Notes on Distributionally Robust Optimization

École Polytechnique Fédérale de Lausanne

Prepared By: Ru Zhang

College of Management of Technology
Station 5,
1015 Lausanne,
Switzerland

Ru.Zhang@epfl.ch

Contents

PART

I

Ambiguity Sets

An ambiguity set $\mathcal{P}$ is a family of probability distributions on a common measurable space. Throughout this paper we assume that $\mathcal{P} \subseteq \mathcal{P}(\mathcal{Z})$, where $\mathcal{P}(\mathcal{Z})$ denotes the entirety of all Borel probability distributions on a closed set $\mathcal{Z} \subseteq \mathbb{R}^d$. This section reviews popular classes of ambiguity sets. For each class, we first give a formal definition and provide historical background information. Subsequently, we exemplify important instances of ambiguity sets and highlight how they are used.

Section 1

Moment Ambiguity Sets

A moment ambiguity set is a family of probability distributions that satisfy finitely many (generalized) moment conditions. Formally, it can thus be represented as

$$\mathcal{P} = \{\mathbb{P} \in \mathcal{P}(\mathcal{Z}) : \mathbb{E}_{\mathbb{P}}[f(Z)] \in \mathcal{F}\}, \tag{1.1}$$

where $f : \mathcal{Z} \to \mathbb{R}^m$ is a Borel measurable moment function, and $\mathcal{F} \subseteq \mathbb{R}^m$ is an uncertainty set. By definition, the moment ambiguity set (1.1) thus contains all probability distributions $\mathbb{P}$ supported on $\mathcal{Z}$ whose generalized moments $\mathbb{E}_{\mathbb{P}}[f(Z)]$ are well-defined and belong to the uncertainty set $\mathcal{F}$. Ambiguity sets of the type (1.1) were first studied by [16, 17] and [19] to establish the sharpness of generalized Chebyshev inequalities. The following subsections review popular instances of the moment ambiguity set.

Subsection 1.1

Support-Only Ambiguity Sets

The support-only ambiguity set contains all probability distributions supported on $\mathcal{Z} \subseteq \mathbb{R}^d$, that is, $\mathcal{P} = \mathcal{P}(\mathcal{Z})$. It can be viewed as an instance of (1.1) with $f(z) = 1$ and $\mathcal{F} = \{1\}$. Any DRO problem with ambiguity set $\mathcal{P}(\mathcal{Z})$ is ostensibly equivalent to a classical robust optimization problem with uncertainty set $\mathcal{Z}$, that is,

$$\inf_{x \in \mathcal{X}} \sup_{\mathbb{P} \in \mathcal{P}(\mathcal{Z})} \mathbb{E}_{\mathbb{P}}[\ell(x, Z)] = \inf_{x \in \mathcal{X}} \sup_{z \in \mathcal{Z}} \ell(x, z).$$

Remark A Dirac distribution, denoted $\delta(z - z^*)$, is a probability distribution that places all its mass at a single point $z^* \in \mathbb{R}^d$. Formally, for any measurable function $f(z)$:

$$\int_{\mathbb{R}^d} f(z)\, \delta(z - z^*)\, dz = f(z^*),$$

with key properties

1. **Support:**
$$\mathrm{Supp}(\delta(z - z^*)) = \{z^*\}.$$
2. **Normalization:**
$$\int_{\mathbb{R}^d} \delta(z - z^*)\, dz = 1.$$

3. **Extreme Point of Probability Space:** Dirac distributions are the extreme points of the space of all probability distributions. Any general distribution $\mathbb{P}$ supported on $\mathcal{Z}$ can be written as a convex combination (integral) of Dirac distributions.

4. **Maximization Property:** For any function $f(z)$:

$$\sup_{\mathbb{P}\in\mathcal{P}(\mathcal{Z})} \mathbb{E}_{\mathbb{P}}[f(Z)] = \sup_{z\in\mathcal{Z}} f(z),$$

where $\mathbb{P} = \delta(z - z^*)$ and $z^* = \arg\max_{z\in\mathcal{Z}} f(z)$.

For any probability distribution $\mathbb{P} \in \mathcal{P}(\mathcal{Z})$, the expected value of the loss function $\ell(x, Z)$ is:

$$\mathbb{E}_{\mathbb{P}}[\ell(x, Z)] = \int_{\mathcal{Z}} \ell(x, z)\, d\mathbb{P}(z),$$

where $\mathbb{P}$ satisfies the constraint $\mathbb{P}(Z \in \mathcal{Z}) = 1$.

