A Distributional Interpretation of Robust Optimization

Huan Xu*, Constantine Caramanis* and Shie Mannor†
*Department of Electrical and Computer Engineering
The University of Texas at Austin, USA
Email: huan.xu@mail.utexas.edu
cmcaram@ece.utexas.edu
† Department of Electrical Engineering
Technion, Israel
Email: shie.mannor@ee.technion.ac.il

Abstract—Motivated by data-driven decision making and sampling problems, we investigate probabilistic interpretations of Robust Optimization (RO). We establish a connection between RO and Distributionally Robust Stochastic Programming (DRSP), showing that the solution to any RO problem is also a solution to a DRSP problem. Specifically, we consider the case where multiple uncertain parameters belong to the same fixed dimensional space, and find the set of distributions of the equivalent DRSP. The equivalence we derive enables us to construct RO formulations that are statistically consistent, and in the process, provides a systematic approach for tuning the uncertainty set.

I. INTRODUCTION

Robust Optimization (RO) considers deterministic (set-based) uncertainty models in optimization, where a (potentially malicious) adversary has a bounded capability to change the parameters of the function the decision-maker seeks to optimize. Thus, the standard optimization problem

\[
\max_v : f(v),
\]

becomes

\[
\max_v : \min_{x \in \mathcal{Z}} f(v, x),
\]

where the vector \( x \) denotes some uncertain parameters of the objective function \( f \), and may take any value in the set \( \mathcal{Z} \).

This approach to uncertainty has a long history in control; in optimization it traces back several decades to the early work in Soyster [1]. Particularly in the last decade since the work of Ben-Tal and Nemirovski [2], [3], Bertsimas and Sim [4], and El Ghaoui and Lebret [5], it has become a common approach in operations research, computer science, engineering, and many other related fields (e.g., Shivaswamy etc [6], Lanckriet etc [7], El Ghaoui and Lebret [5], Ben-Tal etc [8], [9], and Boyd etc [10]); see the recent monograph by Ben-Tal and co-authors [11] for a detailed survey. A key reason for its success has been its computational tractability and the fact that robustified versions of many common optimization classes (linear programming, second order cone programming, among others) remain relatively easy to solve.

A much-researched alternative to RO’s set-based uncertainty, is to represent the uncertain parameter in a probabilistic way, i.e., assume that the parameter \( x \) is a random variable with distribution \( \mu^* \). If we assume the generating distribution, \( \mu^* \), is known, the result is the standard stochastic programming paradigm (e.g., Birge and Louveaux [12], Prékopa [13], and Shapiro [14]). If \( \mu^* \) is not precisely known, and instead \( \mu^* \) is only known to lie in some set of distributions, \( \mathcal{D} \), the resulting optimization formulation is the so-called Distributional Robust Stochastic Program (DRSP), initially proposed in Scarf [15], almost two decades before the first appearance of RO. In DRSP, the decision maker solves the following problem

\[
\max_v : \min_{\mu \in \mathcal{D}} \mathbb{E}_{x \sim \mu} f(v, x).
\]

Since its introduction, DRSP has attracted extensive research (e.g., Kall [16], Dupačová [17], Popescu [18], and Shapiro [19]).

The main focus of this paper is the relationship of these two paradigms. In particular, we show in Section II that RO can be reformulated as a DRSP with respect to a particular class of distributions. For the special case where each uncertain parameter belongs to a different space, such a re-interpretation is a well known folk theorem (also, see Delage and Ye [20]). Yet as we discuss below, in data-driven (or sample-based) optimization problems such as those appearing in stochastic optimization and machine learning, the uncertain parameters belong to the same space.

To develop a framework for these problems, we generalize the equivalence of RO and DRSP to the setting where multiple uncertain parameters \( x_1, \ldots, x_n \) belong to the same space \( \mathbb{R}^m \). Instead of formulating the RO problem for such a problem as a DRSP with respect to a class of distributions supported on the product space \( \mathbb{R}^{m \times n} \), as techniques from the standard literature would require, we seek to find a DRSP interpretation with respect to a class of distributions supported on \( \mathbb{R}^m \). We now elaborate on this.

