Distributionally Robust Optimization based in Wasserstein metric applied to portfolio optimization.

Diego F. Fonseca V. df.fonseca@uniandes.edu.co

Mauricio Junca. P. m.junca20@uniandes.edu.co

Departament of Mathematics
Universidad de los Andes
Bogotá, Colombia

Abstract

We will study the type of optimization problems known as Distributionally Robust Optimization DRO, and how this can be important in portfolio theory. These DRO problems are problems of stochastic optimization formulated from a robust approach, this approach consist on assuming that the true distribution of the random variable, involved in the problem, belongs to a set of distributions called ambiguity set, this set is defined using the Wasserstein metric. Under specific conditions in the objective function, it has been shown that a DRO of this type can be formulated as a semi-infinite optimization program, and, depending on the objective function, this problem can be formulated as a finite convex optimization problem. Our goal is to formulate a robust version of the problem of choosing the optimal weights vector of the mean-variance model. We will also present the advantages that our approach has over some existing methods.

Keywords: Optimization, probability, distributions, Wasserstein

1 Introduction

We will address the problems of stochastic optimization from a robust point of view. These problems are useful to effectively describe many decision-making problems in uncertain environments. Problems of stochastic optimization have their origin problems of optimization of the type

$$\min_{x\in\mathbb{X}} f(x,\xi),$$

where $\mathbb{X}$ is a set of feasible solutions, ξ is a vector of parameters and $f(x,\xi)$ is a cost function. The difficulties emerge when it is assumed that ξ is a random vector. In that sense, if the distribution $\mathbb{P}$ of ξ is known, the previous problem can be formulated as

$$J^* := \min_{x\in\mathbb{X}} \mathbb{E}_P[f(x,\xi)]. \tag{1}$$

However, in practice, $\mathbb{P}$ is not exactly known, so other ways of addressing this last problem emerge, most are data-driven, that is, samples of the random variable ξ are used. One of the first methods is based on approaches via Monte Carlo, it is presented in [22] and [21], but this approximations are computationally expensive and sensitive to outliers in the data. Another option is to take a statistical approach to the problem, and assume that $\mathbb{P}$ has specific characteristics, this approach is perhaps more uncertain than the same problem that we want to solve since a bad choice can lead to a result different from the correct result. Thus, the distributionally robust approach it is an option that can overcome these problems, it considers a set $\mathcal{D}$ of probability distributions such that it contains $\mathbb{P}$. This set is known as *ambiguity set*. A Distributionally Robust Optimization (DRO) problem is formulated as

$$\widehat{J}_{\mathcal{D}} := \min_{x\in\mathbb{X}} \sup_{\mathbb{Q}\in\mathcal{D}} \mathbb{E}_{\mathbb{Q}}[f(x,\xi)]. \tag{2}$$

Note that $J^* \leq \widehat{J}_{\mathcal{D}}$ if $\mathbb{P}\in\mathcal{D}$. In this approach, the objective function becomes the worst expected cost for the choice of a distribution in this set. The origin of this approach is not clear, its origins are attributed to Von

Neumann's Game Theory, but the first work in which these ideas are used is in [18] within the framework of operations research.

The choice of the set $\mathcal{D}$ is a determining factor in the tractability of the problem. There are several ways to define $\mathcal{D}$. For example, in [9] and [20], it is defined as a set of distributions that are supported in a single point, while in [4], [17], [18], and [23], $\mathcal{D}$ is defined as the set of distributions that satisfy specific restrictions in their moments, or distributions belonging to a determinate 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 the empirical distribution[1] 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, it is used in [27]; the Kullback-Leibler divergence, it is employed in [8]; and the Total Variation distance, which is adopted in [24]. In this work, we will use *the Wasserstein distance*, that is, we will define $\mathcal{D}$ as a ball respect to the Wasserstein metric with center in an empirical distribution and radius properly chosen.

Definition 1 (Wassertein distance)**.** *The Wasserstein distance $W_p(\mu,\nu)$ between $\mu,\nu\in\mathcal{P}_p(\Xi)$[2] is defined by*

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

where

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

and d is a metric in Ξ.

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

$$\mathcal{B}_\varepsilon(\mu) := \left\{\nu\in\mathcal{P}(\Xi)\ \middle|\ W_p^p(\mu,\nu)\le\varepsilon^p\right\}. \tag{3}$$

Therefore, given a sample $\widehat{\xi}_1,\ldots,\widehat{\xi}_N$ of a random variable ξ, the ambiguity set is given by $\mathcal{D}:=\mathcal{B}_\varepsilon\left(\widehat{\mathbb{P}}_N\right)$.

One of the first works in which this notion of distance is defined is in [25] 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.

The reformulation of the DRO problem with $\mathcal{D}$ as a ball with respect to the Wasserstein metric centered on the empirical distribution is given in the following theorem.

Theorem 1.1 (Main Theorem)**.** *Assume that $\mathcal{D}:=\mathcal{B}_\varepsilon\left(\widehat{\mathbb{P}}_N\right)$. If f satisfies any of the following conditions:*

1. *f is maximum of concave functions.*
2. *f is Lipschitz respect to ξ and Ξ is compact.*
3. *f is upper semicontinuous.*

then the problem (2) is equivalent to the optimization problem

$$\begin{cases}\inf\limits_{x,\lambda,s} & \lambda\varepsilon^p+\dfrac{1}{N}\displaystyle\sum_{i=1}^N s_i\\ \textit{subject to} & \sup\limits_{\xi\in\Xi}\left(f(x,\xi)-\lambda d^p(\xi,\widehat{\xi}_i)\right)\le s_i\quad\forall i=1,\ldots,N,\\ & \lambda\ge 0.\end{cases} \tag{4}$$

[1] Given a sample $\widehat{\xi}_1,\ldots,\widehat{\xi}_N$ of a random variable ξ, we define the *empirical distribution* of ξ respect to this sample as the probability measure defined by $\widehat{\mathbb{P}}_N := \frac{1}{N}\sum_{i=1}^N \delta_{\widehat{\xi}_i}$ where δ_x is the Dirac delta supported in x.

[2] The p-Wasserstein metric is also defined for distributions outside of $\mathcal{P}_p(\Xi)$, what could probably happen is that in that set the Wasserstein metric is infinite.

In [6], [12] and [2] is shown the Theorem 1.1, specifically, in [6], it is shown assuming condition 1, in [12], it is shown assuming only condition 2, and, in [2], it is shown assuming only condition 3. Condition 3 is the moost general; furthermore, [2] is the most recent work that addresses this problem.

There are theoretical reasons, many exposed in [26], and practical reasons that make this distance very appealing. In fact, the definition of Wasserstein distances makes it convenient to use in optimal transport problems where it is naturally involved. Moreover, if μ a probability measure and $\varepsilon > 0$, we consider two balls $\mathcal{B}_w$ and $\mathcal{B}_\phi$ centered on μ and with radius ε in the space of measures of probability, where $\mathcal{B}_w$ is taken with respect to some Wasserstein metric and $\mathcal{B}_\phi$ with respect to any other notion of distance ϕ such that it is not Wasserstein. A known defect of the notions of distance ϕ is that the ball $\mathcal{B}_\phi$ is not rich in relevant distributions [7], for example, if ϕ is the Kullback-Leibler divergence, and we consider a measurable set A such that $\mu(A) = 0$, then for all $\nu \in \mathcal{B}_\phi$ we have $\nu(A) = 0$, that is, $\mathcal{B}_\phi$ contains only measures that are absolutely continuous with respect to μ, so it only contains measures supported at points where μ is also supported. Furthermore, if μ is discrete then for several notions ϕ the ball $\mathcal{B}_\phi$ does not contain continuous distributions. This situation does not occur in $\mathcal{B}_w$.

Another reason is that the Wasserstein distances have a dual representation. Having an equivalent formulation is always an advantage because it opens the possibility that this formulation, which is an optimization problem, is technically more convenient. Furthermore, the Wasserstein distances are defined by a infimum, this is an advantage since they allow to calculate upper bounds relatively easy; for example, Theorem 6.15 in [26] shows that the Wasserstein metrics are bounded by the Total Variation distance.

Our goal is to use the ideas of Distributionally Robust Optimization (DRO) to address stochastic optimization problems in portfolio optimization, specifically, in *Markowitz mean-variance portfolio theory*. In this contexts, a problem of stochastic optimization was identified, the goal is to propose its DRO version. We will verify sufficient conditions such that the reformulation of the DRO as a semi-infinite optimization problem is valid. Finally, we evaluate the performance of our approach and compare it with other approaches.

In *Markowitz mean-variance portfolio theory* attempts to solve the optimization problem that consists of choosing the portfolio weights such that they minimize the variance of the return rate subject to a constraint on the expected value of the return rate, the formulation is as follows.

$$J := \begin{cases} \min\limits_{x \in \mathbb{R}^m} & \mathrm{Var}_{\mathbb{P}}\left[\langle x, \xi \rangle\right] \\ \text{subject to} & \mathbb{E}_{\mathbb{P}}\left[\langle x, \xi \rangle\right] \geq \mu, \\ & \sum_{i=1}^{m} x_i = 1, \\ & x \geq 0. \end{cases} \tag{5}$$

where m is the number of assets, ξ_i is the return of asset i-th, and x_i the proportion of the initial amount invested in the i-th asset. Given that the returns of each asset are, in practice, random, then $\xi = (\xi_1, \ldots, \xi_m) \in \mathbb{R}^m$ is a *random vector*. Additionally, $x = (x_1, \ldots, x_m) \in \mathbb{R}^m$ is a vector known as the vector of weights, this satisfies the relation $\sum_{i=1}^{m} x_i = 1$. In addition, μ is the level of return admissible.

