Management Science

Publication details, including instructions for authors and subscription information:
http://pubsonline.informs.org

Learning-Based Robust Optimization: Procedures and Statistical Guarantees

L. Jeff Hong, Zhiyuan Huang, Henry Lam

informs
http://pubsonline.informs.org/journal/mnsc

MANAGEMENT SCIENCE
Vol. 67, No. 6, June 2021, pp. 3447–3467
ISSN 0025-1909 (print), ISSN 1526-5501 (online)

Learning-Based Robust Optimization: Procedures and Statistical Guarantees

L. Jeff Hong,[a] Zhiyuan Huang,[b] Henry Lam[c]

[a] School of Management and School of Data Science, Fudan University, Shanghai 200433, China; [b] Department of Industrial and Operations Engineering, University of Michigan, Ann Arbor, Michigan 48109; [c] Department of Industrial Engineering and Operations Research, Columbia University, New York, New York 10027

Contact: hong_liu@fudan.edu.cn, https://orcid.org/0000-0001-7011-4001 (LJH); zhyhuang@umich.edu, https://orcid.org/0000-0003-1284-2128 (ZH); khhlam@gmail.com, https://orcid.org/0000-0002-3193-563X (HL)

Received: April 13, 2017
Revised: June 30, 2018; July 24, 2019
Accepted: February 22, 2020
Published Online in Articles in Advance: December 22, 2020

https://doi.org/10.1287/mnsc.2020.3640

Abstract. Robust optimization (RO) is a common approach to tractably obtain safeguarding solutions for optimization problems with uncertain constraints. In this paper, we study a statistical framework to integrate data into RO based on learning a prediction set using (combinations of) geometric shapes that are compatible with established RO tools and on a simple data-splitting validation step that achieves finite-sample nonparametric statistical guarantees on feasibility. We demonstrate how our required sample size to achieve feasibility at a given confidence level is independent of the dimensions of both the decision space and the probability space governing the stochasticity, and we discuss some approaches to improve the objective performances while maintaining these dimension-free statistical feasibility guarantees.

History: Accepted by Yinyu Ye, optimization.
Funding: This work was supported by the Hong Kong Research Council [Grants GRF11504017, GRF613213, and GRF16203214]. It was also supported by the National Natural Science Foundation of China [Grants 71991470 and 71991473], the National Science Foundation's Division of Civil, Mechanical, and Manufacturing Innovation [Grants 1523453, 1542020, and 1834710], and the National Science Foundation's Division of Information and Intelligent Systems [Grant 1849280].
Supplemental Material: The online appendix is available at https://doi.org/10.1287/mnsc.2020.3640.

Keywords: robust optimization • chance constraint • prediction set learning • quantile estimation

1. Introduction

Many optimization problems in industrial applications contain uncertain parameters in constraints where the enforcement of feasibility is of importance. This paper aims to build procedures to find high-quality solutions for these problems that are tractable and statistically accurate for high-dimensional or limited-data situations.

To locate our scope of study, we consider situations where the uncertainty in the constraints is "stochastic," and a risk-averse modeler wants the solution to be feasible "most of the time," while not making the decision space overly conservative. One common framework to define feasibility in this context is via a chance-constrained program (CCP); that is,

$$\min f(x) \quad \text{subject to} \quad P\big(g(x;\xi) \in \mathcal{A}\big) \geq 1-\epsilon, \qquad (1)$$

where $f(x) \in \mathbb{R}$ is the objective function, $x \in \mathbb{R}^d$ is the decision vector, $\xi \in \mathbb{R}^m$ is a random vector (i.e., the uncertainty) under a probability measure P, and $g(x;\xi) : \mathbb{R}^d \times \mathbb{R}^m \to \Omega$ with $\mathcal{A} \subset \Omega$ for some space Ω. Using existing terminology, we sometimes call $g(x;\xi) \in \mathcal{A}$ the *safety condition* and ϵ the *tolerance level* that controls the violation probability of the safety condition. In this paper, we will consider $g(x;\xi) \in \mathcal{A}$ as linear inequalities, which constitute the most common class of CCPs.

We focus on settings where ξ is observed via a finite amount of data, driven by the fact that in almost every application there is no exact knowledge about the uncertainty and that data are increasingly ubiquitous. Our problem target is to find a solution feasible for (1) with a given statistical confidence (with respect to the data, in a frequentist sense) that has an objective value as small as possible.

First proposed by Charnes et al. (1958), Charnes and Cooper (1959), Miller and Wagner (1965), and Prékopa (1970), the CCP framework (1) has been studied extensively in the stochastic programming literature (see Prékopa (2003) for a thorough introduction), with applications spanning across reservoir system design (Prékopa and Szántai 1978, Prékopa et al. 1978), cash matching (Dentcheva et al. 2004), wireless cooperative networking (Shi et al. 2015), inventory (Lejeune and Ruszczynski 2007), and production management (Murr and Prékopa 2000). Though not always proper (notably when the uncertainty is deterministic or bounded; see e.g., Ben-Tal et al. 2009, pp. 28–29), in many situations it is natural to view uncertainty as "stochastic," and (1) provides a rigorous definition of feasibility under these situations. Moreover, (1) sets a framework to assimilate data in a way that avoids overconservativeness by focusing on the "majority" of the data, as we exploit in this paper.

Our main contribution is a framework to integrate data into robust optimization (RO) as a tool to obtain high-quality solutions feasible in the sense defined by (1). Instead of directly solving (1), which is known to be challenging in general, RO operates by representing the uncertainty via a (deterministic) set, often known as the *uncertainty set* or the *ambiguity set*, and enforces the safety condition to hold for any ξ within it. By suitably choosing the uncertainty set, RO is well known to be a tractable approximation to (1). We revisit these ideas by studying a procedural framework to construct an uncertainty set as a *prediction set* for the data. This consists of approximating a high-probability region via combinations of tractable geometric shapes compatible with RO. As a key development, we propose a simple data-splitting scheme to determine the size of this region, which ensures rigorous statistical performance. This framework is nonparametric and applies under minimal distributional requirements.

In terms of basic statistical properties, our approach satisfies a finite-sample confidence guarantee on the feasibility of the solution in which the minimum required sample size in achieving a given confidence is provably *independent* of the dimensions of both the decision space and the underlying probability space. Whereas finite-sample guarantees are also found in existing sampling-based methods, the dimension-free property of our approach makes it a suitable resort for certain high-dimensional and limited-data situations where previous methods break down.

The dimension-free property, which may appear very strong, needs nonetheless to be complemented with good approaches to curb overconservativeness and maintain tractability. In particular, to reduce conservativeness, a prediction set should accurately trace the shape of the data. By contrast, to retain tractability, the set should be expressible in terms of basic geometric shapes compatible with RO techniques. We present some techniques to construct uncertainty sets that balance these two aspects while simultaneously achieving the basic statistical property. Nonetheless, we caution that these techniques tie conservativeness to the set volume, although often the former is more intricate and depends on the optimization setting at hand (see, e.g., Lagoa and Barmish 2002). Along this line, we also discuss a method to iterate the construction of uncertainty sets that incorporates updated optimality beliefs to improve the objective performance.

Our approach is related to several existing methods for approximating (1). Scenario generation (SG), pioneered by Calafiore and Campi (2005, 2006) and Campi and Garatti (2008, 2011) and independently suggested in the context of Markov decision processes by De Farias and Van Roy (2004), replaces the chance constraint in (1) with a collection of sampled constraints. Related works also include the sample average approximation (SAA) studied by Luedtke and Ahmed (2008), Luedtke et al. (2010), and Luedtke (2014), which restricts the proportion of violated constraints and resembles the discarding approach of Campi and Garatti (2011). SG provides explicit statistical guarantees on the feasibility of the solution obtained in terms of the confidence level, the tolerance level, and the sample size. It directly approximates the chance-constrained optimization without the need of a set-based representation of the uncertainty and hence allows a high geometric flexibility in the resulting set of violation and leads to less conservative solutions. However, in general, the sample size needed to achieve a given confidence grows linearly with the dimension of the decision space, which can be demanding for large-scale problems (as pointed out by, e.g., Nemirovski and Shapiro 2006, p. 971). Recent work reduces dependence on the decision dimension (and its interplay with the tolerance parameter) by, for instance, regularization (Campi and Carè 2013), tighter complexity results in terms of the support rank (Schildbach et al. 2013), solution-dependent number of support constraints (Campi and Garatti 2018), one-off calibration schemes (Carè et al. 2014), sequential validation (Calafiore et al. 2011, Chamanbaz et al. 2016, Calafiore 2017), and hybrid approaches between RO and SG that translate scenario size requirements from decision to stochastic space dimension (Margellos et al. 2014). Among these, our proposed step to tune the set size is closest to the calibration approaches. However, instead of calibrating a solution obtained from a randomized program, we calibrate the coverage of an uncertainty set and control conservativeness and tractability of the resulting RO through proper learning of its shape.

A classical approach to approximating (1) uses safe convex approximation (SCA), by replacing the intractable chance constraint with an inner approximating convex constraint such that a solution feasible for the latter would also be feasible for the former (see, e.g., Ben-Tal and Nemirovski 2000, Nemirovski 2003, Nemirovski and Shapiro 2006). This approach is intimately related to RO, because the approximating constraints are often equivalent to the robust counterparts of RO problems with properly chosen uncertainty sets (see, e.g., Ben-Tal et al. 2009, chapters 2 and 4, and Bertsimas et al. 2001, section 3.1). The statistical guarantees provided by these approximations come from probabilistic deviation bounds, which often rely on the stochastic assumptions and constraint structure on a worst-case basis (see, e.g., Nemirovski and Shapiro 2006; Ben-Tal et al. 2009, chapter 10; Ben-Tal and Nemirovski 1998, 1999; El Ghaoui et al. 1998; Bertsimas and Sim 2004, 2006; Bertsimas et al. 2004; Chen et al. 2007; Calafiore and El Ghaoui 2006). Thus, although the approach carries several advantages (e.g., in handling extraordinarily small tolerance levels), the bounds used can be restrictive to employ in some cases. Moreover, most of the results apply to a single chance constraint; when the safety condition involves several constraints that need to be jointly maintained (known as a *joint chance constraint*), one typically needs to reduce it to individual constraints via the Bonferroni correction, which can add pessimism (there are exceptions, however; see, e.g., Chen et al. 2010). By contrast, these classical results in SCA and RO are capable of constructing uncertainty sets with well-chosen shapes without directly using prediction set properties.

We mention two other lines of work in approximating (1) that can blend with data. Distributionally robust optimization (DRO), an approach dating back to Scarf (1958) and of growing interest in recent years (see, e.g., Delage and Ye 2010, Wiesemann et al. 2014, Goh and Sim 2010, Ben-Tal et al. 2013, Lim et al. 2006), considers using a worst-case probability distribution for ξ within an ambiguity set that represents partial distributional information. The two major classes of sets consist of distance-based constraints (statistical distance from a nominal distribution such as the empirical distribution; see, e.g., Ben-Tal et al. 2013, Wang et al. 2016) and moment-and-support-type constraints—including moments, dispersion, covariance, and/or support (see, e.g., Delage and Ye 2010, Wiesemann et al. 2014, Goh and Sim 2010, Hanasusanto et al. 2017) and shape and unimodality (see, e.g., Popescu 2005, Hanasusanto et al. 2015, Van Parys et al. 2016, Li et al. 2019, Lam and Mottet 2017). To guarantee statistical feasibility, these uncertainty sets need to be properly calibrated from data, either via direct estimation or using the statistical implications from Bayesian (Gupta 2019) or empirical likelihood (Lam and Zhou 2017, Duchi et al. 2016, Blanchet and Kang 2016, Lam 2019) methods. Another line of work takes a Monte Carlo viewpoint and uses sequential convex approximation (Hong et al. 2011, Hu et al. 2013) that stochastically iterates the solution to a Karush–Kuhn–Tucker point, which guarantees local optimality of the convergent solution. This approach can be applied to data-driven situations by viewing the data as Monte Carlo samples.

Finally, some recent RO-based approaches aim to use data more directly. For example, Goldfarb and Iyengar (2003) calibrate uncertainty sets using linear regression under Gaussian assumptions. Bertsimas et al. (2018) study a tight value-at-risk bound on a single constraint and calibrate uncertainty sets via imposing a confidence region on the distributions that govern the bound. Tulabandhula and Rudin (2014) study supervised prediction models to approximate uncertainty sets and suggest using sampling or relaxation to reduce to tractable problems. Our approach follows the general ideas in these works in constructing uncertainty sets that cover the "truth" with high confidence.

The rest of this paper is organized as follows. Section 2 presents our procedural framework and statistical implications. Section 3 discusses some approaches to construct tight and tractable prediction sets. Section 4 reports numerical results and comparisons with existing methods. Additional proofs, numerical results, and useful existing theorems are presented in the online appendix.

2. Basic Framework and Implications

This section lays out our basic procedural framework and implications. First, consider an approximation of (1) via the RO:

$$\min f(x) \quad \text{subject to} \quad g(x;\xi) \in \mathcal{A}, \ \ \forall\, \xi \in \mathcal{U}, \tag{2}$$

where $\mathcal{U} \in \Omega$ is an uncertainty set. Obviously, for any x feasible for (2), $\xi \in \mathcal{U}$ implies $g(x;\xi) \in \mathcal{A}$. Therefore, by choosing $\mathcal{U}$ that covers a $1-\epsilon$ content of ξ (i.e., $\mathcal{U}$ satisfies $P(\xi \in \mathcal{U}) \geq 1-\epsilon$), any x feasible for (2) must satisfy $P(g(x;\xi) \in \mathcal{A}) \geq P(\xi \in \mathcal{U}) \geq 1-\epsilon$, implying that x is also feasible for (1). In other words, we have the following.

Lemma 1. *Any feasible solution of* (2) *using a* $(1-\epsilon)$-*content set* $\mathcal{U}$ *is feasible for* (1).

Note that Ben-Tal et al. (2009, p. 33, discussion point B) point out that it is not necessary for an uncertainty set to contain most values of the stochasticity to induce probabilistic guarantees. Nonetheless, Lemma 1 provides a platform to use a data structure easily and formulate concrete procedures, as we describe.