Relationship to Optimization from Samples: The setup we consider is motivated by sampling problems: in solving

\[
\min_v : \mathbb{E}_{x \sim \mu^*} f(v, x),
\]
a sampled distribution \( (1/n) \sum_{i=1}^{n} \delta_{x_i} \) is often used instead of the true (potentially continuous) distribution \( \mu^* \). This is often done in machine learning because the true distribution is unknown, and the decision-maker has only access to a finite set of samples generated from that distribution. In stochastic programming this is widely applied, either when the distribution is unknown, or when it is known but too complicated to manipulate directly within an optimization problem, and hence the empirical distribution is used instead.

A DRSP interpretation of RO with respect to a class of distributions supported on the same fixed dimensional space would enable us to examine how well or poorly elements of this class of distributions approximate the true (potentially unknown) distribution as the sample size, \( n \), increases. In Section III we explore how such an approach can be used to prove that a robust optimization formulation is asymptotically statistically consistent, and also how to design statistically consistent robust optimization formulations. As a byproduct, we obtain a systematic way to tune the uncertainty set to guarantee consistency.

**Notation:** Throughout the paper, without loss of generality, we assume the unknown parameters belong to \( \mathbb{R}^m \). We use \( \mathcal{P} \) to denote the set of probability distributions of \( \mathbb{R}^m \) (with respect to the Borel set). We use \([1:n]\) to denote the set \( \{1,2,\ldots,n\} \). A Euclidean ball centered at \( x \) with a radius \( r \) is denoted by \( B(x,r) \).

## II. DISTRIBUTIONAL INTERPRETATION FOR ROBUST OPTIMIZATION

In this section we turn our attention to the relationship between RO and DRSP. We consider a general case when multiple uncertain parameters lie in the same space, as opposed to the Cartesian product of the space of each parameter. As discussed above, this setting arises naturally in data-driven problems. We prove a statement that is slightly stronger than the equivalence of RO and DRSP: fixing a candidate solution, the worst case reward of RO is equivalent to the minimal expected reward of DRSP. That is, the inner minimization of RO is equivalent to the inner minimization of DRSP. Hence, in Theorem 1 and the proof, we suppress the decision variable \( u \), in order to reduce unnecessary notation.

**Theorem 1:** Given a measurable function \( f: \mathbb{R}^m \to \mathbb{R} \cup \{-\infty\}, c_1, \ldots, c_n > 0 \) such that \( \sum_{i=1}^{n} c_i = 1 \), and non-empty Borel sets \( Z_1, \ldots, Z_n \subseteq \mathbb{R}^m \), denote

\[
\mathcal{P}_n \triangleq \{ \mu \in \mathcal{P} | \forall S \subseteq [1:n], \mu(\bigcup_{i \in S} Z_i) \geq \sum_{i \in S} c_i \}.
\]

Then the following holds

\[
\inf_{x_i \in Z_i} f(x_i) = \inf_{\mu \in \mathcal{P}_n} \int_{\mathbb{R}^m} f(x) d\mu(x).
\] \hspace{1cm} (4)

Note that the uncertainty sets \( Z_1, \ldots, Z_n \) can have non-empty intersection, or even be identical, as is the case when the points are sampled from the same space.

**Proof:** We can assume without loss of generality that \( \inf_{x_i \in Z_i} f(x_i) > -\infty \) for \( i = 1,2,\ldots,n \), since otherwise both sides equal \(-\infty\) and the theorem holds trivially.

Let \( \tilde{x}_i \) be an \( \epsilon \)-optimal solution to \( \inf_{x_i \in Z_i} f(x_i) \). We first show that the left-hand-side of Equation (4) is larger or equal to the right-hand-side.