In the DRO problem, the inner supremum is:

$$\sup_{\mathbb{P}\in\mathcal{P}(\mathcal{Z})} \mathbb{E}_{\mathbb{P}}[\ell(x, Z)] = \sup_{\mathbb{P}\in\mathcal{P}(\mathcal{Z})} \int_{\mathcal{Z}} \ell(x, z)\, d\mathbb{P}(z).$$

1. **Linearity of Expectation:** Since $\mathbb{E}_{\mathbb{P}}[\ell(x, Z)]$ is a linear functional of the distribution $\mathbb{P}$, the supremum is attained at the extreme points of the convex set $\mathcal{P}(\mathcal{Z})$.

2. **Dirac Distribution:** The extreme points of $\mathcal{P}(\mathcal{Z})$ are Dirac distributions $\delta(z - z^*)$. Thus, the worst-case distribution is:

$$\mathbb{P}^*(z) = \delta(z - z^*), \quad z^* = \arg\max_{z\in\mathcal{Z}} \ell(x, z).$$

3. **Simplification of Expectation:** Substituting $\mathbb{P}^* = \delta(z - z^*)$, the expected value becomes:

$$\mathbb{E}_{\mathbb{P}^*}[\ell(x, Z)] = \int_{\mathcal{Z}} \ell(x, z)\, d\delta(z - z^*) = \ell(x, z^*).$$

Thus:

$$\sup_{\mathbb{P}\in\mathcal{P}(\mathcal{Z})} \mathbb{E}_{\mathbb{P}}[\ell(x, Z)] = \sup_{z\in\mathcal{Z}} \ell(x, z).$$

The DRO problem simplifies as:

$$\inf_{x\in\mathcal{X}} \sup_{\mathbb{P}\in\mathcal{P}(\mathcal{Z})} \mathbb{E}_{\mathbb{P}}[\ell(x, Z)] = \inf_{x\in\mathcal{X}} \sup_{z\in\mathcal{Z}} \ell(x, z).$$

Example

Let's consider a concrete example:
- Support set: $\mathcal{Z} = \{z_1, z_2\}$,
- Loss function:

$$\ell(x, z_1) = (x-1)^2, \quad \ell(x, z_2) = (x+1)^2$$

(1) DRO Calculation:

$$\sup_{\mathbb{P}\in\mathcal{P}(\mathcal{Z})} \mathbb{E}_{\mathbb{P}}[\ell(x, Z)] = \sup_{p_1, p_2 \geq 0, p_1 + p_2 = 1} \left[p_1 \ell(x, z_1) + p_2 \ell(x, z_2)\right].$$

We calculate $\ell(x, z_1) = (x-1)^2$ and $\ell(x, z_2) = (x+1)^2$.
The worst-case distribution $\mathbb{P}$ places all probability on the point that maximizes $\ell(x, z)$:

$$\sup_{\mathbb{P}\in\mathcal{P}(\mathcal{Z})} \mathbb{E}_{\mathbb{P}}[\ell(x,Z)] = \max\left\{\ell\left(x,z_1\right),\ell\left(x,z_2\right)\right\}$$

(2) DRO Becomes Robust Optimization:
Thus, the DRO problem reduces to:

$$\inf_{x\in\mathcal{X}} \sup_{\mathbb{P}\in\mathcal{P}(\mathcal{Z})} \mathbb{E}_{\mathbb{P}}[\ell(x,Z)] = \inf_{x\in\mathcal{X}} \max\left\{\ell\left(x,z_1\right),\ell\left(x,z_2\right)\right\}$$

This is equivalent to:

$$\inf_{x\in\mathcal{X}} \sup_{z\in\mathcal{Z}} \ell(x,z)$$

For a comprehensive review of the theory and applications of robust optimization we refer to [1–8, 12].

