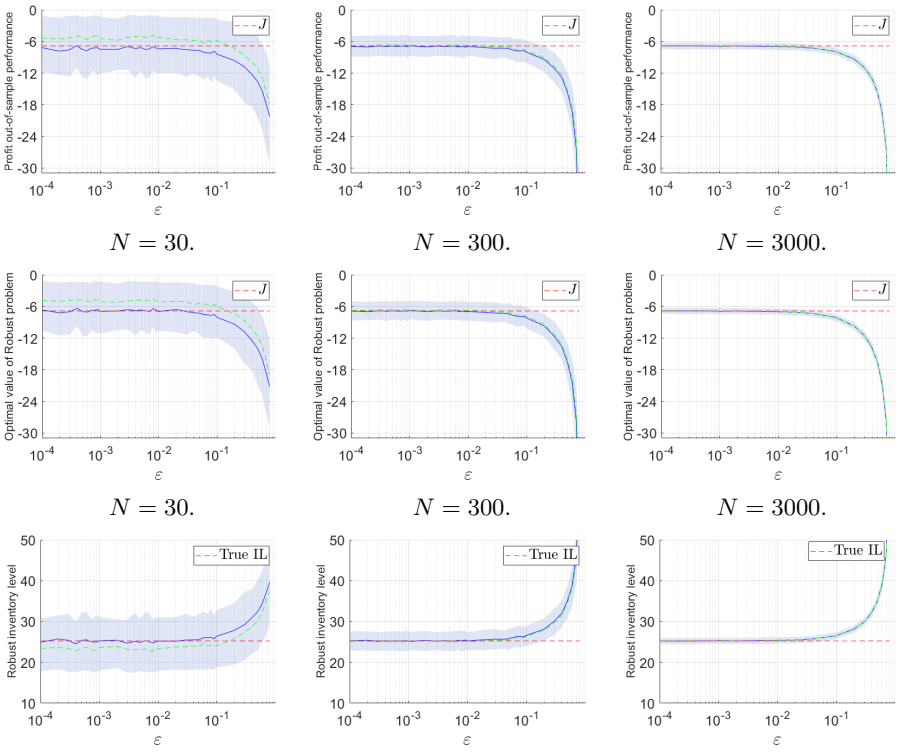
The proof of this proposition can be consulted in A.5. Also, as a result of the Corollary 1, it follows that the problem is feasible if $\varepsilon \leq \min\left\{\bar{\xi}, \alpha\right\}$, where $\bar{\xi} = \frac{1}{N}\sum_i^N \widehat{\xi}_i$. In addition, the reformulation obtained from this proposition is a linear optimization problem.

3.1 Numerical experiments and results

We use synthetically generated data, that is, generated by a known distribution. Specifically, we consider ξ as a random variable exponentially distributed with mean equal to 10. Additionally, we consider the price $p = 2$, the cost $c = 1$ and $\alpha = 0.8$.

Impact of the Wasserstein Radius ε. Our first objective is to analyze the impact of the Wasserstein radius ε on the optimal distributionally robust inventory levels and their out-of-sample performance. Therefore, we solve problem (14) using samples of size $N \in \{30, 300, 3000\}$. Figure 2 shows the tube between

the 20% and 80% quantiles (shaded area), the mean (blue solid line), and the median (green dashed line) of the out-of-sample performance of expected constraint $\mathbb{E}_{\mathbb{P}}\left[(\xi - \widehat{x}_N(\varepsilon))^+\right]$ as a function of ε estimated using 1000 independent simulation runs. We observe that the out-of-sample performance of expected constraint is decreasing as Wasserstein radius ε grows. Furthermore, according to the median of the results, it is observed that, for small N, with the SAA approach ($\varepsilon = 0$) the restriction is not fulfilled in a percentage higher than 50% of the simulations. However, this figure shows that it is possible to find ε such that the constraint is satisfied out-of-sample with high probability, the higher this probability the higher ε. Note that this ε also depends on α (see Corollary 1). This stylized fact was observed consistently across all of simulations and provides an empirical justification for adopting a distributionally robust approach.



$N = 30.$ $N = 300.$ $N = 3000.$

$N = 30.$ $N = 300.$ $N = 3000.$

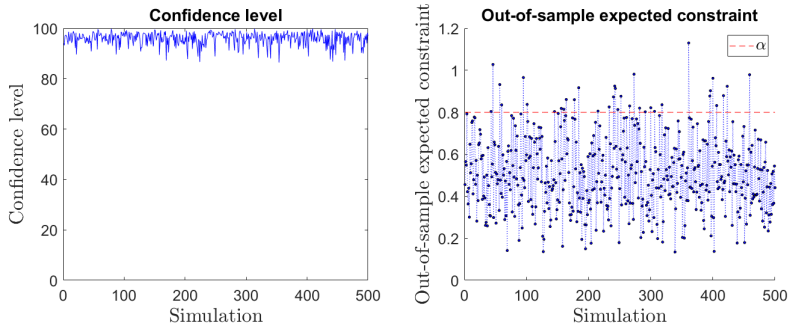
$N = 30.$ $N = 300.$ $N = 3000.$

Fig. 3: Out-of-sample performance of Out-of-sample performance of the profit $\mathbb{E}_{\mathbb{P}}[p \min\{\xi, \widehat{x}_N(\varepsilon)\} - c\widehat{x}_N(\varepsilon)]$, Optimal value $\widehat{J}_N(\varepsilon)$, and robust optimal inventory level (IL) $\widehat{x}_N(\varepsilon)$ as a function of the Wasserstein radius ε and estimated on the basis of 1000 simulations. The blue solid lines are the means, the green dashed lines are the medians, and the blue shaded areas are the tubes between the 20% and 80% quantile of data generated by 1000 simulations. In this case, $\alpha = 0.8$.

Figure 3 shows the profit $\mathbb{E}_{\mathbb{P}}[p\min\{\xi,\widehat{x}_N(\varepsilon)\}-c\widehat{x}_N(\varepsilon)]$, optimal values $\widehat{J}_N(\varepsilon)$, and robust optimal inventory level (IL) $\widehat{x}_N(\varepsilon)$. In the case of the out-of-sample performance of profit $\mathbb{E}_{\mathbb{P}}[p\min\{\xi,\widehat{x}_N(\varepsilon)\}-c\widehat{x}_N(\varepsilon)]$ and optimal value $\widehat{J}_N(\varepsilon)$, are decreasing as Wasserstein radius ε grows. In addition, it is observed that low values of the expected constraint induce low profits. However, the inventory level is high compared to that obtained with the SAA approach. Furthermore, when ε is large, the decision that our approach suggests is to consider high inventory levels. This shows that our strategy prioritizes having all possible demand covered, which is achieved with large inventory levels.

Another aspect that follows from these figures is that if ε is such that $\mathbb{E}_{\mathbb{P}}\left[(\xi-\widehat{x}_N(\varepsilon))^+\right]\leq\alpha$ in a large percentage of the simulations, then $\widehat{J}_N(\varepsilon)\leq J$ in a similar percentage of those simulations. This is consistent with the formulation of problem (7) because if ε is such that $\mathbb{P}^{\widehat{x}_N(\varepsilon),G}\in\mathcal{B}_\varepsilon(\widehat{\mathbb{P}}_N^{\widehat{x}_N(\varepsilon),G})$, then $\widehat{J}_N(\varepsilon)\leq J$ and $\mathbb{E}_{\mathbb{P}}\left[(\xi-\widehat{x}_N(\varepsilon))^+\right]\leq\alpha$. The figures verify for the existence of this ε. Even though this observation was made consistently across all simulations, we were unable to validate it theoretically.

Performance of confidence level of ε. Now, we evaluate the performance of the strategy to determine the confidence level of any given ε, this strategy is based on the Bootstrap method as presented above. For this purpose we generate 500 samples of size $N=300$, for each of these samples we calculate $\widehat{\varepsilon}_N^{\max}(\alpha)$ and consider $\varepsilon=2\widehat{\varepsilon}_N^{\max}(\alpha)/5$. Given this ε we calculate its confidence level obtaining Figure 4(a). This figure shows that the confidence levels are high for this value of ε, specifically, it is around 95.8%. This means that $\mathbb{E}_{\mathbb{P}}\left[(\xi-\widehat{x}_N(\varepsilon))^+\right]$ should be above α by a percentage of around 95.8% of the 500 simulations. Additionally, for each sample we calculate the out-of-sample expected constraint of $\widehat{x}_N(\varepsilon)$, that is, $\mathbb{E}_{\mathbb{P}}\left[(\xi-\widehat{x}_N(\varepsilon))^+\right]$. Figure 4(b) shows that in 92.3% of the simulations, $\mathbb{E}_{\mathbb{P}}\left[(\xi-\widehat{x}_N(\varepsilon))^+\right]\leq\alpha$ was obtained. This shows that our bootstrap-based strategy has a good performance. Moreover, this is consistent with what was observed in the out-of-sample performance of the expected return in Figure 2.



(a) (b)

Fig. 4: In (a), the confidence level of $\varepsilon=2\widehat{\varepsilon}_N^{\max}(\alpha)/5$. In (b), the out-of-sample performance of $\mathbb{E}_{\mathbb{P}}\left[(\xi-\widehat{x}_N(\varepsilon))^+\right]$. 500 simulations with $N=300$ and $\alpha=0.8$.

4 Portfolio optimization

In this case, in (1), we consider Φ as variance, $m = n$ where m is the number of assets, ξ_i is the return of i-th asset, and x_i the proportion of the initial amount invested in the i-th asset. Given that the returns of each asset are random with *unknown* distribution $\mathbb{P}$, $\xi = (\xi_1, \ldots, \xi_m) \in \mathbb{R}^m$ is a random vector. Additionally, $x = (x_1, \ldots, x_m) \in \mathbb{R}^m$ is a portfolio which is vector of weights that satisfies the relation $\sum_{i=1}^m x_i = 1$ and other additional convex constraints admissible for the investor described by the set X, so we have $\mathcal{X} = \{x \in \mathbb{R}^m : \sum_{i=1}^m x_i = 1,\ x \in \mathsf{X}\}$. Additionally, $F = G$ with $F(x, \xi) = \langle x, \xi \rangle$ where $\langle \cdot, \cdot \rangle$ is the Euclidean inner product in $\mathbb{R}^m$. Then, $\langle x, \xi \rangle$ is the return of the portfolio x. Note that F satisfies Assumption 1 because F is a Lipschitz function with respect to ξ with Lipschitz constant $\gamma_{x,F} = \|x\|$. Finally, μ is the minimum level of return admissible for the investor. With these considerations in mind, (1) becomes the Markowitz mean-variance portfolio selection optimization problem. This consists of choosing portfolio weights such that minimize the variance of the return rate subject to a constraint on the expected value of the return rate. From Theorem 3 we obtain the following formulation.

