Division of Operations Research, Osaka Univ.

Data-Driven Distributionally Robust Optimization

Zhicheng Zhang,　D 1
Fujisaki Lab Seminar Camp

September 27, 2021

Optimization under Uncertainty

Robust Optimization

A robust optimization problem can be written as

$$\begin{aligned} \inf_{x \in \mathbb{X}} \quad & c^\top x \\ \text{s.t.} \quad & f(x, \xi) > 0, \quad \forall \xi \in \Xi \end{aligned} \tag{1}$$

where $x \in \mathbb{R}^n$ is the decision variable, c is the given objective direction, $f(x, \xi) : \mathbb{R}^n \times \Xi \mapsto \mathbb{R}$ defining the design constraints and parameterized by uncertainty instances $\xi \in \Xi$;

- Semi-infinite optimization can be solvable under some regularity conditions on uncertainty set and constraints;
- In general, it is intractable; Solution is overly conservatism;
- Take probabilistic (chance) relaxation on constraints.

Chance-constrained Optimization

The general chance-constrained optimization is given by,

$$
\begin{aligned}
(CCO) \quad & \inf_{x \in \mathbb{X}} \quad c^\top x \\
& \text{s.t.} \quad \mathbb{P}\left\{\xi \in \Xi \mid f(x,\xi) > 0\right\} \le \epsilon
\end{aligned}
\tag{2}
$$

where $\mathbb{P}$ is a (known) probability distribution on Ξ, and $\epsilon \in (0,1)$ is the violation of tolerance level.

- Min-max form of CCO: minimize the max cost $\ell(x,\xi)$ with max taken over a reduced uncertainty set $\Xi_\epsilon \subset \Xi$ having probability $\mathbb{P}\{\Xi_\epsilon\} = 1 - \epsilon$, namely,

$$
\inf_{x \in \mathbb{X}} \sup_{\xi \in \Xi_\epsilon} \ell(x,\xi) \tag{3}
$$

- The probability of violation (risk) is defined as: $V(x) = \mathbb{P}\left\{\xi \in \Xi \mid f(x,\xi) > 0\right\} \quad \Rightarrow V(x) \le \epsilon$ (Robustness)

Stochastic Optimization

Stochastic optimization is shown as follows

$$\inf_{x \in \mathbb{X}} \ \mathbb{E}_{\mathbb{P}}\big[\ell(x, \xi)\big] \tag{4}$$

The objective is to find a data-driven solution $\widehat{x}_N$ of (4), constructed using the dataset $\widehat{\Xi} = \{\widehat{\xi}_i\}_{i=1}^N \subset \Xi$, that has a *finite-sample guarantee* given by

$$\mathbb{P}^N\left\{\mathbb{E}_{\mathbb{P}}\big[\ell(x, \xi)\big] \leq \widehat{J}_N\right\} \geq 1 - \beta \tag{5}$$

where $\widehat{J}_N$ might depend on $\widehat{\Xi}$ and $\beta \in (0, 1)$ is the parameter governing $\widehat{x}_N$ and $\widehat{J}_N$. In order to identify the $\widehat{x}_N$ with low $\widehat{J}_N$ and $\beta \in (0, 1)$, the strategy is to design a ambiguity set $\mathcal{P}$.

Distributionally Robust Optimization

The distributionally robust optimization aims to minimize the *worst-case* expected cost according to the well-defined ambiguity set containing all distribution measures, that is

$$\widehat{J}_N : \inf_{x \in \mathbb{X}} \sup_{\mathbb{P} \in \mathcal{P}} \mathbb{E}_{\mathbb{P}}\left[\ell(x, \xi)\right] \tag{6}$$

$$\Longleftrightarrow \inf_{\lambda \geq 0, x \in \mathbb{X}} \left\{ \lambda\gamma^2 + \frac{1}{N} \sum_{i=1}^{N} \max_{\xi \in \Xi} \left(\ell(x, \xi) - \lambda \|\xi - \widehat{\xi}_i\|^2 \right) \right\}$$