If the uncertainty set $\mathcal{Z}$ covers a fraction of $1-\varepsilon$ of the total probability mass of some distribution $\mathbb{P}$, then the worst-case loss $\sup_{z\in\mathcal{Z}} \ell(x,z)$ is guaranteed to exceed the $(1-\varepsilon)$-quantile of $\ell(x,Z)$ under $\mathbb{P}$. This can be achieved by leveraging prior structural information or statistical data from $\mathbb{P}$. For example, $\mathbb{P}(Z\in\mathcal{Z})\geq 1-\varepsilon$ may hold (with certainty) if $\mathcal{Z}$ is an appropriately sized intersection of halfspaces and ellipsoids and if Z has independent, symmetric, unimodal and/or sub-Gaussian components under $\mathbb{P}$ [2, 9, 12, 18, 20]. Alternatively, it may hold (with high confidence) if $\mathcal{Z}$ is constructed from independent samples from $\mathbb{P}$ by using statistical hypothesis tests [10, 11, 21], quantile estimation [15], or learning-based methods [13, 14, 22].

Remark An *uncertainty set* $\mathcal{Z}$ is used to estimate the possible variability of the random variable Z, aiming to minimize the worst-case loss:

$$\sup_{z\in\mathcal{Z}} \ell(x,z),$$

where $\ell(x,z)$ is the loss function, x is the decision variable, and z is the environmental variable.

If the uncertainty set $\mathcal{Z}$ covers at least $1-\varepsilon$ of the probability mass of the distribution $\mathbb{P}$, i.e.,

$$\mathbb{P}(Z\in\mathcal{Z})\geq 1-\varepsilon,$$

then the worst-case loss is guaranteed to exceed the $(1-\varepsilon)$-quantile of the loss $\ell(x,Z)$:

$$\sup_{z\in\mathcal{Z}} \ell(x,z)\geq q_{1-\varepsilon},$$

where $q_{1-\varepsilon}$ is the $(1-\varepsilon)$-quantile of $\ell(x,Z)$, defined as:

$$q_{1-\varepsilon} = \inf\{t\in\mathbb{R} : \mathbb{P}(\ell(x,Z)\leq t)\geq 1-\varepsilon\}.$$

Proof

$$\mathbb{P}(\ell(x,Z)\leq q_{1-\varepsilon})\geq 1-\varepsilon.$$

Then:

$$\mathbb{P}(\ell(x,Z)\leq q_{1-\varepsilon}) = \mathbb{P}(\ell(x,Z)\leq q_{1-\varepsilon}, Z\in\mathcal{Z}) + \mathbb{P}(\ell(x,Z)\leq q_{1-\varepsilon}, Z\notin\mathcal{Z}).$$

Since $\mathbb{P}(Z\notin\mathcal{Z})\leq\varepsilon$, we have:

$$\mathbb{P}(\ell(x,Z)\leq q_{1-\varepsilon}, Z\notin\mathcal{Z})\leq\varepsilon.$$

Thus:

$$\mathbb{P}(\ell(x,Z) \le q_{1-\varepsilon}, Z \in \mathcal{Z}) \ge \mathbb{P}(\ell(x,Z) \le q_{1-\varepsilon}) - \varepsilon \ge 1-\varepsilon-\varepsilon = 1-2\varepsilon.$$

Assume for contradiction that:

$$\sup_{z\in\mathcal{Z}} \ell(x,z) < q_{1-\varepsilon}.$$

Under this assumption, for any $z \in \mathcal{Z}$, $\ell(x,z) < q_{1-\varepsilon}$. This implies:

$$\{Z \in \mathcal{Z}, \ell(x,Z) \ge q_{1-\varepsilon}\} = \emptyset.$$

Thus:

$$\mathbb{P}(\ell(x,Z) \le q_{1-\varepsilon}, Z \in \mathcal{Z}) = \mathbb{P}(Z \in \mathcal{Z}) \ge 1-\varepsilon.$$

This directly contradicts the assumption that $\sup_{z\in\mathcal{Z}} \ell(x,z) < q_{1-\varepsilon}$, as it would imply:

$$\mathbb{P}(\ell(x,Z) \le q_{1-\varepsilon}, Z \in \mathcal{Z}) = 0.$$

Thus, the assumption is false, and we must have:

$$\sup_{z\in\mathcal{Z}} \ell(x,z) \ge q_{1-\varepsilon}.$$

□

If the distribution $\mathbb{P}$ of Z has known structural properties (e.g., independence, symmetry, unimodality, or sub-Gaussian tails), $\mathcal{Z}$ can be designed geometrically. Examples include:

- **Halfspaces**: Linear constraints of the form $a^\top z \le b$,
- **Ellipsoids**: Quadratic constraints such as $(z-\mu)^\top Q^{-1}(z-\mu) \le r^2$.