Proposition 2. *Let M be a matrix with size $m \times N$ where its columns are $\widehat{\xi}_1, \ldots, \widehat{\xi}_N$, and let $\mathbf{e} \in \mathbb{R}^N$ be the column vector with ones. Define*

$$E := \frac{1}{N} M M^T - \frac{1}{N^2} (M\mathbf{e})(M\mathbf{e})^T \quad \textit{and} \quad L := \frac{1}{N} (M\mathbf{e})^T.$$

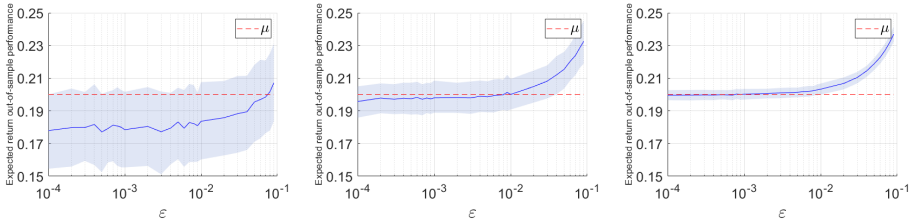
Therefore, (12) is equivalent to the optimization problem

$$\begin{cases} \inf\limits_{x \in \mathbb{R}^m} & \left(\sqrt{x^T E x} + \varepsilon \|x\|\right)^2 \\ \textit{subject to} & Lx - \varepsilon \|x\| \geq \mu, \\ & \mathbf{e}^T x = 1, \\ & x \in \mathsf{X}. \end{cases} \tag{15}$$

Note that E and L are the sample versions of the covariance matrix and the vector of means of ξ respectively, generated by the sample $\widehat{\xi}_1, \ldots, \widehat{\xi}_N$. As a result of the Corollary 2, note that if $\mathsf{X} = \mathbb{R}^m$, that is, when short selling is allowed, then $\widehat{\mu}_N^{\max}(\varepsilon) \to \infty$ when $\varepsilon \to 0$, hence for every μ there is a ε that makes the problem feasible. However, regardless of whether short selling is allowed or not, note that $\widehat{\varepsilon}_N^{\max}(\mu)$ is decreasing with respect to μ, so this reduces the number of ε candidates that we can explore as μ grows.

4.1 Numerical experiments and results

This subsection is divided into two parts. In one we use synthetically generated data, that is, generated by a known distribution. In the other part, we use real data from the financial market. In both cases we consider $\mathsf{X} = \mathbb{R}^m_+$, that is, we do not consider short selling.



$N = 30$. $N = 300$. $N = 3000$.

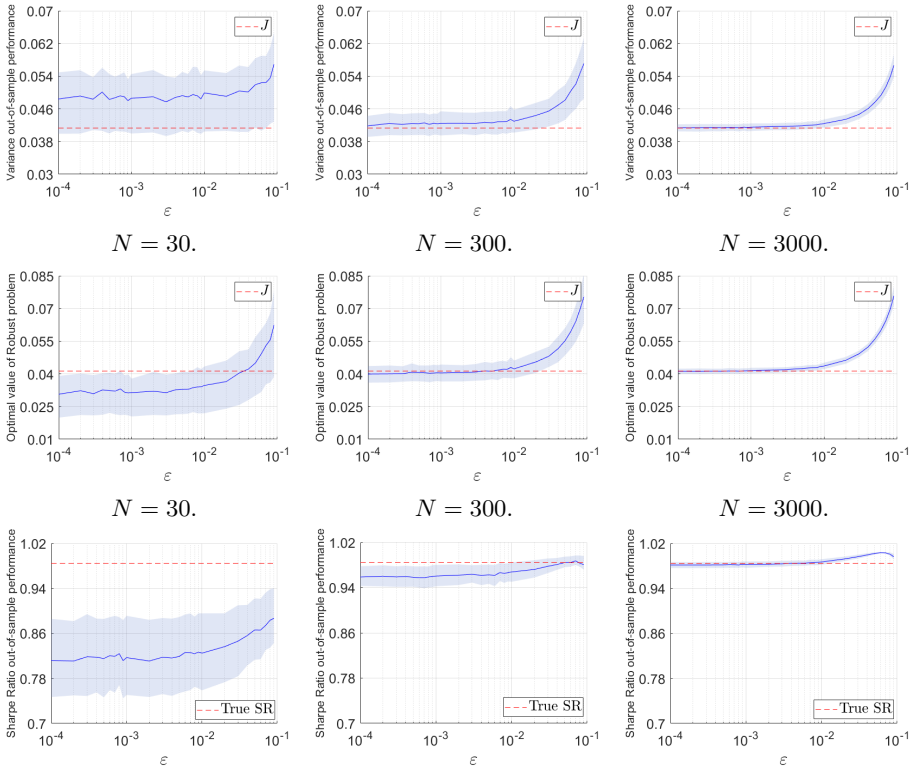
Fig. 5: Out-of-sample performance of expected return $R(\hat{x}_N(\varepsilon))$ as a function of the Wasserstein radius ε and estimated on the basis of 500 simulations. The blue solid line is the mean, and the blue shaded area is the tube between the 20% and 80% quantile of data generated by 500 simulations. In this case, $\mu = 0.2$.

Using simulated data. We consider a market of $m = 10$ assets which returns have the form adopted in [12], that is, we assume that $\xi_i = \psi + \zeta_i$ where ψ and ζ_i are independent, $\psi \sim \mathcal{N}(0, 2\%)$ and $\zeta_i \sim \mathcal{N}(i \times 3\%, i \times 2.5\%)$ for each $i = 1, 2, \ldots, m$. With this assumption, the assets are ordered from the one with the lowest return and volatility to the one with the highest return and volatility. Additionally, we denote by $\mathbf{m}$ the vector of means and Σ the covariance matrix of ξ. In this case, $\mathbf{m}$ and Σ are easy to calculate from the distribution of ξ. Given $x \in \mathbb{R}^m$ we define $R(x) := \mathbf{m}^T x$, the expected return induced by x, and $V(x) := x^T \Sigma x$, the variance induced by x. Because we have all the information about the returns, the vector of optimal weights x^* is known.

Impact of the Wasserstein Radius ε. Our first objective is to analyze the impact of the Wasserstein radius ε on the optimal distributionally robust portfolios and their out-of-sample performance. Therefore, we solve problem (15) using samples of size $N \in \{30, 300, 3000\}$. Figure 5 shows the tube between the 20% and 80% quantiles (shaded area) and the mean (solid line) of the out-of-sample performance of expected returns $R(\widehat{x}_N(\varepsilon))$ as a function of ε estimated using 500 independent simulation runs.We observe that the out-of-sample performance of expected returns $R(\widehat{x}_N(\varepsilon))$ are increasing as Wasserstein radius ε grows. Therefore, as in the previous application, the higher the ϵ the higher the probability of satisfyng the constraint. Note also that this depends on the value of μ.

Regarding the out-of-sample performance of variance $V(\widehat{x}_N(\varepsilon))$ and optimal value $\widehat{J}_N(\varepsilon)$, Figure 6 also shows that this quantities are increasing as Wasserstein radius ε grows. In addition, it is observed that high returns induce large variances. However, this figure also shows that Sharpe ratio is always greater than the one obtained by the SAA strategy, in fact, the Sharpe ratio reaches a maximum value near the end of the ε grid which is close to $\widehat{\varepsilon}_N^{\max}(\mu)$. Furthermore, for large N this maximum Sharpe Ratio exceeds the true Sharpe Ratio, where the latter is defined as $\frac{R(x^*)}{\sqrt{V(x^*)}}$. Overall, from these figures we can extract the same conclusions obtained in the newsvendor case.

Finally, Figure 7 visualizes the corresponding optimal portfolio weights $\widehat{x}_N(\varepsilon)$ as a function of ε, averaged over 500 independent simulation runs. The thin



$N = 30$. $N = 300$. $N = 3000$.

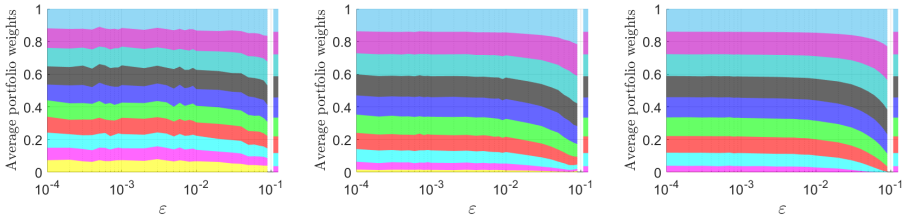
$N = 30$. $N = 300$. $N = 3000$.

$N = 30$. $N = 300$. $N = 3000$.

Fig. 6: Out-of-sample performance of the variance of return $V(\hat{x}_N(\varepsilon))$, optimal value $\widehat{J}_N(\varepsilon)$, and Sharpe Ratio as a function of the Wasserstein radius ε and estimated on the basis of 500 simulations. The blue solid lines are the means, and the blue shaded areas are the tubes between the 20% and 80% quantile of data generated by 500 simulations. In this case, $\mu = 0.2$.

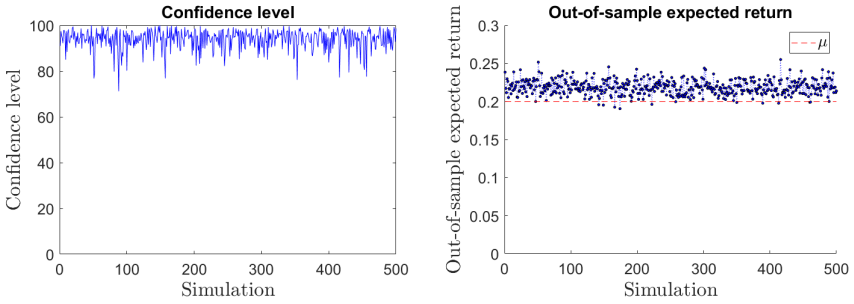
colored bar on the right side of each graph corresponds to x^*. Our numerical results show that the optimal distributionally robust portfolios tends to give little weight to goods with little return even if they have little volatility while it gives more weight to goods with high return even if they have high volatility, all this, as the Wasserstein radius ε increases.

