Data-driven Distributionally Robust Optimization

Division of Operations Research   Zhicheng Zhang

***Abstract*— Data-driven robust optimization is the state-of-the-art decision-making program over uncertainty set that a ambiguity uncertainty is constructed to a family of probability distributions, which generating a finite number of training data. The objective is to find a finite-sample guarantee on the related data-driven solution and optimal value of distributionally robust optimization for the designed ambiguity set.**

I. INTRODUCTION

Robust optimization is a decision-making process to handle with the optimization problem in presence of uncertainty. It is worth noting that uncertainty set is a paramount ingredient in robust optimization schema. The key idea is to define an uncertainty set of all possible realizations of uncertain parameters and then hedge against worst-case realizations within this set. However, the description of uncertainty set is usually ill-defined, and hence it may be overly-conservative. In view of this, *stochastic optimization* and *chance-constrained optimization* is presented to model the randomness in uncertain parameters with probability distributions. Unfortunately, this probabilistic formulation in uncertain constraint is generally computationally intractable. Some methods have been developed to compute chance-constrained optimization, such as sample average approximation, Monte-Carlo sampling, and sequential approximation, to name a few.

In recently, a novel data-driven approach for utilizing data to design uncertainty set for robust optimization has been witnessed. This approach is flexible and widely applicable in machine learning, system identification and control synthesis [19], [20]. Calafiore and EI Ghaoui [15] adopted the data-driven method to analyze the chance-constrained and distributionaly robust problems. Bertsimas [1] proposed a data-driven robust optimization based on data constructed uncertainty set and statistical hypothesis testing. Besides, Campi and Garatti [6] proposed a novel data-driven method called scenario optimization to investigate the chance-constrained program. The key idea for dealing with uncertain counterpart in scenario program by sampling a finite number of uncertain realizations, in which there is no knowledge probability distribution to apply the algorithm. The generated data-driven scenario solution is a feasible solution the original the chance-constrained with high probability.

Besides, robust optimization problems where the uncertainty is a probability distribution are called *distributionally robust optimization*, and the corresponding uncertainty sets are called *ambiguity sets*. The key idea of this is to design a uncertainty set estimated from data, where the distribution is exactly not known, or only partial information is available.

This presentation is in Seminar Camp at Fujisaki Lab, Osaka university

This article aims to adopt the data-driven method (scenario optimization) to deal with dostributionally robust optimization using probabilistic metric, and achieve the finite-sample guarantee on sample complexity.

II. AMBIGUITY SET

The construction of ambiguity set plays an important role in the distributionally robust optimization. When designing ambiguity set, we should make a trade-off between tractability, statistical meaning and performance. In practical, a feasible ambiguity set should be computationally tractable in polynomial time, that is, the reformulated optimization structure (e.g., linear, conic, quadratic or semi-defined programs) from data can be solved with the aid of off-the-shelf software. On the other hand, the desired ambiguity set should be rich enough to contain the true data-generating distribution with high confidence. Meanwhile, the devised ambiguity set should be tight enough to exclude the bad distributions as soon as possible such that could refine the performance of the resulting decision.

Fortunately, many impressive results have been displayed that the effective constructed ambiguity set meets the above advantages and performs the solvable mathematical programs. One commonly used method to devising ambiguity set is *moment-based* ambiguity set, where the data-generating distributions satisfy the certain moment constraints (e.g., first second moments or the mean and the covariance). The ambiguity set contains the support, first and second moment information, is shown as follows

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

where $\mathcal{P}$ is the designed ambiguity set, ξ is the random vector, Ξ represents the uncertainty set, $\mathcal{M}(\Xi)$ denotes the space of all probability distributions $\mathbb{P}$ is supported on Ξ, and $\mathbb{E}_{\mathbb{P}}$ is the expectation with respect to distribution $\mathbb{P}$. Parameters μ and Σ stands for the mean vector and covariance matrix estimated from the uncertainty data, respectively.

Unfortunately, the moment-based ambiguity set may be not guaranteed to converge to the true probability distribution as the number of uncertain data tends to infinity. In view of this, an alternative approach called *metric-based* ambiguity sets have been developed, that is, setting ambiguity set as a statistical or probabilistic distance function or a ball form in the space of probability distribution [5], as shown below,

$$\mathcal{P} \doteq \left\{ \mathbb{P} \in \mathcal{M}(\Xi) \;|\; \rho(\mathbb{P}, \mathbb{P}_0) \leq \gamma \right\} \tag{2}$$

where $\mathbb{P}_0$ stands for the *reference distribution* such as empirical distribution, $\rho(\cdot,\cdot)$ denotes the some probabilistic distance between two distributions, and γ is the confidence level. This means that the uncertainty sets are balls centered around a central measure $\mathbb{P}_0$, that is, (2) is equivalent to

$$\mathcal{P} = \mathcal{B}_\gamma(\mathbb{P}_0) \tag{3}$$

In fact, the distance distance can be further specified as the Kullback-Leibler divergence [3], the relative entropy distance [18], partial order [12], Prohorov metric [2], or Wasserstein metric [4].