The first attempts to solve this problem consider estimates of the vector of means and the covariance matrix of returns, but the resulting portfolios do not perform well out of sample, and they are very sensitive to variation in estimates, this was shown in [3]. 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 problem with this approach is to choose the feasible set of the vector and the matrix, this approach is used in [29], [5], [14], [10], [11], and [28]. Another approach is to consider the distribution of returns as a variable of the optimization problem, and assume that the feasible set of this variable is a set of distributions defined by a notion of distance, for example, this set can be a ball centered on some specific distribution and a radius that must be calibrated. This is a DRO approach, but, in this case, applied to a stochastic problem with probabilistic restrictions. Our interest is in the use of Wasserstein distances. Precisely, the use of Wasserstein distances in portfolio optimization appears for first time in [16]. This tendency to use Wasserstein distances in portfolio optimization is also present in [15], [6] and [1]. The problem addressed in [15] and [6] is different from because the risk measure used in these works is not the variance, they use measures such as Var and CVaR. However, our research in this area shares similar results to those presented in [1], but our work was developed independently. In addition, our objective problem is different and the way as we use the Wasserstein distance is also different. In fact, our work defines the ambiguity set in such a way that it depends on x, this differs from the way in which this set has been defined in the works that have used Wasserstein distances so far.

The organization of this paper is as follows. In Section 2, using Wasserstein distance, we construct a distributionally robust portfolio selection optimization model using variance as the risk measure. In this section, we derive the tractable convex reformulations for the distributionally robust mean-variance model, and we propose

three data-driven techniques for calibrating the parameters. Simulation analysis of the proposed approaches are derived in section 3. In that section, we also report the results of a variety of numerical experiments using real market data. Finally, conclusions are drawn in section 4.

2 Markowitz mean-variance portfolio model using Wasserstein distance

As stated before, in (1), the distribution $\mathbb{P}$ is unknown, to overcome this, we assume that you have access to realizations of the random vector ξ, that is, let $\widehat{\xi}_1, \ldots, \widehat{\xi}_N$ be a sample of ξ, this allows to estimate $\mathbb{P}$ by means of the empirical distribution $\widehat{\mathbb{P}}_N := \frac{1}{N}\sum_{i=1}^{N} \delta_{\widehat{\xi}_i}$. In [13], $\mathbb{P}$ is changed to $\widehat{\mathbb{P}}_N$ in (1). This new optimization problem is the sample version of (1) or Sample Average Approximation (SAA), however, this version does not offer a good out-of-sample performance, which motivates the search for another approach.

Our approach also gives a preponderant role to an empirical distribution, but, in our case, it depends on x, to understand this we establish the following conventions. For $x \in \mathbb{R}^m$ we define $\zeta^x := \langle x, \xi \rangle$, this is random variable. We called $\mathbb{P}^x$ to the distribution of ζ^x, note that it depend of $\mathbb{P}$, so $\mathbb{P}^x$ also is unknown. Additionally, we define $\widehat{\zeta}_i^x := \langle x, \widehat{\xi}_i \rangle$, so $\widehat{\zeta}_1^x, \ldots, \widehat{\zeta}_N^x$ is a sample of ζ^x. This allows us to define the empirical distribution of ζ^x, which is given by $\widehat{P}_N^x := \frac{1}{N}\sum_{i=1}^{N} \delta_{\widehat{\zeta}_i^x}$. Our approach is to solve the following optimization problem.

$$\widehat{J}_N(\varepsilon) := \begin{cases} \min\limits_{x \in \mathbb{R}^m} & \sup\limits_{\mathbb{Q} \in \mathcal{B}_{\|x\|\varepsilon}(\widehat{\mathbb{P}}_N^x)} \operatorname{Var}_{\mathbb{Q}}[\zeta^x] \\ \text{subject to} & \inf\limits_{\mathbb{Q} \in \mathcal{B}_{\|x\|\varepsilon}(\widehat{\mathbb{P}}_N^x)} \mathbb{E}_{\mathbb{Q}}[\zeta^x] \geq \mu, \\ & \sum_{i=1}^{m} x_i = 1. \\ & x \geq 0. \end{cases} \tag{6}$$

where $\mathcal{B}_{\|x\|\varepsilon}(\widehat{\mathbb{P}}_N^x)$ is a ball centered on $\widehat{\mathbb{P}}_N^x$ with radius $\|x\|\,\varepsilon$, the parameter ε must be adjusted. The ball is defined with respect to the 2-Wassertein distance, and $\|\cdot\|$ is the euclidean metric in $\mathbb{R}^n$. In (6), we have the parameter ε, this is chosen prioritizing the performance of the approach. However, some values of ε can make problem (6) infeasible, so these parameters can only take values on a subset of the real numbers which we determine at the end of this section.

The first task is to reformulate (6) as an optimization problem in a context that makes it computationally treatable. In our experiments and results, we will see that solving (6) is a reasonable way to deal with (1) and that it is more aware of the fact that it does not know $\mathbb{P}$; furthermore, our approach shows a better out-of-sample performance than other existing methods that also try to approximate a solution of (1). The reformulation of (6) is achieved in the following theorem.

Theorem 2.1. *The optimization problem (6) is equivalent to the convex optimization problem*

$$\widehat{J}_N(\varepsilon) = \begin{cases} \underset{x \in \mathbb{R}^m}{\text{minimize}} & \left(\sqrt{\frac{1}{N}\sum_{i=1}^{N} \langle x, \widehat{\xi}_i \rangle^2 - \frac{1}{N^2}\left(\sum_{i=1}^{N} \langle x, \widehat{\xi}_i \rangle\right)^2} + \varepsilon \|x\| \right)^2 \\ \textit{subject to} & \sum_{i=1}^{N} \langle x, \widehat{\xi}_i \rangle - \varepsilon \|x\| \geq \mu, \\ & \sum_{i=1}^{m} x_i = 1. \\ & x \geq 0. \end{cases} \tag{7}$$

The optimal solutions of (7) will be denoted by $\widehat{x}_N(\varepsilon)$. The optimization problem that appears in this theorem can be rewritten using matrix notation, that is what the following corollary illustrates.

Corolario 2.1.1. *Let M be a matrix with size $m \times N$ where its columns are vectors of the sample $\widehat{\xi}_1, \ldots, \widehat{\xi}_N$, and let $\mathbf{0}, \mathbf{e} \in \mathbb{R}^N$ column vectors with zeros and ones respectively. From these conventions the following matrices are defined*

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

Therefore, (7) is equivalent to the optimization problem

$$\begin{cases} \inf\limits_{x\in\mathbb{R}^m} & \left(\left\|K^T x\right\| + \varepsilon \left\|x\right\|\right)^2 \\ \textit{sujeto a} & Lx - \varepsilon \left\|x\right\| \geq \mu, \\ & e^T x = 1, \\ & x \geq 0. \end{cases} \tag{8}$$

where K is a matrix that depends on E.

As said before, this problem can be infeasible for some values of μ and ε, the following corollary establishes the values for which this is feasible.

Corolario 2.1.2. *The optimization problem (8) is feasible if μ and ε satisfies the following inequalities*

$$\mu < \widehat{\mu}_N^{\max} := \begin{cases} \sup\limits_{x\in\mathbb{R}^m} & Lx \\ \textit{subject to} & \sum\limits_{i=1}^{m} x_i = 1, \\ & x \geq 0. \end{cases} \qquad \textit{and} \qquad \varepsilon \leq \widehat{\varepsilon}_N^{\max}(\mu) := \begin{cases} \sup\limits_{x\in\mathbb{R}^m} & \dfrac{Lx-\mu}{\|x\|} \\ \textit{subject to} & Lx \geq \mu \\ & \sum\limits_{i=1}^{m} x_i = 1, \\ & x \geq 0. \end{cases}$$

The proofs of Theorem 2.1, and Corollaries 2.1.1 and 2.1.2 can be consulted in Appendix A.

One of the advantages of our approach is that the resulting optimization problem is convex, the other advantages we will see in the section on experiments and results. However, convexity is a desired characteristic in optimization problems since there are several tools to deal with these types of problems. Our intention is to have a balance between performance and treatability, in this sense, treatability is achieved in Theorem 2.1, moreover, a good performance is also achieved, but this will be discussed later in this paper.

2.1 Calibration of the Wasserstein Radius ε

In this part, we present three ways to choose epsilon. The first does not require many details, this consists of considering the largest possible ε, that is, $\varepsilon = \widehat{\varepsilon}_N^{\max}(\mu)$. With this choice we obtain the highest expected return that can be obtained with our approach although the variance will also be high. However, this choice has shown the best performance when the data comes from the stock market.

The following two ways to choose ε are based on the bootstrap method. For this methods, we consider $\mathbb{P}^N$ the distribution of the random vector $\widehat{\Xi} := \left\{\widehat{\xi}_1, \widehat{\xi}_2, \ldots, \widehat{\xi}_N\right\}$. In the first method, given α such that $0 < \alpha < 1$, the goal is to find the smallest $\varepsilon > 0$ such that $\mathbb{P}^N\left(\mathbb{E}_{\mathbb{P}}[\langle \widehat{x}_N(\varepsilon), \xi\rangle] \geq \mu\right) \geq \alpha$. The algorithm for computing this ε is as follows.

