RESEARCH STATEMENT

Diego Fernando Fonseca Valero
df.fonseca@uniandes.edu.co
Universidad de los Andes
Bogotá, Colombia

My research interest is stochastic optimization, so my object of study is optimization problems where the objective function or the constraints depend on a random factor that is generally a random vector, that is, optimization problems with probabilistic constraints. My interest is in the case where the probability distribution that governs the randomness of the problem is unknown. In this case, it is assumed that all we have of the random vector is a sample of data. Therefore, the intention is to find an approximate solution to the stochastic optimization problem using data generated by the unknown probability distribution. The strategy to achieve the latter is to use the Wassertestein distance and the empirical distribution determined by the data sample, this gives rise to the study area known as Distributionally Robust Optimization with Wassertestein metrics (DROW). It is a topic that has had great advances in the last five years. In my current Ph.D., my research has focused on applying the DROW approach to portfolio optimization. Specifically, I have worked on the approach of a DROW version of Markowitz's mean-variance model, and on a chance-constrained optimization approach from a DROW vision. These are my two main research fronts, and my Ph.D. research project is based on them.

1 Distributionally Robust Optimization with Wassertein metric

Stochastic optimization have their origin in 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 was known then the previous problem can be formulated as

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

Usually, $\mathbb{P}$ is unknown, so other ways of addressing this last problem emerge, mostly data-driven, that is, based on samples of the random vector ξ. One of the first methods is based on Monte Carlo approximation presented in [14] and [13], but this is computationally expensive and sensitive to outliers in the data. Thus, the distributionally robust approach arises by considering a set $\mathcal{D}$ of probability distributions in such a way that contains $\mathbb{P}$. This set is known as *ambiguity set*. Hence, 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 $J^* \leq \widehat{J}_{\mathcal{D}}$ when $\mathbb{P}\in\mathcal{D}$. In this approach, the objective function becomes the worst expected cost for the choice of a distribution in this set.

The choice of set $\mathcal{D}$ is a determining factor in the tractability of the problem. There are several ways to define $\mathcal{D}$; for example, [7] and [12] define $\mathcal{D}$ as a set of distributions that are supported in a single point, while [3], [11] and [15] define $\mathcal{D}$ as the set of distributions that satisfy certain restrictions in their moments, or distributions belonging to a certain family of parametric distributions. Another option is to endow the set of probability distributions with a notion of distance, so $\mathcal{D}$ is defined as a ball respect to this distance where the ball is usually 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 problem depends on the notion of distance adopted. Some of the notions of distance used are Burg's entropy used in [18], the Kullback-Leibler divergence used in [6] and the Total Variation distance used in [16]. My interest is focused on using *the Wasserstein distance*, that is, we will define $\mathcal{D}$ as a ball with respect to

[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.

the Wasserstein distance of order[2] p with center in an empirical distribution and radius properly chosen, that is, $\mathcal{D} = \mathcal{B}_\varepsilon\left(\widehat{\mathbb{P}}_N\right)$ where

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

Here $\widehat{\mathbb{P}}_N$ is the empirical distribution determined by a sample $\widehat{\xi}_1, \ldots, \widehat{\xi}_N$ of ξ, Ξ is the support of ξ, and $\mathcal{P}(\Xi)$ is the set of all probability distributions supported in Ξ. One of the most relevant works on this topic is [4]. To highlight the fact that a Wasserstein distance is being used, we use the abbreviation DROW instead of DRO.

There are theoretical reasons, many exposed in [17], and practical reasons that make this distance very appealing. Some of these reasons are the following: the definition of Wasserstein distances makes it convenient to use in optimal transport problems where it is naturally involved, balls defined with Wasserstein distances contain more probability distributions than those defined with other notions of distance, and although Wasserstein distances are defined as a optimization problem, it has a dual representation, this could be technically more convenient.

When $\mathcal{D}$ is defined in terms of Wasserstein distances, the problem (2) is in terms of probability distributions, could be an inconvenience, so it is necessary to find an equivalent formulation of (2). The following Theorem achieves this reformulation.

Theorem 1 (Main Theorem) *Assume that $\mathcal{D} := \mathcal{B}_\varepsilon\left(\widehat{\mathbb{P}}_N\right)$ and assuming that f is upper semicontinuous. Then the problem (2) is equivalent to the optimization problem*

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

This Theorem is formulated and proved in [1]. However, the reformulation (4) was also obtained in [4] and [8] although, in those works, the conditions imposed on (4) are more restrictive.

In many cases, solving problem (4) can be a difficult task. The reason for this is that the suprema that appear in the problem constraints may not be explicitly calculable, and the only way to calculate them is by approximating them by means of a numerical method. Precisely, in my current Ph.d. research, I am working on the formulation of an algorithm that allows solving (4) for this case. Additionally, this algorithm also allows to build the optimal probability distribution of problem (2) from the solution of (4), this is a great advance since in (4) there is no presence of probability distributions.