- In practice, the "true" probability distribution of uncertain model parameters/data may not be known
- Considers a set of probability distributions (ambiguity set).
- Determines decisions that provide hedging against the worst-case distribution by solving a min-max problem.
- An intermediate approach between *stochastic programming* and traditional *robust optimization*

DRO - Choice of ambiguity set

Moment-based ambiguity set

- Typically do not contain the true distribution.
- Conservative solutions: very different distributions can have the same lower moments and the use of higher moments can be impractical
- May be not guaranteed to converge to the true probability distributions the number of uncertain data tends to infinity.

$$\mathcal{P} \doteq \left\{ \mathbb{P} \in \mathcal{M}(\Xi) \;\middle|\; \begin{array}{l} \mathbb{P}[\xi \in \Xi] = 1 \\ \mathbb{E}_{\mathbb{P}}[\xi] = \mu \\ \mathbb{E}_{\mathbb{P}}\left[(\xi - \mu)(\xi - \mu)^T\right] = \Sigma \end{array} \right\}$$

$$\Longleftrightarrow \int_{\Xi} \begin{bmatrix} \xi \\ 1 \end{bmatrix} \begin{bmatrix} \xi \\ 1 \end{bmatrix}^{\top} d\mathbb{P}(\xi) = \begin{bmatrix} \Sigma + \mu\mu^{\top} & \mu \\ \mu^{\top} & 1 \end{bmatrix}$$

Strong duality condition (inf sup = sup inf) $\Rightarrow$ SDP (Solvable)

Statistical or probabilistic metric-based ambiguity set

Given a radius $\gamma \geq 0$. Define

$$\mathcal{P} \doteq \left\{\mathbb{P} \in \mathcal{M}(\Xi) \mid \rho(\mathbb{P}, \mathbb{P}_0) \leq \gamma\right\} \tag{7}$$

where $\mathbb{P}_0$ is the *reference distribution*, $\rho(\cdot, \cdot)$ denotes the some probabilistic distance between two distributions.

E.g. Let $\mathbb{Q}_1$ and $\mathbb{Q}_2$ be two probability distributions over a space $\mathcal{M}$ such that $\mathbb{Q}_1$ is *absolutely continuous* with respect to $\mathbb{Q}_2$. Then, for a convex function ϕ such that $\phi(1) = 0$, the ϕ-divergence of $\mathbb{Q}_1$ from $\mathbb{Q}_2$ is defined as

$$D_\phi(\mathbb{Q}_1 || \mathbb{Q}_2) = \int_{\mathcal{M}} \phi\left(\frac{d\mathbb{Q}_1}{d\mathbb{Q}_2}\right) d\mathbb{Q}_2, \Rightarrow \sum_i \xi_{2,i} \phi\left(\frac{\xi_{1,i}}{\xi_{2,i}}\right)$$

where $\frac{d\mathbb{Q}_1}{d\mathbb{Q}_2}$ is the Radon–Nikodym derivative of $\mathbb{Q}_1$ w.r.t. $\mathbb{Q}_2$.

- ϕ-divergence is asymmetric, i.e., $D_\phi(\mathbb{Q}_1 || \mathbb{Q}_2) \neq D_\phi(\mathbb{Q}_2 || \mathbb{Q}_1)$
- Kullback-Leibler divergence ($\phi_{KL} = tlogt$), total variation ($\phi_{TV} = \frac{1}{2}|t - 1|$), etc.

Wasserstein Metric (Earth Mover's Distance)

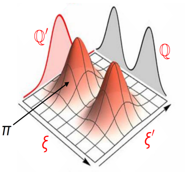
The (p-) Wasserstein metric $\rho_{\mathrm{w}} : \mathcal{M}(\Xi) \times \mathcal{M}(\Xi) \mapsto \mathbb{R}_+$:

$$\rho_{\mathrm{w}}(\mathbb{Q}_1, \mathbb{Q}_2) = \inf_{\Pi \in \mathcal{M}(\Xi^2)} \left\{ \int_{\Xi^2} \|\xi_1 - \xi_2\| \, \Pi(d\xi_1, d\xi_2) \mid \Pi \in \mathcal{H}(\mathbb{Q}_1, \mathbb{Q}_2) \right\}$$