2.1. Learning Uncertainty Sets

Assume a given independent and identically distributed (i.i.d.) data set $D = \{\xi_1, \ldots, \xi_n\}$, where $\xi_i \in \mathbb{R}^m$ are sampled under a continuous distribution P. In view of Lemma 1, our basic strategy is to construct $\mathcal{U} = \mathcal{U}(D)$, that is, a $(1-\epsilon)$-content prediction set for P with a prescribed confidence level $1-\delta$. In other words,

$$\mathbb{P}_D(P(\xi \in \mathcal{U}(D)) \geq 1-\epsilon) \geq 1-\delta, \tag{3}$$

where we use the notation $\mathbb{P}_D(\cdot)$ to denote the probability taken with respect to the data D. Using such a $\mathcal{U}$, any feasible solution of (2) is feasible for (1) with the same confidence level $1-\delta$. Namely, we have the following:

Lemma 2. *Any feasible solution of* (2) *using* $\mathcal{U}$ *that satisfies* (3) *is feasible for* (1) *with confidence* $1-\delta$.

Note that (3) only focuses on the feasibility guarantee for (1) but does not speak much about conservativeness. To alleviate the latter issue, we judiciously choose $\mathcal{U}$ according to two criteria:

1. We prefer a $\mathcal{U}$ that has a smaller volume, which leads to a larger feasible region in (2) and hence a less conservative inner approximation to (1). Note that with a fixed ϵ, a small $\mathcal{U}$ means a $\mathcal{U}$ that contains a high-probability region (HPR) of ξ.

2. We prefer a $\mathcal{U}$ such that $P(\xi \in \mathcal{U}(D))$ is close to, not just larger than, $1-\epsilon$ with confidence $1-\delta$. We also want the coverage probability $\mathbb{P}_D(P(\xi \in \mathcal{U}(D)) \geq 1-\epsilon)$ to be close to, not just larger than, $1-\delta$.

Moreover, $\mathcal{U}$ needs to be chosen to be compatible with tractable tools in RO. Though this tractability depends on the type of safety condition at hand and is problem specific, the general principle is to construct $\mathcal{U}$ as an HPR that is expressed via a basic geometric set or a combination of them.

This discussion motivates us to propose a two-phase strategy in constructing $\mathcal{U}$. We first split the data D into two groups, denoted D_1 and D_2, with sizes n_1 and n_2, respectively. Say that $D_1 = \{\xi_1^1, \ldots, \xi_{n_1}^1\}$ and $D_2 = \{\xi_1^2, \ldots, \xi_{n_2}^2\}$. These two data groups are used as follows.

2.1.1. Phase 1: Shape Learning.

We use D_1 to approximate the shape of an HPR. Two common choices of tractable basic geometric shapes are the following:

1. *Ellipsoid.* Set the shape as $\mathcal{S} = \{(\xi-\mu)'\Sigma^{-1}(\xi-\mu) \leq \rho\}$ for some $\rho > 0$. The parameters can be chosen by, for instance, setting μ as the sample mean of D_1 and Σ as some covariance matrix, for example, the sample covariance matrix, diagonalized covariance matrix, or identity matrix.

2. *Polytope.* Set the shape as $\mathcal{S} = \{\xi : a_i'\xi \leq b_i, i = 1, \ldots, k\}$, where $a_i \in \mathbb{R}^m$ and $b_i \in \mathbb{R}$. For example, for low-dimensional data, this can be obtained from a convex hull (or an approximated version) of D_1 or, alternatively, of the data that leave out $\lfloor n_1\epsilon \rfloor$ of D_1 that are in the "periphery" (i.e., having the smallest Tukey depth; see, e.g., Serfling 2002, Hallin et al. 2010). It can also take the shape of the objective function when it is linear (a case of interest when using the self-improving strategy that we describe later).

We can also combine these two types of geometric sets as follows:

1. *Union of basic geometric sets.* Given a collection of polytopes or ellipsoids $\mathcal{S}_i$, take $\mathcal{S} = \bigcup_i \mathcal{S}_i$.

2. *Intersection of basic geometric sets.* Given a collection of polytopes or ellipsoids $\mathcal{S}_i$, take $\mathcal{S} = \bigcap_i \mathcal{S}_i$.

The choices of ellipsoids and polytopes are motivated from the tractability in the resulting RO, but they may not describe an HPR of ξ to sufficient accuracy. Unions or intersections of these basic geometric sets provide more flexibility in tracking the HPR of ξ. For example, in the case of multimodal distribution, one can group the data into several clusters (Hastie et al. 2009) and then form a union of ellipsoids over the clusters as $\mathcal{S}$. For nonstandard distributions, one can discretize the space into boxes and take the union of boxes that contain at least some data, inspired by the histogram method in the literature on learning minimum volume sets (Scott and Nowak 2006). The intersection of basic sets is useful in handling segments of ξ, where each segment appears in a separate constraint in a joint CCP.

2.1.2. Phase 2: Size Calibration.

We use D_2 to calibrate the size of the uncertainty set so that it satisfies (3) and, moreover, $P(\xi \in \mathcal{U}(D)) \approx 1-\epsilon$ with coverage $\approx 1-\delta$. The key idea is to use quantile estimation on a "dimension-collapsing" transformation of the data. More concretely, first express our geometric shape obtained in Phase 1 in the form $\{\xi : t(\xi) \leq s\}$, where $t(\cdot) : \mathbb{R}^m \to \mathbb{R}$ is a transformation map from the space of ξ to $\mathbb{R}$, and $s \in \mathbb{R}$. For the two geometric shapes that we considered, we have the following:

1. *Ellipsoid.* We set $t(\xi) = (\xi-\mu)'\Sigma^{-1}(\xi-\mu)$. Then the $\mathcal{S}$ described in Phase 1 is equivalent to $\{\xi : t(\xi) \leq \rho\}$.

2. *Polytope.* Find a point, say μ, in $\mathcal{S}^\circ$, the interior of $\mathcal{S}$ (e.g., the Chebyshev center (Boyd and Vandenberghe 2004) of $\mathcal{S}$ or the sample mean of D_1 if it lies in $\mathcal{S}^\circ$). Let $t(\xi) = \max_{i=1,\ldots,k}(a_i'(\xi-\mu))/(b_i - a_i'\mu)$, which is well defined because $\mu \in \mathcal{S}^\circ$. Then the $\mathcal{S}$ defined in Phase 1 is equivalent to $\{\xi : t(\xi) \leq 1\}$.

For the combinations of sets, we suppose that each individual geometric shape $\mathcal{S}_i$ in Phase 1 possesses a transformation map $t_i(\cdot)$. Then we have the following:

1. *Union of the basic geometric sets.* We set $t(\xi) = \min_i t_i(\xi)$ as the transformation map for $\bigcup_i \mathcal{S}_i$. This is because $\bigcup_i\{\xi : t_i(\xi) \leq s\} = \{\xi : \min_i t_i(\xi) \leq s\}$.

2. *Intersection of the basic geometric sets.* We set $t(\xi) = \max_i t_i(\xi)$ as the transformation map for $\bigcap_i \mathcal{S}_i$. This is because $\bigcap_i\{\xi : t_i(\xi) \leq s\} = \{\xi : \max_i t_i(\xi) \leq s\}$.

We overwrite the value of s in the representation $\{\xi : t(\xi) \leq s\}$ as $t(\xi_{(i^*)}^2)$, where $t(\xi_{(1)}^2) < t(\xi_{(2)}^2) < \cdots < t(\xi_{(n_2)}^2)$ are the ranked observations of $\{t(\xi_i^2)\}_{i=1,\ldots,n_2}$, and

$$i^* = \min\left\{r : \sum_{k=0}^{r-1}\binom{n_2}{k}(1-\epsilon)^k\epsilon^{n_2-k} \geq 1-\delta,\ 1 \leq r \leq n_2\right\}. \quad (4)$$

This procedure is valid if such an i^* can be found or, equivalently, $1-(1-\epsilon)^{n_2} \geq 1-\delta$.

2.2. Basic Statistical Guarantees

Phase 1 focuses on criterion 1 in Section 2.1 by learning the shape of an HPR. Phase 2 addresses our basic requirement (3) and criterion 2. The choice of s in

Phase 2 can be explained by the elementary observation that for any arbitrary i.i.d. data set of size n_2 drawn from a continuous distribution, the i^*th ranked observation as defined by (4) is a valid $1-\delta$ confidence upper bound for the $1-\epsilon$ quantile of the distribution.

Lemma 3. *Let $Y_1,\dots,Y_{n_2}$ be i.i.d. data in $\mathbb{R}$ drawn from a continuous distribution. Let $Y_{(1)} < Y_{(2)} < \cdots < Y_{(n_2)}$ be the order statistics. A $1-\delta$ confidence upper bound for the $(1-\epsilon)$-quantile of the underlying distribution is $Y_{(i^*)}$, where*

$$i^* = \min\left\{ r : \sum_{k=0}^{r-1} \binom{n_2}{k} (1-\epsilon)^k \epsilon^{n_2-k} \geq 1-\delta, 1 \leq r \leq n_2 \right\}.$$

If $\sum_{k=0}^{n_2-1}\binom{n_2}{k}(1-\epsilon)^k\epsilon^{n_2-k} < 1-\delta$ or, equivalently, $1-(1-\epsilon)^{n_2} < 1-\delta$, then none of the $Y_{(r)}$ values is a valid confidence upper bound. Similarly, a $1-\delta$ confidence lower bound for the $(1-\epsilon)$-quantile of the underlying distribution is $Y_{(i_*)}$, where

$$i_* = \max\left\{ r : \sum_{k=r}^{n_2} \binom{n_2}{k} (1-\epsilon)^k \epsilon^{n_2-k} \geq 1-\delta,\ 1 \leq r \leq n_2 \right\}.$$

If $\sum_{k=1}^{n_2}\binom{n_2}{k}(1-\epsilon)^k\epsilon^{n_2-k} < 1-\delta$ or, equivalently, $1-\epsilon^{n_2} < 1-\delta$, then none of the $Y_{(r)}$ values is a valid confidence lower bound.

Proof of Lemma 3. Let $q_{1-\epsilon}$ be the $(1-\epsilon)$-quantile, and let $F(\cdot)$ and $\bar{F}(\cdot)$ be the distribution function and tail distribution function, respectively, of Y_i. Consider

$$\begin{aligned}
&P\big(Y_{(r)} \geq q_{1-\epsilon}\big)\\
&\quad = P\big(\leq r-1 \text{ of the data } \{Y_1,\dots,Y_n\} \text{ are } < q_{1-\epsilon}\big)\\
&\quad = \sum_{k=0}^{r-1}\binom{n_2}{k} F\big(q_{1-\epsilon}\big)^k \bar{F}\big(q_{1-\epsilon}\big)^{n_2-k}\\
&\quad = \sum_{k=0}^{r-1}\binom{n_2}{k}(1-\epsilon)^k\epsilon^{n_2-k}
\end{aligned}$$

by the definition of $q_{1-\epsilon}$. Hence, any r such that $\sum_{k=0}^{r-1}\binom{n_2}{k}(1-\epsilon)^k\epsilon^{n_2-k} \geq 1-\delta$ is a $1-\delta$ confidence upper bound for $q_{1-\epsilon}$, and we pick the smallest one. Note that if $\sum_{k=0}^{n_2-1}\binom{n_2}{k}(1-\epsilon)^k\epsilon^{n_2-k} < 1-\delta$, then none of the $Y_{(r)}$ values is a valid confidence upper bound.

Similarly, we have

$$\begin{aligned}
&P\big(Y_{(r)} \leq q_{1-\epsilon}\big)\\
&\quad = P\big(\geq r \text{ of the data } \{Y_1,\dots,Y_n\} \text{ are } \leq q_{1-\epsilon}\big)\\
&\quad = \sum_{k=r}^{n_2}\binom{n_2}{k} F\big(q_{1-\epsilon}\big)^k \bar{F}\big(q_{1-\epsilon}\big)^{n_2-k}\\
&\quad = \sum_{k=r}^{n_2}\binom{n_2}{k}(1-\epsilon)^k\epsilon^{n_2-k},
\end{aligned}$$

by the definition of $q_{1-\epsilon}$. Hence, any r such that $\sum_{k=r}^{n_2}\binom{n_2}{k}(1-\epsilon)^k\epsilon^{n_2-k} \geq 1-\delta$ will be a $1-\delta$ confidence lower bound for $q_{1-\epsilon}$, and we pick the largest one. Note that if $\sum_{k=1}^{n_2}\binom{n_2}{k}(1-\epsilon)^k\epsilon^{n_2-k} < 1-\delta$, then none of the $Y_{(r)}$ values is a valid confidence lower bound. □

Similar results in the aforementioned simple order-statistic calculation can be found in, for example, Serfling (2009, section 2.6.1). A key element of our procedure is that $t(\cdot)$ is constructed using only Phase 1 data D_1, which are independent of Phase 2. Lemma 3 implies that conditional on D_1, $P(t(\xi) \leq t(\xi^2_{(i^*)})) \geq 1-\epsilon$ with a (conditional) confidence $1-\delta$. From this, we can average over the realizations of D_1 to obtain a valid coverage for the resulting uncertainty set in the sense of satisfying (3). This is summarized formally in the following theorem.