A. ϕ-divergence

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

$$D_\phi(\mathbb{Q}_1||\mathbb{Q}_2) = \int_{\mathcal{M}} \phi\left(\frac{d\mathbb{Q}_1}{d\mathbb{Q}_2}\right) d\mathbb{Q}_2, \ \sum_i \xi_{2,i}\phi\left(\frac{\xi_{1,i}}{\xi_{2,i}}\right) \tag{4}$$

where $\frac{d\mathbb{Q}_1}{d\mathbb{Q}_2}$ is the Radon–Nikodym derivative of $\mathbb{Q}_1$ with respect to $\mathbb{Q}_2$. It is clear that the ϕ-divergence is asymmetric, and hence it is not a metric. Precisely, it is a useful statistical function (distance) in machine learning and statistical learning. It is depended on the selected functions ϕ, such as Kullback-Leibler divergence ($\phi_{KL} = tlogt$), total variation ($\phi_{TV} = \frac{1}{2}|t-1|$), etc.

B. Wasserstein Ambiguity Set

Here we introduce some basic concepts about Wasserstein metric, which is defined on the space $\mathcal{M}(\Xi)$ of all probability distributions $\mathbb{Q}$ supported on Ξ with finite (p-th) moment, i.e., $\mathbb{E}_{\mathbb{Q}}[\|\xi\|] = \int_\Xi \|\xi\|\mathbb{Q}(d\xi) < +\infty$.

Definition 1 (The Wasserstein metric): The (p-) Wasserstein metric $\rho_{\rm w} : \mathcal{M}(\Xi)\times\mathcal{M}(\Xi) \mapsto \mathbb{R}_+$:

$$\rho_{\rm w}(\mathbb{Q}_1,\mathbb{Q}_2) = \inf_{\Pi\in\mathcal{M}(\Xi^2)} \left\{ \int_{\Xi^2} \|\xi_1-\xi_2\| \ \Pi(d\xi_1,d\xi_2) \ | \right. \tag{5}$$

$$\left. \Pi \in \mathcal{H}(\mathbb{Q}_1,\mathbb{Q}_2) \right\} \tag{6}$$

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 on $\mathbb{R}^n$. Given $\gamma>0$, denote

$$\mathcal{B}_\gamma(\widehat{\mathbb{P}}_N) := \left\{\mathbb{Q}\in\mathcal{M}(\Xi) \mid \ \rho_{\rm w}(\widehat{\mathbb{P}}_N,\mathbb{Q}) \le \gamma\right\}, \tag{7}$$

which can be viewed as the Wasserstein ball of radius γ centered at the *empirical distribution*

$$\widehat{\mathbb{P}}_N := \frac{1}{N}\sum_{k=1}^N \delta_{\widehat{\xi}_i}, \tag{8}$$

where $\delta_{\widehat{\xi}_i}$ is the unit point mass at $\widehat{\xi}_i$. Intuitively, as $N\to\infty$ this data-generating empirical distribution $\widehat{\mathbb{P}}_N$ should tend to[1] the true unknown distribution $\mathbb{P}$ in stochastic optimization, and is related to the *sample average approximation* (SAA) problem. In [4], it provides a bound for ϵ under the *light-tailed distribution* assumption on true distribution $\mathbb{P}$.