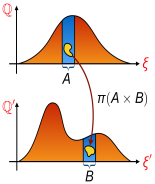
where $\mathcal{H}(\mathbb{Q}_1, \mathbb{Q}_2)$ is the set of all distributions on $\Xi \times \Xi$ with marginals $\mathbb{Q}_1$ and $\mathbb{Q}_2$, and $\|\cdot\|$ represents an *arbitrary* norm. Given $\gamma > 0$, denote $\mathcal{B}_\gamma(\widehat{\mathbb{P}}_N) := \{\mathbb{Q} \in \mathcal{M}(\Xi) \mid \ \rho_{\mathrm{w}}(\widehat{\mathbb{P}}_N, \mathbb{Q}) \leq \gamma\}$, which can be viewed as the Wasserstein ball of radius γ centered at the *empirical distribution* $\widehat{\mathbb{P}}_N := \frac{1}{N}\sum_{k=1}^N \delta_{\widehat{\xi}_i}$. where $\delta_{\widehat{\xi}_i}$ is the unit point mass at $\widehat{\xi}_i$. $\Rightarrow$ $\widehat{\mathbb{P}}_N(\widehat{\xi}_i) = \frac{1}{N}$

- How to guarantee $\mathbb{P}^N\{\rho(\mathbb{P}, \mathbb{P}_0) > \gamma\} \leq \delta$?

  And $\mathbb{P}_0$ can choose empirical distribution $\widehat{\mathbb{P}}_N$
- Convergence: $\mathbb{P}^\infty\{\lim_{N\to\infty} \rho_{\mathrm{w}}(\mathbb{P}, \widehat{\mathbb{P}}_N) = 0\} = 1$.

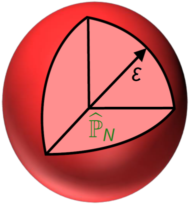


$\Pi(\mathbb{Q}, \mathbb{Q}') =$ set of couplings with marginals $\mathbb{Q}$ and $\mathbb{Q}'$



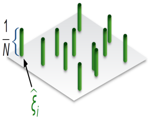
$\pi(A \times B) = \Big\{$ mass moved from source region A to target region B

$\|\xi - \xi'\|^p = \Big\{$ price paid for moving mass from ξ to ξ'



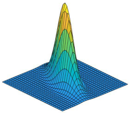
Contains every $\mathbb{Q}$ obtainable by re-shaping $\widehat{\mathbb{P}}_N$ at a cost of at most ε

Non-parametric estimators:



Empirical distribution: $\widehat{\mathbb{P}}_N = \frac{1}{N}\sum_{i=1}^{N} \delta_{\widehat{\xi}_i}$

Parametric estimators:



Elliptical distribution: $\widehat{\mathbb{P}}_N = \mathcal{E}_g(\widehat{\mu}_N, \widehat{\Sigma}_N)$

density generator

Density function: $f(\xi) = C\det(\widehat{\Sigma}_N)^{-1} g((\xi - \widehat{\mu}_N)\widehat{\Sigma}_N^{-1}(\xi - \widehat{\mu}_N))$

Scenario Optimization

Scenario Program (SP)

One considers N i.i.d. random samples of the uncertainty $\{\xi_i\}_{i=1}^N$, and builds a *scenario program* (SP):

$$
\begin{aligned}
(SP) \quad J_N^* = \inf_{x \in \mathbb{X}} \quad & c^\top x \\
\text{s.t.} \quad & f(x, \xi_i) \le 0, \quad i = 1, \cdots, N \\
J_j^* \Leftarrow \ \text{s.t.} \quad & f(x, \xi_i) \le 0, \quad i = \{1, \cdots, N\} \setminus j
\end{aligned}
$$

An optimal solution x_N^* to this problem, if it exists, is a random variable which depends on the multiextraction Ξ^N.