Performance of confidence level of ε. To evaluate the performance of the strategy to determine the confidence level of any given ε we consider $\mu = 0.2$ and perform 500 simulations of size $N = 300$. Using $\varepsilon = 2\widehat{\varepsilon}_N^{\max}(\mu)/5$ and proceeding as in the previous application, Figure 8 shows that the confidence level is around 93,8% while in 95.2% of the simulations, $R(\widehat{x}_N(\varepsilon)) \geq \mu$ was obtained. This shows that the bootstrap-based strategy has a good performance.



$N = 30$. $N = 300$. $N = 3000$.

Fig. 7: Optimal portfolio composition as a function of the Wasserstein radius ε averaged over 500 simulations; the portfolio weights are depicted in ascending order, i.e., the weight of asset 1 at the bottom and that of asset 10 at the top. In this case, $\mu = 0.2$.



(a) (b)

Fig. 8: In (a), the confidence level of $\varepsilon = 2\widehat{\varepsilon}_N^{\max}(\mu)/5$. In (b), the out-of-sample performance of $\widehat{x}_N(\varepsilon)$ for $\varepsilon = 2\widehat{\varepsilon}_N^{\max}(\mu)/5$, that is, $R(\widehat{x}_N(\varepsilon))$. All this for each of the 500 simulations with $N = 300$ and $\mu = 0.2$.

Using real market data We now consider real market data. The data used in this study correspond to the daily returns of 23 companies selected from the S&P 500 index. The selected returns correspond to the companies described below.

AAPL - Apple	INTC - Intel	PG - P&G
AMZN - Amazon	JNJ - Johnson & Johnson	T - AT&T
BAC - Bank of America	JPM - J.P Morgan	UNH -UnitedHealth Group
BRKA - Berkshire Hathaway	KO - Coca Cola	VZ - Verizon
CVX - Chevron	MA - Mastercard	WFC - Wells Fargo
DIS - Disney	MRK - Merck & Co	WMT - Walmart
HD - The Home Depot	XOM - Exxom Mobil	MSFT - Microsoft
PFE - Pfizer	GOOG - Alphabet Google	

These data correspond to the time window between January 1, 2008 to June 30, 2021. In the experiments, we want to analyze the cumulative wealth over time using a rolling horizon procedure with daily rebalancing, that is, we use the corresponding data from January 1 of 2008 to February 13 of 2018 to estimate portfolio vector for February 14, 2018. After that, we use the corresponding data from January 2 of 2008 to February 14 of 2018 to estimate portfolio vector for February 15, 2018, and so on. In summary, this process is continued by removing

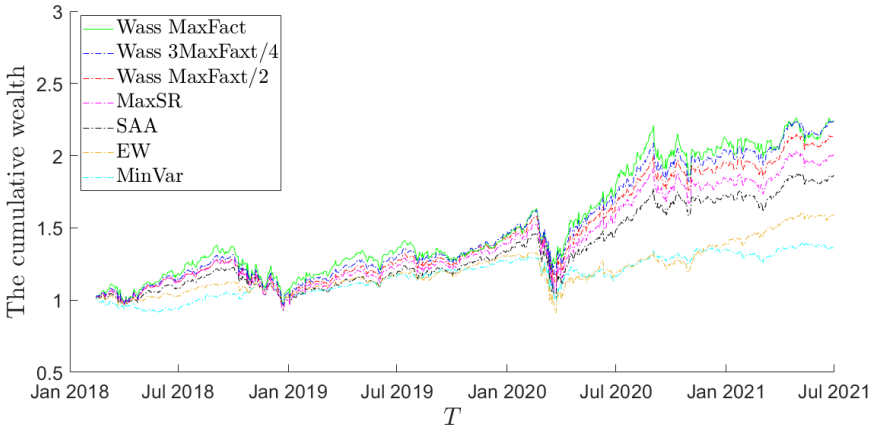


Fig. 9: The cumulative wealth of the trading strategies with $\mu = \min\{0.001, \widehat{\mu}_N^{\max}\}$.

the first return and adding a return to the dataset for the next period until the end of the dataset is reached. The objective is to see how the cumulative wealth evolves in that period of time. In addition, we compare our approach with standard portfolio optimization techniques which are SAA, EW, MinVar, and MaxSR. SAA has already been mentioned before, the other four are explained below.

- Equal Weight (EW): This approach gives equal weight to all assets in the portfolio.
- Minimum variance (MinVar): We use the sample covariance matrix of the returns to find this portfolio.
- Maximum Sharpe ratio (MaxSR): We use sample mean con covariance matrix of the returns to find the portfolio that maximizes the Sharpe ratio.

Now, for a given μ, we consider strategies with $\widehat{\varepsilon}_N^{\max}(\mu)$, called Wasserstein MaxFact, with $\frac{\widehat{\varepsilon}_N^{\max}(\mu)}{2}$ and $\frac{3\widehat{\varepsilon}_N^{\max}(\mu)}{4}$, called Wasserstein MaxFact/2 and Wasserstein 3MaxFact/4 respectively. The reason for including these strategies is to see the influence of ε on the portfolio value. Additionally, we consider $\mu = \min\{0.001, \alpha\widehat{\mu}_N^{\max}\}$ where $\widehat{\mu}_N^{\max} = \widehat{\mu}_N^{\max}(0)$ and $\alpha \leq 1$. Note that μ changes every day because it depends on $\widehat{\mu}_N^{\max}$ which depends on the sample. The rationale for this μ is that the minimum expected return acceptable for the investor is 0.001, but it is possible that this value is very ambitious, which would cause infeasibility. Hence, we correct this with the term $\widehat{\mu}_N^{\max}$. Note that if $\alpha = 1$ we could have that $\widehat{\varepsilon}_N^{\max}(\mu) = 0$.

Figure 9 shows that the strategies based on our approach allows obtaining a portfolio with high cumulative wealth. The wealth of these portfolios exceeds the values obtained with traditional strategies such as SAA, EW, MinVar, and MaxSR. Additionally, the period of time in which we are evaluating the strategies includes the days of the beginning of the COVID-19 pandemic, this phenomenon

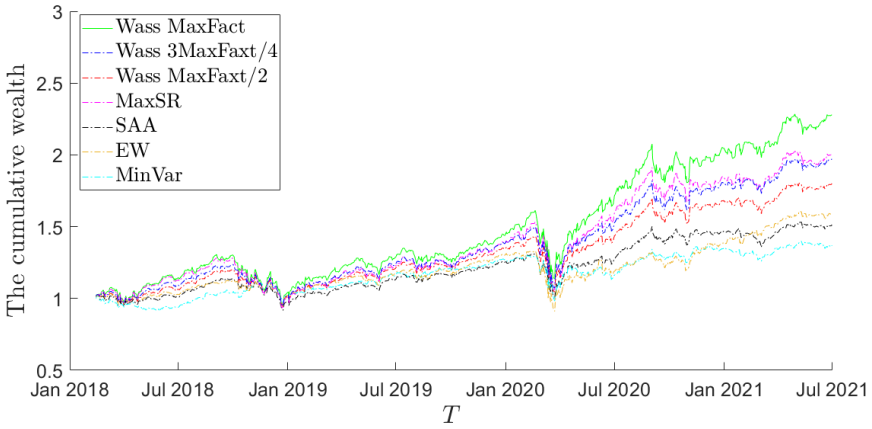


Fig. 10: The cumulative wealth of the trading strategies with $\mu = \min\{0.001, 0.5 \cdot \widehat{\mu}_N^{\max}\}$.

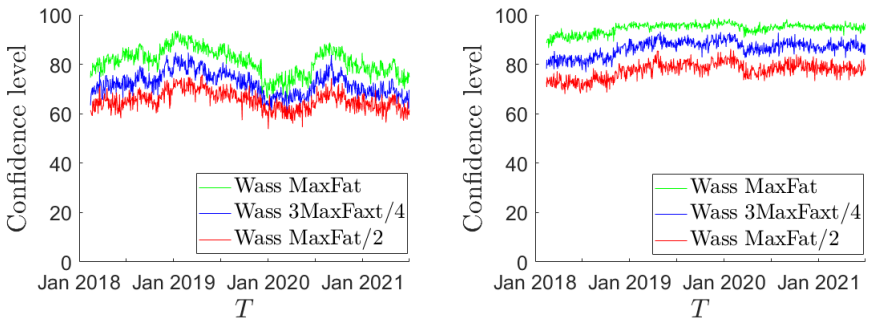
	$\alpha = 1$. The average value of μ was 0.001.						
	Wass MaxFat	Wass 3MaxFact/4	Wass MaxFact/2	MaxSR	EW	MinVar	SAA
Mean	0.0010959	0.0010758	0.0010189	0.0009545	0.00063778	0.00043373	0.00085059
Standard deviation	0.017194	0.015953	0.015645	0.016375	0.01362	0.011263	0.01522
Sharpe Ratio	0.063737	0.067438	0.065125	0.05829	0.046828	0.03851	0.055886
Turnover	0.040606	0.032514	0.03235	0.026879	0.97409	0.0096731	0.040937
Avg. amount of assets	4.4976	6.0106	7.3494	5.3976	23	11.331	8.0188

	$\alpha = 0.5$. The average value of μ was 0.00071337.						
	Wass MaxFat	Wass 3MaxFact/4	Wass MaxFact/2	MaxSR	EW	MinVar	SAA
Mean	0.0011078	0.00090929	0.00078722	0.0009545	0.00063778	0.00043373	0.00056552
Standard deviation	0.016649	0.014813	0.013773	0.016375	0.01362	0.011263	0.012402
Sharpe Ratio	0.066537	0.061384	0.057157	0.05829	0.046828	0.03851	0.045598
Turnover	0.020316	0.022802	0.024214	0.026879	0.97409	0.0096731	0.029465
Avg. amount of assets	7.5882	14.218	14.328	5.3976	23	11.331	10.602

Table 1: Performances of different portfolio strategies.

affected all investment strategies; however, our strategies somehow mitigate this effect on long-term portfolio value. On the other hand, Figure 10 also shows that the highest cumulative wealth is achieved by Wasserstein MaxFact strategy but followed by MaxSR. This situation is due to the influence of α, which is 0.5 in this case. Finally, both figures show that the cumulative wealth of strategies based on Wasserstein strategies increases with respect to ε.