Lemma 1 (Measure Concentration, [4]): If the probability $\mathbb{P}$ is *light-tailed distribution*, then

$$\begin{aligned} &\mathbb{P}^N\left\{\rho_{\rm w}(\mathbb{P},\widehat{\mathbb{P}}_N) > \gamma\right\} \\ &\le \begin{cases} c_1\exp(-c_2N\gamma^{\max\{m,2\}}) & \text{if } \gamma\le 1 \\ c_1\exp(-c_2N\gamma^a) & \text{if } \gamma>1 \end{cases} \end{aligned} \tag{9}$$

for all $N\ge 1$, $m=2$ and $\gamma>0$, where c_1, c_2 are positive constants that depend on light-tailed distribution, that is,

$$A := \mathbb{E}_{\mathbb{P}}\left[\exp(\|\xi\|^a)\right] = \int_\Xi \exp(\|\xi\|^a)\mathbb{P}(d\xi) < +\infty$$

Assumption 1 (Sequence of finite-sample guarantees): Sequences $\{\delta_N\}\in(0,1)$, $N\in\mathbb{N}$ and $\{\gamma_N\}\in(0,\infty)$ are such that $\sum_{N=1}^\infty \delta_N < \infty$,$\lim_{N\to\infty}\gamma_N = 0$, and the following finite-sample guarantee hold

$$\mathbb{P}^N\left\{\rho_{\rm w}(\mathbb{P},\widehat{\mathbb{P}}_N)\le\gamma_N\right\} \ge 1-\delta_N \tag{10}$$

Lemma 2 (Convergence of distributions): If the sequence of finite-sample guarantees holds, then

$$\mathbb{P}^\infty\left\{\lim_{N\to\infty}\rho_{\rm w}(\mathbb{P},\widehat{\mathbb{P}}_N)=0\right\} = 1. \tag{11}$$

III. Review of Optimization Under Uncertainty

A. Robust Optimization

A classic robust optimization problem can be written as

$$\begin{aligned} \inf_{x\in\mathbb{X}} \quad & c^\top x \\ \text{s.t.} \quad & f(x,\xi) > 0, \quad \forall\xi\in\Xi \end{aligned} \tag{12}$$

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 ξ. Such a semi-infinite optimization is possible when the uncertainty set and constraints function satisfy some regularity conditions, and the solution is overly-conservatism. In general, the robust optimization is intractable. In view of this, a probabilistic relaxation is proposed in the following.

B. Chance Constrained Optimization

The general chance-constrained optimization is given by,

$$\begin{aligned} \inf_{x\in\mathbb{X}} \quad & c^\top x \\ \text{s.t.} \quad & \mathbb{P}\left\{\xi\in\Xi \mid \ f(x,\xi)>0\right\} \le \epsilon \end{aligned} \tag{13}$$

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

An alternative to the chance-constrained method is called *min-max* or *worst-case*, that is, 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,

$$\inf_{x\in\mathbb{X}} \ \sup_{\xi\in\Xi_\epsilon} \ell(x,\xi) \tag{14}$$

[1]For fixed measurable $A\subset\mathbb{R}^n$, $\widehat{\mathbb{P}}_N(A)\to\mathbb{P}(A)$: Strong law of large numbers; Uniform convergence of $\widehat{\mathbb{P}}_N$ to $\mathbb{P}$: Vapnik–Chervonenkis theory

C. Stochastic Optimization

The stochastic optimization aims to optimize the expected objective value across all the uncertainty realizations and model the randomness in the uncertain parameter with probability distribution. Then it is given by

$$\inf_{x\in\mathbb{X}} \ \mathbb{E}_{\mathbb{P}}\big[\ell(x,\xi)\big] \tag{15}$$

The objective is to find a data-driven solution $\widehat{x}_N$ of (15), constructed using the dataset $\widehat{\Xi} = \{\widehat{\xi}_i\}_{i=1}^N \subset \Xi$, that has a *finite-sample guarantee* given by

$$\mathbb{P}^N\left\{\mathbb{E}_{\mathbb{P}}\big[\ell(x,\xi)\big] \le \widehat{J}_N\right\} \ge 1-\beta \tag{16}$$

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 $\mathcal{P}$. This is related to the distributionally robust optimization.

D. 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

$$\widehat{J}_N: \ \inf_{x\in\mathbb{X}} \ \sup_{\mathbb{P}\in\mathcal{P}} \ \mathbb{E}_{\mathbb{P}}\big[\ell(x,\xi)\big] \tag{17}$$

(Reformulation of DRO (17)*)* For a finite sample $N \in \mathbb{Z}_{\ge 1}$, the optimal value of (17) with the designed ambiguity set $\mathcal{P} = \mathcal{B}_\gamma(\widehat{\mathbb{P}}_N)$ is equal to the optimum of the problem

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

This optimization problem is convex if $x \mapsto \ell(x,\xi)$ is convex for all $\xi \in \Xi$.