Theorem 1 (Basic Statistical Guarantee). *Suppose that D is an i.i.d. data set drawn from a continuous distribution P on $\mathbb{R}^m$, and we partition D into two sets $D_1 = \{\xi^1_i\}_{i=1,\dots,n_1}$ and $D_2 = \{\xi^2_i\}_{i=1,\dots,n_2}$. Suppose that $n_2 \geq \log\delta/\log(1-\epsilon)$. Consider the set $\mathcal{U} = \mathcal{U}(D) = \{\xi : t(\xi) \leq s\}$, where $t : \mathbb{R}^m \to \mathbb{R}$ is a map constructed from D_1 such that $t(\xi)$, with ξ distributed according to P, is a continuous random variable, and $s = t(\xi^2_{(i^*)})$ is calibrated from D_2 with i^* defined in (4). Then $\mathcal{U}$ satisfies (3). Consequently, an optimal solution obtained from (2) using this $\mathcal{U}$ is feasible for (1) with confidence $1-\delta$.*

Proof of Theorem 1. Because $t(\cdot)$ depends only on D_1 but not on D_2, we have, conditional on any realization of D_1,

$$\begin{aligned}
&\mathbb{P}_{D_2}(P(\xi \in \mathcal{U}(D)) \geq 1-\epsilon | D_1)\\
&\quad = \mathbb{P}_{D_2}\Big(P\Big(t(\xi) \leq t\Big(\xi^2_{(i^*)}\Big)\Big) \geq 1-\epsilon | D_1\Big)\\
&\quad = \mathbb{P}_{D_2}\Big(q_{1-\epsilon} \leq t\Big(\xi^2_{(i^*)}\Big) | D_1\Big) \geq 1-\delta,
\end{aligned} \tag{5}$$

where $q_{1-\epsilon}$ is the $(1-\epsilon)$-quantile of $t(\xi)$ (which depends on D_1). The first equality in (5) follows from the representation of $\mathcal{U} = \{\xi : t(\xi) \leq t(\xi^2_{(i^*)})\}$, the second equality uses the definition of a quantile, and the last inequality follows from Lemma 3 using the condition $1-(1-\epsilon)^{n_2} \geq 1-\delta$ or, equivalently, $n_2 \geq \log\delta/\log(1-\epsilon)$. Note that (5) holds given any realization of D_1. Thus, taking expectation with respect to D_1 on both sides in (5), we have

$$\mathbb{E}_{D_1}\big[\mathbb{P}_{D_2}(P(\xi \in \mathcal{U}(D)) \geq 1-\epsilon | D_1)\big] \geq 1-\delta,$$

where $\mathbb{E}_{D_1}[\cdot]$ denotes the expectation with respect to D_1, which gives

$$\mathbb{P}_D(P(\xi \in \mathcal{U}(D)) \geq 1-\epsilon) \geq 1-\delta.$$

We therefore arrive at (3). Finally, Lemma 2 guarantees that an optimal solution obtained from (2) using the constructed $\mathcal{U}$ is feasible for (1) with confidence $1-\delta$. □

Theorem 1 implies the validity of the approach in giving a feasible solution for CCP (1) with confidence $1-\delta$ for any finite sample size as long as it is large enough that $n_2 \ge \log\delta/\log(1-\epsilon)$. The reasoning of the latter restriction can be seen easily in the proof or, more apparently, from the following argument: in order to get an upper confidence bound for the quantile by choosing one of the ranked statistics, we need the probability of at least one observation to upper-bound the quantile to be at least $1-\delta$. In other words, we need $P(\text{at least one } t(\xi_i^2) \ge (1-\epsilon)\text{-quantile}) \ge 1-\delta$ or, equivalently, $1-(1-\epsilon)^{n_2} \ge 1-\delta$.

We also mention the convenient fact that conditional on D_1,

$$P(\xi \in \mathcal{U}) = P\left(t(\xi) \le t\left(\xi^2_{(i^*)}\right)\right) = F\left(t\left(\xi^2_{(i^*)}\right)\right) \overset{d}{=} U_{(i^*)}, \quad (6)$$

where $F(\cdot)$ is the distribution function of $t(\xi)$, $U_{(i^*)}$ is the i^*th-ranked variable among n_2 uniform variables on $[0,1]$, and $\overset{d}{=}$ denotes equality in distribution. In other words, the theoretical tolerance level induced by our constructed uncertainty set $P(\xi \in \mathcal{U})$ is distributed as the i^*th-order statistic of uniform random variables or, equivalently, $Beta(i^*, n_2 - i^* + 1)$, a beta variable with parameters i^* and $n_2 - i^* + 1$. Note that $P(Beta(i^*,n_2-i^*+1) \ge 1-\epsilon) = P(Bin(n_2,1-\epsilon) \le i^*-1)$, where $Bin(n_2, 1-\epsilon)$ denotes a binomial variable with number of trials n_2 and success probability $1-\epsilon$. This informs an equivalent expression of (4) as

$$\begin{aligned}
&\min\{r : P(Beta(r, n_2 - r + 1) \ge 1-\epsilon) \\
&\quad \ge 1-\delta,\ 1 \le r \le n_2\} \\
&\quad = \min\{r : P(Bin(n_2, 1-\epsilon) \le r-1) \\
&\quad \ge 1-\delta,\ 1 \le r \le n_2\}.
\end{aligned}$$

To address criterion 2 in Section 2.1, we use the following asymptotic behavior as $n_2 \to \infty$.

Theorem 2 (Asymptotic Tightness of Tolerance and Confidence Levels)**.** *Under the same assumptions as in Theorem 1, we have the following, conditional on D_1:*

a. *$P(\xi \in \mathcal{U}) \to 1-\epsilon$ in probability (with respect to D_2) as $n_2 \to \infty$;*

b. *$\mathbb{P}_{D_2}(P(\xi \in \mathcal{U}) \ge 1-\epsilon | D_1) \to 1-\delta$ as $n_2 \to \infty$.*

Theorem 2 confirms that $\mathcal{U}$ is tightly chosen in the sense that the tolerance level and the confidence level are held asymptotically exact. This can be shown by using (6) together with an invocation of the Berry–Essen theorem (Durrett 2010) applied on the normal approximation to a binomial distribution. Online Appendix EC.1 shows the details of the proof, which use techniques similar to those of Li and Liu (2008) and Serfling (2009, section 2.6). In fact, one could further obtain that our choice of i^* satisfies $\sqrt{n_2}(i^*/n_2 - (1-\epsilon)) \to \sqrt{(1-\epsilon)\epsilon}\Phi^{-1}(1-\delta)$ as $n_2 \to \infty$. As a result, the theoretical tolerance level $P(\xi \in \mathcal{U})$ given D_1 concentrates at $1-\epsilon$ by being approximately $(1-\epsilon) + Z/\sqrt{n_2}$, where $Z \sim N(\sqrt{\epsilon(1-\epsilon)}\Phi^{-1}(1-\delta), \epsilon(1-\epsilon))$. For further details, see Online Appendix EC.1.

Note that because of the discrete nature of our quantile estimate, the theoretical confidence level is not a monotone function of the sample size, and there is no guarantee on an exact confidence level at $1-\delta$ using a finite sample (see Online Appendix EC.2). By contrast, part (b) of Theorem 2 guarantees that asymptotically our construction can achieve an exact confidence level.

The idea of using a dimension-collapsing transformation map $t(\cdot)$ resembles the notion of data depth in the literature of generalized quantile (Li and Liu 2008, Serfling 2002). In particular, the data depth of an observation is a positive number that measures the position of the observation from the "center" of the data set. The larger the data depth, the closer the observation is to the center. For example, the half-space depth is the minimum number of observations on one side of any line passing through the chosen observation (Hodges 1955, Tukey 1975), and the simplicial depth is the number of simplices formed by different combinations of observations surrounding an observation (Liu 1990). Other common data depths include the ellipsoidally defined Mahalanobis depth (Mahalanobis 1936) and projection-based depths (Donoho and Gasko 1992, Zuo 2003). Instead of measuring the position of the data relative to the center as in the data-depth literature, our transformation map is constructed to create uncertainty sets with good geometric and tractability properties.

2.3. Dimension-Free Sample-Size Requirement

Theorem 1 and the associated discussion state that we need at least $n_2 \ge \log\delta/\log(1-\epsilon)$ observations in Phase 2 to construct an uncertainty set that guarantees a feasible solution for (1) with confidence $1-\delta$. From a *purely* feasibility viewpoint, this lower bound on n_2 is the minimum total sample size we need: regardless of what shape we generate in Phase 1, as long as we can express it in terms of the $t(\cdot)$ and have $\log\delta/\log(1-\epsilon)$ Phase 2 observations, the basic feasibility guarantee (3) is attained. This number does not depend on the dimension of the decision space or the probability space. It does, however, depend roughly linearly on $1/\epsilon$ for small ϵ, a drawback that is also common among sampling-based approaches, including both SG and SAA, and gives more edge to using safe convex approximation when applicable.

We should caution, however, that if we take $n_1 = 0$ or choose an arbitrary shape in Phase 1, then the resulting solution is likely extremely conservative in terms of objective performance. To combat this issue, it is thus recommended to set aside some data for Phase 1 with the help of established methods borrowed from statistical learning (Section 3 and Sections EC.4 and EC.5 in the online appendix discuss these).

2.4. Enhancing Optimality Performance via Self-Improving Reconstruction

We propose a mechanism, under the framework in Section 2.2, to improve the performance of an uncertainty set by incorporating updated optimality belief.

2.4.1. An Elementary Explanation. As indicated at the beginning of this section, the RO we construct is a conservative approximation to the CCP. A question is whether there is an "optimal" uncertainty set, in the sense that it is a $(1-\epsilon)$-level prediction set and at the same time gives rise to the same solution between the RO and the CCP. As a first observation, the uncertainty set $\mathcal{U} = \{\xi : g(x^*;\xi) \in \mathcal{A}\}$, where x^* is an optimal solution to the CCP, satisfies both properties: by the definition of x^*, this set contains $(1-\epsilon)$-content of P. Moreover, when we use this $\mathcal{U}$ in (2), x^* is trivially a feasible solution. Because this RO is an inner approximation to CCP, x^* is optimal for both the RO and the CCP. The catch, of course, is that in reality we do not know what is x^*. Our suggestion is to replace x^* with some approximate solution $\hat{x}$, leading to a set $\{\xi : g(\hat{x},\xi) \in \mathcal{A}\}$.

Alternatively, the conservativeness of the RO can be reasoned from the fact that $\xi \in \mathcal{U}$, independent of what the obtained solution $\hat{x}$ is in (2), implies that $g(\hat{x};\xi) \in \mathcal{A}$. Thus, our target tolerance probability $P(g(\hat{x};\xi) \in \mathcal{A})$ satisfies $P(g(\hat{x};\xi) \in \mathcal{A}) \geq P(\xi \in \mathcal{U})$ and, in the presence of data, makes the actual confidence level (namely $\mathbb{P}_D(P(g(\hat{x};\xi) \in \mathcal{A}) \geq 1-\epsilon)$) potentially overconservative. However, this inequality becomes an equality if $\mathcal{U}$ is exactly $\{\xi : g(\hat{x};\xi) \in \mathcal{A}\}$. This suggests again that on a high level, an uncertainty set that resembles the form $g(\hat{x};\xi) \in \mathcal{A}$ is less conservative and preferable.

Using this intuition, a proposed strategy is as follows. Consider finding a solution for (1). In Phase 1, find an approximate HPR of the data (using some suggestions in Section 3) with a reasonably chosen size (e.g., just enough to cover $1-\epsilon$ of the data points). Solve the RO problem using this HPR to obtain an initial solution $\hat{x}_0$. Then reshape the uncertainty set as $\{\xi : g(\hat{x}_0;\xi) \in \mathcal{A}\}$. Finally, conduct Phase 2 by tuning the size of this reshaped set; say we get $\{\xi : g(\hat{x}_0;\xi) \in \tilde{\mathcal{A}}\}$, where $\tilde{\mathcal{A}}$ is size tuned. The final RO is

$$\min f(x) \text{ subject to } g(x,\xi) \in \mathcal{A},\ \forall\, \xi : g(\hat{x}_0;\xi) \in \tilde{\mathcal{A}}. \quad (7)$$

Evidently, if the tuning step can be done properly (i.e., the set $\{\xi : g(\hat{x}_0;\xi) \in \mathcal{A}\}$ can be expressed in the form $\{\xi : t(\xi) \leq s\}$ and s is calibrated using the method in Section 2.1), then the procedure retains the overall statistical confidence guarantees presented in Theorems 1 and 2. For convenience, we call the RO (7) created from $\hat{x}_0$ and the discussed procedure a *reconstructed* RO.

More explicitly, consider the safety condition $g(x;\xi) \in \mathcal{A}$ in the form of linear inequalities $Ax \leq b$, where $A \in \mathbb{R}^{l\times d}$ is stochastic and $b \in \mathbb{R}^l$ is constant. After we obtain an initial solution $\hat{x}_0$, we set the uncertainty set as $\mathcal{U} = \{A : A\hat{x}_0 \leq b + sk\}$, where $k = (k_i)_{i=1,\ldots,l} \in \mathbb{R}^l$ is some positive vector and $s \in \mathbb{R}$. The value of s is calibrated by letting $t(A) = \max_{i=1,\ldots,l}\{(a_i'\hat{x}_0 - b_i)/k_i\}$, where a_i' is the ith row of A, b_i is the ith entry of b, and s is chosen as $t(A^2_{(i^*)})$, the order statistic of Phase 2 data as defined in Section 2.1. Using the uncertainty set $\mathcal{U}$, the constraint $Ax \leq b\ \forall\ A \in \mathcal{U}$ becomes $\max_{a_i'\hat{x}_0 \leq b_i + sk_i} a_i'x \leq b_i, i = 1,\ldots,l$, via constraint-wise projection of the uncertainty set, which can be reformulated into linear constraints by using standard RO machinery (see, e.g., Theorem EC.2 in the online appendix).

2.4.2. Properties of Self-Improving Reconstruction. We formalize the discussion in Section 2.4.1 by showing some properties of the optimization problem (7). We focus on the setting of inequality-based safety conditions

$$\min f(x) \text{ subject to } P\big(g(x;\xi) \leq b\big) \geq 1-\epsilon, \quad (8)$$

where $g(x;\xi) = (g_j(x;\xi))_{j=1,\ldots,l} \in \mathbb{R}^l$ and $b = (b_j)_{j=1,\ldots,l} \in \mathbb{R}^l$. Suppose that $\hat{x}_0$ is a given solution (not necessarily feasible). Suppose for now that there is a way to compute quantiles exactly for functions of ξ, and consider the reconstructed RO

$$\min f(x) \text{ subject to } g(x,\xi) \leq b, \quad \forall\ \xi : g(\hat{x}_0;\xi) \leq b + \rho k, \quad (9)$$

where $k = (k_j)_{j=1,\ldots,l} \in \mathbb{R}^l$ is a positive vector, and $\rho = \rho(\hat{x}_0)$ is the $(1-\epsilon)$-quantile of $\max_{j=1,\ldots,l}\{(g_j(\hat{x}_0;\ \xi) - b_j)/k_j\}$. A useful observation is the following.