2 A DROW approach of the mean-variance model

The objective is to find a vector of weights x that guarantees a high expected return $\mathbb{E}_{\mathbb{P}}[\langle x,\xi\rangle]$, but with a low volatility $\mathrm{Var}_{\mathbb{P}}[\langle x,\xi\rangle]$. Therefore, a minimum expected return level μ is established, this means that the only vectors of weights x that are taken into account are those that guarantee an expected return greater than or equal to μ. This view is represented in the following model known as the mean-variance model introduced in 1952 by [9]:

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

where $\xi = (\xi_1,\ldots,\xi_m) \in \mathbb{R}^m$ is the returns vector $x = (x_1,\ldots,x_m) \in \mathbb{R}^m$ is the vector of weights, m is a number of assets, and $\langle\cdot,\cdot\rangle$ the usual inner product of $\mathbb{R}^m$.

This stochastic optimization problem is complicated when the distribution $\mathbb{P}$ is unknown, and instead we have realizations of the random vector ξ, that is, we have a sample $\widehat{\xi}_1,\ldots,\widehat{\xi}_N$ of ξ. This allows us to estimate $\mathbb{P}$ through the empirical distribution $\widehat{\mathbb{P}}_N := \frac{1}{N}\sum_{i=1}^{N}\delta_{\widehat{\xi}_i}$. Precisely, in [9], $\mathbb{P}$ is exchanged for $\widehat{\mathbb{P}}_N$ in (5), this new optimization problem is known as the sample version of (5) or Sample Average Approximation (SAA); however, this version does not offer good out-of-sample performance, which motivates the search for another approach.

[2]The Wasserstein distance depends on a parameter called the order, this parameter influences the complexity of the calculation of this distance. An exact definition of the Wasserstein distance is given in [4].

Working on my Ph.d. research, I have proposed a DROW approach for (5). This approach also gives a preponderant role to an empirical distribution, but, in this case, it depends on x. To understand this, the following conventions are established. For $x \in \mathbb{R}^m$, $\zeta^x := \langle x, \xi \rangle$ which is a random variable. $\mathbb{P}^x$ is the distribution of ζ^x, note that it depends on $\mathbb{P}$, so $\mathbb{P}^x$ is also unknown. Additionally, given $\widehat{\xi}_1, \ldots, \widehat{\xi}_N$ a sample of $\mathbb{P}$, then $\widehat{\zeta}_1^x, \ldots, \widehat{\zeta}_N^x$, defined by $\widehat{\zeta}_i^x := \langle x, \widehat{\xi}_i \rangle$, is a sample of ζ^x. This allows defining the empirical distribution $\widehat{\mathbb{P}}_N^x$ of ζ^x, which is given by $\widehat{P}_N^x := \frac{1}{N}\sum_{i=1}^N \delta_{\widehat{\zeta}_i^x}$. My approach is to solve the optimization problem

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

where the ambiguity set is $\mathcal{B}_{\|x\|\varepsilon}(\widehat{\mathbb{P}}_N^x)$, which is a ball centered in $\widehat{\mathbb{P}}_N^x$ with radius $\|x\|\,\varepsilon$ with respect to Wasserstein distance of order 2, and where $\|\cdot\|$ is the euclidean metric in $\mathbb{R}^n$.

There are other DRO approaches to (5), but none using Wasserstein distances. However, this distance is used in problems similar to (5), the difference is that the risk measure used in these works is not the variance, they use measures such as Var and CVaR. Some of the works that do this are [10] and [4]. Also, the way they use the Wasserstein distance is different from how it is used in my approach because, in (6), the ambiguity set depends on x.

As a result of my research, it was possible to obtain a reformulation of (6) where the variables to be optimized are not probability distributions. Precisely, I showed that (6) can be formulated as a convex optimization problem whose variables to be optimized are finite-dimensional. This result is presented in the following theorem.

Theorem 2 *Let M be the matrix of size $m \times N$ whose columns are the sample vectors $\widehat{\xi}_1, \ldots, \widehat{\xi}_N$, and let $\mathbf{0}, \mathbf{e} \in \mathbb{R}^N$ be the column vectors of zeros and ones respectively. From these conventions, the following matrices are defined*

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

Therefore, (6) is equivalent to the optimization problem

$$\begin{cases} \inf\limits_{x \in \mathbb{R}^m} & \left(\left\|K^T x\right\| + \varepsilon \|x\|\right)^2 \\ \textit{subject to} & Lx - \varepsilon \|x\| \geq \mu, \\ & e^T x = 1. \end{cases} \tag{7}$$

where K is a matrix that depends on E.

Currently and in the time remaining to finish my Ph.d., I am evaluating the out-of-sample performance of the decision made from solving (6), I am also comparing this performance with that shown with other existing methods that also try approximate a solution of (5), and I have found that performance is good for specific values of ε. Indeed, one of the objectives is to establish criteria to choose ε in the model (6), and determine with theoretical justification its influence on the performance of the selected portfolio.

