DISTRIBUTIONALLY ROBUST CHANCE-CONSTRAINED OPTIMIZATION*

Abstract. This document provides an overview of decisionmaking under uncertainty within a distributionally robust chance-constrained (DRCC) framework. The DRCC framework sits between two well-known optimization paradigms: (1) robust optimization (RO) and (2) stochastic optimization (SO). We use a third framework, (3) data-driven optimization (and two numerical studies), to bridge RO and SO and contextualize a computationally tractable form of the DRCC model.

1. Introduction. Optimization has primarily explored problems involving uncertainty in two complementary directions: robust optimization (RO) and stochastic optimization (SO). In standard problems such as

$$\begin{aligned} \underset{x}{\text{minimize}} \quad & f_0(x) \\ \text{subject to} \quad & f_i(x) \leq 0, \quad i = 1, \ldots, m \end{aligned} \tag{1}$$

often we lack precise information about the objective and constraints. For example, in a linear constraint of the form $f_1(x) = Ax - b \leq 0$, the problem data A and b may not be known fully. Robust optimization and stochastic optimization address this uncertainty in different but related ways.

Broadly, RO takes the view that the true problem data are unknown (e.g., possibly due to measurement error) but that this uncertainty is deterministic in the sense that it is representable with a set $\mathcal{U}$ that entirely describes the problem uncertainty. In this setting, a (robust-) feasible solution is one that is guaranteed to remain feasible for all possible data realizations in $\mathcal{U}$. The uncertainty set may may directly bound the variation in each entry A_{ij} and b_i (e.g., box constraints), or it may specify that A and b reside in some ellipsoidal set. Common to both approaches and characteristic of RO broadly are explicit uncertainty set assumptions that allow formulation of computationally tractable programs [1].

On the other hand, stochastic optimization assigns inherent randomness to the data, e.g., the measurement error has a known (or at least estimable) distribution. This means that we can think of the quantities $f_i(x)$ as dependent on additional uncertain factors. For example, we might know how $f_i(x)$ behaves in various "states" of the world, and we might parametrize its behavior in these states by a vector $\omega \in \Omega \subseteq \mathbb{R}^d$ which contains the necessary information about the environment to determine $f_i(x)$. Under this framework, a feasible solution is one that satisfies constraints of the form $f_i(x, \omega) \leq 0$ where ω can be described by some distribution D. In this view, $f_i(x, \omega)$ becomes the basic object of interest, and in our linear constraint case, this may be represented as $f_1(x, \omega) \coloneqq A(\omega)x - b(\omega) \leq 0$.

In contrast to RO, SO may only require that the constraints hold for $\omega \in \Omega$ with high probability. This means that a solution which satisfies a probabilistic constraint may break the deterministic version of the constraint. In this way, there is a natural interpretation of SO as a relaxation of RO. However, it is also possible to construct an uncertainty set $\mathcal{U}$ that replicates the SO uncertainty, though in general this is more difficult. This direction corresponds to (1) identifying the particular set of states $\hat{\Omega} \subset \Omega$ where the SO problem satisfies the constraint; and (2) equating $\mathcal{U}$ with $\hat{\Omega}$. The difficulty of translating between RO and SO formulations depends on the problem, but access to prior data aids in the process of determining $\mathcal{U}$ or estimating

*July 23, 2018

the distribution over Ω. In this way, a data-driven approach helps interpolate between RO and SO formulations.

Organization. In this paper, we briefly review robust, stochastic, and data-driven optimization frameworks (section 2) before discussing chance-constrained programming (section 3) and its distributionally robust extension interpreted as both an SO and RO method when utilizing prior data. We then present results from two numerical studies (section 4): (1) a portfolio optimization problem where we compare standard RO and SO approaches with a data-driven DRCC method; and (2) a semi-infinite, continuously indexed chance-constrained toy problem where we characterize the problem's feasible set and worst-case distribution on a bounded support in one dimension. These examples highlight the similarities and differences among the RO, SO, and data-driven approaches. Finally, we discuss the relative merits of the data-driven DRCC approach in light of the simulation studies (section 5).