Theorem 3 (Feasibility Guarantee for Reconstruction)**.** *Given any solution $\hat{x}_0$, if ρ is the $(1-\epsilon)$-quantile of $\max_{j=1,\ldots,l}\{(g_j(\hat{x}_0;\xi) - b_j)/k_j\}$, then any feasible solution of* (9) *is also feasible for* (8).

Proof of Theorem 3. Because $\{\xi : g(\hat{x}_0;\xi) \leq b + \rho k\}$ is by construction a $(1-\epsilon)$-content set for ξ under P, Lemma 1 concludes the theorem immediately. □

Note that Theorem 3 holds regardless of whether $\hat{x}_0$ is feasible for (8). That is, (9) is a way to output a feasible solution from the input of a possibly infeasible $\hat{x}_0$. What is more, in the case that $\hat{x}_0$ is feasible, (9) is guaranteed to give a solution at least as good.

Theorem 4 (Monotonic Objective Improvement)**.** *Under the same assumption as Theorem 3, an optimal solution $\hat{x}$ of (9) is feasible for (8). Moreover, if $\hat{x}_0$ is feasible for (8), then $\hat{x}$ satisfies $f(\hat{x}) \le f(\hat{x}_0)$.*

Proof of Theorem 4. Note that if $\hat{x}_0$ is feasible for (8), then we must have $\rho \le 0$ (or else the chance constraint does not hold), and hence $\hat{x}_0$ must be feasible for (9). By the optimality of $\hat{x}$ for (9), we must have $f(\hat{x}) \le f(\hat{x}_0)$. The theorem concludes by invoking Theorem 3, which implies that $\hat{x}$ is feasible for (8). □

Together, Theorems 3 and 4 give a mechanism to improve any input solution in terms of either feasibility or optimality for (8): if $\hat{x}_0$ is infeasible, then (9) corrects the infeasibility and gives a feasible solution; if $\hat{x}_0$ is feasible, then (9) gives a feasible solution that has an objective value at least as good.

Similar statements hold if the quantile ρ is only calibrated under a given statistical confidence. To link our discussion to the procedure in Section 2.1, suppose that a solution $\hat{x}_0$ is obtained from an RO formulation (or, in fact, any other procedures) using only Phase 1 data. We have the following corollaries.

Corollary 1 (Feasibility Guarantee for Reconstruction Under Statistical Confidence)**.** *Given any solution $\hat{x}_0$ obtained using Phase 1 data, suppose that ρ is the upper bound of the $(1-\epsilon)$-quantile of $\max_{j=1,\dots,l}\{(g_j(\hat{x}_0;\xi) - b_j)/k_j\}$ with confidence level $1-\delta$ generated under Phase 2 data. Any feasible solution of (9) is also feasible for (8) with the same confidence.*

Corollary 2 (Improvement from Reconstruction Under Statistical Confidence)**.** *Under the same assumptions as Corollary 1, an optimal solution $\hat{x}$ of (9) is feasible for (8) with confidence $1-\delta$. Moreover, if $\rho \le 0$, then $\hat{x}$ satisfies $f(\hat{x}) \le f(\hat{x}_0)$.*

The proofs of Corollaries 1 and 2 are the same as those of Theorems 3 and 4, except that Lemma 2 is invoked instead of Lemma 1. Note that $\rho \le 0$ in Corollary 2 implies that $\hat{x}_0$ is feasible for (8) with confidence $1-\delta$. However, the case $\rho > 0$ in Corollary 2 does not directly translate to a conclusion that $\hat{x}_0$ is infeasible under confidence $1-\delta$ because ρ is a confidence upper bound, instead of a lower bound, for the quantile. This implies a possibility that $\hat{x}_0$ is feasible and close to the boundary of the feasible region. There is no guarantee of objective improvement under the reconstructed RO in this case, but there is still a guarantee that the output $\hat{x}$ is feasible with confidence $1-\delta$.

Our numerical experiments in Section 4 show that when applicable, such reconstructions frequently lead to notable improvements. Nonetheless, we caution that depending on the constraint structure, the reconstruction step does not always lead to a significant or a strict improvement even if $\rho \le 0$, and in these cases, some transformation of the constraint is needed. For example, in the case of a single linear chance constraint in the form (8) with $l = 1$ and a bilinear $g(x;\xi)$, the reconstructed uncertainty set consists of one linear constraint. Consequently, the dualization of the RO (see Theorem EC.2 in the online appendix) consists of one dual variable, which optimally scales $\hat{x}_0$ by a scalar factor. When b in (8) (with $l = 1$) is also a stochastic source, no scaling adjustment is allowed because the "decision variable" associated with b (viewing b as a random coefficient in the linear constraint) is constrained to be one. Thus, the proposed reconstruction will show no strict improvement. However, this behavior could be avoided by suitably reexpressing the constraint. When b is, say, positively distributed (or very likely so), one can divide both sides of the inequality by b to obtain an equivalent inequality with the right-hand side fixed to be one. This equivalent constraint is now improvable by our reconstruction (and the new stochasticity now comprises the ratios of the original variables, which can still be observed from the data).

3. Constructing Uncertainty Sets

Our proposed strategy in Section 2 requires constructing an uncertainty set that is tractable for RO and recommends tracing the shape of an HPR as much as possible. Regarding tractability, linear RO with the uncertainty set shapes mentioned in Section 2.1 can be reformulated into standard optimization formulations. For convenience, we document some of these results in Section EC.3 in the online appendix, along with some explanation on how to identify $t(\cdot)$ for the size calibration in our procedure.

Because taking unions or intersections of basic sets gives more capability to trace HPR, we highlight the following two immediate observations. First is that unions of basic sets preserve the tractability of the robust counterpart associated with each union component, with a linear growth of the number of constraints against the number of components.

Lemma 4 (Reformulating Unions of Sets)**.** *The constraint*

$$g(x;\xi) \in \mathcal{A}, \ \ \forall\ \xi \in \mathcal{U},$$

where $\mathcal{U} = \bigcup_{i=1}^{k} \mathcal{U}^i$ is equivalent to the joint constraints

$$g(x;\xi) \in \mathcal{A}, \ \ \forall\ \xi \in \mathcal{U}^i, \ \ i = 1, \dots, k.$$

Second, in the special case of intersections of sets where each intersection component is on the portion of the stochasticity associated with each of multiple constraints, the projective separability property of uncertainty sets (see, e.g., Ben-Tal et al. 2009) gives the following.

Lemma 5 (Reformulating Intersections of Sets). *Let $\xi \in \mathbb{R}^m$ be a vector that can be represented as $\xi = (\xi^i)_{i=1,\dots,k}$, where $\xi^i \in \mathbb{R}^{m^i}, i = 1, \dots, k$, are vectors such that $\sum_{i=1}^k m^i = m$. Suppose that $\mathcal{U} = \prod_{i=1}^k \mathcal{U}^i$, where each $\mathcal{U}^i$ is a set on the domain of ξ^i. The set of constraints*

$$g(x; \xi^i) \in \mathcal{A}^i, i = 1, \dots, k, \ \ \forall \ \xi \in \mathcal{U},$$

is equivalent to

$$g(x; \xi^i) \in \mathcal{A}^i, \ \ \forall \ \xi^i \in \mathcal{U}^i, \ \ i = 1, \dots, k.$$

Note that in approximating a joint CCP, all the $\mathcal{U}^i$ in Lemma 5 need to be jointly calibrated statistically to account for the simultaneous estimation error (which can be conducted by introducing a max operation for the intersection of sets). Intuitively, with weakly correlated data across the constraints, it fares better to use a separate $\mathcal{U}^i$ to represent the uncertainty of each constraint rather than using a single $\mathcal{U}$ and projecting it. Section EC.4 in the online appendix provides a formal statement to support this intuition by arguing a lower level of conservativeness in using individual ellipsoids rather than a single aggregated block-diagonal ellipsoid.

In addition, we can borrow the following statistical tools to more tightly trace an HPR, that is, a smaller-volume prediction set:

1. When data appear in multimodal form, we can use clustering. Label the data into different clusters (using k-means, Gaussian mixture models, or any other techniques), form a simple set $\mathcal{U}_i$ like a ball or an ellipsoid for each cluster, and use the union $\cup_i \mathcal{U}_i$ as the final shape.

2. If the high-dimensional data set has an intrinsic low-dimensional representation, we can use dimension-reduction tools such as principal component analysis. Suppose that $\tilde{\xi} = M\xi + N$, where $M \in \mathbb{R}^{r\times m}$ and $N \in \mathbb{R}^r$ is a low-dimensional representation of a raw random vector $\xi \in \mathbb{R}^m$. Then we can use an uncertainty set in the form

$$\mathcal{U} = \{(M\xi - \mu)'\Sigma^{-1}(M\xi - \mu) \le s\}, \tag{10}$$

where μ is the sample mean of $\tilde{\xi}$, and Σ is a covariance estimate of $\tilde{\xi}$. Tractability is preserved by a straightforward use of existing RO results (see Theorem EC.4 in the online appendix).

3. In situations of unstructured data where clustering or dimension-reduction techniques do not apply, one approach is to view each data point as a "cluster" by forming a ball surrounding a data point, followed by taking a union of those balls. Intriguingly, this scheme coincides with the one studied in Erdoğan and Iyengar (2006) to approximate an ambiguous CCP where the underlying distribution is within a neighborhood of some baseline measure.

We provide further illustrations of these tools in Section EC.5 of the online appendix.

4. Numerical Examples

We present numerical examples to illustrate the performance of our RO approach:

1. We set $\epsilon = 0.05$ and $\delta = 0.05$.

2. For each setting, we repeat the experimental run 1,000 times, each time generating a new independent data set.

3. We define $\hat{\epsilon}$ to be the estimated expected violation probability of the obtained solution. In other words, $\hat{\epsilon} = \hat{E}_D[P_{violation}]$, where $\hat{E}_D[\cdot]$ refers to the empirical expectation taken among the 1,000 data sets, and $P_{violation}$ denotes the probability $P(g(\hat{x}(D); \xi) \notin \mathcal{A})$. For single linear CCPs with Gaussian-distributed ξ, $P_{violation}$ can be computed analytically. In other cases, $P_{violation}$ is estimated using 10,000 new independent realizations of ξ. For approaches that do not depend on data, for example, SCA, we set $\hat{\epsilon} = P_{violation}$ directly.

4. We define $\hat{\delta} = \hat{P}_D(P_{violation} > \epsilon)$, where $\hat{P}_D(\cdot)$ refers to the empirical probability with respect to the 1,000 data sets, and $P_{violation}$ is similarly defined as for $\hat{\epsilon}$. For approaches that do not depend on data, the chance constraint is always satisfied, and therefore, we have $\hat{\delta} = 0$.

5. We denote "Obj. Val." as the average optimal objective value of the 1,000 solutions generated from the independent data sets.

6. When the reconstruction technique described in Section 2.4 is applied, the initial guessed solution is obtained from an uncertainty set with size calibrated to be just enough to cover $1 - \epsilon$ of the Phase 1 data.

Recall that d is the decision-space dimension, n is the total sample size, and n_1 and n_2 are the sample sizes for Phases 1 and 2. These numbers differ across the examples for the purpose of illustration.

Moreover, we compare our RO approaches with several methods:

1. Scenario approaches, including the classical SG (Campi and Garatti 2008) described in the Introduction and its variant Fast Algorithm for the Scenario Technique (FAST; Carè et al. 2014). FAST was introduced to reduce the sample-size requirement of the classical SG. It consists of two steps, each step using n_1 and n_2 samples, respectively (the notations are unified with our method for easy comparisons). The first step of FAST is similar to SG, which solves a sampled program with n_1 constraints and obtains a tentative solution. The second step is a detuning step to adjust the tentative solution with the help of a *robust feasible solution*, that is, a solution feasible for any possible ξ. The adjusted solution is a convex combination of the tentative solution and the robust feasible solution so that the final solution satisfies the other n_2 sampled constraints. In our comparison, we use the minimum required sample sizes in the detuning step suggested in Carè et al. (2014) so that the total

Table 1. Optimality and Feasibility Performances on a Single d = 11-Dimensional Linear CCP with Gaussian Distribution for Several Methods, Using Sample Size n = 120

Method	RO	Recon	SG	FAST	DRO Mo	DRO KL	SCA
n	120	120	120	120	120	120	—
n_1	60	60	—	61	—	60	—
n_2	60	60	—	59	—	60	—
Obj. Val.	–1,189.31	–1,194.87	–1,196.60	–1,193.53	–1,187.35	0	–1,195.07
$\hat{\epsilon}$	1.34×10^{-5}	0.0164	0.090	0.0164	2.55×10^{-8}	0	0.0072
$\hat{\delta}$	0	0.048-5	0.957	0.043	0	0	0

Notes. The true optimal value is –1,196.7. RO, robust optimization; Recon, reconstructed RO; SG, scenario generation; FAST, Fast Algorithm for the Scenario Technique; DRO Mo, distributionally robust optimization (DRO) with ellipsoidal moment set; DRO KL, DRO with KL-divergence set; SCA, safe convex approximation; Obj. Val., average optimal objective value of the 1,000 solutions generated from the independent data sets.

required sample size is precisely the given overall size. We compare with FAST here because the latter elicits a small sample size requirement with the help of a validation-type scheme that is similar to our approaches applied to the RO setting.

2. DRO with first- and second-moment information, where the moments lie in an ellipsoidal joint confidence region. First, supposing that we are given exact first and second moments, we can reformulate a distributionally robust linear chance constraint into a quadratic constraint as suggested in El Ghaoui et al. (2003). By contrast, using the delta method suggested in Marandi et al. (2019), we can construct ellipsoidal confidence regions for the vectorized mean and covariance matrix. Combining the quadratic constraint in El Ghaoui et al. (2003) and the ellipsoidal set in Marandi et al. (2019), we can use theorem 1(II) and example 4 in Marandi et al. (2019) to reformulate the DRO with an ellipsoidal moment set into a semidefinite program. We provide further details of this reformulation in Section EC.6 in the online appendix.