E. Ambiguous Chance Constrained Optimization

The ambiguous (distributionally) chance constrained optimization is presented in the following form [15], [16]

$$\begin{aligned} &\inf_{x\in\mathbb{X}} && c^\top x \\ &\text{s.t.} && \sup_{\mathbb{P}\in\mathcal{P}} \ \mathbb{P}\big\{\xi\in\Xi \mid \ f(x,\xi) > 0\big\} \le \epsilon \\ &\quad \text{or} && \inf_{\mathbb{P}\in\mathcal{P}} \ \mathbb{P}\big\{\xi\in\Xi \mid \ f(x,\xi) < 0\big\} \ge 1-\epsilon \end{aligned} \tag{18}$$

F. Scenario Program

Campi and Garatti [6] proposed a novel data-driven method that utilizes a dataset with N scenarios $\{\xi_i\}_{i=1}^N$ to approximate the chance-constrained program (13), which is referred to as the scenario optimization

$$\begin{aligned} &\inf_{x\in\mathbb{X}} && c^\top x \\ &\text{s.t.} && f(x,\xi_i) > 0, \quad i = 1,\cdots,N \end{aligned} \tag{19}$$

In some cases, the scenario program also can be written as the min-max form, that is, min-max of a finite random sample of i.i.d. convex functions, as follows

$$\inf_{x\in\mathbb{X}} \ \sup_{i=1,\cdots,N} \ \ell(x,\xi_i) \tag{20}$$

where $\ell(\cdot,\cdot)$ is a object function.

In this scenario program setup, the objective is to find a data-driven solution of (19), denoted as $x_N^* \in \mathbb{X}$, estimated from the uncertainty set Ξ through a finite number of randomly samples ξ_i, $i = 1,\cdots,N$, such that hold a guarantee

$$\mathbb{P}^N\big\{V(x_N^*) < \epsilon\big\} \ge 1-\beta \tag{21}$$

where the $V(x_N^*)$ is the risk or the probability of violation with a robust level $\epsilon \in (0,1)$, denoted as

$$V(x_N^*) = \mathbb{P}\big\{\xi\in\Xi \mid f(x,\xi) > 0\big\} \tag{22}$$

A natural question in scenario approach is: *How many number of scenarios N should be sampled ?* In other words, what is the smallest N such that $V(x_N^*) < \epsilon$ with high probability. This is related to the *feasibility guarantees*, which can be classified into two categories: *a-priori* and *a-posteriori* guarantees.

Case 1 (A-Priori Guarantee): The a priori ones typically provide priori condition on SP and the number of samples $\{\xi_i\}_{i=1}^N$, the feasibility of corresponding solution x_N^* is guaranteed *before* obtaining x_N^*. In [6], it provides a prior guarantees for SP, that is,

$$\mathbb{P}^N\big\{V(x_N^*) > \epsilon\big\} \le \sum_{i=1}^{n-1}\binom{N}{i}\epsilon^i(1-\epsilon)^{N-i} \le \beta \tag{23}$$

The probability $\mathbb{P}^N$ is taken with respect to N random samples $\{\xi_i\}_{i=1}^N$, and the inequality is tight for fully-supported problems (the cardinality of support scenarios set $\mathcal{S}$ is exactly n, i.e., $s_N^* = |\mathcal{S}| = n$.)

Meanwhile, Calafiore [7] examined the Helly's dimension case of scenario program that for a smallest integer h

$$\operatorname{ess\,sup}_{\xi\in\Xi^N} |\mathcal{S}(\xi)| \le h < \bar{h} \tag{24}$$

holds for any finite $N \ge 1$, where $\mathcal{S}(\xi)$ is the support scenarios on ξ. Therefore, the similar theoretical guarantees on sample complexity can be derived easily by replacing n with Helly's dimension h in (23). However, Helly's dimension is usually hard to compute, while identifying the upper bounds $\bar{h}$ on Helly's dimension h is much easier to calculate.

Based on these, we conclude a procedure for a priori feasibility guarantee on scenario optimization (three steps):

1) Exploring the problem structure of SP and obtain the number of support scenarios $|\mathcal{S}| = s_N^*$;
2) Select the sample complexity $N(\epsilon,\beta,s_N^*)$ using (21);
3) Obtain optimal solution x_N^* and optimal objective value J_N^*.

From the sample complexity viewpoint, we note that when the accuracy level ϵ of violation probability is very small, the number of design parameters is large, then the sample size N may be large. Even if the above a priori approach is work for scenario program, it may be difficult to perform. In order to overcome the *large-size* scenario optimization, a *sequential randomized algorithm* called *sequential probabilistic validation* (SPV) algorithm was proposed to generate a "validation sample set" and to verify the risk of "temporary solution" [8]. In the iteration k-step, the update law is designed as

$$\sum_{i=1}^{n-1} \binom{N}{i} \epsilon^i (1-\epsilon)^{N-i} \leq \frac{\beta}{2} \tag{25}$$

This sequential method claimed that it does not provide a priori bounds on sample complexity.