Consider the following distribution

\[
\hat{\mu}(\{\tilde{x}_i\}) = c_i + \sum_{j \neq i} c_j.
\]

Observe that \( \hat{\mu} \) belongs to \( \mathcal{P}_n \), and

\[
\sum_{i=1}^{n} c_i f(\tilde{x}_i) = \int_{\mathbb{R}^m} f(x) d\hat{\mu}(x).
\]

Thus we have

\[
\sum_{i=1}^{n} c_i f(\tilde{x}_i) \geq \inf_{\mu \in \mathcal{P}_n} \int_{\mathbb{R}^m} f(x) d\mu(x).
\]

Since \( \epsilon \) can be arbitrarily close to zero, we have

\[
\sum_{i=1}^{n} c_i f(\tilde{x}_i) \geq \inf_{\mu \in \mathcal{P}_n} \int_{\mathbb{R}^m} f(x) d\mu(x).
\] \hspace{1cm} (5)

We next prove the reverse inequality to complete the proof. By re-indexing if necessary, assume

\[
f(\tilde{x}_1) \geq f(\tilde{x}_2) \geq \cdots \geq f(\tilde{x}_n).
\] \hspace{1cm} (6)

Now construct the following function

\[
\hat{f}(x) \triangleq \begin{cases} \max_{i} f(x_i) & \text{if } x \in \bigcup_{j=1}^{n} Z_j; \\ f(x) & \text{otherwise.} \end{cases}
\]

Observe that \( f(x) \geq \hat{f}(x) - \epsilon \) for all \( x \).

Furthermore, fixing a \( \mu \in \mathcal{P}_n \), we have

\[
\int_{\mathbb{R}^m} f(x) d\mu(x) + \epsilon \geq \int_{\mathbb{R}^m} \hat{f}(x) d\mu(x) = \sum_{k=1}^{n} \mu(\bigcup_{i=1}^{k} Z_i) - \mu(\bigcup_{i=1}^{k-1} Z_i).
\]

Here the inequality holds because \( f(x) \geq \hat{f}(x) - \epsilon \), and the equality holds because \( f(\tilde{x}_1) \geq f(\tilde{x}_2) \geq \cdots \geq f(\tilde{x}_n) \).

Define \( \alpha_k \triangleq \mu(\bigcup_{i=1}^{k} Z_i) - \mu(\bigcup_{i=1}^{k-1} Z_i) \). Then, by telescoping and the fact that \( \mu \in \mathcal{P}_n \), we have

\[
\sum_{k=1}^{t} \alpha_k = \mu(\bigcup_{i=1}^{t} Z_i) \geq \sum_{k=1}^{t} c_k; \quad \sum_{k=1}^{n} \alpha_k = \mu(\bigcup_{i=1}^{n} Z_i) = 1.
\]

Hence, since \( f(\tilde{x}_1) \geq f(\tilde{x}_2) \geq \cdots \geq f(\tilde{x}_n) \), we have

\[
\sum_{k=1}^{n} c_k f(\tilde{x}_k) \geq \sum_{k=1}^{n} c_k f(\tilde{x}_k).
\]

Thus, for any \( \mu \in \mathcal{P}_n \), we have

\[
\int_{\mathbb{R}^m} f(x) d\mu(x) + \epsilon \geq \sum_{k=1}^{n} c_k f(\tilde{x}_k) \geq \sum_{k=1}^{n} c_k \inf_{x_i \in Z_i} f(x_i).
\]

Since \( \epsilon \) and \( \mu \) are arbitrary, this leads to

\[
\inf_{\mu \in \mathcal{P}_n} \int_{\mathbb{R}^m} f(x) d\mu(x) \geq \sum_{k=1}^{n} c_k \inf_{x_i \in Z_i} f(x_i).
\]
Combining this with Equation (5), we establish the theorem.

We observe that if \( Z_i \) are disjoint, then the equivalent distributional set \( \mathcal{P}_n \) has the following form:

\[
\mathcal{P}_n = \{ \mu \in \mathcal{P} | \mu(Z_i) = c_i, \ i = 1, \ldots, n \}.
\]

Taking \( n = 1 \), this reduces to the following theorem, well known in the literature (e.g., Delage and Ye [20]), stating that the solution to a robust optimization problem is the solution to a special DRSP problem, where the distributional set is the one that contains all distributions whose support is contained in the uncertainty set, in the Cartesian product of the space of each parameter.