Bootstrapping procedure prioritizing probability $\geq \alpha$ (BtPα procedure)

Step 1. Consider a list $\mathcal{E}$ of candidate to be ε. For each $\varepsilon \in \mathcal{E}$, do the following two steps.

Step 2. Partition $\widehat{\xi}_1, \widehat{\xi}_2, \ldots, \widehat{\xi}_N$ into a training dataset of size N_T and a validation dataset of size $N_V = N - N_T$. Using only the training dataset, solve (8) to obtain $\widehat{x}_{N_T}(\varepsilon)$. Use the validation dataset to estimate $\mathbb{E}_{\mathbb{P}}[\langle \widehat{x}_{N_T}(\varepsilon), \xi\rangle]$ via the sample average approximation, that is, $\mathbb{E}_{\widehat{\mathbb{P}}_{N_V}}[\langle \widehat{x}_{N_T}(\varepsilon), \xi\rangle]$.

Step 3. Repeat **Step 2** B times (for example, $B = 1000$), and count the number of times that $\mathbb{E}_{\widehat{\mathbb{P}}_{N_V}}[\langle \widehat{x}_{N_T}(\varepsilon), \xi\rangle] \geq \mu$ occurred, we call this number b_ε. Note that $\frac{b_\varepsilon}{B}$ acts as an estimate of $\mathbb{P}^N\left(\mathbb{E}_{\mathbb{P}}[\langle \widehat{x}_N(\varepsilon), \xi\rangle] \geq \mu\right)$.

Step 4. Set $\widehat{\varepsilon}_N^{\mathrm{BtP}\alpha}$ to the smallest $\varepsilon \in \mathcal{E}$ that satisfies $\frac{b_\varepsilon}{B} \geq \alpha$.

Step 5. Report $\widehat{x}_N^{\mathrm{BtP}\alpha}$ as the data-driven solution of (8) with $\varepsilon = \widehat{\varepsilon}_N^{\mathrm{BtP}\alpha}$, that is, $\widehat{x}_N^{\mathrm{BtP}\alpha} = \widehat{x}_{N_T}\left(\widehat{\varepsilon}_N^{\mathrm{BtP}\alpha}\right)$.

This way of choosing ε guarantees high expected returns and $\alpha \times 100$ percent of the time above μ, all this for portfolio vector $\widehat{x}_N^{\mathrm{BtP}\alpha}$. However, high variance could be present. Additionally, this procedure cannot always be carried out because the ε that satisfies **Step 4** might not be found. This occurs when μ is large or for certain samples of the returns.

In the second method, the goal is to find a $\varepsilon > 0$ such that $\mathbb{E}_{\mathbb{P}^N}\left[\mathbb{E}_{\mathbb{P}}[\langle \widehat{x}_N(\varepsilon), \xi\rangle]\right]$ is very close to μ. Note that $\mathbb{E}_{\mathbb{P}^N}\left[\mathbb{E}_{\mathbb{P}}[\langle \widehat{x}_N(\varepsilon), \xi\rangle]\right] = \mathbb{E}_{\mathbb{P}}[\langle \mathbb{E}_{\mathbb{P}^N}\left[\widehat{x}_N(\varepsilon)\right], \xi\rangle]$ where $\mathbb{E}_{\mathbb{P}^N}\left[\widehat{x}_N(\varepsilon)\right]$ is a vector of expected values that is also a vector of weights. The algorithm for computing this ε is as follows.

Bootstrapping procedure prioritizing expected return close to μ (BtE procedure)

Step 1. Consider a list $\mathcal{E}$ of candidate to be ε. For each $\varepsilon \in \mathcal{E}$, do the following two steps.

Step 2. This step is the same as **Step 2** of BtPα procedure.

Step 3. Repeat **Step 2** B times (for example, $B = 1000$), and set s_ε to the sum of all $\mathbb{E}_{\widehat{\mathbb{P}}_{N_V}}[\langle \widehat{x}_{N_T}(\varepsilon), \xi \rangle]$ obtained in the B repetitions. Note that $\frac{s_\varepsilon}{B}$ acts as an estimate of $\mathbb{E}_{\mathbb{P}^N}[\mathbb{E}_{\mathbb{P}}[\langle \widehat{x}_N(\varepsilon), \xi \rangle]]$. Additionally, set $\widehat{x}_N^{sum}(\varepsilon)$ to the component-by-component sum of all the $\widehat{x}_{N_T}(\varepsilon)$ vectors obtained. In this case, $\widehat{x}_N^{sum}(\varepsilon)$ acts as an estimate of $\mathbb{E}_{\mathbb{P}^N}[\widehat{x}_N(\varepsilon)]$.

Step 4. Set $\widehat{\varepsilon}_N^{\mathrm{BtE}}$ to $\varepsilon \in \mathcal{E}$ that satisfies $\frac{s_\varepsilon}{B} \approx \mu$.

Step 5. Report $\widehat{x}_N^{\mathrm{BtE}} = \widehat{x}_N^{sum}\left(\widehat{\varepsilon}_N^{\mathrm{BtE}}\right)$ as the data-driven solution.

This way of choosing ε guarantee expected returns close to μ for portfolio vector $\widehat{x}_N^{\mathrm{BtP}\alpha}$. However, this procedure cannot always be carried out for the same reasons that this occurs in BtPα procedure.

3 Experiments and results

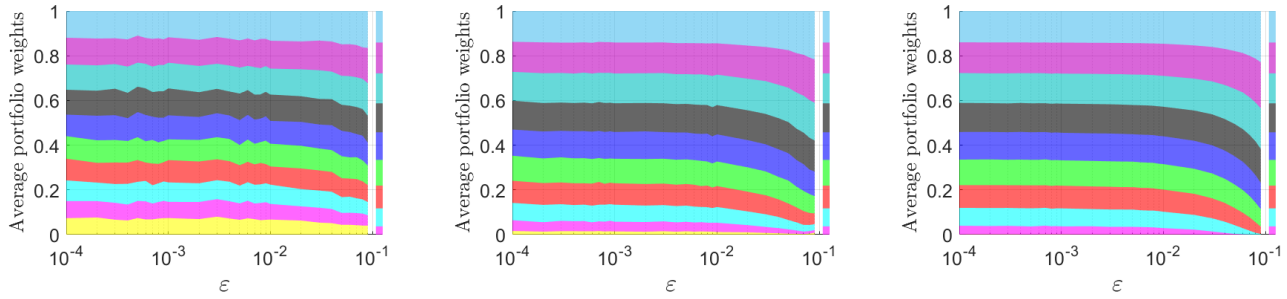
This section is divided into two parts, in one we use synthetically generated data, that is, generated by a known distribution. In the other part, we use real data from the financial market.

3.1 Using synthetic data

Our experiments consider a market of $m = 10$ assets. We assume that the returns have the form adopted in [6], that is, if $\xi = [\xi_1, \ldots, \xi_m]$ is the random vector representing the returns, we assume that $\xi_i = \psi + \zeta_i$ where ψ and ζ_i are independent; furthermore, $\psi \sim \mathcal{N}(0, 2\%)$ 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. To evaluate the performance of our approach, we must define some concepts. Given $x \in \mathbb{R}^m$ we define $R(x) := \mathbf{b}^T x$, this is the *expected return induced by* x. Moreover, we define $V(x) := x^T \Sigma x$, this is *the variance induced by* x. Because we have all the information about the returns, the vector of optimal weights x^* is known. We will use this vector to compare the performance of our approach.

3.1.1 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 (8) using samples of cardinality $N \in \{30, 300, 3000\}$. Figure 3.1 visualizes the corresponding optimal portfolio weights $\widehat{x}_N(\varepsilon)$ as a function of ε, averaged over 500 independent simulation runs. The thin colored bar that is separated and located 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.