3. DRO with an uncertainty set defined by a neighborhood surrounding a reference distribution measured by a ϕ-divergence. We use the reformulation in Jiang and Guan (2016) that transforms such a distributionally robust chance constraint into an ordinary chance constraint, under the reference distribution, with an adjusted tolerance level ϵ^*, which then allows us to resort to SG or SAA using Monte Carlo samples (as we will see momentarily; whichever method to resort to does not quite matter in our experiments). We use the Kullback–Leibler (KL) divergence and construct the reference distribution using kernel-density estimation (with Gaussian kernel). We set the size of the KL-divergence ball by estimating the divergence using the k-nearest-neighbor (k-NN) estimator, a provably consistent estimator proposed in Wang et al. (2009) and Póczos et al. (2012) (other related estimators and theoretical results are found in Moon and Hero (2014), Liu et al. (2012), Pál et al. (2010), and Póczos and Schneider (2012). We use $k = 1$ in our experiments because the experimental results indicate that the bias increases significantly as k increases. Moreover, to estimate the divergence properly, we split the data into two portions, n_1 and n_2—the first portion is used to construct the reference kernel density, and the

Table 2. Optimality and Feasibility Performances on a Single d = 100 Dimensional Linear CCP with Gaussian Distribution for Several Methods, Using Sample Size n = 120

Method	RO	Recon	SG	FAST	DRO Mo	DRO KL	SCA
n	120	120	120	120	120	120	—
n_1	60	60	—	61	—	60	—
n_2	60	60	—	59	—	60	—
Obj. Val.	–832.12	–1112.11	Unbounded	Unbounded	–1,193.21	0	–1,193.0
$\hat{\epsilon}$	0	0.0158	—	—	0.195	0	0.0072
$\hat{\delta}$	0	0.041	—	—	1	0	0

Notes. The true optimal value is –1,195.3. Results on moment-based DRO are based on 30 replications because of the high computational demand. RO, robust optimization; Recon, reconstructed RO; SG, scenario generation; FAST, Fast Algorithm for the Scenario Technique; DRO Mo, distributionally robust optimization (DRO) with ellipsoidal moment set; DRO KL, DRO with KL-divergence set; SCA, safe convex approximation; Obj. Val., average optimal objective value of the 1,000 solutions generated from the independent data sets.

Table 3. Optimality and Feasibility Performances on a Single $d = 11$-Dimensional Linear CCP with Gaussian Distribution for Several Methods, Using Sample Size $n = 336$

Method	RO	Recon	SG	FAST	DRO Mo	DRO KL	SCA
n	336	336	336	336	336	336	—
n_1	212	212	—	318	—	168	—
n_2	124	124	—	18	—	168	—
Obj. Val.	–1,190.33	–1,195.82	–1,195.67	–1,195.14	–1,188.48	0	–1,195.07
$\hat{\epsilon}$	3.47×10^{-6}	0.0247	0.0331	0.0259	2.19×10^{-8}	0	0.0072
$\hat{\delta}$	0	0.04	0.056	0.043	0	0	0

Notes. The true optimal value is –1,196.7. RO, robust optimization; Recon, reconstructed RO; SG, scenario generation; FAST, Fast Algorithm for the Scenario Technique; DRO Mo, distributionally robust optimization (DRO) with ellipsoidal moment set; DRO KL, DRO with KL-divergence set; SCA, safe convex approximation; Obj. Val., average optimal objective value of the 1,000 solutions generated from the independent data sets.

second portion is used for the k-NN divergence estimation. The reason for this split is that otherwise the estimation of the reference distribution and the divergence would depend on and interfere with each other, leading to estimation accuracy so poor that the divergence estimate becomes negative all the time. We provide further implementation details in Section EC.7.3 of the online appendix.

4. SCA. We will state the underlying a priori distributional assumptions in using the considered SCA, which differ case by case.

When applying moment-based DRO and SCA to joint CCPs, we use the Bonferroni correction (more details in the relevant examples). We also make two additional remarks. First, when comparing the objective values from different methods, because we can always translate or scale the problem by adding/multiplying constants to distort the apparent magnitudes, we mostly focus our comparisons on the direction (bigger or smaller), which is invariant under the aforementioned distortions. Second, even though we report only the point estimates of the mean objective values and ϵ and δ, our conclusions in comparing the objective values and constraint violation probabilities remain unchanged, even if we consider the 95% confidence intervals of these estimates (from the 1,000 experimental repetitions), and we do not report the confidence intervals for the sake of succinctness. Finally, our codes are available at https://github.com/zhyhuang/Learningbased-RO.

4.1. Test Case 1: Multivariate Gaussian on a Single Chance Constraint

We consider a single linear CCP

$$\min c'x \quad \text{subject to} \quad P(\xi'x \le b) \ge 1-\epsilon, \qquad (11)$$

where $x \in \mathbb{R}^d$ is the decision vector, and $c \in \mathbb{R}^d$ and $b \in \mathbb{R}$ are arbitrarily chosen constants. The random vector $\xi \in \mathbb{R}^d$ is drawn from a multivariate Gaussian distribution with an arbitrary mean (here we set it to $-c$) and an arbitrarily chosen positive-definite covariance matrix. Because (11) is exactly solvable when the Gaussian distribution is known, we can verify that it has a bounded optimal solution.

We consider $d = 11$ and 100 as the dimension of the decision vector. Tables 1 and 2 show these two cases with a small sample size $n = 120$, whereas Tables 3 and 4 show these cases with a bigger sample size

Table 4. Optimality and Feasibility Performances on a Single $d = 100$-Dimensional Linear CCP with Gaussian Distribution for Several Methods, Using Sample Size $n = 2331$

Method	RO	Recon	SG	FAST	DRO Mo	DRO KL	SCA
n	2,331	2,331	2,331	2,331	2,331	2,331	—
n_1	1,318	1,318	—	2,326	—	1,166	—
n_2	1,013	1,013	—	5	—	1,165	—
Obj. Val.	–1,168.35	–1,194.76	–1,194.13	–1,193.85	–1,175.48	0	–1,193.0
$\hat{\epsilon}$	0	0.0395	0.0428	0.0386	8.76×10^{-14}	0	0.0072
$\hat{\delta}$	0	0.051	0.039	0.052	0	0	0

Notes. The true optimal value is –1,195.3. Results on moment-based DRO are based on 30 replications because of the high computational demand. RO, robust optimization; Recon, reconstructed RO; SG, scenario generation; FAST, Fast Algorithm for the Scenario Technique; DRO Mo, distributionally robust optimization (DRO) with ellipsoidal moment set; DRO KL, DRO with KL-divergence set; SCA, safe convex approximation; Obj. Val., average optimal objective value of the 1,000 solutions generated from the independent data sets.

Table 5. Optimality and Feasibility Performances on a Single $d = 10$-Dimensional Linear CCP with the Beta-Perturbation Model for Several Methods, Using Sample Size $n = 120$

Method	RO	Recon	SG	FAST	DRO Mo	DRO KL	SCA
n	120	120	120	120	120	120	—
n_1	60	60	—	61	—	60	—
n_2	60	60	—	59	—	60	—
Obj. Val.	–988.78	–1,087.85	–1,114.57	–1,071.77	–968.30	0	–815.06
$\hat{\epsilon}$	1.02×10^{-5}	0.0161	0.0643	0.0171	0	0	0
$\hat{\delta}$	0	0.037	0.723	0.063	0	0	0

Note. RO, robust optimization; Recon, reconstructed RO; SG, scenario generation; FAST, Fast Algorithm for the Scenario Technique; DRO Mo, distributionally robust optimization (DRO) with ellipsoidal moment set; DRO KL, DRO with KL-divergence set; SCA, safe convex approximation; Obj. Val., average optimal objective value of the 1,000 solutions generated from the independent data sets.

(336 and 2,331, respectively) so that the classical SG provides provable feasibility guarantees. In each table, we show the results for our RO using an ellipsoidal uncertainty set ("RO"), our reconstructed RO ("Recon"), SG ("SG"), FAST ("FAST"), DRO with ellipsoidal moment set ("DRO Mo"), DRO with KL-divergence set ("DRO KL"), and SCA ("SCA"). The last approach does not need the data and instead assumes partial a priori distributional information.

For our RO approaches, we use ellipsoidal uncertainty sets with estimated covariance matrices for the case $d = 11$ (Tables 1 and 3) and diagonalized ellipsoidal sets (i.e., using only variance estimates) for $d = 100$ (Tables 2 and 4) to stabilize our estimates because n_1 is smaller than d in the latter case. The tables show that the solutions from our plain RO tend to be conservative, because $\hat{\delta} = 0$. Nonetheless, the reconstructed RO is less conservative across all settings, reflected by the better average optimal values and $\hat{\delta}$ close to the target confidence level 0.05. In all cases, both the plain RO and the reconstructed RO give valid (i.e., confidently feasible) solutions.

We compare our ROs with scenario approaches. When the sample size is small (Tables 1 and 2), SG cannot obtain a valid solution. In the case $d = 11$, it gives $\hat{\delta}$ much greater than 0.05. Furthermore, in the case $d = 100$, SG gives unbounded solutions in all 1,000 replications because the number of sampled constraints is very close to the decision dimension. For FAST, because b is chosen to be positive, we can use the origin to be the robust feasible solution. Table 1 shows that when $d = 11$, FAST gives confidently feasible solutions. The average optimal value from reconstructed RO (−1,194.87) is (slightly) better than the value from FAST (−1,193.53), whereas RO using ellipsoidal sets is more conservative (−1,189.31). However, when $d = 100$ (Table 2), the first-step problem of FAST is unbounded in all $1,000$ replications.

When the sample size is adequate (Tables 3 and 4), the values of $\hat{\delta}$ from SG being less than or close to 0.05 confirms the validity of the solutions. Note that in these cases, FAST gives more conservative solutions than SG. (This is a general consequence of the construction of FAST, which is designed to have a smaller feasible region than SG under the same data set.) RO with ellipsoidal sets obtains more conservative solutions than SG, as shown by the zero $\hat{\delta}$ values and worse average objective values. By using reconstruction, however, the $\hat{\delta}$ values become very close to the desired confidence level $\delta = 0.05$, and the average objective values are almost identical to (and slightly better than) those obtained from SG.

These results reveal that when the sample size is large enough, SG can perform better than our RO

Table 6. Optimality and Feasibility Performances on a Joint Linear CCP with Gaussian Distribution for Several Methods, Using Sample Size $n = 120$

Method	RO	Recon	SG	FAST	DRO Mo	DRO KL	SCA
n	120	120	120	120	120	120	—
n_1	60	60	—	61	—	60	—
n_2	60	60	—	59	—	60	—
Obj. Val.	–6,956.49	–7,920.12	–9,283.35	–8,925.74	–3,996.87	0	–8,927.71
$\hat{\epsilon}$	3.46×10^{-5}	0.0161	0.0581	0.0169	0	0	0.026
$\hat{\delta}$	0	0.044	0.607	0.045	0	0	0

Note. RO, robust optimization; Recon, reconstructed RO; SG, scenario generation; FAST, Fast Algorithm for the Scenario Technique; DRO Mo, distributionally robust optimization (DRO) with ellipsoidal moment set; DRO KL, DRO with KL-divergence set; SCA, safe convex approximation; Obj. Val., average optimal objective value of the 1,000 solutions generated from the independent data sets.

Table 7. Optimality and Feasibility Performances on a Joint Linear CCP with Gaussian Distribution for Several Methods, Using Sample Size $n = 336$

Method	RO	Recon	SG	FAST	DRO Mo	DRO KL	SCA
n	336	336	336	336	336	336	—
n_1	212	212	—	318	—	168	—
n_2	124	124	—	18	—	168	—
Obj. Val.	−7,146.54	−8,029.83	−9,130.95	−9,081.81	−4,209.86	0	−8,927.71
$\hat{\epsilon}$	7.32×10^{-5}	0.0235	0.0223	0.0185	0	0	0.026
$\hat{\delta}$	0	0.038	0.005	0.002	0	0	0

Note. RO, robust optimization; Recon, reconstructed RO; SG, scenario generation; FAST, Fast Algorithm for the Scenario Technique; DRO Mo, distributionally robust optimization (DRO) with ellipsoidal moment set; DRO KL, DRO with KL-divergence set; SCA, safe convex approximation; Obj. Val., average optimal objective value of the 1,000 solutions generated from the independent data sets.

using basic uncertainty sets. By contrast, our RO can provide feasibility guarantees in small sample situations where SG may fail. FAST is valid in small sample situations but is more likely to have unbounded solutions in high-dimensional problems than our RO. Thus, generally, our RO appears most useful for small sample sizes when compared with scenario approaches, a benefit postulated in previous sections. It also appears that using reconstruction can boost our performance to a level comparable to that of SG (and hence also FAST) in situations where the latter is applicable in the shown examples. Note that our reconstruction by design can improve the objective performance compared with plain RO, whereas FAST is used primarily to reduce the sample size requirement and is necessarily more conservative than SG in terms of achieved objective value. Finally, we note that unbounded solutions in SG can potentially be avoided by adding artificial constraints. In this regard, we show in Section EC.7.1 of the online appendix the same example but with additional nonnegativity constraints to illustrate the comparisons further.

Next, we compare with moment-based DRO. In low-dimensional cases with $d = 11$, moment-based DRO gives solutions that are more conservative than RO using ellipsoidal sets, as shown by the larger objective values, that is, −1,187.35 (DRO) versus −1,189.31 (RO) in the small sample case (Table 1) and −1,184.48 (DRO) versus −1,190.33 (RO) in the large sample case (Table 3). The conservativeness of moment-based DRO is also revealed in the small $\hat{\epsilon}$ and $\hat{\delta} = 0$ in both cases. For high-dimensional problems with $d = 100$, we present the performance of moment-based DRO with only 30 replications (instead of 1,000) because of the large program size and consequently the demanding computational effort when solving the reformulated semidefinite programs (although the replication size is smaller, conclusions can still be drawn rigorously; i.e., the confidence intervals of the estimated $\hat{\epsilon}$ and $\hat{\delta}$ turn out to either lie completely under or above 0.05). In the small sample size case (Table 2), moment-based DRO fails to provide feasible solutions ($\hat{\delta} = 1$; i.e., obtained solutions violate the chance constraint in all 30 replications). This can be attributed to a poor estimation of the moment confidence region with small data and high dimension. (Note that forming an ellipsoidal first- and second-moment set for moment-based DRO requires estimating a covariance matrix of size $(3d + d^2)/2 \times (3d + d^2)/2$ because it uses the estimation variances of the first and second moments that involve even higher-order moments, in contrast to a size of $d \times d$ in our ellipsoidal RO.) When the sample size is larger (Table 3), moment-based DRO provides

Table 8. Optimality and Feasibility Performances on a Joint Linear CCP with Beta Distribution for Several Methods, Using Sample Size $n = 120$

Method	RO	Recon	SG	FAST	DRO Mo	DRO KL	SCA
n	120	120	120	120	120	120	—
n_1	60	60	—	61	—	60	—
n_2	60	60	—	59	—	60	—
Obj. Val.	−1,241.05	−1,796.74	−2,105.77	−1,732.73	−230.74	0	−361.079
$\hat{\epsilon}$	6.90×10^{-5}	0.0138	0.0577	0.0170	0	0	0
$\hat{\delta}$	0	0.022	0.576	0.045	0	0	0

Note. RO, robust optimization; Recon, reconstructed RO; SG, scenario generation; FAST, Fast Algorithm for the Scenario Technique; DRO Mo, distributionally robust optimization (DRO) with ellipsoidal moment set; DRO KL, DRO with KL-divergence set; SCA, safe convex approximation; Obj. Val., average optimal objective value of the 1,000 solutions generated from the independent data sets.