**Corollary 1**: Given a measurable function \( f : \mathbb{R}^m \rightarrow \mathbb{R} \cup \{ -\infty \} \), and a non-empty Borel set \( Z \subset \mathbb{R}^m \), the following holds:

\[
\inf_{x' \in Z} f(x') = \inf_{\mu \in \mathcal{P} | \mu(Z) = 1} \int_{\mathbb{R}^m} f(x) d\mu(x).
\]

**III. Consistency of Robust Optimization**

Besides establishing a theoretical equivalence, Theorem 1 has algorithmic consequences as well. In this section, we consider sampled stochastic optimization problems, and use the equivalence relationship stated in Theorem 1 to construct a sequence of robust optimization problems that are asymptotically consistent in a statistical sense. En route, this construction provides a systematic approach to choosing the appropriate size of the uncertainty set.

The main theorem of the section states that as long as the utility function \( f(\cdot, \cdot) \) is bounded, and satisfies a mild continuity condition, then an explicitly stated robust optimization formulation for the sampled stochastic optimization, is asymptotically consistent. That is, it recovers the optimal solution to

\[
\max_{x} \mathbb{E}_{x \sim \mu}[f(v, x)],
\]

where \( \mu \) is given only through a sequence of i.i.d. samples. We note that these conditions are weaker than those required for consistency of sampled stochastic programs, e.g., as in King and Wets [21]. Thus, Theorem 2 provides a computationally tractable avenue for developing algorithms for solution of stochastic programming with stronger consistency guarantees (that is, they require weaker assumptions on the problem). We provide several examples of this at the end of the section.

**Theorem 2**: Suppose that \( x_1, \ldots, x_n, \ldots \) are i.i.d. samples of a distribution \( h^* \) on \( \mathbb{R}^m \), and the utility function \( f(\cdot, \cdot) \) satisfies

1. **Boundedness** \( \max_{v, x} |f(v, x)| \leq C \).
2. **Equicontinuity** \( d(\epsilon) \downarrow 0 \) where

\[
d(\epsilon) \triangleq \max_{v, x : ||\delta||_\infty \leq \epsilon} |f(v, x) - f(v, x + \delta)|.
\]

If \( \{\epsilon(n)\} \) satisfies

\[
\epsilon(n) \downarrow 0; \quad nc(n)^m \uparrow \infty,
\]

then the sequence of optimal solutions to the RO formulation

\[
v(n) \triangleq \arg\max_v \sum_{i=1}^n \frac{1}{n} \inf_{\delta_i} f(v, x_i + \delta_i)
\]

satisfies

\[
\lim_{n \to \infty} \int_{\mathbb{R}^m} f(v(n), x) h^*(x) dx = \inf_v \int_{\mathbb{R}^m} f(v, x) h^*(x) dx.
\]

That is, the RO formulation is consistent.

**Proof**: The key to the proof rests on the equivalence established in Theorem 1. This equivalence then allows us to show that the RO formulation given in the theorem statement, is equivalent to a DRSP, whose distribution set contains a Kernel Density Estimator (KDE). From here, the proof is essentially immediate: exploiting the fact that a KDE converges to the generating distribution in the \( \ell_1 \) sense, one can easily conclude that the sequence of solutions to the RO problems given, are consistent. For completeness, we include a brief introduction to Kernel Density Estimators in the Appendix.

Consider a fixed \( n \). Define the sets

\[
Z_i \triangleq \{ x_i + \delta : ||\delta||_\infty \leq \epsilon(n) \}.
\]

Thus, the RO formulation is to maximize \( \sum_{i=1}^n \frac{1}{n} \inf_{x'} f(v, x') \). By Theorem 1, we know that for any \( v \), the following holds:

\[
\sum_{i=1}^n \frac{1}{n} \inf_{x \in Z_i} f(v, x_i) = \inf_{\mu \in \mathcal{P}_n} \int_{\mathbb{R}^m} f(v, x) d\mu(x);
\]

where:

\[
\mathcal{P}_n = \{ \mu \in \mathcal{P} | \forall S \subseteq [1 : n] : \mu(\bigcup_{i \in S} Z_i) \geq |S|/n \}.
\]