2. Review of optimization under uncertainty.

In this section, we introduce robust, stochastic, and data-driven optimization, and motivate relevant definitions and background.

2.1. Robust optimization.

Robust optimization is a framework that extends standard optimization approaches to handle uncertainty via an explicit, deterministic set $\mathcal{U}$ that "contains" the problem's uncertainty. That is, $\mathcal{U}$ is often a polyhedral or ellipsoidal set which contains the uncertain parameters involved in the problem. Thus, $\mathcal{U}$ summarizes our knowledge of the unknown in a (semi) non-parametric way. We can represent a standard problem as

$$\underset{x}{\text{minimize}} \left\{ \sup_{f_i \in \mathcal{U}} f_0(x) : f_i(x) \leq 0, \quad i = 1, \ldots, m, \quad \forall f_i \in \mathcal{U} \right\} \tag{2}$$

where $f_i \in \mathcal{U}$ means that $\mathcal{U}$ contains the uncertain problem data that defines the relationship $f_i(x)$. However, it is also useful to think of the above problem more parametrically such that f_i is defined on both x and u where $u \in \mathcal{U}$ is a vector representing the uncertain problem data. In this way, we can also think of (2) as

$$\underset{x}{\text{minimize}} \left\{ \sup_{u \in \mathcal{U}} f_0(x, u) : f_i(x, u) \leq 0, \quad i = 1, \ldots, m, \quad \forall u \in \mathcal{U} \right\}. \tag{3}$$

In this form, we see that RO seeks the minimax optimal over all possible outcomes specified by the uncertainty set. Problems (2) and (3) are formulations of the original, uncertain problem, but the program we solve numerically, called the *robust counterpart*, can be framed as

$$\underset{x,t}{\text{minimize}} \{ t : f_0(x, u) \leq t, \ f_i(x, u) \leq 0, \quad i = 1, \ldots, m, \quad \forall u \in \mathcal{U} \}. \tag{4}$$

Without loss of generality, (4) can be simplified by the following assumptions [9]:

- the objective is certain (it is always possible to reformulate in epigraph form)
- the RHS constraint is certain (it is always possible to add an additional variable to the LHS with a random coefficient to adjust the inequality)
- $\mathcal{U}$ is compact and convex (it is always possible to replace $\mathcal{U}$ with its convex hull)
- the uncertainty is constraint-wise (uncertainty in the data is equivalent to its 1-dimensional projections, e.g., we convert a linear constraint into its robust counterpart as $Ax \leq b \iff a_i^T x \leq b_i \, \forall [a_i, b_i] \in \mathcal{U}_i$)

When solving RO problems, it is often possible to cast the formulation into a standard convex problem. Frequently, we assume that the uncertainty is specified by a system of linear, conic, or matrix inequality constraints, which lead to linear, SOCP, and SDP problems, respectively [14].

2.2. Stochastic optimization. In the previous section, we described uncertainty through a deterministic set $\mathcal{U}$, and our formulations sought robustness against every possible uncertainty realization $u \in \mathcal{U}$. Stochastic optimization provides a different view of the feasible set and allows us to exchange guaranteed feasibility for a less conservative solution through a probabilistic interpretation. In particular, the outcome is a random variable, and a decision can only affect the probabilities of achieving certain outcomes rather than specify certain outcomes themselves [15]. However, given this flexibility, stochastic optimization problem (in their general form) are often more difficult to solve.

Recalling the parameter vectors ω, if we assume that Ω is known, a next step might be to choose *a priori* a measure $D : \Omega \to [0, 1]$ to govern the probability of each state ω. This assumption reformulates the general problem (1) as a stochastic program where robustness can be quantified probabilistically

$$\begin{aligned} \underset{x}{\text{minimize}} \quad & F_0^D(x) = \int_\Omega f_0(x, \omega)\ D(d\omega) \\ \text{subject to} \quad & F_i^D(x) = \int_\Omega f_i(x, \omega)\ D(d\omega) \leq 0, \quad i = 1, \ldots, m \end{aligned} \tag{5}$$

where F_i^D represent integrals of functions f_i against the distribution (measure) D [7]. For example, we may be interested in the average objective value across all possible states, which we could represent as

$$F_0^D(x) = \int_\Omega f_0(x, \omega)\ D(d\omega) = \mathbb{E}_D\left[f_0(x, \omega)\right].$$