- Min-max form of SP: $\inf_{x \in \mathbb{X}} \ \sup_{i=1,\cdots,N} \ell(x, \xi_i), \quad \xi_i \in \Xi$
- Support constraints: $J_j^* < J_N^*$ (its removal changes solution)
- Goal: Find a data-driven solution $x_N^* \in \mathbb{X}$, such that hold a guarantee $\mathbb{P}^N\{V(x_N^*) < \epsilon\} \ge 1 - \beta$

A Priori Guarantee (*before* obtaining x_N^*)

Given $N \geq n$, a prior guarantees for SP is

$$\mathbb{P}^N\left\{V(x_N^*) > \epsilon\right\} \leq \sum_{i=1}^{n-1} \binom{N}{i} \epsilon^i (1-\epsilon)^{N-i} \leq \beta \tag{8}$$

- Fully-supported problems (the cardinality of support scenarios set $\mathcal{S}$ is exactly n, i.e., $s_N^* = |\mathcal{S}| = n$.)
- Helly's dimension: $\operatorname{ess\,sup}_{\xi \in \Xi^N} |\mathcal{S}(\xi)| \leq h < \bar{h}$, the upper bounds $\bar{h}$ on h is much easier to calculate.
- Sample size N is sufficient (If Large-size: sequential)

Conclude a procedure for a priori feasibility guarantee on SP:

1. Exploring the problem structure of SP and obtain $|\mathcal{S}| = s_N^*$;
2. Select the sample complexity $N(\epsilon, \beta, s_N^*)$ using (8);
3. Obtain optimal solution x_N^* and optimal objective value J_N^*.

If the resulting risk $V(\beta, s_N^*, N) > \epsilon$, we repeat this process with more scenarios until reaching $\epsilon(\beta, s_N^*, N) \leq \epsilon$. If the number of available scenarios is *limited*, then it might be impossible to obtain a solution x_N^* such that $V(x_N^*) \leq \epsilon$. The objective is

$$\mathbb{P}^N\left\{V(x_N^*) < \epsilon(s_N^*)\right\} \geq 1 - \beta. \tag{9}$$

A Posteriori Guarantee (*after* obtaining x_N^*)

A *wait-and-judge* method to give a *a-posteriori* guarantee on sample complexity, which is based on a polynomial equation in variable t, for any $k = 1, 2, \cdots, n$,

$$\frac{\beta}{N+1} \sum_{i=1}^{n-1} \binom{i}{k} t^{i-k} \binom{N}{k} t^{N-i} = 0. \tag{10}$$

has exactly one solution $\epsilon(k) \in (0, 1)$.

- It is not fully-support thus difficult to calculate *a priori* bounds on number of support scenarios($s_N^* = |\mathcal{S}| <?$).
- Sampling point is *insufficient*, it is difficult to meet the sample complexity from the a-priori guarantees.

SP is also related to the *sample average approximation* (SAA). A *sampling-and-discarding* method is that draw N scenarios and discard any k of them, then use the SP with remaining $N-k$ samples, and the associated solution is denoted as $x^*_{N,k}$.
This removal method can be any algorithm that

$$\mathcal{A}\{\xi_1, \cdots, \xi_N\} = \{i_1, \cdots, i_k\}$$

of the k indexes of the k discarded constraints.
The objective is to guarantee that

$$\mathbb{P}^N\left\{V(x^*_{N,k}) < \epsilon\right\} \geq 1 - \beta. \tag{11}$$

Sampling and Discarding (Non-degeneracy)

Given parameters N, ϵ and β, and find the largest removal constraints k such that

$$\binom{k+n-1}{k} \sum_{i=1}^{k+n-1} \binom{N}{i} \epsilon^i (1-\epsilon)^{N-i} \leq \beta \tag{12}$$

holds, then the solution to SAA with $\epsilon = N/k$ is feasible to chance constrained optimization with probability at least $1 - \beta$.