Table 1 shows some out-of-sample indicators of the different strategies. Recall that in our experiment μ depends on the sample and changes on each day. The first thing to note is that the mean of all Wasserstein strategies exceeds the average value of μ. This is remarkable because the SAA approach does not achieve this despite the sample size is considerably large. It is known that if the



(a) $\alpha = 1$ (b) $\alpha = 0.5$

Fig. 11: Confidence level target mean return, $\mu = \min\{0.001, \alpha\widehat{\mu}_N^{\max}\}$.

sample size is large enough, then the SAA strategy gives a portfolio very close to the one obtained by solving (1) if the distribution of returns were known, and therefore this portfolio will satisfy the constraint $\mathbb{E}_{\mathbb{P}}[\langle x, \xi\rangle] \geq \mu$ with very high probability. However, in this case, this is not achieved because not all data in the sample comes from the same distribution, which implies that the average returns obtained with the SAA strategy do not exceed the average value of μ. In contrast, Wasserstein-based strategies manage to overcome this situation.

Another important indicator is the standard deviation. The standard deviations of the Wasserstein approaches are among the largest, but the difference with respect to the SAA approach is not so significant, this becomes evident when we see the Sharpe ratios. We also observe that the highest Sharpe ratio among Wasserstein-based strategies is not always attained at Wasserstein MaxFact, as oppose to the mean. This agrees with what is observed in Figure 6 for simulated data, where it was evidenced that the highest Sharpe ratio is not always obtained in Wasserstein MaxFact.

Finally, regarding the other indicators, we can see that the turnover increases with ε but in all cases is lower that the one obtained in SAA. On the other hand, the number of assets on average that make up the portfolio of the Wasserstein-based strategies decreases with ε and, compared to the other strategies, the situation is different for $\alpha = 1$ (among the smallest) than for $\alpha = 0.5$ (among the highest). This indicator can be useful to identify the most promising companies within the portfolio. In summary, if we combined the information of Table 1 with the fact that Wasserstein approaches produce portfolios that generate more cumulative wealth as time progresses, then Wasserstein approaches are a good option for making investment decisions with real market data.

We aslo evaluate the confidence level of each of the Wasserstein-based strategies. Figure 11 shows that this confidence level is high for all Wasserstein strategies and, as expected increases with ε. Clearly, the hardest is the constraint ($\alpha = 1$), the lower the confidence level. Note that in this case determining exactly whether that confidence level is satisfied out-of-sample is not possible because the true vector of means is not known, nevertheless, as we said before, Table 1 gives us indications that this level is satisfied in our financial market

data because the mean of the Wasserstein strategies always exceeded the average value of μ.

5 Conclusions and future work

In this work we have shown that the Wasserstein distance-based approach (7) has an equivalent finite dimensional formulation, where, in some cases, it is a convex formulation. Furthermore, we established theoretical results that characterize the values of μ and ε for which the Wasserstein approach (7) is valid and feasible. As future work, we want to extend the results presented in this work for a set of functions F and G larger than the set determined by the Lipschitz functions respect to ξ. Additionally, we want to explore other types of Φ functions, for example, consider Φ as a quantile in order to consider Value-at-Risk as risk function.

We implemented our strategy in two applications, a newsvendor problem and a portfolio optimization problem. The experiments with synthetic data show that it is possible to find ε such that, out-of-sample, the constraint is satisfied with high probability. This is most evident for small sample sizes. In both contexts, the proposed bootstrap-based method for choosing this parameter performed well. The experiment with real market data showed that our approach has higher expected returns, higher Sharpe Ratios, and a reduction in turnover compared to the SAA approach.

A Proofs of Lemmas and Theorems

We present proofs of the results presented in this work. Section A.1 present the proofs of Lemmas 1 and 2. Section A.2 explores worst-case expectation problems with expected value constraints and the corresponding dual formulation. Therorem 4 is an important result on its own. Section A.3 presents a distributionally robust estimation of the variance under known mean, a key result to prove Theorem 3. Section A.4 shows the proofs of Theorems 2 and 3 and its corollaries. Finally, Section A.5 contains the proofs of Propositions 1 and 2.

A.1 Proofs of Lemmas 1 and 2

Proof (Lemma 1). Let $\widetilde{\xi}_1, \ldots, \widetilde{\xi}_M$ be another sample of ξ, then let $\widetilde{\mathbb{P}}_M$ be the empirical distribution generated by this sample. This last sample of ξ induces the sample $\widetilde{\zeta}_1^{x,F}, \ldots, \widetilde{\zeta}_M^{x,F}$ of $\zeta^{x,F}$ given by $\widetilde{\zeta}_i^{x,F} := F\left(x, \widetilde{\xi}_i\right)$, so we consider $\widetilde{\mathbb{P}}_M^{x,F}$, the empirical distribution generated by the sample $\left\{\widetilde{\zeta}_i^{x,F}\right\}_{i=1}^{M}$. Because $\widetilde{\mathbb{P}}_M \to \mathbb{P}$ and $\widetilde{\mathbb{P}}_M^{x,F} \to \mathbb{P}^{x,F}$ weakly as M goes to ∞, by Corollary 6.11 en [34] we have that

$$W_p\left(\widehat{\mathbb{P}}_N, \widetilde{\mathbb{P}}_M\right) \overset{M\to\infty}{\longrightarrow} W_p\left(\widehat{\mathbb{P}}_N, \mathbb{P}\right) \quad \text{and} \quad W_p\left(\widehat{\mathbb{P}}_N^x, \widetilde{\mathbb{P}}_M^{x,F}\right) \overset{M\to\infty}{\longrightarrow} W_p\left(\widehat{\mathbb{P}}_N^{x,F}, \mathbb{P}^{x,F}\right). \tag{16}$$

Additionally, for each M we get that

$$
\begin{aligned}
W_p^p\left(\widehat{\mathbb{P}}_N^{x,F}, \widetilde{\mathbb{P}}_M^{x,F}\right) &= \inf\left\{ \sum_{i=1}^{N}\sum_{j=1}^{M} \lambda_{i,j} \left|\widehat{\zeta}_i^{x,F} - \widetilde{\zeta}_j^{x,F}\right|^p \;\middle|\; \begin{array}{l} \sum_{i=1}^{N} \lambda_{i,j} = \frac{1}{M}, \\ \sum_{j=1}^{M} \lambda_{i,j} = \frac{1}{N}, \\ \lambda_{i,j} \geq 0, \\ i = 1, \ldots, N, \\ j = 1, \ldots, M \end{array} \right\} \\
&= \inf\left\{ \sum_{i=1}^{N}\sum_{j=1}^{M} \lambda_{i,j} \left|F\left(x, \widehat{\xi}_i\right) - F\left(x, \widetilde{\xi}_j\right)\right|^p \;\middle|\; \begin{array}{l} \sum_{i=1}^{N} \lambda_{i,j} = \frac{1}{M}, \\ \sum_{j=1}^{M} \lambda_{i,j} = \frac{1}{N}, \\ \lambda_{i,j} \geq 0, \\ i = 1, \ldots, N, \\ j = 1, \ldots, M \end{array} \right\} \\
&\leq \inf\left\{ \sum_{i=1}^{N}\sum_{j=1}^{M} \lambda_{i,j} \gamma_{x,F}^p \left\|\widehat{\xi}_i - \widetilde{\xi}_j\right\|^p \;\middle|\; \begin{array}{l} \sum_{i=1}^{N} \lambda_{i,j} = \frac{1}{M}, \\ \sum_{j=1}^{M} \lambda_{i,j} = \frac{1}{N}, \\ \lambda_{i,j} \geq 0, \\ i = 1, \ldots, N, \\ j = 1, \ldots, M \end{array} \right\} \text{ by Assumption 1} \\
&= \gamma_{x,F}^p W_p^p\left(\widehat{\mathbb{P}}_N, \widetilde{\mathbb{P}}_M\right).
\end{aligned}
$$

Therefore, by (16) we conclude

$$
W_p^p\left(\widehat{\mathbb{P}}_N^{x,F}, \mathbb{P}^{x,F}\right) \leq \gamma_{x,F}^p W_p^p\left(\widehat{\mathbb{P}}_N, \mathbb{P}\right).
$$

Proof (Lemma 2). Let $\varepsilon > 0$ such that $\mathbb{P} \in \mathcal{B}_\varepsilon(\widehat{\mathbb{P}}_N)$ and (7) is feasible, then, by Lemma 1, we have that $\mathbb{P}^{\widehat{x}_N(\varepsilon),G} \in \mathcal{B}_{\varepsilon\gamma_{\widehat{x}_N(\varepsilon),G}}\left(\widehat{\mathbb{P}}_N^{\widehat{x}_N(\varepsilon),G}\right)$. Hence, because $\inf_{\mathbb{Q}\in\mathcal{B}_{\varepsilon\gamma_{\widehat{x}_N(\varepsilon),G}}\left(\widehat{\mathbb{P}}_N^{\widehat{x}_N(\varepsilon),G}\right)} \mathbb{E}_{\mathbb{Q}}[\zeta] \geq \mu$, we obtain $\mathbb{E}_{\mathbb{P}^{\widehat{x}_N(\varepsilon),G}}[\zeta] \geq \mu$. However, note that $\mathbb{E}_{\mathbb{P}^{\widehat{x}_N(\varepsilon),G}}[\zeta] = \mathbb{E}_{\mathbb{P}^{\widehat{x}_N(\varepsilon),G}}[\zeta^{\widehat{x}_N(\varepsilon),G}] = \mathbb{E}_{\mathbb{P}}[G\left(\widehat{x}_N(\varepsilon), \xi\right)]$. Therefore, we conclude that $\mathbb{E}_{\mathbb{P}}\left[G\left(\widehat{x}_N(\varepsilon), \xi\right)\right] \geq \mu$.

A.2 Duality of worst-case expectation problems with expected value constraints

When we refer to worst-case expectation problems with restrictions on the expected value, we are referring to problems of the form

$$
\begin{cases} \sup\limits_{\mathbb{Q}\in\mathcal{B}_\varepsilon(\widehat{\mathbb{P}}_N)} & \mathbb{E}_{\mathbb{Q}}\left[h(\xi)\right] \\ \text{subject to} & \mathbb{E}_{\mathbb{Q}}\left[g_i(\xi)\right] = b_i.\ \forall\, i = 1, \ldots, k. \end{cases} \tag{17}
$$

where $b_i \in \mathbb{R}$ and $g_1, \ldots, g_k$ are integrable functions respect to each measure in $\mathcal{B}_\varepsilon(\widehat{\mathbb{P}}_N)$. This problem is important for the following section and to prove Thereom 5.