On the other hand, it is necessary to analyze the *insufficient samples* case, that is, the amount of data is quite limited, since it is more general in many real-world applications, such as, cross validation in machine learning, system identification in data-driven model, and quantitative finance. To this end, the scenario approach is extended to *a posterior* guarantee on sample size.

Case 2 (A Posterior Guarantee): If the resulting violation probability $V(\beta, s_N^*, N)$ is greater than the tolerance level ϵ, where s_N^* is the number of support scenarios. 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 to hold that

$$\mathbb{P}^N \{ V(x_N^*) < \epsilon(s_N^*) \} \geq 1 - \beta. \tag{26}$$

According these analysis, Campi and Garatti [9] presented 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 $k = 1, 2, \cdots, n$,

$$\frac{\beta}{N+1} \sum_{i=1}^{n-1} \binom{i}{k} t^{i-k} \binom{N}{k} t^{N-i} = 0. \tag{27}$$

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

The wait-and-judge is particularly useful in the following cases: (i) the problem is *not* fully-support thus difficult to calculate *a priori* bounds on number of support scenarios; or (ii) only a moderate or small amount of data points is available, it is difficult to meet the sample complexity from the a-priori guarantees.

Similar with the *a-priori* guarantees, we can also perform some steps for a posteriori feasibility guarantee on scenario optimization (three steps):

1) Given a finite scenarios $\{\xi_i\}_{i=1}^N$, solve the scenario problem SP and obtain x_N^*;
2) Find support scenarios in dataset $\{\xi_i\}_{i=1}^N$, whose number is denoted as s_N^*;
3) Calculate the posterior violation probability $\epsilon(\beta, s_N^*, N) \leq \epsilon$ using (27).

In fact, the scenario optimization is also related to the sample average approximation (SAA). In [10], it presented a *sampling-and-discarding* method, that is, draw N scenarios and discard any k of them, then use the scenario approach 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

$$\mathbb{P}^N \{ V(x_{N,k}^*) < \epsilon \} \geq 1 - \beta. \tag{28}$$

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

$$\binom{k+n-1}{k} \sum_{i=1}^{k+n-1} \binom{N}{i} \epsilon^i (1-\epsilon)^{N-i} \leq \beta \tag{29}$$

holds, then the solution to SAA with $\epsilon = N/k$ is feasible to chance constrained optimization with probability at least $1-\beta$. For large-size on sampling data points, the left-hand side of update for SPV in (25) can be replaced by (29).

Note that the above results are all related to the convex scenario optimization, for the non-convex case, *a posterioi* guarantee is much suitable than a priori guarantee on sample complexity. Hence, we still focus on the goal of guarantee which is well-defined in (26).

The main difference between the sampling and discarding in convex SP and non-convex SP is that, the former is asked to satisfy the *non-degeneracy*, while the latter is phrased as: *with probability 1, the problem has a unique irreducible support subsample*[2], *consisting precisely of support constraints*. Therefore, in [11], the authors proved that 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

$$\sum_{k=0}^{N} \binom{N}{k} (1 - \epsilon(k))^{N-k} = \beta, \quad \epsilon(N) = 1. \tag{30}$$

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

$$\mathbb{P}^N \{ V(x_N^*) > \epsilon(s_N^*) \} \leq \beta. \tag{31}$$

Whenever β is split equally among the N terms, it can derive

$$\epsilon(k) = \begin{cases} 1 & \text{if } k = N, \\ 1 - \sqrt[N-k]{\frac{\beta}{N\binom{N}{k}}} & \text{otherwise.} \end{cases} \tag{32}$$

G. Distributionally Robust Scenario Optimization

This motivates us to approximate the distributionally chance constrained optimization through scenario approach, we call this as distributionally robust scenario optimization [2], and it can be defined as

$$\begin{aligned} \inf_{x \in \mathbb{X}} \quad & c^\top x \\ \text{s.t.} \quad & f(x, q) > 0, \\ \forall q \quad & \|q - \xi_i\| \leq \gamma, \quad \forall \xi_i \in \Xi, \quad i = 1, \cdots, N \end{aligned} \tag{33}$$

where ξ_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 (2).

