Distributionally Robust Optimization

David Morton and Ivilina Popova

April 2020

1 Introduction

Consider a nominal stochastic optimization problem:

$$\max_{x \in X} \sum_{\sigma \in \Sigma} q^\sigma f(x, \xi^\sigma). \tag{1}$$

To make this concrete, model (1) could represent a stochastic capital budgeting problem in which we prioritize project selection subject to uncertainty in costs, and the net present value (NPV) of each project as well as uncertainty in resource availability. The goal is to prioritize so as to maximize expected NPV, assuming that the nominal distribution, specified by the probability mass function q^σ, $\sigma \in \Sigma$, is correct.

Suppose that ξ is a discrete random variable with finite sample space Ω, so that ξ^ω, $\omega \in \Omega$, are the only possible realizations. Further suppose that we only have observations of ξ^σ, for $\sigma \in \Sigma \subset \Omega$, which may arise in a data-driven setting with, for example, probability mass $q^\sigma = 1/|\Sigma|$.

A distributionally robust optimization variant of model (1) is then given by:

$$\max_{x \in X} \min_{p \in \mathcal{P}} \sum_{\omega \in \Omega} p^\omega f(x, \xi^\omega). \tag{2}$$

Here, we are playing a "game" against nature. First, we select x, and then knowing x, nature selects a worst-case probability distribution, $p \in \mathcal{P}$, to minimize the expected NPV. We will make precise what we mean by $\mathcal{P}$ in the next section, but it will represent a neighborhood of probability distributions centered on the given probability mass function, q, with the radius of the neighborhood specified by parameter ε. If $\varepsilon = 0$ then model (2) reduces to model (1). If ε is very large then nature will select the single worst-case scenario; e.g., the scenario with lowest budgets, highest costs, and lowest NPVs. This is too conservative to be useful. However, with moderate values of ε we obtain solutions that hedge against deviations from q without being excessively conservative.

Importantly, we do not view nature as malevolent, despite occasional evidence to the contrary. Rather, we use "$\min_{p \in \mathcal{P}}$" to combat over-adapting our solution to a specific assumption about the probability distribution. In this sense, it plays the role of a "regularizer" to combat over-fitting that is analogous to regularizers in high-dimensional statistics and statistical machine learning.

2 Wasserstein Distance

We define $d_{\sigma,\omega} = \text{dist}(\xi^\sigma, \xi^\omega)$, $\sigma \in \Sigma, \omega \in \Omega$, where $\text{dist}(\cdot,\cdot)$ could, for example, be the two-norm distance, or a more general η-norm distance, between the two vectors, ξ^σ and ξ^ω, $\text{dist}(\cdot,\cdot) = \|\xi^\sigma - \xi^\omega\|_\eta$.

There are multiple ways to measure the "distance" between two probability distributions, which include the notions of Kolmogorov-Smirnov distance, KL-divergence, Chi-squared distances, total variation, more general ϕ-divergences, etc. The Wasserstein distance—based on the idea of optimal transport—is a particularly useful way to measure such a distance in the context of distributionally robust optimization. For a distribution with known finite support, the Wasserstein distance, $D(q,p)$, between a given distribution, q^σ, $\sigma \in \Sigma$, and a given candidate robust distribution, p^ω, $\omega \in \Omega$, is given by the optimal value of the transportation problem:

$$\begin{aligned}
D(q,p) = \quad \min_z \quad & \sum_{\sigma \in \Sigma, \omega \in \Omega} d_{\sigma,\omega} z_{\sigma,\omega} && \text{(3a)}\\
\text{s.t.} \quad & \sum_{\omega \in \Omega} z_{\sigma,\omega} = q^\sigma, \quad \sigma \in \Sigma && \text{(3b)}\\
& \sum_{\sigma \in \Sigma} z_{\sigma,\omega} = p^\omega, \quad \omega \in \Omega && \text{(3c)}\\
& z_{\sigma,\omega} \geq 0, \quad \sigma \in \Sigma, \omega \in \Omega. && \text{(3d)}
\end{aligned}$$

The intuition behind this measure concerns the magnitude of probability mass, q^σ that must be transported distance $d_{\sigma,\omega}$ from vector ξ^σ to vector ξ^ω via variable $z_{\sigma,\omega}$. As one extreme case, if $\Omega = \Sigma$, $p^\omega = q^\omega$, and $d_{\omega,\omega} = 0$ for all $\omega \in \Omega$ then $D(q,p) = 0$.