Next we show that the set of distributions, \( \mathcal{P}_n \), contains a kernel density estimator. Consider a distribution \( h_n \) defined as

\[
h_n(x) = (nc(n)^m)^{-1} \sum_{i=1}^n K \left( \frac{x - x_i}{\epsilon(n)} \right);
\]

where:

\[
K(z) \triangleq \mathbb{1}(||z||_\infty \leq 1).
\]

Indeed, observe that \( h_n \) is a kernel density estimator. Now, for any \( S \subseteq [1 : n] \), we have

\[
\int_{\mathbb{R}^m} 1(x \in \bigcup_{j \in S} Z_j) h_n(x) dx
\]

\[
= \int_{\mathbb{R}^m} 1(x \in \bigcup_{j \in S} Z_j) (nc(n)^m)^{-1} \sum_{i=1}^n K \left( \frac{x - x_i}{\epsilon(n)} \right) dx
\]

\[
\geq \int_{\mathbb{R}^m} 1(x \in \bigcup_{j \in S} Z_j) (nc(n)^m)^{-1} \sum_{i \in S} K \left( \frac{x - x_i}{\epsilon(n)} \right) dx
\]

\[
= \sum_{i \in S} \int_{\mathbb{R}^m} 1(x \in Z_j) (nc(n)^m)^{-1} K \left( \frac{x - x_i}{\epsilon(n)} \right) dx
\]

\[
\geq \sum_{i \in S} \int_{\mathbb{R}^m} (nc(n)^m)^{-1} K \left( \frac{x - x_i}{\epsilon(n)} \right) dx = |S|/n.
\]
Here, the second-to-last equality, (a), holds because, due to definition of $K$ and $Z_i$, $K((x - x_i)/\varepsilon(n))$ is non-zero only when $x \in Z_i$. Hence, $h_n \in \mathcal{P}_n$, \footnote{More precisely, $h_n$ is the density function of a probability measure that belongs to $\mathcal{P}_n$.} which by Equation (8) implies

$$\sum_{i=1}^{n} \frac{1}{n} \inf_{x' \in Z_i} f(v, x') \leq \int_{\mathbb{R}^m} f(v, x) h_n(x) dx.$$ 

Since $h_n$ is a kernel density estimator, it is well-known (e.g., see Devroye and Győrfi \cite{22}) that under the condition that $\varepsilon(n) \to 0$ and $n\varepsilon(n)^m \to \infty$ the following holds,

$$\int_{\mathbb{R}^m} |h_n(x) - h^*(x)| dx \xrightarrow{\mathbb{P}} 0.$$ 

Therefore, since $|f(v, x)| \leq C$, there exists $\{M_n\} \to 0$ such that the following holds for all $v$,

$$\int_{\mathbb{R}^m} f(v, x) h_n(x) dx \leq \int_{\mathbb{R}^m} f(v, x) h^*(x) dx + M_n,$$

which leads to for all $v$,

$$\sum_{i=1}^{n} \frac{1}{n} \inf_{x' \in Z_i} f(v, x') - M_n \leq \int_{\mathbb{R}^m} f(v, x) h^*(x) dx.$$ 

By symmetry, we also have

$$\int_{\mathbb{R}^m} f(v, x) h^*(x) dx \leq \sum_{i=1}^{n} \frac{1}{n} \sup_{x' \in Z_i} f(v, x') + M_n.$$ 

Further note that

$$\sup_{x \in Z_i} f(v, x) - \inf_{x \in Z_i} f(v, x) \leq d(2\varepsilon(n)).$$

Thus we have for all $v$

$$\sum_{i=1}^{n} \frac{1}{n} \inf_{x' \in Z_i} f(v, x') - M_n \leq \int_{\mathbb{R}^m} f(v, x) h^*(x) dx \leq \sum_{i=1}^{n} \frac{1}{n} \inf_{x' \in Z_i} f(v, x') + M_n + d(2\varepsilon(n)).$$

Since both $M_n$ and $d(2\varepsilon(n))$ go to zero, the theorem follows easily.