3 A chance-constrained optimization approach from a DROW vision

My research interest is focused on optimization problems with probabilistic constraints, the chance constrained programs illustrate this type of problem very well. A *chance constrained program* is a optimization problem that has the following form:

$$J := \begin{cases} \min\limits_{x \in \mathbb{R}^m} & \langle x, c \rangle \\ \text{subject to} & \mathbb{P}(\xi \in \mathcal{S}(x)) \geq 1 - \beta, \\ & x \in \mathcal{X}. \end{cases} \tag{8}$$

where $c \in \mathbb{R}^m$ is a constant vector, $\xi \in \mathbb{R}^m$ is a random vector and $\mathcal{S}(x) \subset \mathbb{R}^m$ is a set that depend on x.

The problem (8) is a difficult problem because if $\mathbb{P}$ is discrete, then it can be viewed as a mixed-integer optimization problem, and if $\mathbb{P}$ is continuous, then in general it becomes a non-convex optimization problem that can have several local minima. Additionally, in practice, $\mathbb{P}$ is unknown, so one of the first

attempts to estimate a solution is to change $\mathbb{P}$ to its empirical version $\widehat{\mathbb{P}}_N$ supported on a sample of the random vector ξ. But the downside to this attempt is that it leads to a mixed-integer optimization problem and the out-of-sample performance is not good.

In my research, I have formulated a DROW version of (8) for the case where $\mathcal{S}(x) := \{\xi \in \mathbb{R}^m \mid \langle \xi, x\rangle \geq \rho\}$ and $\mathcal{X} = \{x \in \mathbb{R}^m \mid \sum_{i=1}^m x_i = 1\ ,\ x \geq 0\}$, this formulation is as follows:

$$\widehat{J}_N := \begin{cases} \min\limits_{x\in\mathbb{R}^m} & \langle x, c\rangle \\ \text{sujeto a} & \mathbb{Q}_x\left(\widehat{\zeta}^x \geq \rho\right) \geq 1-\beta,\ \ \forall\, \mathbb{Q}_x \in \mathcal{B}_{\|x\|\varepsilon}(\widehat{\mathbb{P}}_N^x) \\ & \sum\limits_{i=1}^m x_i = 1, \\ & x_i \geq 0, \end{cases} \tag{9}$$

where $\widehat{\zeta}^x$, $\widehat{\mathbb{P}}_N^x$ and $\mathcal{B}_{\|x\|\varepsilon}(\widehat{\mathbb{P}}_N^x)$ are the same expressions that were defined in Section 2 to present (6). Recently, we showed that (9) can be expressed as a convex optimization problem, this is stated in the following theorem.

Theorem 3 *The DROW (9) is equivalent to the convex optimization problem*

$$\widehat{J}_N = \begin{cases} \min\limits_{x,\lambda,s,\alpha} & \langle c, \xi\rangle & & \\ \textit{subject to} & \lambda\varepsilon\,\|x\| + \dfrac{1}{N}\sum\limits_{i=1}^N s_i \leq \beta - 1, & & \\ & \alpha_i(\rho - \langle \widehat{\xi}_i, x\rangle) \leq s_i, & \forall\, i = 1,\ldots,N, \\ & -1 \leq s_i, & \forall\, i = 1,\ldots,N, \\ & 0 \leq \alpha_i \leq \lambda, & \forall\, i = 1,\ldots,N, \\ & \sum\limits_{i=1}^m x_i = 1, & \\ & x_i \geq 0. & \end{cases} \tag{10}$$

This approach is not the only one that uses Wasserstein distances to solve (8). In [20], [2], [5], and [19], this type of distance is also used in this problem although it is used differently from how it is used in my approach. In (9), the ambiguity set is $\mathcal{B}_{\|x\|\varepsilon}(\widehat{\mathbb{P}}_N^x)$, and it depends on x. In contrast, in these works, the ambiguity set is $\mathcal{B}_\varepsilon(\widehat{\mathbb{P}}_N)$, and does not depend on x. Additionally, [20], [2], [5], and [19], a reformulation is also posed, but this is a mixed-integer optimization problem. It is known that mixed-integer optimization problems are computationally more complex than convex optimization problems. Therefore, my approach is more computationally tractable.

Currently, I am working on obtaining convex formulations like (10) for more general cases of $\mathcal{S}(x)$. In addition, I am investigating the influence of the parameter ε on the out-of-sample performance of the decision obtained from solving (9), and I try to establish theoretically supported criteria to choose this parameter.

References

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

[2] Zhi Chen, Daniel Kuhn, and Wolfram Wiesemann. Data-driven chance constrained programs over wasserstein balls. -, 2018. Available from Optimization Online.

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

[4] 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.

[5] A.R. Hota, A. Cherukuri, and J. Lygeros. Data-Driven Chance Constrained Optimization under Wasserstein Ambiguity Sets. *2019 American Control Conference (ACC)*, pages 1501–1506, 2019.

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

[7] 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.

[8] 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.

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

[10] 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.

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

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

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

[14] 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.

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

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

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

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

[19] W. Xie. On distributionally robust chance constrained programs with Wasserstein distance. *Mathematical Programming*, 2019.

[20] W. Xie and S. Ahmed. Bicriteria approximation of chance constrained covering problems. *Operations Research*, 68:516–533, 2020.