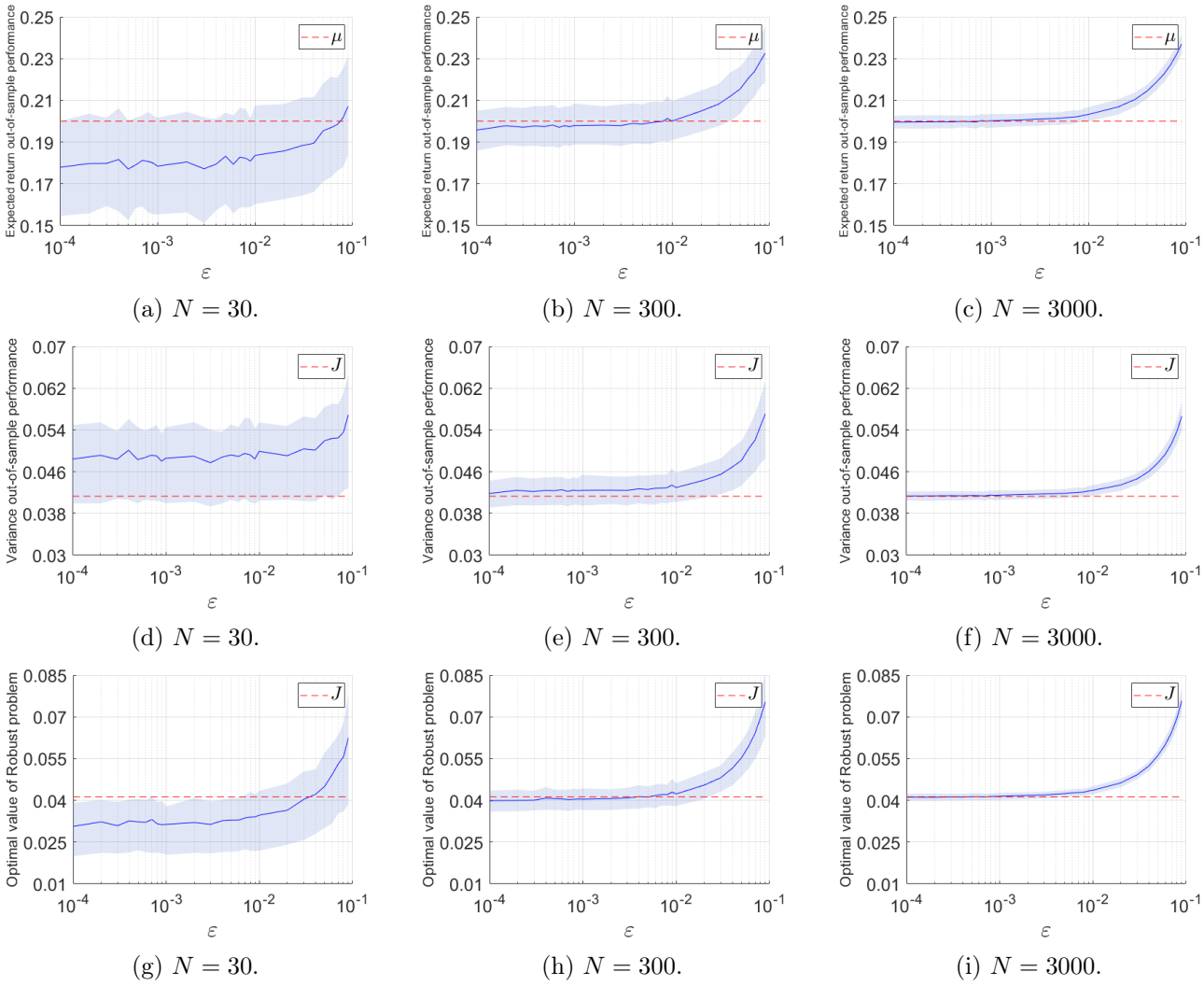


(a) $N = 30$. (b) $N = 300$. (c) $N = 3000$.

Figure 3.1: 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, $\mu = 0.2$.

Figure 3.2 shows the tubes between the 20% and 80% quantiles (shaded areas) and the means (solid lines) of the out-of-sample performance of expected returns $R(\widehat{x}_N(\varepsilon))$, variance returns $V(\widehat{x}_N(\varepsilon))$, and optimal values $\widehat{J}_N(\varepsilon)$, all these are presented 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. We consider a critical Wasserstein radius ε_{crit} as the smallest ε that guarantees that the mean of the expected returns exceeds μ. The existence of ε_{crit} depends on μ. This makes sense because it is consistent with the fact that we cannot put just any μ, we should put one that is achievable (see Corollary 2.1.2). This stylized fact was observed consistently across all of simulations and provides an empirical justification for adopting a distributionally robust approach.

In the case of the out-of-sample performance of variance $V(\widehat{x}_N(\varepsilon))$ and optimal value $\widehat{J}_N(\varepsilon)$, Figure 3.2 also shows that this expressions are increasing as Wasserstein radius ε grows. In addition, it is observed that high returns induce large variances. Another aspect that follows from these graphs is that if ε is such that $R(\widehat{x}_N(\varepsilon)) \geq \mu$ in a large percentage of the simulations, then $\widehat{J}_N(\varepsilon) \geq J$ in an even larger percentage of those simulations. This is consistent with the formulation of problem (6) because if ε is such that $\mathbb{P}^{\widehat{x}_N(\varepsilon)} \in \mathcal{B}_{\|\widehat{x}_N(\varepsilon)\|\varepsilon}(\widehat{\mathbb{P}}_N^{\widehat{x}_N(\varepsilon)})$, then $\widehat{J}_N(\varepsilon) \geq J$ and $R(\widehat{x}_N(\varepsilon)) \geq \mu$. The graph verifies for the existence of this ε. Even though this observation was made consistently across all simulations, we were unable to validate it theoretically.



(a) $N = 30$. (b) $N = 300$. (c) $N = 3000$.

(d) $N = 30$. (e) $N = 300$. (f) $N = 3000$.

(g) $N = 30$. (h) $N = 300$. (i) $N = 3000$.

Figure 3.2: Out-of-sample performance of expected return $R(\hat{x}_N(\varepsilon))$, Out-of-sample performance of variance return $V(\hat{x}_N(\varepsilon))$, and Optimal value $\widehat{J}_N(\varepsilon)$ 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% quantil of data generated by 500 simulations. In this case, $\mu = 0.2$.

3.1.2 Out-of-sample performance of asset allocation strategies

In this subsection, we compare the distributionally robust approach based on the Wasserstein ambiguity set with the classical sample average approximation (SAA). For this purpose, out-of-sample performance of expected return $R(\hat{x}_N(\varepsilon))$ and out-of-sample performance of variance return $V(\hat{x}_N(\varepsilon))$ are drawn for each of the three ways of choosing ε (see Section 2.1). Figure 3.3 presents boxplots for each Wasserstein approach, that is, for each way of choosing ε which are BtPα procedure (Wass BtPα), BtE procedure (Wass BtE) and considering ε as $\widehat{\varepsilon}_N^{\max}(\mu)$ (Wass MaxFact). In bootstrap-based Wasserstein approaches, 80% of the data are used for training and 20% for validation. In each boxplot, the green point corresponds to the mean of the data that determines the boxplot. Additionally, each boxplot is determined by 200 data generated by 200 independent simulation runs.

Figure 3.3(a) shows boxplots of $R(\hat{x}_N(\varepsilon))$ for each of the approaches we are comparing. Given an approach to solve (1), we want the portfolio vector produced by it to generate expected returns above or close to μ most of the time when they are evaluated out of sample. These boxplots show that, in a high percentage, out of the sample the portfolio vectors obtained through the SAA approach give expected returns below μ. This goes against what is unconsciously expected when this approach is considered. However, Wasserstein BtE approach shows results where its expected returns are closer to μ, it could be said that Wasserstein BtE is a corrected version of SAA. On the other hand, most of the time, Wasserstein BtPα and Wasserstein MaxFat

(Wass MaxFat) show expected returns greater than μ. Therefore, if getting an expected return greater than μ is the priority, then Wasserstein approaches are ideal for this purpose.

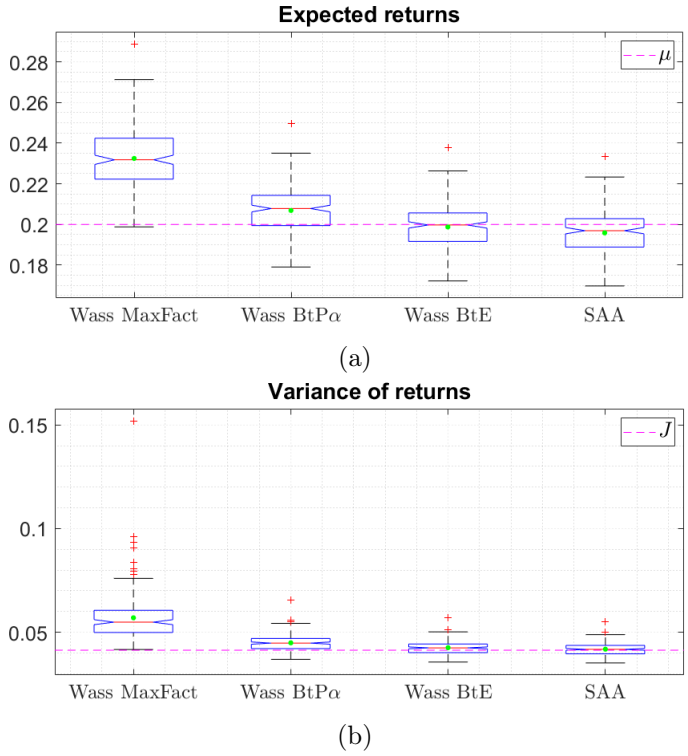


Figure 3.3: Boxplots of out-of-sample performance of expected return $R(\hat{x}_N(\varepsilon))$ and out-of-sample performance of variance return $V(\hat{x}_N(\varepsilon))$ estimated on the basis of 200 simulations using each of the methods for choosing ε proposed in Section 2.1. In this case, $\mu = 0.2$, $N = 300$ and $\alpha = 0.8$.

The out-of-sample variance induced by the portfolio vector is an important aspect. Precisely, Figure 3.3(b) shows boxplots of $V(\hat{x}_N(\varepsilon))$ for each of the approaches. These boxplots show that the variances of the returns generated by Wasserstein BtE are slightly larger than the variances generated by SAA, this difference is not evident at first glance. Therefore, unlike SAA, Wasserstein BtE allows us to obtain portfolio vectors with expected returns closer to μ, but with variances similar to those obtained from SAA. Consequently, we can reaffirm that Wasserstein BtE is an improved version of SAA. Additionally, since we know the distribution of the returns, we know J the optimal value of (1). The red dashed line represents this value. Hence, we observe that the variances obtained from Wasserstein MaxFat and BtPα are not close to J; furthermore, they are even larger than J. This is not surprising because high returns imply high volatilities.