Theorem 4. *Assume that the optimal value of the problem (17) is finite and that any of the following conditions are satisfied*

i) The point $(b_1, \ldots, b_k, 1)$ *is a interior point of the set*

$$\left\{ \lambda \left(\int g_1(\xi)\mathbb{Q}(d\xi), \ldots, \int g_k(\xi)\mathbb{Q}(d\xi), 1 \right) \;\middle|\; \lambda > 0,\ \mathbb{Q} \in \mathcal{B}_\varepsilon(\widehat{\mathbb{P}}_N) \right\},$$

ii) The set of optimal distributions of (17) is not empty and bounded.

Then (17) satisfies strong duality, that is, the optimal value of (17) is equal to

$$\inf_{a_1,\ldots,a_k} \left\{ \sum_{i=1}^{k} a_i b_i + \sup_{\mathbb{Q}\in\mathcal{B}_\varepsilon(\widehat{\mathbb{P}}_N)} \int_{\Xi} \left(h(\xi) - \sum_{i=1}^{k} a_i g_i(\xi) \right) \mathbb{Q}(d\xi) \right\}.$$

Proof. The strategy is to characterize (17) as a Linear Conic Problem (LCP) and then to apply the strong duality results from [29]. First, we consider $\overline{\mathcal{P}}(\Xi)$ as the set of non-negative probability measures in the measurable space $(\Xi, \mathcal{E})$ such that $h, g_1, g_2, \ldots, g_k$ are integrable with respect to each measure in $\overline{\mathcal{P}}(\Xi)$, where Ξ is the support of ξ and $\mathcal{E}$ is a σ-algebra that contains all singletons in Ξ, i.e., $\{\xi\} \in \mathcal{E}$, $\forall \xi \in \Xi$. Also, we called X to the linear space (over $\mathbb{R}$) of signed measures generated by $\overline{\mathcal{P}}(\Xi)$. Additionally, we define X' as the linear space of functions $\ell : \Xi \to \mathbb{R}$ generated by $h, g_1, \ldots, g_k$, that is, the functions that can be expressed as a linear combination of $h, g_1, \ldots, g_k$. We define a bilinear form between X and X' given by

$$\begin{aligned} \langle \cdot, \cdot \rangle_X : X' \times X &\longrightarrow \mathbb{R} \\ (\ell, \mathbb{Q}) &\longmapsto \textstyle\int_\Xi \ell(\xi)\mathbb{Q}(d\xi). \end{aligned}$$

Moreover, we consider $Y = \mathbb{R}^{k+1}$ and Y' as the dual space of Y, note that Y' is equal to $\mathbb{R}^{k+1}$. Also, $\langle \cdot, \cdot \rangle_Y$ is the traditional bilinear form between a space and its dual space, in this case, it is the Euclidean inner product in $\mathbb{R}^{k+1}$. Continuing with the characterizations, we establish $K = \{0\}$ where 0 is zero vector of $\mathbb{R}^{k+1}$, G as the convex cone generated by $\mathcal{B}_\varepsilon(\widehat{\mathbb{P}}_N)$. Because $\mathcal{B}_\varepsilon(\widehat{\mathbb{P}}_N)$ is convex, we have $G = \cup_{\lambda>0}\lambda\mathcal{B}_\varepsilon(\widehat{\mathbb{P}}_N)$ (see [4]). In addition, we called $\mathbf{b}$ as a vector in $\mathbb{R}^{k+1}$ such that $\mathbf{b}_i = b_i$ for each $i = 1, \ldots, k$ and $\mathbf{b}_{k+1} = 1$, y we define $\psi_i = g_i$ for $i = 1, \ldots, k$ and $\psi_{k+1}(x) = 1$ for all $x \in \Xi$. Finally, we define the linear application $A : X \to Y$ given by $A(\mathbb{Q}) = (\langle \psi_1, \mathbb{Q} \rangle_X, \ldots, \langle \psi_k, \mathbb{Q} \rangle_X, \langle \psi_{k+1}, \mathbb{Q} \rangle_X)$. Taking into account these characterizations, we have the following LCP

$$\begin{cases} \max\limits_{\mathbb{Q}\in G} & \langle h, \mathbb{Q} \rangle_X \\ \text{subject to} & A(\mathbb{Q}) - \mathbf{b} \in K \end{cases} = \begin{cases} \sup\limits_{\mathbb{Q}\in G} & \langle h, \mathbb{Q} \rangle_X \\ \text{subject to} & \langle g_i, \mathbb{Q} \rangle_X = b_i \ \forall i \le k \\ & \langle 1, \mathbb{Q} \rangle_X = 1. \end{cases} \tag{18}$$

However, because $G = \cup_{\lambda>0}\lambda\mathcal{B}_\varepsilon(\widehat{\mathbb{P}}_N)$ and $\langle 1, \mathbb{Q}\rangle_X = 1$ for all $\mathbb{Q} \in \mathcal{B}_\varepsilon(\widehat{\mathbb{P}}_N)$, we have (18) is equal to the following

$$\begin{cases} \sup\limits_{\mathbb{Q}\in\mathcal{B}_\varepsilon(\widehat{\mathbb{P}}_N)} & \lambda\langle h, \mathbb{Q}\rangle_X \\ \text{subject to} & \lambda\langle g_i, \mathbb{Q}\rangle_X = b_i \ \forall i \le k \\ & \lambda\langle 1, \mathbb{Q}\rangle_X = 1 \\ & \lambda > 0. \end{cases} = \begin{cases} \sup\limits_{\mathbb{Q}\in\mathcal{B}_\varepsilon(\widehat{\mathbb{P}}_N)} & \langle h, \mathbb{Q}\rangle_X \\ \text{subject to} & \langle g_i, \mathbb{Q}\rangle_X = b_i \ \forall i \le k. \end{cases}$$

Note that the problem to the right of the above equality is (17), so we can conclude that (17) is equal to (18). Therefore, by conditions i) and ii), and Theorem 2.8 in [29], we have that (18) satisfies strong duality and its dual formulation is

$$\begin{cases} \min\limits_{a\in -K^*} & \langle a, \mathbf{b}\rangle_Y \\ \text{subject to} & A^*a - h \in G^*, \end{cases} \tag{19}$$

where $A^* : Y^* \to X^*$ is the adjoint map of A and G^* and K^* are the polar cones of G and K respectively. Because $\mathcal{B}_\varepsilon(\widehat{\mathbb{P}}_N)$ is convex, G^* can be expressed as

$$G^* = \left\{ f \in X' \mid \lambda\langle f, \mathbb{Q}\rangle_X \ge 0,\ \mathbb{Q} \in \mathcal{B}_\varepsilon(\widehat{\mathbb{P}}_N),\ \lambda > 0 \right\} = \left\{ f \in X' \mid \langle f, \mathbb{Q}\rangle_X \ge 0,\ \mathbb{Q} \in \mathcal{B}_\varepsilon(\widehat{\mathbb{P}}_N) \right\}.$$

Moreover, note that $-K^* = \mathbb{R}^{k+1}$. Therefore, we have that (19) can be expressed as

$$\begin{aligned} & \begin{cases} \inf\limits_{a\in\mathbb{R}^{k+1}} & \langle a, \mathbf{b}\rangle_Y \\ \text{subject to} & \sum\limits_{i=1}^k a_i \langle g_i, \mathbb{Q}\rangle_X + a_{k+1} \ge \langle h, \mathbb{Q}\rangle_X \ \forall \mathbb{Q} \in \mathcal{B}_\varepsilon(\widehat{\mathbb{P}}_N). \end{cases} \\ = & \begin{cases} \inf\limits_{a\in\mathbb{R}^{k+1}} & \sum\limits_{i=1}^k a_i b_i + a_{k+1} \\ \text{subject to} & \sum\limits_{i=1}^k a_i \int_\Xi g_i(\xi)\mathbb{Q}(d\xi) + a_{k+1} \ge \int_\Xi h(\xi)\mathbb{Q}(d\xi) \ \forall \mathbb{Q} \in \mathcal{B}_\varepsilon(\widehat{\mathbb{P}}_N). \end{cases} \\ = & \begin{cases} \inf\limits_{a\in\mathbb{R}^{k+1}} & \sum\limits_{i=1}^k a_i b_i + a_{k+1} \\ \text{subject to} & \int_\Xi \left(h(\xi)\mathbb{Q}(d\xi) - \sum\limits_{i=1}^k a_i g_i(\xi) \right) \mathbb{Q}(d\xi) \le a_{k+1} \ \forall \mathbb{Q} \in \mathcal{B}_\varepsilon(\widehat{\mathbb{P}}_N). \end{cases} \\ = & \inf_{a_1,\dots,a_k} \left\{ \sum_{i=1}^k a_i b_i + \sup_{\mathbb{Q}\in\mathcal{B}_\varepsilon(\widehat{\mathbb{P}}_N)} \int_\Xi \left(h(\xi) - \sum_{i=1}^k a_i g_i(\xi) \right) \mathbb{Q}(d\xi) \right\}. \end{aligned}$$

Using Theorem 1 we can further reformulate (17) as a semi-infinite optimization problem.

Corollary 3. *Suppose that the function $F_a(\xi) := h(\xi) - \sum_{i=1}^k a_i(g_i(\xi) - b_i)$ satisfies the hypotheses of the Theorem 1 for all $a \in \mathbb{R}^k$, and satisfies any of the conditions i) and ii) of Theorem 4, then the problem (17) can be rewritten as*

$$\begin{cases} \inf\limits_{a_1,\ldots,a_k,\lambda} & \lambda\varepsilon + \dfrac{1}{N}\displaystyle\sum_{i=1}^{N} s_i \\ \textit{subject to} & \sup\limits_{\xi\in\Xi}\left(h(\xi) - \displaystyle\sum_{i=1}^{k} a_i(g_i(\xi)-b_i) - \lambda d^p(\xi,\widehat{\xi}_i)\right) \leq s_i \;\forall\, i=1,\ldots,N, \\ & \lambda \geq 0. \end{cases} \tag{20}$$

A.3 Distributionally robust estimation of the variance of a random variable with known mean

In this part, we will formulate a robust distributional version of the problem of estimating the variance of a random variable with known mean, and we will demonstrate that the obtained optimization problem admits an explicit solution.