If $\mathbb{P}$ is unknown but can be estimated from samples, $\mathcal{Z}$ can be constructed using:

- **Hypothesis Tests**: Define regions consistent with the observed data,
- **Quantile Estimation**: Use empirical quantiles to estimate regions containing $1-\varepsilon$ probability,
- **Learning-Based Methods**: Apply machine learning models to infer the high-probability region.

PART

II

Bibliograph

[1] Ben-Tal, A., den Hertog, D., and Vial, J.-P. (2015). Deriving robust counterparts of nonlinear uncertain inequalities. *Mathematical Programming*, 149(1):265–299.

[2] Ben-Tal, A., El Ghaoui, L., and Nemirovski, A. (2009). *Robust Optimization*. Princeton University Press.

[3] Ben-Tal, A. and Nemirovski, A. (1998). Robust convex optimization. *Mathematics of Operations Research*, 23(4):769–805.

[4] Ben-Tal, A. and Nemirovski, A. (1999). Robust solutions of uncertain linear programs. *Operations Research Letters*, 25(1):1–13.

[5] Ben-Tal, A. and Nemirovski, A. (2000). Robust solutions of linear programming problems contaminated with uncertain data. *Mathematical Programming*, 88(4):411–424.

[6] Ben-Tal, A. and Nemirovski, A. (2002). Robust optimization–methodology and applications. *Mathematical Programming*, 92(3):453–480.

[7] Bertsimas, D., Brown, D. B., and Caramanis, C. (2011). Theory and applications of robust optimization. *SIAM Review*, 53(3):464–501.

[8] Bertsimas, D. and den Hertog, D. (2022). *Robust and Adaptive Optimization*. Dynamic Ideas.

[9] Bertsimas, D., den Hertog, D., and Pauphilet, J. (2021). Probabilistic guarantees in robust optimization. *SIAM Journal on Optimization*, 31(4):2893–2920.

[10] Bertsimas, D., Gupta, V., and Kallus, N. (2018a). Data-driven robust optimization. *Mathematical Programming*, 167(2):235–292.

[11] Bertsimas, D., Gupta, V., and Kallus, N. (2018b). Robust sample average approximation. *Mathematical Programming*, 171(1-2):217–282.

[12] Bertsimas, D. and Sim, M. (2004). The price of robustness. *Operations Research*, 52(1):35–53.

[13] Goerigk, M. and Kurtz, J. (2023). Data-driven robust optimization using deep neural networks. *Computers & Operations Research*, 151:Article 106087.

[14] Han, B., Shang, C., and Huang, D. (2021). Multiple kernel learning-aided robust optimization: Learning algorithm, computational tractability, and usage in multi-stage decision-making. *European Journal of Operational Research*, 292(3):1004–1018.

[15] Hong, L. J., Huang, Z., and Lam, H. (2021). Learning-based robust optimization: Procedures and statistical guarantees. *Management Science*, 67(6):3447–3467.

[16] Isii, K. (1960). The extrema of probability determined by generalized moments (I) Bounded random variables. *Annals of the Institute of Statistical Mathematics*, 12(2):119–134.

[17] Isii, K. (1962). On sharpness of Tchebycheff-type inequalities. *Annals of the Institute of Statistical Mathematics*, 14(1):185–197.

[18] Janak, S. L., Lin, X., and Floudas, C. A. (2007). A new robust optimization approach for scheduling under uncertainty: II. Uncertainty with known probability distribution. *Computers & Chemical Engineering*, 31(3):171–195.

[19] Karlin, S. and Studden, W. J. (1966). *Tchebycheff Systems: With Applications in Analysis and Statistics*. Interscience Publishers.

[20] Li, Z., Ding, R., and Floudas, C. A. (2011). A comparative theoretical and computational study on robust counterpart optimization: I. Robust linear optimization and robust mixed integer linear optimization. *Industrial & Engineering Chemistry Research*, 50(18):10567–10603.

[21] Postek, K., den Hertog, D., and Melenberg, B. (2016). Computationally tractable counterparts of distributionally robust constraints on risk measures. *SIAM Review*, 58(4):603–650.

[22] Wang, I., Becker, C., Van Parys, B. P., and Stellato, B. (2023). Learning decision-focused uncertainty sets in robust optimization. *arXiv:2305.19225*.