3.2 Using real market data

This part of our work presents the results from the simulation of our approach applied to real market data. The data used in this study correspond to the daily returns of 23 companies selected from those that make up the (S&P 500). The selected returns correspond to the companies described below.

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

These data correspond to the time window between January 1, 2008 to May 29, 2020. In the experiments, we want to analyze the portfolio value over time, that is, we use the corresponding data from January 2008 to April 28 of 2017 to estimate portfolio vector, we keep this investment decision for the following days until May 29, 2020. The objective is to see how the portfolio value evolves in that period of time. Precisely, Figure 3.4 shows this evolution. Before analyzing the results obtained, it is important to bear in mind that we used 5 benchmarks for comparison of our approach with standard portfolio optimization techniques. The 5 benchmarks used were SAA, EW, MinVar, MaxSR, and S&P 500. 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): Using the available data, that is, $\widehat{\xi}_1, \ldots, \widehat{\xi}_N$, the sample estimate of the covariance matrix of the returns is calculated, which is noted with the expression $\widehat{\Sigma}_N$. The vector x that solves the optimization problem $\min_{x \in \mathcal{X}} x^T \widehat{\Sigma} x$ is the portfolio vector obtained with this approach, where $\mathcal{X} = \{x \in \mathbb{R}^m \ : \ \sum_i^m x_i = 1, \ x_i \geq 0, \ \forall i\}$.
- Maximizing Sharpe ratio (MaxSR): Using the available data, the sample estimates of the covariance matrix and mean vector of the returns are calculated, this estimates are noted with the expressions $\widehat{\Sigma}_N$ and $\widehat{\mathrm{m}}$ respectively. The vector x that solves the optimization problem $\max_{x \in \mathcal{X}} \frac{\widehat{\mathrm{m}} x}{x^T \widehat{\Sigma} x}$ is the portfolio vector obtained with this approach.
- S&P 500: It is a stock index that compiles the 500 largest companies in the United States. It is pertinent to use this index as a reference point in our analysis since the returns that we use in our experiments are from companies that are part of the composition of that index.

Wasserstein's approaches that are based on Bootstrap methods could not be evaluated because their execution was not successful. We think that this situation is due to aspects related to the sample of returns. To make up for this absence, we include in this study the performance of the Wasserstein approach with $\varepsilon = 0.855 \times 10^{-4}$, this value of ε is less than that obtained in Wasserstein MaxFact. The reason for including this strategy is to see the influence of ε on the portfolio value. Additionally, we consider $\mu = 0.001$ in the experiments, this is because 0.001 is the expected return in-sample obtained from portfolio vector generated by the MaxSR approach.

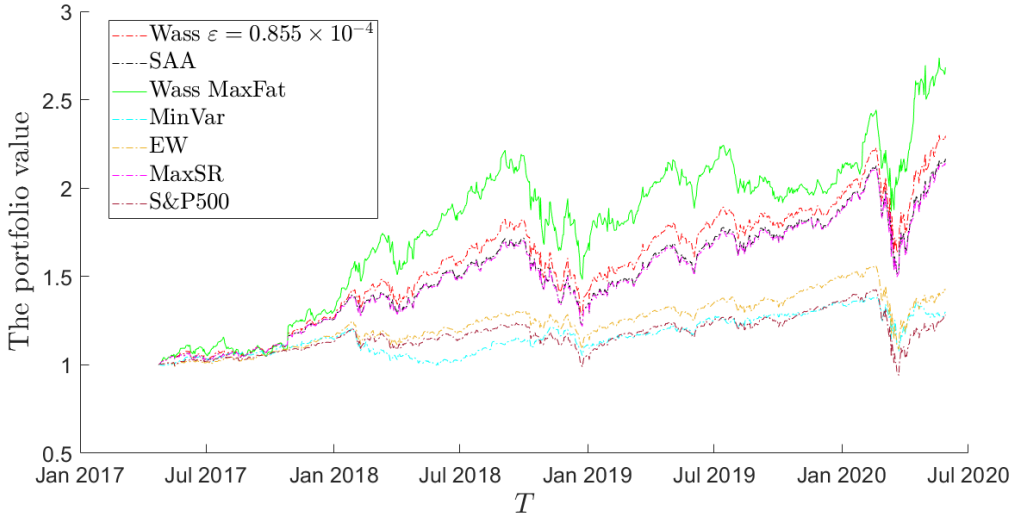


Figure 3.4: The portfolio value of the trading strategies over the period April 28, 2017–May 29, 2020 for target mean return, $\mu = 0.001$. The evolution of S&P 500 is also provided for reference purposes.

Figure 3.4 shows that the strategy based on Wasserstein allows obtaining a portfolio with high value, this value increases as time passes. The value of these portfolios exceeds the values obtained with traditional strategies such as SAA, EW, MinVar, and MaxSR. Furthermore, an investment based on the Wasserstein approach is going to be more valuable than that investment in S&P 500. Additionally, the period of time in

which we are evaluating the strategies includes the days of the start of the COVID-19 pandemic, this phenomenon affected all investment strategies; however, the Wasserstein approaches somewhat mitigate this effect on long-term portfolio value. Although the value of the portfolio experiences a fall, this does not make its value less than the amount invested at the beginning.

	Wass MaxFat	Wass $\varepsilon = 0.855 \times 10^{-4}$	MaxSR	EW	MinVar	SAA
Mean	0.0014656	0.001234	0.001148	0.00051467	0.0043067	0.0011583
Standard deviation	0.018782	0.016067	0.015904	0.013323	0.011407	0.015926
Sharpe Ratio	0.078033	0.076811	0.072186	0.038629	0.037755	0.07273

Table 1: Performances of different portfolio strategies.

Table 1 shows the out-of-sample results of indicators such as the mean, standard deviation, and Sharpe Ratio of each of the strategies evaluated. These indicators were calculated for the period from April 28, 2017 to May 29, 2020. Regarding the mean, the Wassertein approaches show that the expected returns are greater than 10, especially, Wassesrtein MaxFact shows the highest expected return among all the strategies. However, the standard deviations of the Wasserstein approaches are the largest, but the difference with respect to the SAA approach is not so significant, this becomes evident when we see the Sharpe ratios. Indeed, we observe that the Wasserstein MaxFat strategy has the highest Sharpe ratio; furthermore, all the strategies based on Wasserstein show the highest Sharpe ratios. If we add to this the fact that Wasserstein approaches produce portfolio vectors that generate more valuable portfolios as time progresses, then Wasserstein approaches are a good option for making investment decisions in real life.

4 Conclusions and future works

In this work, we have shown that the Wasserstein distance-based approach (6) has an equivalent convex formulation, and we have proposed methods to choose the size of the ambiguity set. Furthermore, we established theoretical results that characterize the values of μ and ε for which the Wasserstein approach is valid. The experiments in synthetic data and real market data have shown aspects such as the behavior of the portfolio vectors with respect to ε which, as ε grows, tend to give more weight to goods that give high returns. Also, we observe a good performance with respect to expected return, standard deviation, and Sharpe ratio. Precisely, the values obtained for the Sharpe ratio show a good balance between expected return and standard deviation. As future work, it remains to carry out experiments with daily rebalancing to analyze other indicators such as turnover.

References

[1] J. Blanchet, Chen L., and X. Y. Zhou. Distributionally Robust Mean-Variance Portfolio Selection with Wasserstein Distances. *arXiv:1802.04885v1.*, 2018.

[2] J. Blanchet and K. Murthy. Quantifying distributional model risk via optimal transport. *Mathematics of Operations Research*, 44(2):565–600, 2019.

[3] V.K. Chopra and Ziemba W.T. The effect of errors in means, variances and covariances on optimal portfolio choice. *Journal of Portfolio Management*, 19(2):6–11, 1993.

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

[5] L. El Ghaoui, M. Oks, and F. A. Oustry. Worst-case value-at-risk and robust portfolio optimization: a conic programming approach. *Operations Research*, 51(4):543–553, 2003.

[6] PM. Esfahani and D. Kuhn. Data-driven distributionally robust optimization using the wasserstein metric: Performance guarantees and tractable reformulations. *Mathematical Programming*, 171:115–166, 2018.

[7] R. Gao and AJ. Kleywegt. Distributionally Robust Stochastic Optimization with Wasserstein Distance. *arXiv preprint arXiv:1604.02199v2*, 2016.

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

[9] C. M. Lagoa and R. B. Barmish. Distributionally robust Monte Carlo simulation. *In Proceedings of the International Federation of Automatic Control World Congress*, pages 1–12, 2002.

[10] S. Lotf, M. Salahi, and F. Mehrdoust. Adjusted robust mean-value-at-risk model: less conservative robust portfolios. *Optim Eng*, 18(2):467–497, 2017.

[11] S. Lotf and S. Zenios. Robust VaR and CVaR optimization under joint ambiguity in distributions, means, and covariances. *European Journal of Operational Research*, 269(2):556–576, 2018.

[12] F. Luo and S Mehrotra. 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.

[13] H. Markowitz. Portfolio selection. *Journal of Finance*, 7(1):77–91, 1952.

[14] K. Natarajan, M. Sim, and J. Uichanco. Tractable robust expected utility and risk models for portfolio optimization. *Math Finance*, 18(2):695–731, 2010.