W.p. 1, the SP has a unique irreducible support subsample [1]*, consisting precisely of support constraints.*

Generalization

For an algorithm $\mathcal{A}_N$, let $\beta \in (0,1)$ and $\epsilon : \{0, 1, \cdots, N\} \mapsto [0,1]$ be a function such that

$$\sum_{k=0}^{N} \binom{N}{k} (1 - \epsilon(k))^{N-k} = \beta, \quad \epsilon(N) = 1. \tag{13}$$

Then, for any $\mathcal{A}_N$, $\mathcal{G}_N$, and probability $\mathbb{P}$, it holds that

$$\mathbb{P}^N \{ V(x_N^*) > \epsilon(s_N^*) \} \leq \beta. \tag{14}$$

[1] A *support subsample* $S = (\xi_{i_1}, \cdots, \xi_{i_k})$ with $i_1 < i_2 < \cdots < i_k$ for $(\xi_1, \cdots, \xi_N)$ is a k-tuple of elements extracted from $(\xi_1, \cdots, \xi_N)$, which yields the same solution as the full sample, that is, $\mathcal{A}_k(\xi_{i_1}, \cdots, \xi_{i_k}) = \mathcal{A}_N(\xi_1, \cdots, \xi_N)$. A support subsample is said to to be *irreducible* if no element can be further removed from S without changing the solution. Meanwhile, suppose that exists another $\mathcal{G}_N : \Xi^N \mapsto \{i_1, \cdots, i_k\}$ with $s_N^* = |\mathcal{G}_N(\xi_1, \cdots, \xi_N)|$.

This motivates us to approximate the distributionally CCO

$$\begin{aligned} \inf_{x \in \mathbb{X}} \quad & c^\top x \\ \text{s.t.} \quad & \sup_{\mathbb{P} \in \mathcal{P}} \mathbb{P}\left\{\xi \in \Xi \mid f(x,\xi) > 0\right\} \leq \epsilon \end{aligned} \tag{15}$$

through SP, we call this as DRO via SP, and it can be defined as

$$\begin{aligned} \inf_{x \in \mathbb{X}} \quad & c^\top x \\ \text{s.t.} \quad & f(x,q) > 0, \\ \forall q \quad & \|q - \xi_i\| \leq \gamma, \quad \forall \xi_i \in \Xi, \quad i = 1, \cdots, N \end{aligned} \tag{16}$$

where $\xi_i, i = 1, \cdots, N$ is i.i.d. samples drawn according to the central measure $\mathbb{P}_0$ and the *arbitrary* norm on the space $\mathcal{M}(\Xi)$ is related to the probability metric $\rho(\cdot,\cdot)$.

In min-max case, it is written as

$$\inf_{x\in\mathbb{X}} \sup_{i=1,\cdots,N} \ell(x,\xi_i) \tag{17}$$

where $\{\xi_i\}_i^N$ is an i.i.d. sequence of scenarios randomly sampled from a reference distribution $\mathbb{P}_0$, and the true distribution $\mathbb{P} \in \mathcal{B}_\gamma(\mathbb{P}_0) := \left\{\mathbb{P} \in \mathcal{M}(\Xi) \mid \rho(\mathbb{P},\mathbb{P}_0) \leq \gamma\right\}$.

Based on finite-sample guarantee,

$$\mathbb{P}^N\left\{\rho_{\mathrm{W}}(\mathbb{P},\widehat{\mathbb{P}}_N) \leq \gamma_N\right\} \geq 1-\delta_N$$

consider the relation between

$$\begin{aligned}
J_N^{SP} &= \inf_{x\in\mathbb{X}} \sup_{i=1,\cdots,N} \ell(x,\xi_i) \\
J^{CCO} &= \inf_{x\in\mathbb{X}} \sup_{\xi\in\Xi_\epsilon} \ell(x,\xi) \\
J_N^{DRCCO} &= \inf_{x\in\mathbb{X}} \sup_{\mathbb{P}\in\mathcal{P}} \ell(x,\xi), \quad \mathcal{P} = \mathcal{B}_{\gamma_N(\delta_N)}(\widehat{\mathbb{P}}_N)
\end{aligned}$$