Table 9. Optimality and Feasibility Performances on a Joint Linear CCP with Beta Distribution for Several Methods, Using Sample Size $n = 336$

Method	RO	Recon	SG	FAST	DRO Mo	DRO KL	SCA
n	336	336	336	336	336	336	—
n_1	212	212	—	318	—	168	—
n_2	124	124	—	18	—	168	—
Obj. Val.	−1,304.89	−1,911.36	−1,881.69	−1,828.98	−251.69	0	−361.079
$\hat{\epsilon}$	1.20×10^{-4}	0.0199	0.0229	0.0192	0	0	0
$\hat{\delta}$	0	0.023	0.004	0.003	0	0	0

Note. RO, robust optimization; Recon, reconstructed RO; SG, scenario generation; FAST, Fast Algorithm for the Scenario Technique; DRO Mo, distributionally robust optimization (DRO) with ellipsoidal moment set; DRO KL, DRO with KL-divergence set; SCA, safe convex approximation; Obj. Val., average optimal objective value of the 1,000 solutions generated from the independent data sets.

valid feasible solutions ($\hat{\delta} = 0$). The average objective (−1,175.48) is less conservative than our plain RO (−1,168.35) but is more conservative than our reconstructed RO (−1,194.76).

These observations show that when the moment information is well estimated (i.e., the sample size is sufficient relative to the dimension), moment-based DRO provides solutions with a similar conservative level as our RO using ellipsoidal sets. However, when the sample size is too small to get reasonable estimates for the moments, moment-based DRO can fail to obtain feasible solutions. Reconstructed RO appears to outperform moment-based DRO generally. The benefits of our RO approaches in small samples and the boosted performance of reconstructed RO compared with moment-based DRO are in line with our comparisons with scenario approaches.

DRO with an estimated KL-divergence set suffers from general setbacks in the experiments. In all cases we considered, the kernel-density estimator cannot provide a good enough reference distribution f_0, so the size of the divergence ball is too big and subsequently results in conservative solutions. The construction of f_0 is poor because of the curse of dimensionality in kernel-density estimation, whose accuracy deteriorates exponentially with the dimension, because we have a relatively high dimension compared with the data size. By contrast, the performance of DRO, which relies on using the adjusted tolerance level ϵ^*, appears sensitive to the divergence ball size and demands a high accuracy in estimating f_0. Subsequently, the big divergence ball size leads to a zero ϵ^* in all replications, which, in turn, forces us to choose a solution x that satisfies the safety condition $\xi' x \leq b$ for all $\xi \in \mathbb{R}^d$. The origin is then output as the only such feasible solution, and the objective is zero, as shown in Tables 1–4. This indicates that DRO with KL divergence, calibrated using a density estimator and the divergence estimation technique suggested in the literature, gives overly conservative solutions for our considered problems.

Lastly, we compare with SCA. Consider a perturbation model for ξ given by $\xi = a_0 + \sum_{i=1}^{L} \zeta_i a_i$, where $a_i \in \mathbb{R}^d$ for all $i = 0, 1, \ldots, L$, and $\zeta_i \in \mathbb{R}$ are independent Gaussian variables with mean μ_i and variance s_i^2 such

Table 10. Optimality and Feasibility Performances on a Single $d = 11$-Dimensional Linear CCP with t-Distribution for Several Methods, Using Sample Size $n = 120$

Method	RO	Recon	SG	FAST	DRO Mo	DRO KL
n	120	120	120	120	120	120
n_1	60	60	—	61	—	60
n_2	60	60	—	59	—	60
Obj. Val.	−1,112.75	−1,166.52	−1,182.20	−1,158.27	−1,134.38	0
$\hat{\epsilon}$	0.000252	0.0161	0.0910	0.0172	0.000461	0
$\hat{\delta}$	0	0.046	0.961	0.064	0	0

Note. RO, robust optimization; Recon, reconstructed RO; SG, scenario generation; FAST, Fast Algorithm for the Scenario Technique; DRO Mo, distributionally robust optimization (DRO) with ellipsoidal moment set; DRO KL, DRO with KL-divergence set; Obj. Val., average optimal objective value of the 1,000 solutions generated from the independent data sets.

Table 11. Optimality and Feasibility Performances on a Single $d = 11$-Dimensional Linear CCP with t-Distribution for Several Methods, Using Sample Size $n = 336$

Method	RO	Recon	SG	FAST	DRO Mo	DRO KL
n	336	336	336	336	336	336
n_1	212	212	—	318	—	168
n_2	124	124	—	18	—	168
Obj. Val.	−1,126.66	−1,175.64	−1,175.04	−1,170.35	−1,137.19	0
$\hat{\epsilon}$	0.00023	0.024	0.0334	0.0259	0.000407	0
$\hat{\delta}$	0	0.055	0.069	0.04	0	0

Note. RO, robust optimization; Recon, reconstructed RO; SG, scenario generation; FAST, Fast Algorithm for the Scenario Technique; DRO Mo, distributionally robust optimization (DRO) with ellipsoidal moment set; DRO KL, DRO with KL-divergence set; Obj. Val., average optimal objective value of the 1,000 solutions generated from the independent data sets.

that $\mu_i \in [\mu_i^-, \mu_i^+]$ and $s_i^2 \le \sigma_i^2$. A safe approximation of (11) is in Ben-Tal et al. (2009):

$$\min c'x \quad \text{subject to} \quad (a_0'x - b) + \sum_{i=1}^{L} \max\left[a_i'x\mu_i^-, a_i'x\mu_i^+\right] + \sqrt{2\log(1/\epsilon)}\sqrt{\sum_{i=1}^{L} \sigma_i^2 (a_i'x)^2} \le 0.$$

To apply this SCA to (11), we set ζ_i to be independent $N(0,1)$ variables, $a_0 = \mu$ and a_i to be the ith column of $\Sigma^{1/2}$, and $\mu_i^- = \mu_i^+ = 0$ and $\sigma_i^2 = 1$ for $i = 1, \ldots, d$. This, in fact, assumes knowledge on the mean and covariance of the Gaussian vector ξ, thus giving an upper hand to SCA.

Tables 1–4 all show that the optimal objective values obtained from SCA (−1,195.07 and −1,193.0, respectively, for $d = 11, 100$) are close to the true optimal values (−1,196.7 and −1,195.3) compared with other methods. Our ROs using ellipsoidal sets obtain more conservative solutions generally. The relative conservativeness also shows up in reconstructed ROs with small sample sizes (Tables 1 and 2), but with more samples (Tables 3 and 4), our reconstructed RO outperforms the considered SCA.

Note that in this example, the normality and mean and covariance information used in the SCA makes the latter perform very well. Our RO using estimated ellipsoidal sets does not achieve this level of preciseness. However, the reconstructed RO can still outperform this SCA when the sample size is large enough. Note that the performance of SCA depends on the true distribution (because it is related to the tightness of the SCA constraint in approximating the chance constraint). In the next example, we consider an alternative underlying distribution where SCA does not perform as well.

4.2. Test Case 2: Beta Models on a Single Chance Constraint

We consider the single linear CCP in (11), where each component of ξ is now bounded. We use a perturbation model for ξ given by $\xi = a_0 + \sum_{i=1}^{L} \zeta_i a_i$, where $a_i \in \mathbb{R}^d$ for all $i = 0, 1, \ldots, L$, and $\zeta_i \in \mathbb{R}$ are independent random variables, each with mean zero and bounded in $[-1, 1]$, where $d = 10$, $L = 10$, and $a_i \in \mathbb{R}^{10}$ are known arbitrarily chosen vectors. This allows the use of an SCA. In particular, we set each ζ_i to be a beta distribution with parameters $\alpha = 10$ and $\beta = 10$ that are multiplied by two and shifted by one. Similar to

Table 12. Optimality and Feasibility Performances on a Single $d = 100$-Dimensional Linear CCP with t-Distribution for Several Methods, Using Sample Size $n = 120$

Method	RO	Recon	SG	FAST	DRO Mo	DRO KL
n	120	120	120	120	120	120
n_1	60	60	—	61	—	60
n_2	60	60	—	59	—	60
Obj. Val.	−1,077.56	−1,184.45	Unbounded	Unbounded	−1,190.70	0
$\hat{\epsilon}$	6.00×10^{-14}	0.0156	—	—	0.22	0
$\hat{\delta}$	0	0.045	—	—	1	0

Notes. Results on moment-based DRO are based on 30 replications because of the high computational demand. RO, robust optimization; Recon, reconstructed RO; SG, scenario generation; FAST, Fast Algorithm for the Scenario Technique; DRO Mo, distributionally robust optimization (DRO) with ellipsoidal moment set; DRO KL, DRO with KL-divergence set; Obj. Val., average optimal objective value of the 1,000 solutions generated from the independent data sets.

Table 13. Optimality and Feasibility Performances on a Joint $d = 11$-Dimensional Linear CCP with t-Distribution for Several Methods, Using Sample Size $n = 120$

Method	RO	Recon	SG	FAST	DRO Mo	DRO KL
n	120	120	120	120	120	120
n_1	60	60	—	61	—	60
n_2	60	60	—	59	—	60
Obj. Val.	−4,229.6	−6,499.93	−8,313	−7,220.37	−3,888.63	0
$\hat{\epsilon}$	0.00108	0.00847	0.0404	0.0152	4.17×10^{-4}	0
$\hat{\delta}$	0	0.002	0.284	0.048	0	0

Note. RO, robust optimization; Recon, reconstructed RO; SG, scenario generation; FAST, Fast Algorithm for the Scenario Technique; DRO Mo, distributionally robust optimization (DRO) with ellipsoidal moment set; DRO KL, DRO with KL-divergence set; Obj. Val., average optimal objective value of the 1,000 solutions generated from the independent data sets.

Section 4.1, we set c to be the negative of the mean of ξ, and $b \in \mathbb{R}$ is an arbitrarily chosen positive constant.

Regarding the comparison with SCA, this problem is supplementary to the Gaussian cases in Section 4.1 in that it presents performances of SCA when we use less information about ξ. Suppose that we have chosen a correct perturbation model in the SCA (i.e., knowledge of d, L, a_i and the boundedness on $[-1, 1]$). We use the Hoeffding inequality to replace the chance constraint with

$$\eta\sqrt{\sum_{i=1}^{L}(a_i'x)^2} \le b - a_0'x,$$

where $\eta \ge \sqrt{2\log(1/\epsilon)}$. This SCA is equivalent to an RO imposing an uncertainty set $\mathcal{U} = \{\zeta : \|\zeta\|_2 \le \eta\}$, where $\zeta = (\zeta_i)'_{i=1,\dots,L}$ is the vector of perturbation random variables (Ben-Tal et al. 2009, section 2.3).

Table 5 shows the results from different approaches with sample size $n = 120$. Our RO performs better than SCA in terms of achieved objective values (−988.78 versus −815.06), the latter appearing more conservative than the example in Section 4.1, as shown by $\hat{\epsilon} = 0$. Also, as in the preceding example, reconstruction boosts further our RO performance (from −988.78 to −1,087.85). Our RO here performs better than SCA because the latter, derived on a worst-case basis, does not tightly apply to the "truth" in this example; that is, the Hoeffding bound does not lead to tight performance guarantees on the scaled beta distribution (putting aside the assumed knowledge of d, L, a_i and the boundedness on $[-1, 1]$ when applying the SCA). Note that because SCA also has an RO interpretation, our observations show the superiority of our geometry or size selection of the uncertainty set. Our fully nonparametric approach shows a full-fledged advantage over SCA in this example.

We also report the outcomes of SG, which breaks down as shown by $\hat{\delta}$ being much bigger than 0.05 because 120 observations are not enough to achieve the needed feasibility confidence. FAST obtains valid solutions and outperforms our RO with ellipsoidal sets but underperforms our reconstructed RO in terms of achieved objective value. Moment-based DRO also obtains valid solutions but is conservative, as shown by $\hat{\delta} = 0$ and $\hat{\epsilon} = 0$. Its objective value underperforms our RO approaches. For divergence-based DRO, the poor construction of a reference distribution again leads to a large divergence ball size, which renders the adjusted tolerance level ϵ^* to be zero in all but one of 1,000 replications (for the one replication, where ϵ^* is nonzero, it is $\epsilon^* = 1.10 \times 10^{-11}$)

Table 14. Optimality and Feasibility Performances on a Joint $d = 11$-imensional Linear CCP with t-Distribution for Several Methods, Using Sample Size $n = 336$

Method	RO	Recon	SG	FAST	DRO Mo	DRO KL
n	336	336	336	336	336	336
n_1	212	212	—	318	—	168
n_2	124	124	—	18	—	168
Obj. Val.	−5,778.44	−7,562.60	−7,387.98	−7,173.97	−3,891.83	0
$\hat{\epsilon}$	0.00248	0.0133	0.0144	0.0126	3.97×10^{-4}	0
$\hat{\delta}$	0	0	0	0	0	0

Note. RO, robust optimization; Recon, reconstructed RO; SG, scenario generation; FAST, Fast Algorithm for the Scenario Technique; DRO Mo, distributionally robust optimization (DRO) with ellipsoidal moment set; DRO KL, DRO with KL-divergence set; Obj. Val., average optimal objective value of the 1,000 solutions generated from the independent data sets.