Let ζ be a random variable with unknown distribution $\mathbb{P}$ with support $\Xi \subseteq \mathbb{R}$, we assume that the expected value of ζ is known, specifically, we assume that $\mathbb{E}_{\mathbb{P}}[\zeta] = \eta$. Also, we consider a sample $\widehat{\zeta}_1, \ldots, \widehat{\zeta}_N$ of ζ. Let $\widehat{\mathbb{P}}_N := \frac{1}{N}\sum_{i=1}^{N} \delta_{\widehat{\zeta}_i}$ be the empirical distribution, and denote by $\bar{\zeta} := \mathbb{E}_{\widehat{\mathbb{P}}_N}[\zeta] = \frac{1}{N}\sum_{i=1}^{N}\widehat{\zeta}_i$ and $\widehat{\sigma}^2 := \mathbb{E}_{\widehat{\mathbb{P}}_N}[(\zeta-\eta)^2] = \frac{1}{N}\sum_{i=1}^{N}(\widehat{\zeta}_i - \eta)^2$, the empirical mean and variance respectively. Let $\mathcal{B}_\varepsilon(\widehat{\mathbb{P}}_N)$ be the 2-Wasserstein ball with center in $\widehat{\mathbb{P}}_N$ and radius ε. We call the following problem distributionally robust estimate of the variance of ζ:

$$\begin{cases} \sup\limits_{\mathbb{Q}\in\mathcal{B}_\varepsilon(\widehat{\mathbb{P}}_N)} & \mathbb{E}_{\mathbb{Q}}\left[(\zeta-\eta)^2\right] \\ \text{subject to} & \mathbb{E}_{\mathbb{Q}}\left[\zeta\right] = \eta. \end{cases} \tag{21}$$

However, for some values of ε, this problem may be not feasible as the following result shows.

Proposition 3. *If* $\varepsilon < \left|\eta - \bar{\zeta}\right|$ *then*

$$\mathcal{B}_\varepsilon(\widehat{\mathbb{P}}_N) \cap \{\mathbb{Q} \in \mathcal{P}(\mathbb{R}) \mid \mathbb{E}_{\mathbb{Q}}[\zeta] = \eta\} = \emptyset.$$

Proof. Let $\mathbb{Q} \in \mathcal{B}_\varepsilon(\widehat{\mathbb{P}}_N)$, we must show that $\mathbb{E}_{\mathbb{Q}}[\zeta] \neq \eta$. Indeed, by Observation 6.6 in [34] we know that if $p \leq q$ then $W_p \leq W_q$. In particular, we have that $W_1 \leq W_2$ and this implies that

$$W_1\left(\mathbb{Q}, \widehat{\mathbb{P}}_N\right) \leq W_2\left(\mathbb{Q}, \widehat{\mathbb{P}}_N\right) \leq \varepsilon < \left|\eta - \bar{\zeta}\right|. \tag{22}$$

Therefore, defining $\mathcal{S}(\mathbb{Q}, \widehat{\mathbb{P}}_N)$ as the set of couplings between $\mathbb{Q}$ and $\widehat{\mathbb{P}}_N$, there exists $\mathit{\Pi} \in \mathcal{S}(\mathbb{Q}, \widehat{\mathbb{P}}_N)$ such that

$$\int_{\Xi\times\Xi} |\zeta - \delta| \mathit{\Pi}(d\xi, d\zeta) < \left|\eta - \bar{\zeta}\right|.$$

We also have

$$\int_{\Xi\times\Xi}\zeta\Pi(d\zeta,d\delta)=\int_{\Xi}\zeta\mathbb{Q}(d\zeta)=\mathbb{E}_{\mathbb{Q}}[\zeta]\quad\text{ and }\quad\int_{\Xi\times\Xi}\delta\Pi(d\zeta,d\delta)=\int_{\Xi}\delta\widehat{\mathbb{P}}_N(d\delta)=\mathbb{E}_{\widehat{\mathbb{P}}_N}[\delta].$$

Then

$$\left|\mathbb{E}_{\mathbb{Q}}[\zeta]-\bar{\zeta}\right|=\left|\int_{\Xi\times\Xi}(\zeta-\delta)\Pi(d\xi,d\zeta)\right|\leq\int_{\Xi\times\Xi}|\zeta-\delta|\Pi(d\xi,d\zeta).$$

In consequence, we obtain

$$\left|\mathbb{E}_{\mathbb{Q}}[\zeta]-\bar{\zeta}\right|<\left|\eta-\bar{\zeta}\right|.$$

From the above and the inverse triangular inequality follows

$$|\eta-\mathbb{E}_{\mathbb{Q}}[\zeta]|=\left|\eta-\bar{\zeta}-\left(\mathbb{E}_{\mathbb{Q}}[\zeta]-\bar{\zeta}\right)\right|\geq\left|\left|\eta-\bar{\zeta}\right|-\left|\mathbb{E}_{\mathbb{Q}}[\zeta]-\bar{\zeta}\right|\right|>0.$$

This allows us to conclude that $\mathbb{E}_{\mathbb{Q}}[\zeta]\neq\eta$.

The following theorem establishes an explicit expression for the optimal value of the optimization problem to the right of (21).

Theorem 5. *Let $\varepsilon>0$ with $\varepsilon^2\geq(\bar{\zeta}-\eta)^2$, and $\Xi=\mathbb{R}$. Then, the optimal value of* (21) *is equal to*

$$\left(\sqrt{\widehat{\sigma}^2-\left(\bar{\zeta}-\eta\right)^2}+\sqrt{\varepsilon^2-\left(\bar{\zeta}-\eta\right)^2}\right)^2.$$

Proof. By Theorem 4 we have that (21) satisfies strong duality and its optimal value is equal to

$$\inf_{\beta}\sup_{\mathbb{Q}\in\mathcal{B}_\varepsilon(\widehat{\mathbb{P}}_N)}\mathbb{E}_{\mathbb{Q}}\left[(\zeta-\eta)^2-\beta\zeta+\beta\eta\right].\tag{23}$$

Note that $\zeta\mapsto(\zeta-\eta)^2-\beta\zeta+\beta\eta$ satisfies the hypotheses of Corollary 3, therefore this last formulation is equivalent to the semi-infinite optimization program

$$\begin{cases}\inf\limits_{\beta,\,\lambda,\,s_i} & \lambda\varepsilon^2+\dfrac{1}{N}\displaystyle\sum_{i=1}^{N}s_i\\ \text{subject to}\ \sup\limits_{\zeta\in\Xi} & \left((\zeta-\eta)^2-\beta\zeta+\beta\eta-\lambda\left|\zeta-\widehat{\zeta}_i\right|^2\right)\leq s_i\ \forall i=1,\ldots,N\\ & \lambda\geq 0.\end{cases}\tag{24}$$

If $\lambda<1$, then λ is not a optimal value of (24) because, in this case, the set

$$\left\{(\zeta-\eta)^2-\beta\zeta+\beta\eta-\lambda\left|\zeta-\widehat{\zeta}_i\right|^2\ |\ \zeta\in\Xi\right\}$$

is not bounded. On the other hand, if $\lambda\geq 1$, then

$$\sup_{\zeta\in\mathbb{R}}\left((\zeta-\eta)^2-\beta\zeta+\beta\eta-\lambda\left|\zeta-\widehat{\zeta}_i\right|^2\right)$$

can be calculated explicitly because the function is a concave quadratic polynomial. The unique maximum is attained at $\widehat{\varphi}_i = \frac{2\eta+\beta-2\lambda\widehat{\zeta}_i}{2(1-\lambda)}$ and (24) is equivalent to

$$\begin{aligned}
&\begin{cases} \inf\limits_{\beta,\,\lambda,\,s_i} & \lambda\varepsilon^2 + \dfrac{1}{N}\displaystyle\sum_{i=1}^{N} s_i \\ \text{subject to} & \dfrac{\beta^2}{4(\lambda-1)} + \dfrac{\lambda}{\lambda-1}\left(\beta(\eta-\widehat{\zeta}_i) + (\eta-\widehat{\zeta}_i)^2\right) \le s_i \;\forall i = 1,\ldots,N \\ & \lambda \ge 1. \end{cases} \\
=&\begin{cases} \inf\limits_{\lambda,\,\beta} & \lambda\varepsilon^2 + \dfrac{\beta^2}{4(\lambda-1)} + \dfrac{\lambda}{\lambda-1}\left(\dfrac{\beta}{N}\displaystyle\sum_{i=1}^{N}(\eta-\widehat{\zeta}_i) + \dfrac{1}{N}\sum_{i=1}^{N}(\eta-\widehat{\zeta}_i)^2\right) \\ \text{subject to} & \lambda \ge 1. \end{cases} \\
=&\begin{cases} \inf\limits_{\lambda,\,\beta} & \lambda\varepsilon^2 + \dfrac{\beta^2}{4(\lambda-1)} + \dfrac{\lambda}{\lambda-1}\left(\beta(\eta-\bar{\zeta}) + \widehat{\sigma}^2\right) \\ \text{subject to} & \lambda \ge 1. \end{cases}
\end{aligned} \tag{25}$$

This previous problem can be simplified by analyzing the objective function with respect to λ. For a fixed $\beta \in \mathbb{R}$, first note that the function goes to infinity when $\lambda \to 1^+$ or $\lambda \to \infty$. Now, its second derivative is given by

$$\frac{\beta^2 + 4\beta(\eta-\bar{\zeta}) + 4\widehat{\sigma}^2}{2(\lambda-1)^3}.$$

Since $\lambda \ge 1$, the sign of the last expression is determined by the sign of its numerator, which, in terms of β, is a polynomial with discriminant given by $\left(\eta-\bar{\zeta}\right)^2 - \widehat{\sigma}^2$. As a consequence of Cauchy-Schwarz inequality this discriminant is always negative which implies that the polynomial is always positive. Therefore, the objective function in (25) is convex and has a unique minimum value in the region $\lambda \ge 1$. This minimum is reached at $\lambda^* = 1 + \frac{1}{\varepsilon}\sqrt{\frac{\beta^2}{4} + \beta(\eta-\bar{\zeta}) + \widehat{\sigma}^2}$ and (25) can be rewritten as

$$\inf_{\beta\in\mathbb{R}} \left(\varepsilon^2 + \beta(\eta-\bar{\zeta}) + \widehat{\sigma}^2 + \varepsilon\sqrt{\beta^2 + 4\beta(\eta-\bar{\zeta}) + 4\widehat{\sigma}^2}\right). \tag{26}$$

Since the objective function differentiable for all $\beta \in \mathbb{R}$, after some calculations we obtain that the infimum is attained at $\beta^* = 2(\bar{\zeta}-\eta) + 2(\bar{\zeta}-\eta)\sqrt{\dfrac{\left(\widehat{\sigma}^2 - (\bar{\zeta}-\eta)^2\right)}{\varepsilon^2 - (\bar{\zeta}-\eta)^2}}$ and the optimal value of (21) is

$$\left(\sqrt{\widehat{\sigma}^2 - \left(\bar{\zeta}-\eta\right)^2} + \sqrt{\varepsilon^2 - \left(\bar{\zeta}-\eta\right)^2}\right)^2.$$

A.4 Proofs of Theorems 2 and 3 and Corollaries

Before proceeding with the proof of Theorems 2 and 3, we need the following lemma which allows us to express the feasible set of problem (7) in terms of finite dimensional variables.