Given distribution q, we can then define $\mathcal{P} = \{p : D(p,q) \leq \varepsilon, \sum_{\omega \in \Omega} p^\omega = 1, p^\omega \geq 0, \omega \in \Omega\}$ for a given radius ε. Here, we think of $\mathcal{P}$ as a ball, or neighborhood, of probability distributions centered on q, where the neighborhood has radius ε. With $\Sigma \subset \Omega$, if $\varepsilon = 0$ then $\mathcal{P}$ is the singleton $\{q\}$, and larger values of ε lead to increasingly large neighborhoods. In the context of robust optimization, if $\varepsilon = 0$ then we will simply be solving the nominal stochastic optimization model, and as ε grows large we will consider increasingly conservative models.

We can now represent the set $\mathcal{P}$ via the extended-variable set of constraints:

$$\begin{aligned}
& \sum_{\sigma \in \Sigma} \sum_{\omega \in \Omega} d_{\sigma,\omega} z_{\sigma,\omega} \leq \varepsilon && \text{(4a)}\\
& \sum_{\omega \in \Omega} z_{\sigma,\omega} = q^\sigma, \quad \sigma \in \Sigma && \text{(4b)}\\
& \sum_{\sigma \in \Sigma} z_{\sigma,\omega} - p^\omega = 0 \quad \omega \in \Omega && \text{(4c)}\\
& z_{\sigma,\omega} \geq 0, \quad \sigma \in \Sigma, \omega \in \Omega. && \text{(4d)}
\end{aligned}$$

In other words, $\mathcal{P} = \{p : \exists z \text{ satisfying (4a)–(4d)}\}$.

3 Towards a Computationally Tractable Reformulation

Due to the max-min construct in (2), the model is not amenable to direct solution via optimization software, and so we reformulate the model to facilitate computation. For the moment let $x \in X$ be fixed so that $f(x, \xi^\omega)$ is just a known numerical value for each $\omega \in \Omega$. Then, nature's problem may written:

$$\begin{aligned}
\min_{p,z} \quad & \sum_{\omega\in\Omega} p^\omega f(x,\xi^\omega) && \text{(5a)}\\
\text{s.t.} \quad & \sum_{\sigma\in\Sigma}\sum_{\omega\in\Omega} d_{\sigma,\omega} z_{\sigma,\omega} \le \varepsilon \;:\; [-\gamma] && \text{(5b)}\\
& \sum_{\omega\in\Omega} z_{\sigma,\omega} = q^\sigma, \quad \sigma\in\Sigma \;:\; [\nu^\sigma] && \text{(5c)}\\
& \sum_{\sigma\in\Sigma} z_{\sigma,\omega} - p^\omega = 0, \quad \omega\in\Omega \;:\; [\beta^\omega] && \text{(5d)}\\
& z_{\sigma,\omega} \ge 0, \quad \sigma\in\Sigma, \omega\in\Omega. && \text{(5e)}
\end{aligned}$$

Note that in formulation (4), q and p are given, and z is the only free variable. In model (5), q^ω and ξ^ω $\forall\omega \in \Omega$ are given, and x is (temporarily) given, too, and in the model nature optimizes over z *and* over p to select a worst-case distribution within radius ε of q.

Taking the dual of the linear program (5), and substituting out the dual variable $\beta^\omega = -f(x, \xi^\omega)$, we obtain the following:

$$\begin{aligned}
\max_{\gamma,\nu} \quad & -\gamma\varepsilon + \sum_{\sigma\in\Sigma} \nu^\sigma q^\sigma && \text{(6a)}\\
\text{s.t.} \quad & -\gamma d_{\sigma,\omega} + \nu^\sigma \le f(x,\xi^\omega), \quad \sigma\in\Sigma, \omega\in\Omega && \text{(6b)}\\
& \gamma \ge 0. && \text{(6c)}
\end{aligned}$$