[15] G. Pflug, A. Pichler, and D. Wozabal. The 1/N investment strategy is optimal under high model ambiguity. *Journal of Banking & Finance*, 36(2):410–417, 2012.

[16] G. Pflug and D. Wozabal. Ambiguity in portfolio selection. *Quantitative Finance.*, 7(4):435–447, 2007.

[17] I. Popescu. Robust mean-covariance solutions for stochastic optimization. *Oparations Research*, 55(1):98–112, 2007.

[18] H. Scarf, K. Arrow, and S. Karlin. A min-max solution of an inventory problem. *Studies in the Mathematical Theory of Inventory and Production*, 10:201–209, 1958.

[19] A. Shapiro. On duality theory of conic linear problems. *In: Goberna M.Á., López M.A. (eds) Semi-Infinite Programming. Nonconvex Optimization and Its Applications*, pages 135–365, 2001.

[20] A. Shapiro. Worst-case distribution analysis of stochastic programs. *Mathematical Programming*, 107(1):91–96, 2006.

[21] A. Shapiro and D. Dentcheva. Lectures on Stochastic programming: modeling and theory. *SIAM*, 2016.

[22] A. Shapiro and T. Homem-de Mello. On the rate of convergence of optimal solutions of Monte Carlo approximations of stochastic programs . *SIAM Journal on Optimization*, 11(1):70–86, 2000.

[23] A. Shapiro and A. Kleywegt. Minimax analysis of stochastic problems. *Optimizations Methods and Software*, 17(3):523–542, 2002.

[24] H. Sun and H. Xu. Convergence analysis for distributionally robust optimization and equilibrium problems. *Mathematicas of Operations Research*, 2015.

[25] L. N. Vasershtein. Markov processes over denumerable products of spaces describing large system of automata. *Probl. Peredachi Inf.*, 5(3):64–72, 1969.

[26] C. Villani. *Optimal transport: old and new*, volume 338. Springer Science & Business Media, 2003.

[27] Z. Wang, PW. Glynn, and Y. Ye. Likelihood robust optimization for data-driven problems. *Computational Magmagement Science*, pages 1–21, 2015.

[28] J. Won and S. Kim. Robust trade-off portfolio selection. *Optim Eng*, 21:867–904, 2020.

[29] S. Zymler, B. Rustem, and D. Kuhn. Robust portfolio optimization with derivative insurance guarantees. *European Journal of Operational Research*, 210(2):410–424, 2011.

A Proof of Theorem 2.1

For the proof of Theorem 2.1, it is necessary to establish some previous results. The strategy in this proof is to reduce the problem to a one-dimensional context since most of the following result applies for random variables. These results are presented in sections in order to make the reading less dense. If the reader wants to jump to the proof, he can turn to section A.3.

A.1 Duality of distributionally robust problems with restrictions on the expected value

When we refer to distributionally robust problems with restrictions on the expected value, we are referring to problems of the form

$$\begin{cases} \sup\limits_{\mathbb{Q}\in\mathcal{B}_\varepsilon(\widehat{\mathbb{P}}_N)} & \mathbb{E}_{\mathbb{Q}}\left[h(\xi)\right] \\ \text{subject to} & \mathbb{E}_{\mathbb{Q}}\left[g_i(\xi)\right] = b_i. \quad \forall\, i = 1,\dots,k. \end{cases} \tag{9}$$

where $b_i \in \mathbb{R}$ and $g_1,\dots,g_k$ integrable functions respect to each measure in $\mathcal{B}_\varepsilon(\widehat{\mathbb{P}}_N)$. This problem is important because versions of it emerge in different contexts, for that reason it is important to know when this problem satisfies strong duality, and how is this dual formulation.

Theorem A.1. *Assume that the optimal value of the problem (9) is finite. If any of the following conditions are satisfied*

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

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

ii) The set of optimal distributions[3] of (9) is not empty and bounded.

then (9) satisfies strong duality, that is, the optimal value of (9) is equal to

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

Having strong duality is not important if the dual formulation is not in a context of optimization problems where there are more tools to address it. In this case, the dual representation is in a known context, this is achieved with the following corollary.

Corolario A.1.1. *Assume that f and g_i are functions of (9). Suppose that function $F_a(\xi) := h(\xi) - \sum_{i=1}^k a_i(g_i(\xi) - b_i)$ satisfies the hypotheses of the Theorem 1.1 for all $a\in\mathbb{R}^k$, and satisfies any of the conditions i) and ii) of Theorem A.1, then the problem (9) can be rewritten as*

$$\begin{cases} \inf\limits_{a_1,\dots,a_k,\lambda} & \lambda\varepsilon + \dfrac{1}{N}\displaystyle\sum_{i=1}^N s_i \\ \textit{subject to} & \sup\limits_{\xi\in\Xi}\left(h(\xi) - \displaystyle\sum_{i=1}^k a_i(g_i(\xi)-b_i) - \lambda d^p(\xi,\widehat{\xi}_i)\right) \le s_i \quad \forall\, i=1,\dots,N, \\ & \lambda \ge 0. \end{cases} \tag{10}$$

This corollary is a consequence of the Theorem 1.1 and the Theorem A.1. The strategy for the proof of the Theorem A.1 consists in identifying that (9) can be seen as a linear conic problem, and, being in this context, we use results that allow us to infer strong duality, some of these are exposed in [19].

[3] Distributions such that they reach the optimal value.

A.2 Distributionally Robust estimation of the variance of a random variable with known mean using Wasserstein distance

This section is important to proofs of the results that are presented in the following sections of this appendix. 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 optimization problem obtained in this reformulation can be solved, which allows obtaining an explicit optimal value.

Let ζ be a random variable with unknown distribution $\mathbb{P}$ with support $\Xi = \mathbb{R}$, 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. We consider $\widehat{\mathbb{P}}_N := \sum_{i=1}^N \delta_{\widehat{\zeta}_i}$ as the empirical distribution induced by the previous sample, and let $\mathcal{B}_\varepsilon(\widehat{\mathbb{P}}_N)$ be the ball with center in $\widehat{\mathbb{P}}_N$ and radius ε, this ball is defined with respect to 2-Wasserstein distance with cost function d as the euclidean distance in $\mathbb{R}$. Note that if we choose ε such that $\mathbb{P} \in \mathcal{B}_\varepsilon(\widehat{\mathbb{P}}_N)$, then we have

$$\mathrm{Var}_{\mathbb{P}}[\xi] \leq \begin{cases} \sup\limits_{\mathbb{Q} \in \mathcal{B}_\varepsilon(\widehat{\mathbb{P}}_N)} & \mathbb{E}_{\mathbb{Q}}\left[(\zeta - \eta)^2\right] \\ \text{subject to} & \mathbb{E}_{\mathbb{Q}}\left[\zeta\right] = \eta. \end{cases} \tag{11}$$

The problem on the right is what we call the robust distributional estimate of the variance of ζ. However, for some values of ε, the problem on the right may be not feasible, the following statement sets those values.

Proposición A.1. *If* $\varepsilon < \left|\eta - \frac{1}{N}\sum_{i=1}^N \widehat{\zeta}_i\right|$ *then*

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

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[4] in [26] we know that

$$p \leq q \quad \Longrightarrow \quad W_p \leq W_q.$$

In particular, we have that $W_1 \leq W_2$, this implies that

$$W_1\left(\mathbb{Q}, \widehat{\mathbb{P}}_N\right) \leq W_2\left(\mathbb{Q}, \widehat{\mathbb{P}}_N\right) \leq \varepsilon < \left|\eta - \frac{1}{N}\sum_{i=1}^N \widehat{\zeta}_i\right|. \tag{12}$$

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

$$\int_{\Xi\times\Xi} |\zeta - \delta| \Pi(d\xi, d\zeta) < \left|\eta - \frac{1}{N}\sum_{i=1}^N \widehat{\zeta}_i\right|.$$

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

$$\left|\mathbb{E}_{\mathbb{Q}}[\zeta] - \frac{1}{N}\sum_{i=1}^N \widehat{\zeta}_i\right| = \left|\int_{\Xi\times\Xi} (\zeta - \delta)\Pi(d\xi, d\zeta)\right| \leq \int_{\Xi\times\Xi} |\zeta - \delta| \Pi(d\xi, d\zeta).$$

In consequence, For this last and (12) we obtain

$$\left|\mathbb{E}_{\mathbb{Q}}[\zeta] - \frac{1}{N}\sum_{i=1}^N \widehat{\zeta}_i\right| < \left|\eta - \frac{1}{N}\sum_{i=1}^N \widehat{\zeta}_i\right|.$$

From the immediately above and the inverse triangular inequality follows

$$|\eta - \mathbb{E}_Q[\zeta]| = \left|\eta - \tfrac{1}{N}\textstyle\sum_{i=1}^N \widehat{\zeta}_i - \left(\mathbb{E}_Q[\zeta] - \tfrac{1}{N}\textstyle\sum_{i=1}^N \widehat{\zeta}_i\right)\right| \geq \left|\left|\eta - \tfrac{1}{N}\textstyle\sum_{i=1}^N \widehat{\zeta}_i\right| - \left|\mathbb{E}_{\mathbb{Q}}[\zeta] - \tfrac{1}{N}\textstyle\sum_{i=1}^N \widehat{\zeta}_i\right|\right| > 0.$$

Which allows us to conclude that $\mathbb{E}_Q[\zeta] \neq \eta$. □

[4] This is a consequence of Hölder's inequality.