Lemma 3. *Assume the same setting as in the previous section.*

1. *1-Wasserstein distance. If* $\Xi = [A, B]$ *then*

$$\inf_{\mathbb{Q}\in\mathcal{B}_\varepsilon(\widehat{\mathbb{P}}_N)} \mathbb{E}_{\mathbb{Q}}[\zeta] = \max\left\{\bar{\zeta} - \varepsilon, A\right\} \quad \textit{and} \quad \sup_{\mathbb{Q}\in\mathcal{B}_\varepsilon(\widehat{\mathbb{P}}_N)} \mathbb{E}_{\mathbb{Q}}[\zeta] = \min\{\bar{\zeta} + \varepsilon, B\}.$$

2. *2-Wasserstein distance. Then*

$$\inf_{\mathbb{Q}\in\mathcal{B}_\varepsilon(\widehat{\mathbb{P}}_N)} \mathbb{E}_{\mathbb{Q}}[\zeta] = \bar{\zeta} - \varepsilon \;\; \textit{if } \Xi = (-\infty, B]$$

and

$$\sup_{\mathbb{Q}\in\mathcal{B}_\varepsilon(\widehat{\mathbb{P}}_N)} \mathbb{E}_{\mathbb{Q}}[\zeta] = \bar{\zeta} + \varepsilon \;\; \textit{if } \Xi = [A, \infty).$$

Proof. We only show the first equality of each case since the second is analogous. For p-Wasserstein distances, by Theorem 1 we have that

$$\inf_{\mathbb{Q}\in\mathcal{B}_\varepsilon(\widehat{\mathbb{P}}_N)} \mathbb{E}_{\mathbb{Q}}[\zeta] = -\begin{cases} \inf\limits_{\lambda\geq 0} & \lambda\varepsilon^2 + \frac{1}{N}\sum_{i=1}^N s_i \\ \text{subject to } \sup_{\zeta\in\Xi}\left(-\zeta - \lambda\left|\zeta - \widehat{\zeta}_i\right|^p\right) \leq s_i \;\forall\, i = 1, \ldots, N, \end{cases} \tag{27}$$

In case 1 we have that

$$\begin{aligned} \sup_{\zeta\in[A,B]}\left(-\zeta - \lambda\left|\zeta - \widehat{\zeta}_i\right|\right) &= \sup_{\zeta\in[A,B]}\left(-\zeta - \max_{|z_i|\leq\lambda} z_i\left(\zeta - \widehat{\zeta}_i\right)\right) \\ &= \min_{|z_i|\leq\lambda}\sup_{\zeta\in[A,B]}\left(-\zeta - z_i\left(\zeta - \widehat{\zeta}_i\right)\right) \qquad (28) \\ &= \min_{|z_i|\leq\lambda}\max\left\{(\widehat{\zeta}_i - A)z_i - A, (\widehat{\zeta}_i - B)z_i - B\right\} \\ &= \begin{cases} -\widehat{\zeta}_i & \text{if } \lambda \geq 1, \\ -(\widehat{\zeta}_i - A)\lambda - A & \text{if } \lambda < 1. \end{cases} \end{aligned}$$

Equality (28) is guaranteed by Von Neumann's minimax theorem (see [2]). Therefore, we obtain

$$\begin{aligned} \inf_{\mathbb{Q}\in\mathcal{B}_\varepsilon(\widehat{\mathbb{P}}_N)} \mathbb{E}_{\mathbb{Q}}[\zeta] &= -\min\left\{\inf_{\lambda\geq 1}\left(\lambda\varepsilon - \frac{1}{N}\sum_{i=1}^N \widehat{\zeta}_i\right), \inf_{\lambda<1}\left(\lambda\varepsilon - \frac{1}{N}\sum_{i=1}^N(-(\widehat{\zeta}_i - A)\lambda - A)\right)\right\} \\ &= \max\left\{\bar{\zeta} - \varepsilon, A\right\}. \end{aligned}$$

In case 2 we have that

$$\sup_{\zeta\in(-\infty,B]}\left(-\zeta - \lambda\left|\zeta - \widehat{\zeta}_i\right|^2\right) = -\widehat{\zeta}_i + \frac{1}{4\lambda}.$$

Therefore,

$$\inf_{\mathbb{Q}\in\mathcal{B}_\varepsilon(\widehat{\mathbb{P}}_N)} \mathbb{E}_{\mathbb{Q}}[\zeta] = -\inf_{\lambda\geq 0}\left(\lambda\varepsilon - \bar{\zeta} + \frac{1}{4\lambda}\right) = \bar{\zeta} - \varepsilon.$$

Proof (Theorem 2). Let $\mathbb{X}$ the feasible set of (7), then, by Lemma 3-1, we have that

$$\mathbb{X} = \left\{ x \in \mathbb{R}^m \;\middle|\; \max\left\{ \frac{1}{N}\sum_{i=1}^{N} G\left(x, \widehat{\xi}_i\right) - \varepsilon\gamma_{x,G}, A_G(x) \right\} \geq \mu,\ x \in \mathcal{X} \right\}.$$

Therefore, (7) is equivalent to

$$\widehat{J}_N := \min_{x\in\mathbb{X}} \sup_{\mathbb{Q}\in\mathcal{B}_{\varepsilon\gamma_{x,F}}(\widehat{\mathbb{P}}_N^{x,F})} \mathbb{E}_{\mathbb{Q}}\left[\zeta\right],$$

but, again, by Lemma 3-1, we have that

$$\widehat{J}_N = \min_{x\in\mathbb{X}} \min\left\{ \frac{1}{N}\sum_{i=1}^{N} F\left(x, \widehat{\xi}_i\right) + \varepsilon\gamma_{x,F}, B_F(x) \right\},$$

which is equivalent to (8). Analogously, to prove (9), we use Lemma 3-2.

Proof (Theorem 3). Let $\mathbb{X}$ the feasible set of (7), then, by Lemma 3-1, we have that

$$\mathbb{X} = \left\{ x \in \mathbb{R}^m \;\middle|\; \frac{1}{N}\sum_{i=1}^{N} G\left(x, \widehat{\xi}_i\right) - \varepsilon\gamma_{x,G} \geq \mu,\ x \in \mathcal{X} \right\}.$$

Therefore, by Proposition 3 (7) is equivalent to

$$\begin{aligned}
\widehat{J}_N &:= \min_{x\in\mathbb{X}} \sup_{\mathbb{Q}\in\mathcal{B}_{\varepsilon\gamma_{x,F}}(\widehat{\mathbb{P}}_N^{x})} \mathrm{Var}_{\mathbb{Q}}\left[\zeta\right] \\
&= \min_{x\in\mathbb{X}} \sup_{\left(\eta - \frac{1}{N}\sum_{i=1}^{N}\widehat{\zeta}_i^{x,F}\right)^2 \leq \varepsilon^2\gamma_{x,F}^2} \begin{cases} \sup_{\mathbb{Q}\in\mathcal{B}_{\varepsilon\gamma_{x,F}}(\widehat{\mathbb{P}}_N^{x})} \mathrm{Var}_{\mathbb{Q}}\left[\zeta\right] \\ \text{subject to} \quad \mathbb{E}_{\mathbb{Q}}\left[\zeta\right] = \eta. \end{cases}
\end{aligned} \tag{29}$$

By Theorem 5, the maximization problem in (29) which has a variance as its objective function, can be rewritten. Then (29) is equivalent to

$$\widehat{J}_N = \min_{x\in\mathbb{X}} \begin{cases} \sup & \left(\sqrt{\frac{1}{N}\sum_{i=1}^{N} F\left(x,\widehat{\xi}_i\right)^2 - \frac{1}{N^2}\left(\sum_{i=1}^{N} F\left(x,\widehat{\xi}_i\right)\right)^2} + \sqrt{\varepsilon^2\gamma_{x,F}^2 - \left(\eta - \frac{1}{N}\sum_{i=1}^{N} F\left(x,\widehat{\xi}_i\right)\right)^2} \right)^2 \\ \text{subject to} & \left(\eta - \frac{1}{N}\sum_{i=1}^{N}\widehat{\zeta}_i^{x,F}\right)^2 \leq \varepsilon^2\gamma_{x,F}^2 \end{cases} \tag{30}$$

But, note that the internal maximization problem of (30) can be explicitly solved. Actually, this problem reaches its optimal value in $\eta^* = \frac{1}{N}\sum_{i=1}^{N}\widehat{\zeta}_i^{x,F}$. Therefore, (30) can be rewritten as

$$\widehat{J}_N(\varepsilon) = \underset{x\in\mathbb{X}}{\text{minimize}} \left(\sqrt{\frac{1}{N}\sum_{i=1}^{N} F\left(x,\widehat{\xi}_i\right)^2 - \frac{1}{N^2}\left(\sum_{i=1}^{N} F\left(x,\widehat{\xi}_i\right)\right)^2} + \varepsilon\gamma_{x,F} \right)^2$$

$$= \begin{cases} \underset{x\in\mathbb{R}^m}{\text{minimize}} & \left(\sqrt{\frac{1}{N}\sum_{i=1}^{N} F\left(x,\widehat{\xi}_i\right)^2 - \frac{1}{N^2}\left(\sum_{i=1}^{N} F\left(x,\widehat{\xi}_i\right)\right)^2} + \varepsilon\gamma_{x,F} \right)^2 \\ \text{subject to} & \frac{1}{N}\sum_{i=1}^{N} G\left(x,\widehat{\xi}_i\right) - \varepsilon\gamma_{x,G} \geq \mu, \\ & x \in \mathcal{X}. \end{cases}$$

A.5 Proofs of Propositions 1 and 2.