Remark 1: Observe from the proof that if we relax the requirement of equicontinuity of $\{f(v, \cdot)\}$, then RO is essentially maximizing an asymptotic lower bound of the true expected reward. Furthermore, if we instead only require the equicontinuity of $\{f(v(n), \cdot)\}_{n=1,2,\cdots}$, then the consistency result still holds. In fact, as $v(n)$ is the optimal solution of a robust optimization, this condition is much easier to satisfy than the equicontinuity of $\{f(v, \cdot)\}$.

Remark 2: Theorem 2 suggests a methodology to choosing an appropriate size $\varepsilon(n)$ of the uncertainty set. Previous work on RO (e.g., Bertsimas and Sim \cite{4}) considers the setting where the observed parameter is the result of corruption of the true parameter (via additive noise). Consequently, the decision maker tunes the size of the uncertainty set used in the RO formulation, to satisfy some probabilistic bounds or risk measures constraints, given a priori information of the noise and in particular the noise magnitude.

Theorem 2 provides a different paradigm for the design of the uncertainty set: when the uncertainty is due to inherent randomness of the parameters, then to approximately solve the stochastic program, the decision maker can use RO, with the size of uncertainty set slowly decreasing in the number of samples. To be more specific, the theorem shows that if the size scales like $o(1)$ and $\omega(n^{-1/m})$, then we achieve asymptotic consistency. Finally, also note that it is straightforward to modify the uncertainty set from a $\ell^\infty$ ball to other parameterized uncertainty sets. Indeed, the only requirement is that the family of distributions obtained using Theorem 1, contain a kernel density estimator.

Remark 3: The consistency of robust optimization is in fact exploited implicitly by many widely used learning algorithms. In our previous work \cite{23} and \cite{24}, we showed that the classical and much used algorithms, namely Support Vector Machines and Lasso, are both special cases of the more general approach outlined in Theorem 2.

IV. CONCLUSION

We showed that robust optimization problems can be reformulated as distributionally robust stochastic problems. I.e., robust optimization is equivalent to maximizing the worst-case expected value over a class of distributions. While such an equivalence is well known in the special case where each uncertain parameter belongs to a different space, we generalize it to the case where multiple parameters belong to the same fixed dimensional space. This setting arises naturally in stochastic problems that are attacked via sampling (as in machine learning or stochastic programming). Using this reformulation, we show how to construct robust optimization problems that are statistically consistent, even when the original empirical optimization is not.

Acknowledgement

The research of Constantine Caramanis was partially supported by NSF grants EFRI-0735905, CNS-0721532, CNS-0831580, and DTRA grant HDTRA1-08-0029. The research of Shie Mannor was partially supported by the Israel Science Foundation under contract 890015 and by a Horev Fellowship.

REFERENCES

Kernel Density Estimator

The kernel density estimator for a density \( h \) in \( \mathbb{R}^d \), originally proposed in Rosenblatt [25] and Parzen [26], is defined by

\[
h_n(x) = (nc_n)^{-d} \sum_{i=1}^{n} K \left( \frac{x - \hat{x}_i}{c_n} \right),
\]

where \( \{c_n\} \) is a sequence of positive numbers, \( \hat{x}_i \) are i.i.d. samples generated according to \( h \), and \( K \) is a Borel measurable function (kernel) satisfying \( K \geq 0, \int K = 1 \). See Devroye and Györfi [22], Scott [27], and the reference therein for detailed discussions. Figure 1 illustrates a kernel density estimator using Gaussian kernel for a randomly generated sample-set. A celebrated property of a kernel density estimator is that it converges in \( L^1 \) to \( h \), i.e.,

\[
\int_{\mathbb{R}^d} |h_n(x) - h(x)| \, dx \to 0
\]

when \( c_n \downarrow 0 \) and \( nc_n^d \uparrow \infty \) (Devroye and Györfi [22]).