The following theorem establishes an explicit expression for the optimal value of the optimization problem to the right of (11).

Theorem A.2. *Let* $\varepsilon > 0$ *with*[5] $\varepsilon^2 \geq \left(\frac{1}{N}\sum_{i=1}^{N}(\eta - \widehat{\zeta}_i)\right)^2$, *and such that* $\mathbb{P} \in \mathcal{B}_\varepsilon(\widehat{\mathbb{P}}_N)$. *Then, the inequality (11) is satisfied, and the optimal value of*

$$\begin{cases} \sup\limits_{\mathbb{Q}\in\mathcal{B}_\varepsilon(\widehat{\mathbb{P}}_N)} & \mathbb{E}_{\mathbb{Q}}\left[(\zeta-\eta)^2\right] \\ \textit{subject to} & \mathbb{E}_{\mathbb{Q}}\left[\zeta\right] = \eta. \end{cases} \tag{13}$$

is equal to

$$\left(\sqrt{\frac{1}{N}\sum_{i=1}^{N}(\eta-\widehat{\zeta}_i)^2 - \left(\frac{1}{N}\sum_{i=1}^{N}(\eta-\widehat{\zeta}_i)\right)^2} + \sqrt{\varepsilon^2 - \left(\frac{1}{N}\sum_{i=1}^{N}(\eta-\widehat{\zeta}_i)\right)^2}\right)^2.$$

We need the following proposition to prove Theorem A.2.

Proposición A.2. *Let* $\{a_i\}_{i=1}^N \subset \mathbb{R}$, *then*

$$\left(\sum_{i=1}^{N} a_i\right)^2 \leq N\sum_{i=1}^{N} a_i^2.$$

This proposition is a consequence of the Cauchy–Schwarz inequality.

Proof of Theorem A.2. By the Theorem A.1 we have that (13) satisfies strong duality, and its optimal value is equal to

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

Note that $g(\zeta) := (\zeta-\eta)^2 - \beta\zeta + \beta\eta$ satisfies the hypotheses of Theorem 1.1; therefore, this last formulation is equivalent to the semi-infinite optimization program

$$\begin{cases} \inf\limits_{\beta,\,\lambda,\,s_i} & \lambda\varepsilon^2 + \dfrac{1}{N}\displaystyle\sum_{i=1}^{N} s_i \\ \text{subject to} & \sup\limits_{\zeta\in\mathbb{R}}\left((\zeta-\eta)^2 - \beta\zeta + \beta\eta - \lambda\left|\zeta-\widehat{\zeta}_i\right|^2\right) \leq s_i \quad \forall i = 1,\ldots,N \\ & \lambda \geq 0. \end{cases} \tag{14}$$

If $\lambda \leq 1$, then λ is not a optimal value, this is because, in this case, the set

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

is not bounded. For other hand, if $\lambda > 1$, 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 it is supremum of a concave quadratic polynomial, so (14) is equivalent to

$$\begin{cases} \inf\limits_{\beta,\,\lambda,\,s_i} & \lambda\varepsilon^2 + \dfrac{1}{N}\displaystyle\sum_{i=1}^{N} s_i \\ \text{subject to} & \dfrac{\beta^2}{4(\lambda-1)} + \dfrac{\lambda}{\lambda-1}\left(\beta(\eta-\widehat{\zeta}_i) + (\eta-\widehat{\zeta}_i)^2\right) \leq s_i \quad \forall i = 1,\ldots,N \\ & \lambda \geq 1. \end{cases} \tag{15}$$

[5] Imposing this condition on ε is natural since, by the Proposition A.1, it guarantees that (13) is feasible.

In this formulation, we can suppress the variables s_i, which leads to the following optimization problem

$$\begin{cases} \inf\limits_{\lambda,\,\beta} & \lambda\varepsilon^2 + \dfrac{\beta^2}{4(\lambda-1)} + \dfrac{\lambda}{\lambda-1}\left(\dfrac{\beta}{N}\displaystyle\sum_{i=1}^{N}(\eta-\widehat{\zeta}_i) + \dfrac{1}{N}\displaystyle\sum_{i=1}^{N}(\eta-\widehat{\zeta}_i)^2\right) \\ \text{subject to} & \lambda \geq 1. \end{cases} \tag{16}$$

This previous problem can be simplified by analyzing the objective function. Indeed, for a fixed $\beta \in \mathbb{R}$, from Proposition A.2 it follows that next function that has λ as variable is given by

$$\lambda\varepsilon^2 + \frac{\beta^2}{4(\lambda-1)} + \frac{\lambda}{\lambda-1}\left(\frac{\beta}{N}\sum_{i=1}^{N}(\eta-\widehat{\zeta}_i) + \frac{1}{N}\sum_{i=1}^{N}(\eta-\widehat{\zeta}_i)^2\right) \tag{17}$$

is convex for $\lambda \geq 1$. In fact, To demonstrate this convexity it is enough to show that the second derivative with respect to λ of (17) is positive for $\lambda \geq 1$; in that sense, we have that the second derivative is given for

$$\frac{\beta^2 + 4\frac{\beta}{N}\sum_{i=1}^{N}(\eta-\widehat{\zeta}_i) + \frac{4}{N}\sum_{i=1}^{N}(\eta-\widehat{\zeta}_i)^2}{2(\lambda-1)^3}.$$

Since $\lambda \geq 1$, the sign of the last expression is determined by the sign of $\beta^2 + 4\frac{\beta}{N}\sum_{i=1}^{N}(\eta-\widehat{\zeta}_i) + \frac{4}{N}\sum_{i=1}^{N}(\eta-\widehat{\zeta}_i)^2$; however, this expresion is positive for all β because, in terms of β, this is a polynomial with negative discriminant given by

$$\left(\frac{1}{N}\sum_{i=1}^{N}(\eta-\widehat{\zeta}_i)\right)^2 - \frac{1}{N}\sum_{i=1}^{N}(\eta-\widehat{\zeta}_i)^2,$$

It is negative by Proposition A.2, so the polynomial that appears in the second derivative is always positive with respect to β. In addition, note that expression in (17) tends to infinity when $\lambda \to 1^+$ or $\lambda \to \infty$; in consequence, all of the above guarantees that the function with respect to λ in (17) has a minimum value in the region $\lambda \geq 1$. This value is calculated in the usual way, that is, deriving to determine the critical points, then the critical points that are in the region of interest are identified, and these points are evaluated in the objective function to determine which one acts as a minimum; thus, after developing these steps, we have that the minimum value is given by

$$\varepsilon^2 + \frac{\beta}{N}\sum_{i=1}^{N}(\eta-\widehat{\zeta}_i) + \frac{1}{N}\sum_{i=1}^{N}(\eta-\widehat{\zeta}_i)^2 + \sqrt{\varepsilon^2\left(\beta^2 + 4\frac{\beta}{N}\sum_{i=1}^{N}(\eta-\widehat{\zeta}_i) + \frac{4}{N}\sum_{i=1}^{N}(\eta-\widehat{\zeta}_i)^2\right)}.$$

Therefore, (16) can be rewritten as

$$\inf_{\beta\in\mathbb{R}}\left(\varepsilon^2 + \frac{\beta}{N}\sum_{i=1}^{N}(\eta-\widehat{\zeta}_i) + \frac{1}{N}\sum_{i=1}^{N}(\eta-\widehat{\zeta}_i)^2 + \sqrt{\varepsilon^2\left(\beta^2 + 4\frac{\beta}{N}\sum_{i=1}^{N}(\eta-\widehat{\zeta}_i) + \frac{4}{N}\sum_{i=1}^{N}(\eta-\widehat{\zeta}_i)^2\right)}\right). \tag{18}$$

But this infimum can be calculated explicitly by analyzing the objective function because this function is differentiable for all $\beta \in \mathbb{R}$. Calculating this infimum is important since this is a concrete expression of the optimal value of (16) which in turn is optimal value of (13); therefore, to simplify the notation $A := \frac{1}{N}\sum_{i=1}^{N}(\eta-\widehat{\zeta}_i)$ y $B := \frac{1}{N}\sum_{i=1}^{N}(\eta-\widehat{\zeta}_i)^2$, then from the calculation of the infimum in (18) it follows that the optimal value of (13) is

$$\begin{cases} \left(\sqrt{B-A^2} + \sqrt{\varepsilon^2-A^2}\right)^2 & \text{if } A^2 \leq \varepsilon^2 \\ -\infty & \text{if } A^2 > \varepsilon^2. \end{cases} \tag{19}$$

But, by hypothesis, we know that the case $A^2 > \varepsilon$ is not possible; therefore, we have the desired result. □

A.3 The Proof

Before proceeding with the proof of Theorem 2.1, we need the following lemma, this allows us to express the feasible set of problem (6) in terms of finite dimensional variables and not in terms of probability measures.