Proof (Proposition 1). According to Theorem 2-1, we have that (13) is

$$\begin{aligned} & - \begin{cases} \underset{x}{\text{minimize}} & \min\left\{ cx - \frac{p}{N}\sum_{i=1}^{N} \min\{\widehat{\xi}_i, x\} + p\varepsilon, cx \right\} \\ \text{subject to} & \max\left\{ \frac{1}{N}\sum_{i}^{N} \min\left\{x - \widehat{\xi}_i, 0\right\} - \varepsilon, -\infty \right\} \geq -\alpha \\ & x \geq 0. \end{cases} \\ = & \begin{cases} \underset{x}{\text{maximize}} & \max\left\{ \frac{p}{N}\sum_{i=1}^{N} \min\{\widehat{\xi}_i, x\} - p\varepsilon, 0 \right\} - cx \\ \text{subject to} & \min\left\{ -\frac{1}{N}\sum_{i}^{N} \min\left\{x - \widehat{\xi}_i, 0\right\} + \varepsilon, \infty \right\} \leq \alpha \\ & x \geq 0. \end{cases} \\ = & \begin{cases} \underset{x}{\text{maximize}} & \frac{p}{N}\sum_{i=1}^{N} \min\{\widehat{\xi}_i, x\} - p\varepsilon - cx \\ \text{subject to} & -\frac{1}{N}\sum_{i}^{N} \min\left\{x - \widehat{\xi}_i, 0\right\} + \varepsilon \leq \alpha \\ & \frac{p}{N}\sum_{i=1}^{N} \min\{\widehat{\xi}_i, x\} - p\varepsilon \geq 0, \\ & x \geq 0. \end{cases} \\ = & \begin{cases} \underset{x}{\text{maximize}} & \frac{p}{N}\sum_{i=1}^{N} \min\{\widehat{\xi}_i, x\} - p\varepsilon - cx \\ \text{subject to} & -\frac{1}{N}\sum_{i}^{N} \min\left\{x, \widehat{\xi}_i\right\} + \frac{1}{N}\sum_{i}^{N} \widehat{\xi}_i + \varepsilon \leq \alpha \\ & \frac{p}{N}\sum_{i=1}^{N} \min\{\widehat{\xi}_i, x\} - p\varepsilon \geq 0, \\ & x \geq 0. \end{cases} \end{aligned}$$

$$= \begin{cases} \underset{x,z\in\mathbb{R}^N}{\text{maximize}} & \frac{p}{N}\sum_{i=1}^{N} z_i - p\varepsilon - cx \\ \text{subject to} & -\frac{1}{N}\sum_{i}^{N} x_i + \frac{1}{N}\sum_{i}^{N}\widehat{\xi}_i + \varepsilon \leq \alpha \\ & \frac{p}{N}\sum_{i=1}^{N} x_i - p\varepsilon \geq 0, \\ & z_i \leq \widehat{\xi}_i \quad \forall i = 1,\ldots,N, \\ & z_i \leq x \quad \forall i = 1,\ldots,N, \\ & x \geq 0. \end{cases}$$

Proof (Proposition 2). According to Theorem 3, we have that (7) is

$$\begin{cases} \underset{x\in\mathbb{R}^m}{\text{minimizar}} & \left(\sqrt{\frac{1}{N}\sum_{i=1}^{N}\langle x,\widehat{\xi}_i\rangle^2 - \frac{1}{N^2}\left(\sum_{i=1}^{N}\langle x,\widehat{\xi}_i\rangle\right)^2} + \varepsilon\|x\|\right)^2 \\ \text{subject to} & \frac{1}{N}\sum_{i=1}^{N}\langle x,\widehat{\xi}_i\rangle - \varepsilon\|x\| \geq \mu, \\ & \sum_{i=1}^{m} x_i = 1. \\ & x \in \mathcal{X} \end{cases} \tag{31}$$

Note that

$$\frac{1}{N}\sum_{i=1}^{N}\langle x,\widehat{\xi}_i\rangle^2 - \frac{1}{N^2}\left(\sum_{i=1}^{N}\langle x,\widehat{\xi}_i\rangle\right)^2 = x^T E x \quad \text{and} \quad \frac{1}{N}\sum_{i=1}^{N}\langle x,\widehat{\xi}_i\rangle = Lx.$$

Therefore, (15) follows.

References

1. Akhtar, Z., Bedi, A.S., Rajawat, K.: Conservative stochastic optimization with expectation constraints. IEEE Transactions on Signal Processing **69**, 3190–3205 (2021)
2. Bertsekas, D.: Convex Optimization Theory. Athena Scientific (2009)
3. Blanchet, J., Murthy, K.: Quantifying distributional model risk via optimal transport. Mathematics of Operations Research **44**(2), 565–600 (2019)
4. Boyd, S., Vandenberghe, L.: Convex optimization. Cambridge University Press (2009)
5. Chen, F., Federgruen, A.: Mean - variance analysis of basic inventory models. Technical manuscript, Columbia University (2000)
6. Choi, T.M., Li, D., Yan, H.: Mean–variance analysis for the newsvendor problem. IEEE Transactions on Systems, Man, and Cybernetics - Part A: Systems and Humans **38**(5), 1169–1180 (2008)

7. Chopra, V., W.T., Z.: The effect of errors in means, variances and covariances on optimal portfolio choice. Journal of Portfolio Management **19**(2), 6–11 (1993)
8. Delage, E., Ye, Y.: Distributionally robust optimizatión under moment uncertainty with application to data-driven problems. Oparations Research **58**(3), 595–612 (2010)
9. Dentcheva, D., Ruszczyńsk, A.: Optimization with stochastic dominance constraints. SIAM J. Opti **14**(2), 548–566 (2003)
10. Efron, B., Tibshirani, R.: An Introduction to the Bootstrap. Chapman & Hall/CRC Monographs on Statistics & Applied Probability, Taylor & Francis (1994)
11. El Ghaoui, L., Oks, M., Oustry, F.A.: Worst-case value-at-risk and robust portfolio optimization: a conic programming approach. Operations Research **51**(4), 543–553 (2003)
12. Esfahani, P., Kuhn, D.: Data-driven distributionally robust optimization using the wasserstein metric: Performance guarantees and tractable reformulations. Mathematical Programming **171**, 115–166 (2018)
13. Jiang, R., Guan, Y.: Data-driven chance costrained stochastic program. Mathematical Programming **158**, 291–327 (2016)
14. Lagoa, C.M., Barmish, R.B.: Distributionally robust Monte Carlo simulation. In Proceedings of the International Federation of Automatic Control World Congress pp. 1–12 (2002)
15. Lan, G., Zhou, Z.: Algorithms for stochastic optimization with function or expectation constraints. Comput Optim Appl **76**, 461–498 (2020)
16. Li, X., Xu, Q., Chen, C.: Designing a hierarchical decentralized system for distributing large-scale, cross-sector, and multipollutant control accountabilities. IEEE Systems Journal **11**(4), 2774–2783 (2017)
17. Lotf, S., Salahi, M., Mehrdoust, F.: Adjusted robust mean-value-at-risk model: less conservative robust portfolios. Optim Eng **18**(2), 467–497 (2017)
18. Lotf, S., Zenios, S.: Robust VaR and CVaR optimization under joint ambiguity in distributions, means, and covariances. European Journal of Operational Research **269**(2), 556–576 (2018)
19. Luo, F., Mehrotra, S.: Decomposition algorithm for distributionally robust optimization using Wasserstein metric with an application to a class of regression models. European Journal of Operational Research **278**(1), 20–35 (2019)
20. Markowitz, H.: Portfolio selection. Journal of Finance **7**(1), 77–91 (1952)
21. Miller, B.L., Wagner, H.M.: Chance constrained programming with joint constraints. Operations Research **13**(6), 930–945 (1965)
22. Mu, Y., Liu, W., Liu, X., Fan, W.: Stochastic gradient made stable: A manifold propagation approach for large-scale optimization. IEEE Transactions on Knowledge and Data Engineering **29**(2), 458–471 (2017)
23. Natarajan, K., Sim, M., Uichanco, J.: Tractable robust expected utility and risk models for portfolio optimization. Math Finance **18**(2), 695–731 (2010)
24. Popescu, I.: Robust mean-covariance solutions for stochastic optimization. Oparations Research **55**(1), 98–112 (2007)
25. Rigollet, P., Tong, X.: Neyman-pearson classification, convexity and stochastic constraints. Journal of machine learning research **12**(3), 2831–2855 (2011)
26. Rockafellar, R., Uryasev, S.: Optimization of conditional value-at-risk. J. Risk **2**, 21–42 (2000)
27. Rubio-Herrero, J., Baykal-Gürsoy, M., Jaśkiewicz, A.: A price-setting newsvendor problem under mean-variance criteria. European Journal of Operational Research **247**(2), 575–587 (2015)

28. Scarf, H., Arrow, K., Karlin, S.: A min-max solution of an inventory problem. Studies in the Mathematical Theory of Inventory and Production **10**, 201–209 (1958)
29. Shapiro, A.: On duality theory of conic linear problems. In: Goberna M.Á., López M.A. (eds) Semi-Infinite Programming. Nonconvex Optimization and Its Applications pp. 135–365 (2001)
30. Shapiro, A.: Worst-case distribution analysis of stochastic programs. Mathematical Programming **107**(1), 91–96 (2006)
31. Shapiro, A., Kleywegt, A.: Minimax analysis of stochastic problems. Optimizations Methods and Software **17**(3), 523–542 (2002)
32. Sun, H., Xu, H.: Convergence analysis for distributionally robust optimization and equilibrium problems. Mathematics of Operations Research **41**(2), 377–401 (2015)
33. Vasershtein, L.N.: Markov processes over denumerable products of spaces describing large system of automata. Probl. Peredachi Inf. **5**(3), 64–72 (1969)
34. Villani, C.: Optimal transport: old and new, vol. 338. Springer Science & Business Media (2003)
35. Wang, Z., Glynn, P., Ye, Y.: Likelihood robust optimization for data-driven problems. Computational Magmagement Science **13**, 241–261 (2016)
36. Won, J., Kim, S.: Robust trade-off portfolio selection. Optim Eng **21**, 867–904 (2020)
37. Xiao, X.: Penalized stochastic gradient methods for stochastic convex optimization with expectation constraints. Optimization-online (2019)
38. Zhang, J., Sethi, S.P., Choi, T., Cheng, T.C.E.: Supply chains involving a mean-variance-skewness-kurtosis newsvendor: Analysis and coordination. Production and Operations Management **29**(6), 1397–1430 (2020)
39. Zymler, S., Rustem, B., Kuhn, D.: Robust portfolio optimization with derivative insurance guarantees. European Journal of Operational Research **210**(2), 410–424 (2011)