To gain intuition regarding model (6), consider two extreme cases, $\varepsilon = 0$ and $\varepsilon = \infty$. If $\varepsilon = 0$ then there is no penalty in the objective function for allowing γ to grow large. As γ grows large, constraint (6b) becomes vacuous for all $\sigma \ne \omega$; however, for $\sigma = \omega$ we have $d_{\sigma,\sigma} = \|\xi^\sigma - \xi^\sigma\| = 0$, and hence the constraint reduces to $\nu^\sigma \le f(x, \xi^\sigma)$, and coupled with the objective function the optimal value reduces to $\sum_{\sigma\in\Sigma} q^\sigma f(x, \xi^\sigma)$, as it must with $\varepsilon = 0$. In the other extreme, as ε grows sufficiently large we must have $\gamma = 0$ to avoid a huge penalty in the objective function. Thus, for each $\sigma \in \Sigma$, constraint (6b) reduces to $\nu^\sigma \le f(x, \xi^\omega)$, $\forall\omega \in \Omega$; i.e., $\nu^\sigma = \min_{\omega\in\Omega} f(x, \xi^\omega)$ so that ν^σ takes on the lowest NPV for all σ, and the objective function reduces to

$$\sum_{\sigma\in\Sigma} \min_{\omega\in\Omega} f(x,\xi^\omega) q^\sigma = \min_{\omega\in\Omega} f(x,\xi^\omega) \sum_{\sigma\in\Sigma} q^\sigma = \min_{\omega\in\Omega} f(x,\xi^\omega).$$

Again, this matches what it must: If $\varepsilon = \infty$ then nature has enough latitude to place probability one on the single worst-case scenario.

4 A Computationally Tractable Reformulation

In Section 3, we fixed the primary decision variables, x, and took the dual of problem (5). The reason for doing so was to overcome the max-min structure in model (2). With the inner min reformulated as a max we can now formulate a single large maximization problem as follows:

$$\begin{aligned}
\max_{x,\gamma,\nu} \quad & -\gamma\varepsilon + \sum_{\sigma\in\Sigma} \nu^{\sigma} q^{\sigma} && \text{(7a)}\\
\text{s.t.} \quad & x \in X && \text{(7b)}\\
& -\gamma d_{\sigma,\omega} + \nu^{\sigma} \leq f(x, \xi^{\omega}), \quad \sigma \in \Sigma, \omega \in \Omega && \text{(7c)}\\
& \gamma \geq 0. && \text{(7d)}
\end{aligned}$$

Model (7) provides the general formulation of a distributionally robust optimization model, which applies to any stochastic optimization model of form (1). and we have our tractable model. However, often $f(x, \xi^{\omega})$ and $x \in X$ are shorthand for constructs in another model, and so in the next section we specify this for the stochastic capital budgeting model.

5 Tractable Reformulation Specialized to Stochastic Capital Budgeting

The non-robust version of our stochastic capital budgeting model is specified below.

Indexes and sets:

$t \in T$	time periods (years)
$i, i^{'}, i^{''} \in I$	candidate projects
$j, j^{'} \in J_i$	options for selecting project i
$i^{'}, j^{'} \in IJ_{ij}$	piggybacking situations
$k \in K$	types of resources
$\omega \in \Omega$	scenarios

Data:

a_{ij}^{ω}	reward of selecting project i via option j under scenario ω
b_{kt}^{ω}	available budget for a resource of type k in year t under scenario ω
c_{ijkt}^{ω}	consumption of resource of type k in year t if project i is performed via option j under scenario ω

Decision variables:

x_{ij}^{ω}	1 if project i is selected via option j under scenario ω; 0 otherwise
$s_{ii'}$	1 if project i has no lower priority than project $i^{'}$; 0 otherwise
y_i^{ω}	1 if project i is selected for *some* option under scenario ω; 0 otherwise
z_{ij}	1 if project i is selected via option j under *some* scenario; 0 otherwise