Lema A.1. *Let ζ be a random variable with unknown distribution $\mathbb{P}$, and let $\widehat{\zeta}_1, \ldots, \widehat{\zeta}_N$ be a sample of ζ. Given $\varepsilon > 0$ we consider $\mathcal{B}_\varepsilon(\widehat{\mathbb{P}}_N)$ as the ball centered in empirical distribution $\widehat{\mathbb{P}}_N$ defined by previous sample, and with radius ε; this ball is defined respect to 2-Wasserstein metric. Then, we have*

$$\inf_{\mathbb{Q}\in\mathcal{B}_\varepsilon(\widehat{\mathbb{P}}_N)} \mathbb{E}_{\mathbb{Q}}[\zeta] = \frac{1}{N}\sum_{i=1}^{N}\widehat{\xi}_i - \varepsilon \quad \textit{and} \quad \sup_{\mathbb{Q}\in\mathcal{B}_\varepsilon(\widehat{\mathbb{P}}_N)} \mathbb{E}_{\mathbb{Q}}[\zeta] = \frac{1}{N}\sum_{i=1}^{N}\widehat{\xi}_i + \varepsilon$$

Proof. We show only the first equality because the second equality can be obtained from the first by a change of variable. Focusing on showing the first equality, note that

$$\inf_{\mathbb{Q}\in\mathcal{B}_\varepsilon(\widehat{\mathbb{P}}_N)} \mathbb{E}_{\mathbb{Q}}[\zeta] = -\sup_{\mathbb{Q}\in\mathcal{B}_\varepsilon(\widehat{\mathbb{P}}_N)} \mathbb{E}_{\mathbb{Q}}[-\zeta].$$

So, we turn our attention to the problem on the right of equality above. In that problem, the objective function is $g(\zeta) := -\zeta$, this function satisfies the hypothesis of Theorem 1.1; therefore, we have

$$\begin{aligned}
\inf_{\mathbb{Q}\in\mathcal{B}_\varepsilon(\widehat{\mathbb{P}}_N)} \mathbb{E}_{\mathbb{Q}}[\zeta] &= -\left\{\begin{array}{ll} \inf\limits_{\lambda\geq 0} & \lambda\varepsilon^2 + \frac{1}{N}\sum_{i=1}^{N} s_i \\ \text{subject to} & \sup_{\zeta\in\mathbb{R}}\left(-\zeta - \lambda\left(\zeta - \widehat{\zeta}_i\right)^2\right) \leq s_i \quad \forall\, i = 1,\ldots,N, \end{array}\right. \\
&= -\left\{\begin{array}{ll} \inf\limits_{\lambda\geq 0} & \lambda\varepsilon^2 + \frac{1}{N}\sum_{i=1}^{N} s_i \\ \text{subject to} & \frac{1}{4\lambda} - \widehat{\zeta}_i \leq s_i \qquad \forall\, i = 1,\ldots,N, \end{array}\right. \\
&= -\inf_{\lambda\geq 0}\left(\lambda\varepsilon^2 + \frac{1}{4\lambda} - \frac{1}{N}\sum_{i=1}^{N}\widehat{\zeta}_i\right) \\
&= -\left(\varepsilon - \frac{1}{N}\sum_{i=1}^{N}\widehat{\zeta}_i\right) \\
&= \frac{1}{N}\sum_{i=1}^{N}\widehat{\zeta}_i - \varepsilon.
\end{aligned}$$

□

Proof of Theorem 2.1. The strategy is to rewrite the feasible set and the objective function of (6). In fact, let $\mathbb{X}$ the feasible set of (6), then, by Lemma A.1, we have

$$\begin{aligned}
\mathbb{X} &= \left\{ x\in\mathbb{R}^m \;\middle|\; \sum_{i=1}^{m} x_i = 1,\ x_i \geq 0,\ \frac{1}{N}\sum_{i=1}^{N}\widehat{\zeta}_i^x - \varepsilon\,\|x\| \geq \mu \right\} \\
&= \left\{ x\in\mathbb{R}^m \;\middle|\; \sum_{i=1}^{m} x_i = 1,\ x_i \geq 0,\ \frac{1}{N}\sum_{i=1}^{N}\langle x, \widehat{\xi}_i\rangle - \varepsilon\,\|x\| \geq \mu \right\}
\end{aligned} \tag{20}$$

Therefore, (6) is equal to

$$\begin{aligned}
\widehat{J}_N &:= \operatorname*{minimize}_{x\in\mathbb{X}} \sup_{\mathbb{Q}\in\mathcal{B}_{\|x\|\varepsilon}(\widehat{\mathbb{P}}_N^x)} \operatorname{Var}_{\mathbb{Q}}\left[\zeta^x\right] \\
&= \operatorname*{minimize}_{x\in\mathbb{X}} \sup_{\eta\geq\mu}\left\{\begin{array}{ll} \sup_{\mathbb{Q}\in\mathcal{B}_{\|x\|\varepsilon}(\widehat{\mathbb{P}}_N^x)} & \operatorname{Var}_{\mathbb{Q}}\left[\zeta^x\right] \\ \text{subject to} & \mathbb{E}_{\mathbb{Q}}\left[\zeta^x\right] = \eta. \end{array}\right.
\end{aligned} \tag{21}$$

But, by Proposition A.1, we have

$$\widehat{J}_N = \underset{x\in\mathbb{X}}{\text{minimizar}} \quad \sup_{\substack{\eta\geq\mu,\\ \left(\eta-\frac{1}{N}\sum_{i=1}^N \widehat{\zeta}_i^x\right)^2\leq\varepsilon^2\|x\|^2}} \quad \begin{cases} \sup_{\mathbb{Q}\in\mathcal{B}_{\|x\|\varepsilon}(\widehat{\mathbb{P}}_N^x)} & \mathrm{Var}_{\mathbb{Q}}\left[\zeta^x\right] \\ \text{subject to} & \mathbb{E}_{\mathbb{Q}}\left[\zeta^x\right]=\eta. \end{cases} \tag{22}$$

By Theorem A.2, the first maximization problem in (22), which has a variance as its objective function, can be rewritten. This allows us to affirm that (22) is equivalent to

$$\widehat{J}_N = \underset{x\in\mathbb{X}}{\text{minimizar}} \begin{cases} \sup\limits_{\eta\geq\mu} & \left(\sqrt{\frac{1}{N}\sum\limits_{i=1}^N\langle x,\widehat{\xi}_i\rangle^2-\frac{1}{N^2}\left(\sum\limits_{i=1}^N\langle x,\widehat{\xi}_i\rangle\right)^2}+\sqrt{\varepsilon^2\|x\|^2-\left(\mu-\frac{1}{N}\sum\limits_{i=1}^N\langle x,\widehat{\xi}_i\rangle\right)^2}\right)^2 \\ \text{subject to} & \left(\eta-\frac{1}{N}\sum\limits_{i=1}^N\widehat{\zeta}_i^x\right)^2\leq\varepsilon^2\|x\|^2 \end{cases} \tag{23}$$

But, note that the internal maximization problem of (23) can be explicitly solved; actually, this problem reaches its optimal value at $\eta^*=\frac{1}{N}\sum_{i=1}^N\widehat{\zeta}_i^x$. Therefore, (23) can be rewritten as

$$\begin{aligned}\widehat{J}_N &= \underset{x\in\mathbb{X}}{\text{minimizar}} \left(\sqrt{\frac{1}{N}\sum_{i=1}^N\langle x,\widehat{\xi}_i\rangle^2-\frac{1}{N^2}\left(\sum_{i=1}^N\langle x,\widehat{\xi}_i\rangle\right)^2}+\varepsilon\|x\|\right)^2 \\ &= \begin{cases} \underset{x\in\mathbb{R}^m}{\text{minimizar}} & \left(\sqrt{\frac{1}{N}\sum\limits_{i=1}^N\langle x,\widehat{\xi}_i\rangle^2-\frac{1}{N^2}\left(\sum\limits_{i=1}^N\langle x,\widehat{\xi}_i\rangle\right)^2}+\varepsilon\|x\|\right)^2 \\ \text{subject to} & \sum\limits_{i=1}^N\langle x,\widehat{\xi}_i\rangle-\varepsilon\|x\|\geq\mu, \\ & \sum\limits_{i=1}^m x_i=1. \\ & x\geq 0. \end{cases}\end{aligned}$$

□

Proof of Corollary 2.1.1. After some extensive calculations, but without difficulty, we obtain the following

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

By Proposition A.2, we have that E is positive semidefinite, then E has LDL factorization, that is, there are an upper triangular matrix L, a diagonal matrix D, and a permutation matrix P, such that $E=(P^{-1})^TLDL^TP^{-1}$. We define [6] $K:=(P^{-1})^TLD^{1/2}$. Therefore, (7) is equivalent to (8). □

Proof of Corollary 2.1.2. Note that (8) is feasible if there are μ, ε and x such that $\sum_{i=1}^m x_i=1$, $x\geq 0$, $\varepsilon>0$, and $\mu\geq Lx-\varepsilon\|x\|$. But this last inequality is equivalent to $\varepsilon\leq\frac{Lx-\mu}{\|x\|}$, this implies $Lx-\mu>0$. Therefore, (8) is feasible if

$$\mu<\widehat{\mu}_N^{\max}:=\begin{cases}\sup\limits_{x\in\mathbb{R}^m} & Lx \\ \text{subject to} & \sum\limits_{i=1}^m x_i=1, \\ & x\geq 0.\end{cases} \quad \text{and} \quad \varepsilon\leq\widehat{\varepsilon}_N^{\max}(\mu):=\begin{cases}\sup\limits_{x\in\mathbb{R}^m} & \frac{Lx-\mu}{\|x\|} \\ \text{subject to} & Lx\geq\mu \\ & \sum\limits_{i=1}^m x_i=1, \\ & x\geq 0.\end{cases}$$

□

[6] If E were positive definite, then we would have the Cholesky factorization.