Table 15. Optimality and Feasibility Performances on a Single d = 11-imensional Linear CCP with Log-Normal Distribution for Several Methods, Using Sample Size n = 120

Method	RO	Recon	SG	FAST	DRO Mo	DRO KL
n	120	120	120	120	120	120
n_1	60	60	—	61	—	60
n_2	60	60	—	59	—	60
Obj. Val.	–294.00	–588.58	–784.27	–510.38	–418.30	0
$\hat{\epsilon}$	1.45×10^{-4}	0.0164	0.0902	0.0159	5.11×10^{-4}	0
$\hat{\delta}$	0	0.041	0.961	0.048	0	0

Note. RO, robust optimization; Recon, reconstructed RO; SG, scenario generation; FAST, Fast Algorithm for the Scenario Technique; DRO Mo, distributionally robust optimization (DRO) with ellipsoidal moment set; DRO KL, DRO with KL-divergence set; Obj. Val., average optimal objective value of the 1,000 solutions generated from the independent data sets.

and essentially outputs the origin as the solution all the time. In this example, our reconstructed RO performs the best among all considered approaches.

4.3. Test Case 3: Multivariate Gaussian on Joint Chance Constraints

We consider a joint CCP with $d = 11$ variables and $l = 15$ constraints in the form

$$\min c'x \quad \text{subject to} \quad P(Ax \le b) \ge 1 - \epsilon,\ x \ge 0, \qquad (12)$$

where $c \in \mathbb{R}^{11}$ and $b \in \mathbb{R}^{15}$ are arbitrary constants, and b is positive in each element. The random vector $\xi = vec(A)$ is generated from a multivariate Gaussian distribution with mean $vec(\bar{A})$ and covariance matrix Σ, where $\bar{A} \in \mathbb{R}^{15\times11}$ is arbitrary, and $\Sigma \in \mathbb{R}^{165\times165}$ is also an arbitrary positive-definite matrix.

Tables 6 and 7 present the experimental results using two different sample sizes on the same problem. We use diagonalized ellipsoids in our RO and conduct reconstruction with scaling parameters k_i described in Section EC.4.3 of the online appendix. To use DRO and SCA, we apply the Bonferroni correction to decompose the joint CCP by evenly dividing the tolerance level into ϵ/m to create individual chance constraints. For each individual chance constraint, we construct DRO and SCA constraints following the scheme in Section 4.1.

Comparing with scenario approaches, we see that much like the examples in Sections 4.1 and 4.2, SG fails with a small sample size (confirmed by $\hat{\delta}$ much larger than 0.05 in Table 6) but obtains valid solutions as the sample size grows (confirmed by $\hat{\delta} < 0.05$ in Table 7). Although reconstruction improves the optimal values for RO in both cases, SG (and FAST as well) gives better optimal value (–9,130.95) than the reconstructed RO (–8,029.83) under a big sample size. Moment-based DRO appears very conservative for both small and large sample cases because the obtained average objective values (–3,996.87 and –4,209.86) are much greater than other approaches, including our ROs, and the associated $\hat{\epsilon}$ and $\hat{\delta}$ are zero. As in the preceding experiments, divergence-based DRO outputs the origin as the solution and gives objective value zero because of oversized uncertainty sets. By contrast, SCA obtains a better solution than our ROs, thanks to the tightness of the approximation for Gaussian distributions.

4.4. Test Case 4: Beta Models on Joint Chance Constraints

We consider the joint CCP in (12) with a bounded random vector ξ. We use the perturbation model described in Section 4.2, where $d = 165$ and $L = 165$, and $a_i \in \mathbb{R}^{165}, i = 1, \ldots, L$, are arbitrarily chosen vectors,

Table 16. Optimality and Feasibility Performances on a Single d = 11-Dimensional Linear CCP with Log-Normal Distribution for Several Methods, Using Sample Size n = 336

Method	RO	Recon	SG	FAST	DRO Mo	DRO KL
n	336	336	336	336	336	336
n_1	212	212	—	318	—	168
n_2	124	124	—	18	—	168
Obj. Val.	–354.10	–685.01	–683.60	–646.83	–429.75	0
$\hat{\epsilon}$	8.07×10^{-14}	0.0243	0.0333	0.0261	3.33×10^{-14}	0
$\hat{\delta}$	0	0.057	0.052	0.033	0	0

Note. RO, robust optimization; Recon, reconstructed RO; SG, scenario generation; FAST, Fast Algorithm for the Scenario Technique; DRO Mo, distributionally robust optimization (DRO) with ellipsoidal moment set; DRO KL, DRO with KL-divergence set; Obj. Val., average optimal objective value of the 1,000 solutions generated from the independent data sets.

Table 17. Optimality and Feasibility Performances on a Single $d = 100$-Dimensional Linear CCP with Log-Normal Distribution for Several Methods, Using Sample Size $n = 120$

Method	RO	Recon	SG	FAST	DRO Mo	DRO KL
n	120	120	120	120	120	120
n_1	60	60	—	61	—	60
n_2	60	60	—	59	—	60
Obj. Val.	−309.93	−784.24	Unbounded	Unbounded	−1,030.52	0
$\hat{\epsilon}$	6.00×10^{-14}	0.0174	—	—	0.2772	0
$\hat{\delta}$	0	0.063	—	—	1	0

Notes. Results on moment-based DRO are based on 30 replications because of the high computational demand. RO, robust optimization; Recon, reconstructed RO; SG, scenario generation; FAST, Fast Algorithm for the Scenario Technique; DRO Mo, distributionally robust optimization (DRO) with ellipsoidal moment set; DRO KL, DRO with KL-divergence set; Obj. Val., average optimal objective value of the 1,000 solutions generated from the independent data sets.

and the same random variables for ζ_i are as in Section 4.2. Again, we apply the Bonferroni correction to invoke DRO and SCA as in Section 4.3 and the corresponding schemes for each individualized chance constraint as in Section 4.2.

Tables 8 and 9 show our experimental results. The major difference with Section 4.3 is that now our reconstructed RO outperforms all other methods, including SG and SCA. It gives smaller objective values than FAST under both small and large sample sizes. It also gives smaller objective values than SG under a large sample size, whereas SG does not give valid solutions under a small sample size. SCA is very conservative in this case, and DROs (both moment and divergence based) continue to be very conservative and are significantly outperformed by our RO.

4.5. Test Case 5: *t*-Distributions and Log-Normal Distributions

We consider problems with two heavier-tailed distributions, namely, *t*-distributions and log-normal distributions. We test both the single CCP (11) and the joint CCP (12) with different dimensions and sample sizes. Because the considered SCA does not apply to these distributions, we do not include it in our comparisons here.

Table 18. Optimality and Feasibility Performances on a Joint $d = 11$-Dimensional Linear CCP with Log-Normal Distribution for Several Methods, Using Sample Size $n = 120$

Method	RO	Recon	SG	FAST	DRO Mo	DRO KL
n	120	120	120	120	120	120
n_1	60	60	—	61	—	60
n_2	60	60	—	59	—	60
Obj. Val.	−0.1284	−1.1166	−4.5359	−1.0369	−0.8360	0
$\hat{\epsilon}$	0.00228	0.0157	0.0598	0.0165	0.0131	0
$\hat{\delta}$	0	0.043	0.646	0.044	0.006	0

Note. RO, robust optimization; Recon, reconstructed RO; SG, scenario generation; FAST, Fast Algorithm for the Scenario Technique; DRO Mo, distributionally robust optimization (DRO) with ellipsoidal moment set; DRO KL, DRO with KL-divergence set; Obj. Val., average optimal objective value of the 1,000 solutions generated from the independent data sets.

Tables 10–12 show the comparisons among different approaches for the single CCP, and Tables 13 and 14 show the counterparts for joint CCP when ξ is generated from a multivariate *t*-distribution with degree of freedom five and an arbitrary positive-definite dispersion matrix. The comparisons are largely consistent with the Gaussian and beta cases shown in previous subsections. Compared with SG, our ROs output feasible solutions in the small sample case ($n = 120$), whereas SG struggles to obtain feasible solutions ($\hat{\delta}$ much greater than 0.05 in Tables 10 and 13). In the large sample case ($n = 336$), SG gains enough feasibility and outperforms our plain RO in average objective value (−1,175.04 versus −1,126.66 in the single CCP case in Table 11 and −7,387.98 versus −5,778.44 in the joint CCP case in Table 14) but underperforms our reconstructed RO (−1,175.64 and −7,562.60 for single and joint CCPs, respectively). FAST remedies the infeasibility issue of SG in the small sample cases and outperforms our plain RO. By contrast, our reconstructed RO performs competitively against FAST. Among all four cases where $d = 11$, the reconstructed RO outperforms FAST in three cases but underperforms in the case of the small sample joint CCP (average objective values −1,166.52, −1,175.64, and −7,562.60 versus −1,158.27, −1,170.35, and −7,173.97 in Tables 10, 11, and 14, respectively, and −6,499.93 versus −7,220.37 in Table 13). Note that when the dimension is large ($d = 100$ in Table 12), SG and FAST output unbounded solutions in all 1,000 experimental replications, whereas plain and reconstructed RO output feasible bounded solutions.

As in previous subsections, our reconstructed RO outperforms moment-based DRO in all cases. When the dimension is large ($d = 100$ in Table 12), moment-based DRO fails to obtain feasible solutions in all 30 replications, attributed to the difficulty in estimating

Table 19. Optimality and Feasibility Performances on a Joint $d = 11$-Dimensional Linear CCP with Log-Normal Distribution for Several Methods, Using Sample Size $n = 336$

Method	RO	Recon	SG	FAST	DRO Mo	DRO KL
n	336	336	336	336	336	336
n_1	212	212	—	318	—	168
n_2	124	124	—	18	—	168
Obj. Val.	−0.0844	−1.9373	−1.7135	−1.4058	−1.2021	0
$\hat{\epsilon}$	0.0074	0.0239	0.0238	0.0197	0.0131	0
$\hat{\delta}$	0	0.05	0.011	0.007	0.026	0

Note. RO, robust optimization; Recon, reconstructed RO; SG, scenario generation; FAST, Fast Algorithm for the Scenario Technique; DRO Mo, distributionally robust optimization (DRO) with ellipsoidal moment set; DRO KL, DRO with KL-divergence set; Obj. Val., average optimal objective value of the 1,000 solutions generated from the independent data sets.

valid moment confidence regions. Compared with our plain RO, moment-based DRO outperforms in a single CCP (−1,134.38 and −1,137.19 versus −1,112.75 and −1,126.66 in Tables 10 and 11, respectively) but underperforms in a joint CCP (−3,888.63 and −3,891.83 versus −4,229.6 and −5,778.44 in Tables 13 and 14, respectively). Lastly, divergence-based DRO is once again very conservative, resulting in zero objective values all the time.

Next, we consider ξ generated from log-normal distributions with arbitrarily chosen means and covariance matrices. Tables 15–17 show the results for the single CCP, and Tables 18 and 19 show those for the joint CCP. The comparisons are quite similar to the t-distribution cases. SG in a small sample outputs invalid solutions ($\hat{\delta}$ much greater than 0.05) and in a large sample outputs solutions with average objective values (e.g., −683.60 in Table 16) better than our plain RO (−354.10) but worse than our reconstructed RO (−685.01). FAST remedies the infeasibility issue of SG in the small sample cases but underperforms our reconstructed RO in all cases. Moment-based DRO outperforms our plain RO but underperforms our reconstructed RO in all cases, and it continues to struggle in obtaining feasible solutions for high-dimensional problems ($\hat{\delta} = 1$ in Table 17). Lastly, divergence-based DRO continues to be conservative and outputs zero objective values. In all considered settings, reconstructed RO appears the best among all compared methods in terms of feasibility and optimality.

4.6. Summary of the Experiment Results

From the results in this section (and additional ones in Section EC.7 of the online appendix), we highlight the following situations where our method is the most recommended. The competitiveness of our method compared with scenario approaches is most seen in small sample situations. Classical SG needs a much larger sample size than ours to achieve feasibility. FAST is capable of obtaining feasible solutions in small sample cases but appears more susceptible than RO in generating unbounded solutions. With reconstruction, our approach tends to work as well as SG and FAST for large samples (when they are all applicable). Moreover, our reconstruction has the capability to improve the optimality over plain RO, whereas FAST is by design always more conservative than SG in terms of optimality. Nonetheless, we should mention that some constraint-removal approaches such as sampling and discarding (Campi and Garatti 2011) can improve SG performances in large sample situations.

Compared with our ROs, moment-based DRO can generate infeasible solutions when the problem dimension is high compared with data size (e.g., $d = 100$ and $n = 120$), which is attributed to the difficulty in constructing valid moment confidence regions. In cases where moment-based DRO generates valid solutions, the solution performances seem to be sometimes better, sometimes worse than our plain RO, but in all considered instances, they perform worse than our reconstructed RO. KL-divergence-based DRO appears to perform poorly in the experiments because of the challenge in obtaining a small enough divergence ball size. (To get a further sense of this behavior, we investigate a very low-dimensional problem ($d = 3$) with a sufficient sample size in Section EC.7.3 of the online appendix, where divergence-based DRO provides nontrivial but still conservative solutions.)

Lastly, compared with SCA, our performance is best seen when the data are nonnormal. In this case, the approximate constraint in SCA may not tightly approximate the original chance constraint and tends to be significantly more conservative than our approach. Moreover, SCA generally requires at least some partial distributional knowledge (e.g., moments, support) in deriving the needed relaxing constraint, in contrast to our approach, which is fully data driven and nonparametric.

Acknowledgment

The authors wish to acknowledge the funding agencies that supported them.

References

Ben-Tal A, Nemirovski A (1998) Robust convex optimization. *Math. Oper. Res.* 23(4):769–805.

Ben-Tal A, Nemirovski A (1999) Robust solutions of uncertain linear programs. *Oper. Res. Lett.* 25(1):1–13.

Ben-Tal A, Nemirovski A (2000) Robust solutions of linear programming problems contaminated with uncertain data. *Math. Programming* 88(3):411–424.

Ben-Tal A, El Ghaoui L, Nemirovski A (2009) *Robust Optimization* (Princeton University Press, Princeton, NJ).