Model formulation:

$$\max_{s,x,y,z} \quad \sum_{\omega\in\Omega} q^{\omega} \sum_{i\in I} \sum_{j\in J_i} a^{\omega}_{ij} x^{\omega}_{ij} \tag{8a}$$

$$\text{s.t.} \quad s_{ii'} + s_{i'i} \geq 1,\ i < i', i, i' \in I \tag{8b}$$

$$y^{\omega}_{i} \geq y^{\omega}_{i'} + s_{ii'} - 1,\ i \neq i', i, i' \in I, \omega \in \Omega \tag{8c}$$

$$\sum_{i\in I} \sum_{j\in J_i} c^{\omega}_{ijkt} x^{\omega}_{ij} \ \leq b^{\omega}_{kt},\ k \in K, t \in T, \omega \in \Omega \tag{8d}$$

$$\sum_{j\in J_i} x^{\omega}_{ij} = y^{\omega}_{i}, i \in I,\ \omega \in \Omega \tag{8e}$$

$$y^{\omega}_{i} = 1,\ i \in I, \omega \in \Omega \tag{8f}$$

$$x^{\omega}_{i'j'} \leq x^{\omega}_{ij}, \quad \left(i', j'\right) \in IJ_{ij}, j \in J_i, i \in I \tag{8g}$$

$$s_{ii'} \ + s_{i'i''} + s_{i''i} \leq 2,\ i \neq i', i' \neq i'', i'' \neq i,\ i,\ i', i'' \in I \tag{8h}$$

$$s_{ii'} + s_{i'i} \leq 1,\ i < i', i, i' \in I \tag{8i}$$

$$x^{\omega}_{i'j} + s_{ii'} \ - 1 \leq \sum_{j' \in J_i : j' \leq j} x^{\omega}_{ij}\ ,\ \ i \neq i'\ i,\ i' \in I,\ j \in J_{i'},\ \omega \in \Omega \tag{8j}$$

$$\sum_{\omega\in\Omega} x^{\omega}_{ij} \leq |\Omega|\ z_{ij}, i \in I, j \in J_i \tag{8k}$$

$$\sum_{j\in J_i} z_{ij} \leq \ 1,\ i \in I \tag{8l}$$

$$s_{ii'}, x^{\omega}_{ij}, y^{\omega}_{i}, z_{ij} \in \{0, 1\}, i \neq i', i, i' \in I, j \in J_i, \omega \in \Omega \tag{8m}$$

For simplicity, in what follows we will say that variable $s_{ii'} = 1$ means that project i is higher priority than i' even though the variable definition allows for ties, i.e., the projects being the same priority. Constraint (8b) indicates that either project i is higher priority than project i' or vice versa, and further allows both (i.e., a tie). Constraint (8c) indicates that if project i is higher priority than project i', i.e., $s_{ii'} = 1$, then if we select the lower priority project *under some option* then we must also select the higher priority project; if $s_{ii'} = 0$ or if $y^{\omega}_{i'} = 0$ then the constraint is vacuous. Constraint (8d) requires that we be within budget in each time period, for each resource type, and under each scenario. Constraint (8e) defines binary variable y^{ω}_{i} and simultaneously ensures that we select project i via at most one option. Constraint (8f) ensures that we select all must-do projects. We note that this illustrates the alternative to the situation in which we must include the *Do Nothing* option among the alternatives for optional projects. Constraint (8g) captures piggybacking conditions. Constraints (8i)-(8h) require that we produce a total ordering of the projects rather than allowing for ties. If we remove constraints (8i)-(8h) then it will not change the optimal NPV that we obtain, but including the constraints can facilitate easier parsing of the solutions. Constraint (8j) is a type of consistency constraint with respect to the notion of options; the constraint matters only when project i is higher priority than project i' $s_{ii'} = 1$. In this case, if we select Plan A for the lower priority project then we must select plan A for the higher priority project. If we select Plan B for the lower priority project, then we can select Plan A or Plan B for the higher priority project. And, if we select Plan C for the lower priority project then we can select Plan A, B, or C for the higher priority project. Inclusion of constraint (8j) is optional and reflects how the decision maker prefers to interpret the notion of priorities. Constraints (8k) and (8l) taken together indicate that, for each project separately, we cannot mix use of Plans A, B, and C across different scenarios.