In min-max case, we consider

$$\inf_{x \in \mathbb{X}} \sup_{i=1,\cdots,N} \ell(x, \xi_i) \tag{34}$$

[2]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 the cardinality $s_N^* = |\mathcal{G}_N(\xi_1, \cdots, \xi_N)|$.

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 $\mathbb{P}$ falls into an ambiguity set $\mathcal{B}_\gamma(\mathbb{P}_0) := \big\{\mathbb{P} \in \mathcal{M}(\Xi) \mid \ \rho(\mathbb{P}, \mathbb{P}_0) \leq \gamma\big\}$.

IV. CONCLUSIONS

This article discusses some novel data-driven method to deal with robust optimization. In particular, we concluded a priori and a posteriori guarantee on sample complexity for scenario optimization. In the future work, we want to use the scenario approach to establish finite sample guarantee on distributionally robust optimization under a probabilistic metric-based ambiguity set.

REFERENCES

[1] D. Bertsimas, V. Gupta and N. Kallus, "Data-driven robust optimization," *Math. Program.* 167, 235–292 (2018).

[2] E. Erdoğan and G. Iyengar, "Ambiguous chance constrained problems and robust optimization", *Math. Program,* 107, 37–61, 2006.

[3] Z. Hu and L. J. Hong, "Kullback-Leibler divergence constrained distributionally robust optimization", *Available at Optimization Online*, 2013.

[4] P. M. Esfahani and D. Kuhn, "Data-driven distributionally robust optimization using the Wasserstein metric: performance guarantees and tractable reformulations", *Math. Program*, 171, 115–166, 2018.

[5] A. L. Gibbs and F. E. Su, "On choosing and bounding probability metrics", *Intl. Stat. Rew.*, 7(3), 419–435, 2002.

[6] M. C. Campi and S. Garatti, "The exact feasibility of randomized solutions of uncertain convex programs," *SIAM J. Optim* 19, 1211–1230, 2008.

[7] G. C. Calafiore, "Random convex programs," *SIAM Journal on Optimization* 20, 3427–3464, 2010.

[8] M. Chamanbaz, F. Dabbene, R. Tempo, V. Venkataramanan and Q. Wang, "Sequential randomized algorithms for convex optimization in the presence of uncertainty," *IEEE Trans. Automatic Control*, vol. 61, no. 9, pp. 2565–2571, Sept. 2016,

[9] M. C. Campi and S. Garatti, "Wait-and-judge scenario optimization," *Math. Program*, 1–35, 2016.

[10] M. C. Campi and S. Garatti, " A sampling-and-discarding approach to chance-constrained optimization: feasibility and optimality," *Journal of Optimiz. Theo. Appli,* 148, 257–280, 2011.

[11] M. C. Campi, S. Garatti and F. A. Ramponi, "A general scenario theory for nonconvex optimization and decision making," *IEEE Trans. Automatic Control*, vol. 63, no. 12, pp. 4067–4078, Dec. 2018,

[12] A. Nemirovski and A. Shapiro, "Scenario approximations of chance constraints," *Probabilistic and Randomized Methods for Design under Uncertainty*. Springer, 2006.

[13] R. Gao, A. J. Kleywegt, "Distributionally robust stochastic optimization with wasserstein distance," arXiv:1604.02199v2.

[14] Alexander Shapiro, "Distributionally robust stochastic programming," *SIAM J. Optim*, 227:4, 2258–2275, 2017.

[15] G. C. Calafiore and L. EI Ghaoui, "On Distributionally Robust Chance-Constrained Linear Programs," *J Optim Theory Appl* 130, 1—22 (2006).

[16] S. Y. Guo, H. F. Xu and L. W. Zhang "Convergence Analysis for Mathematical Programs with Distributionally Robust Chance Constraint," *SIAM J. Optim* 27:2, 784-816, 2017.

[17] H. Rahimian and S. Mehrotra "Distributionally robust optimization: A review," arXiv:1908.05659

[18] B. P. G Van Parys, P. M Esfahani and D. Kuhn, "From Data to Decisions: Distributionally Robust Optimization is Optimal" *Management Science*, 2021.

[19] B. P. G. Van Parys, D. Kuhn, P. J. Goulart and M. Morari, "Distributionally robust control of constrained stochastic systems," *IEEE Trans. Automatic Control*, vol. 61, no. 2, pp. 430-442, Feb. 2016.

[20] I. Yang, "Wasserstein distributionally robust stochastic control: A data-driven approach," *IEEE Trans. Automatic Control*, vol. 66, no. 8, pp. 3863-3870, Aug. 2021.