Ben-Tal A, Den Hertog D, De Waegenaere A, Melenberg B, Rennen G (2013) Robust solutions of optimization problems affected by uncertain probabilities. *Management Sci.* 59(2): 341–357.

Bertsimas D, Sim M (2004) The price of robustness. *Oper. Res.* 52(1):35–53.

Bertsimas D, Sim M (2006) Tractable approximations to robust conic optimization problems. *Math. Programming* 107(1–2):5–36.

Bertsimas D, Brown DB, Caramanis C (2011) Theory and applications of robust optimization. *SIAM Rev.* 53(3):464–501.

Bertsimas D, Gupta V, Kallus N (2018) Data-driven robust optimization. *Math. Programming* 167(2):235–292.

Bertsimas D, Pachamanova D, Sim M (2004) Robust linear optimization under general norms. *Oper. Res. Lett.* 32(6):510–516.

Blanchet J, Kang Y (2016) Sample out-of-sample inference based on Wasserstein distance. Preprint, submitted May 4, https://arxiv.org/abs/1605.01340.

Boyd S, Vandenberghe L (2004) *Convex Optimization* (Cambridge University Press, Cambridge, UK).

Calafiore GC (2017) Repetitive scenario design. *IEEE Trans. Automatic Control* 62(3):1125–1137.

Calafiore G, Campi MC (2005) Uncertain convex programs: Randomized solutions and confidence levels. *Math. Programming* 102(1):25–46.

Calafiore GC, Campi MC (2006) The scenario approach to robust control design. *IEEE Trans. Automatic Control* 51(5):742–753.

Calafiore GC, El Ghaoui L (2006) On distributionally robust chance-constrained linear programs. *J. Optim. Theory Appl.* 130(1):1–22.

Calafiore GC, Dabbene F, Tempo R (2011) Research on probabilistic methods for control system design. *Automatica* 47(7):1279–1293.

Campi MC, Carè A (2013) Random convex programs with L_1-regularization: Sparsity and generalization. *SIAM J. Control Optim.* 51(5):3532–3557.

Campi MC, Garatti S (2008) The exact feasibility of randomized solutions of uncertain convex programs. *SIAM J. Optim.* 19(3): 1211–1230.

Campi MC, Garatti S (2011) A sampling-and-discarding approach to chance-constrained optimization: Feasibility and optimality. *J. Optim. Theory Appl.* 148(2):257–280.

Campi MC, Garatti S (2018) Wait-and-judge scenario optimization. *Math. Programming* 167(1):155–189.

Carè A, Garatti S, Campi MC (2014) FAST: Fast algorithm for the scenario technique. *Oper. Res.* 62(3):662–671.

Chamanbaz M, Dabbene F, Tempo R, Venkataramanan V, Wang QG (2016) Sequential randomized algorithms for convex optimization in the presence of uncertainty. *IEEE Trans. Automatic Control* 61(9):2565–2571.

Charnes A, Cooper WW (1959) Chance-constrained programming. *Management Sci.* 6(1):73–79.

Charnes A, Cooper WW, Symonds GH (1958) Cost horizons and certainty equivalents: An approach to stochastic programming of heating oil. *Management Sci.* 4(3):235–263.

Chen X, Sim M, Sun P (2007) A robust optimization perspective on stochastic programming. *Oper. Res.* 55(6):1058–1071.

Chen W, Sim M, Sun J, Teo CP (2010) From CVaR to uncertainty set: Implications in joint chance-constrained optimization. *Oper. Res.* 58(2):470–485.

De Farias DP, Van Roy B (2004) On constraint sampling in the linear programming approach to approximate dynamic programming. *Math. Oper. Res.* 29(3):462–478.

Delage E, Ye Y (2010) Distributionally robust optimization under moment uncertainty with application to data-driven problems. *Oper. Res.* 58(3):595–612.

Dentcheva D, Lai B, Ruszczyński A (2004) Dual methods for probabilistic optimization problems. *Math. Methods Oper. Res.* 60(2): 331–346.

Donoho DL, Gasko M (1992) Breakdown properties of location estimates based on halfspace depth and projected outlyingness. *Ann. Statist.* 20(4):1803–1827.

Duchi J, Glynn P, Namkoong H (2016) Statistics of robust optimization: A generalized empirical likelihood approach. Preprint, submitted October 11, https://arxiv.org/abs/1610.03425.

Durrett R (2010) *Probability: Theory and Examples* (Cambridge University Press, Cambridge, UK).

El Ghaoui L, Oks M, Oustry F (2003) Worst-case value-at-risk and robust portfolio optimization: A conic programming approach. *Oper. Res.* 51(4):543–556.

El Ghaoui L, Oustry F, Lebret H (1998) Robust solutions to uncertain semidefinite programs. *SIAM J. Optim.* 9(1):33–52.

Erdoğan E, Iyengar G (2006) Ambiguous chance constrained problems and robust optimization. *Math. Programming* 107(1–2): 37–61.

Goh J, Sim M (2010) Distributionally robust optimization and its tractable approximations. *Oper. Res.* 58(4, part 1):902–917.

Goldfarb D, Iyengar G (2003) Robust portfolio selection problems. *Math. Oper. Res.* 28(1):1–38.

Gupta V (2019) Near-optimal ambiguity sets for distributionally robust optimization. *Management Sci.* 65(9):3949–4450.

Hallin M, Paindaveine D, Šiman M, Wei Y, Serfling R, Zuo Y, Kong L, Mizera I (2010) Multivariate quantiles and multiple-output regression quantiles: From *L_*1 optimization to halfspace depth. *Ann. Statist.* 38(2):635–669.

Hanasusanto GA, Roitch V, Kuhn D, Wiesemann W (2015) A distributionally robust perspective on uncertainty quantification and chance constrained programming. *Math. Programming* 151(1): 35–62.

Hanasusanto GA, Roitch V, Kuhn D, Wiesemann W (2017) Ambiguous joint chance constraints under mean and dispersion information. *Oper. Res.* 65(3):751–767.

Hastie T, Tibshirani R, Friedman J (2009) Unsupervised learning. *The Elements of Statistical Learning: Data Mining, Inference, and Prediction* (Springer, New York), 485–585.

Hodges JL (1955) A bivariate sign test. *Ann. Math. Statist.* 26(3):523–527.

Hong LJ, Yang Y, Zhang L (2011) Sequential convex approximations to joint chance constrained programs: A Monte Carlo approach. *Oper. Res.* 59(3):617–630.

Hu Z, Hong LJ, Zhang L (2013) A smooth Monte Carlo approach to joint chance-constrained programs. *IIE Trans.* 45(7):716–735.

Jiang R, Guan Y (2016) Data-driven chance constrained stochastic program. *Math. Programming* 158(1–2):291–327.

Lagoa CM, Barmish BR (2002) Distributionally robust Monte Carlo simulation: A tutorial survey. Camacho EF, Basanez L, de la Puente JA, eds. *Proc. IFAC World Congress*, vol. 35, issue 1, subvolume E (IFAC, New York), 1–12.

Lam H (2019) Recovering best statistical guarantees via the empirical divergence-based distributionally robust optimization. *Oper. Res.* 67(4):1090–1105.

Lam H, Mottet C (2017) Tail analysis without parametric models: A worst-case perspective. *Oper. Res.* 65(6):1696–1711.

Lam H, Zhou E (2017) The empirical likelihood approach to quantifying uncertainty in sample average approximation. *Oper. Res. Lett.* 45(4):301–307.

Lejeune MA, Ruszczynski A (2007) An efficient trajectory method for probabilistic production-inventory-distribution problems. *Oper. Res.* 55(2):378–394.

Li B, Jiang R, Mathieu JL (2019) Ambiguous risk constraints with moment and unimodality information. *Math. Programming* 173(1–2): 151–192.

Li J, Liu RY (2008) Multivariate spacings based on data depth: I. Construction of nonparametric multivariate tolerance regions. *Ann. Statist.* 36(3):1299–1323.

Lim AE, Shanthikumar JG, Shen ZM (2006) Model uncertainty, robust optimization, and learning. Johnson MP, Norman B, Secomandi N, eds. *Models, Methods, and Applications for Innovative Decision Making*, TutORials in Operations Research (INFORMS, Catonsville, MD), 66–94.

Liu H, Wasserman L, Lafferty JD (2012) Exponential concentration for mutual information estimation with application to forests. Pereira F, Burges CJC, Bottou L, Weinberger KQ, eds. *Advances in Neural Information Processing Systems*, vol. 25 (Curran Associates, Red Hook, NY), 2537–2545.

Liu RY (1990) On a notion of data depth based on random simplices. *Ann. Statist.* 18(1):405–414.

Luedtke J (2014) A branch-and-cut decomposition algorithm for solving chance-constrained mathematical programs with finite support. *Math. Programming* 146(1–2):219–244.

Luedtke J, Ahmed S (2008) A sample approximation approach for optimization with probabilistic constraints. *SIAM J. Optim.* 19(2): 674–699.

Luedtke J, Ahmed S, Nemhauser GL (2010) An integer programming approach for linear programs with probabilistic constraints. *Math. Programming* 122(2):247–272.

Mahalanobis PC (1936) On the generalized distance in statistics. *Proc. National Inst. Sci.* 2(1):49–55.

Marandi A, Ben-Tal A, den Hertog D, Melenberg B (2019) Extending the scope of robust quadratic optimization. Preprint, submitted September 4, https://arxiv.org/abs/1909.01762.

Margellos K, Goulart P, Lygeros J (2014) On the road between robust optimization and the scenario approach for chance constrained optimization problems. *IEEE Trans. Automatic Control* 59(8):2258–2263.

Miller BL, Wagner HM (1965) Chance constrained programming with joint constraints. *Oper. Res.* 13(6):930–945.

Moon K, Hero A (2014) Multivariate *f*-divergence estimation with confidence. Ghahramani Z, Welling M, Cortes C, eds. *Advances in Neural Information Processing Systems*, vol. 27 (Curran Associates, Red Hook, NY), 2420–2428.

Murr MR, Prékopa A (2000) Solution of a product substitution problem using stochastic programming. Uryasev SP, ed. *Probabilistic Constrained Optimization* (Springer, Boston), 252–271.

Nemirovski A (2003) On tractable approximations of randomly perturbed convex constraints. *Proc. 42nd IEEE Conf. Decision Control*, vol. 3 (IEEE, New York), 2419–2422.

Nemirovski A, Shapiro A (2006) Convex approximations of chance constrained programs. *SIAM J. Optim.* 17(4):969–996.

Pál D, Póczos B, Szepesvári C (2010) Estimation of Rényi entropy and mutual information based on generalized nearest-neighbor graphs. Lafferty JD, Williams CKI, Shawe-Taylor J, Zemel RS, Culotta A, eds. *Advances in Neural Information Processing Systems*, vol. 23 (Curran Associates, Red Hook, NY), 1849–1857.

Póczos B, Schneider JG (2012) Nonparametric estimation of conditional information and divergences. Lawrence ND, Girolami M, eds. *Proc. 15th Internat. Conf. Artificial Intelligence Statist.*, vol. 22 (PMLR), 914–923.

Póczos B, Xiong L, Schneider J (2012) Nonparametric divergence estimation with applications to machine learning on distributions. Preprint, submitted February 14, https://arxiv.org/abs/1202.3758.

Popescu I (2005) A semidefinite programming approach to optimal-moment bounds for convex classes of distributions. *Math. Oper. Res.* 30(3):632–657.

Prékopa A (1970) On probabilistic constrained programming. Kuhn HW, ed. *Proc. Princeton Sympos. Math. Programming* (Princeton University Press, Princeton, NJ), 113–138.

Prékopa A (2003) Probabilistic programming. Ruszczynski A, Shapiro A, eds. *Stochastic Programming* (Elsevier, Amsterdam), 267–351.

Prékopa A, Szántai T (1978) Flood control reservoir system design using stochastic programming. Balinski ML, Lemarechal C, eds. *Mathematical Programming in Use* (Springer, Berlin), 138–151.

Prékopa A, Rapcsák T, Zsuffa I (1978) Serially linked reservoir system design using stochastic programing. *Water Resources Res.* 14(4):672–678.

Scarf H (1958) A min-max solution of an inventory problem. Arrow K, Karlin S, Scarf H, eds. *Studies in the Mathematical Theory of Inventory and Production* (Stanford University Press, Stanford, CA), 201–209.

Schildbach G, Fagiano L, Morari M (2013) Randomized solutions to convex programs with multiple chance constraints. *SIAM J. Optim.* 23(4):2479–2501.

Scott CD, Nowak RD (2006) Learning minimum volume sets. *J. Machine Learn. Res.* 7:665–704.

Serfling R (2002) Quantile functions for multivariate analysis: Approaches and applications. *Statistica Neerlandica* 56(2):214–232.

Serfling RJ (2009) *Approximation Theorems of Mathematical Statistics* (Wiley, New York).

Shi Y, Zhang J, Letaief KB (2015) Optimal stochastic coordinated beamforming for wireless cooperative networks with CSI uncertainty. *IEEE Trans. Signal Processing* 63(4):960–973.

Tukey JW (1975) Mathematics and the picturing of data. James RD, ed. *Proc. Internat. Congress Mathematicians* (Canadian Mathematical Congress, Ottawa), vol. 2, 523–531.

Tulabandhula T, Rudin C (2014) Robust optimization using machine learning for uncertainty sets. Preprint, submitted July 4, https://arxiv.org/abs/1407.1097.

Van Parys BP, Goulart PJ, Kuhn D (2016) Generalized Gauss inequalities via semidefinite programming. *Math. Programming* 156(1–2):271–302.

Wang Q, Kulkarni SR, Verdú S (2009) Divergence estimation for multidimensional densities via k -nearest-neighbor distances. *IEEE Trans. Inform. Theory* 55(5):2392–2405.

Wang Z, Glynn PW, Ye Y (2016) Likelihood robust optimization for data-driven problems. *Comput. Management Sci.* 13(2): 241–261.

Wiesemann W, Kuhn D, Sim M (2014) Distributionally robust convex optimization. *Oper. Res.* 62(6):1358–1376.

Zuo Y (2003) Projection-based depth functions and associated medians. *Ann. Statist.* 31(5):1460–1490.