So, model (8) has considerable detail, complications including integer constraints, etc. This, however, serves to illustrate how our generic distributionally robust model of form (7) applies. In particular, the x in model (7) now involves (s, x, y, z); the $x \in X$ in constraint (7b) now involves constraints (8b)-(8m); and, the $f(x, \xi^\omega)$ on the right-hand side of (7c) is replaced by the expression in (8a). This yields:

$$\max_{s,x,y,z,\gamma,\nu} \quad -\gamma\varepsilon + \sum_{\sigma\in\Sigma} \nu^\sigma q^\sigma \tag{9a}$$

$$\text{s.t.} \quad -\gamma d_{\sigma,\omega} + \nu^\sigma \le \sum_{i\in I} \sum_{j\in J_i} a_{ij}^\omega x_{ij}^\omega, \quad \sigma \in \Sigma, \omega \in \Omega \tag{9b}$$

$$s_{ii'} + s_{i'i} \ge 1,\ i < i^{'}, i, i^{'} \in I \tag{9c}$$

$$y_i^\omega \ge y_{i'}^\omega + s_{ii'} - 1,\ i \ne i^{'}, i, i^{'} \in I, \omega \in \Omega \tag{9d}$$

$$\sum_{i\in I} \sum_{j\in J_i} \ c_{ijkt}^\omega x_{ij}^\omega \ \le b_{kt}^\omega,\ k \in K, t \in T, \omega \in \Omega \tag{9e}$$

$$\sum_{j\in J_i} \ x_{ij}^\omega = y_i^\omega, i \in I,\ \omega \in \Omega \tag{9f}$$

$$y_i^\omega = 1,\ i \in I, \omega \in \Omega \tag{9g}$$

$$x_{i'j'}^\omega \le x_{ij}^\omega, \quad \left(i^{'}, j^{'}\right) \in IJ_{ij}, j \in J_i, i \in I \tag{9h}$$

$$s_{ii'} \ + s_{i'i''} + s_{i''i} \le 2,\ i \ne i^{'}, i^{'} \ne i^{''}, i^{''} \ne i,\ i,\ i^{'}, i^{''} \in I \tag{9i}$$

$$s_{ii'} + s_{i'i} \le 1,\ i < i^{'}, i, i^{'} \in I \tag{9j}$$

$$x_{i'j}^\omega + s_{ii'} \ - 1 \le \sum_{j^{'}\in J_i : j^{'} \le j} x_{ij}^\omega \ , \quad i \ne i^{'}\ i,\ i^{'} \in I,\ j \in J_{i'},\ \omega \in \Omega \tag{9k}$$

$$\sum_{\omega\in\Omega} \ x_{ij}^\omega \le |\Omega| \ z_{ij}, i \in I, j \in J_i \tag{9l}$$

$$\sum_{j\in J_i} z_{ij} \le \ 1,\ i \in I \tag{9m}$$

$$s_{ii'}, x_{ij}^\omega, y_i^\omega, z_{ij} \in \{0, 1\}, i \ne i^{'}, i, i^{'} \in I, j \in J_i, \omega \in \Omega \tag{9n}$$

$$\gamma \ge 0. \tag{9o}$$

6 Notes

- How do we select the radius ε? One way is to do out-of-sample testing in the spirit of what Congjian ran last year. In particular, we may have Σ represent the sample space we use to solve the DRO model, and we could use $\Sigma = \Omega$ in that model. However, we really have a very large out-of-sample set of scenarios, and we can assess performance of a given prioritization scheme in an out-of-sample manner using simulation. If things work well, as ε grows from zero, we may have improved performance out-of-sample, but as ε grows too large out-of-sample performance will degrade because our selection is too conservative.

- So, the Wasserstein metric allows us to capture the distance between ξ^σ and ξ^ω when moving the probability mass. And, it also allows us to not be restricted to the situation in which we only have one set of realizations. Here, we can't have Ω be a continuum of realizations and still have computational tractability, but we can have $\Omega \supset \Sigma$ and in principle we could have $|\Omega| \gg |\Sigma|$.

- There are a number of references, and I haven't tried to enumerate them here. Rahimian and Mehrotra [2] have an excellent review, and we have some related work [1]. References cited within these [1, 2] are enough to pull the relevant thread of the literature.

References

[1] D. Duque and D. P. Morton. *Distributionally Robust Stochastic Dual Dynamic Programming*, 2019. Available at http://www.optimization-online.org/DB_HTML/2019/12/7539.html.

[2] H. Rahimian and S. Mehrotra. *Distributionally Robust Optimization: A Review*, 2019. Available at https://arxiv.org/abs/1908.05659.