On the other hand, we may want a constraint in our original problem (1) to hold with at least an α-level of probability instead of for every uncertainty realization. Constraints of this form are called *chance-constraints* (CC). Chance-constraints are a basic building block of more sophisticated, robust, and data-driven approaches, as we will see later. We can represent their dependence on ω as

$$f_1(x, \omega) = \begin{cases} \alpha - 1, & \text{if } f_1(x) \leq 0 \text{ in state } \omega \\ \alpha, & \text{else} \end{cases}$$

so that $\mathbb{P}_D\left[f_1(x, \omega) \leq 0\right] \geq \alpha$. In the expectation objective and chance-constraint cases, the result of integrating out ω leaves the problem in terms of a complicated function of x which might not be convex.

Unfortunately, numerically solving problems such as (5) involves approximating integrals (either with quadrature or sampling methods) if the distributions are not "nice". As a solution technique, SO uses probabilistic tools and relaxations. Since the outcome is a random variable, it can be characterized in terms of its distribution function, and likewise, it is natural to use the quantile function to describe the impact of a particular decision. A standard formulation in risk-modeling is to model the true relationship of interest $f(x) \leq 0$ by an approximate relationship $\mathcal{R}(f(x)) \leq 0$ where $\mathcal{R}$ is a measure of risk that assigns a real-valued "risk" to the random variable $f(x)$.

When considering the space of possible risk measures, an important consideration is to identify those measures that preserve useful structure from the original problem. An important structure to preserve for theoretical and practical reasons is convexity. If the problem outcome is convex with respect to the decision, i.e., $f(x)$ convex in x, then a risk measure is called "coherent" if $\mathcal{R}(f(x))$ is convex in x [15]. This set of risk measures satisfies

- translation invariance: $\mathcal{R}(Z) = Z$ if Z is constant
- subadditivity: $\mathcal{R}(Z_1 + Z_2) \leq \mathcal{R}(Z_1) + \mathcal{R}(Z_2)$
- positive homogeneity: $\mathcal{R}(tZ) = t\mathcal{R}(Z)$ for $t > 0$
- monotonicity: if $Z_1 \leq Z_2$ almost surely, then $\mathcal{R}(Z_2) \leq \mathcal{R}(Z_1)$

In a risk-modeling framework, quantiles (or VaR, see Definition 3.2) are a natural way to describe the risk of an outcome, but as a risk measure, the quantile is incoherent (it violates sub-additivity). The notion of a "superquantile" (or CVaR, see Definition 3.3) was introduced as a particular convexification of the quantile function [15]. From this, we see that compared to RO, SO directly utilizes probability theory to aid in the problem modeling.

2.3. (Robust) data-driven optimization. Traditional RO and SO frameworks take an analytic approach toward framing optimization problems where the tractability of a model often depends on a successful appeal to the theory of convex optimization and its existing solvers. On the other hand, data-driven optimization (DDO) makes use of computational tools like hypothesis testing, Monte Carlo simulation, and statistical estimation to numerically approximate optimization problems involving uncertainty [3]. In particular, DDO extends RO problems through the empirically guided design of an appropriate uncertainty set from finite sample data. Likewise, it extends SO problems by generating (or analyzing an existing) finite sample of problem data and numerically approximating the stochastic objective and constraint functions with sample average approximations (SAA). These approaches increase the class of computationally efficient models provided that the functions are still convex in the appropriate arguments so that convex solvers can still be used.

In the data-driven framework, RO and SO approaches often end up sharing many features. A common RO approach for introducing an uncertainty set based on a finite sample is to consider a function $f(x, U)$ where we assume that U is a random variable from an unknown distribution. We can then learn structural information about the distribution through a finite sample of U [2]. Such approaches are often called *distributionally robust optimization* methods. In these models, constraints are often relaxed to hold with high probability over the whole uncertainty rather than with certainty for every possible realization. This is a step towards the chance-constrained programming framework of traditional SO but with uncertain distribution. The uncertainty set over distributions are often called "ambiguity" sets in this context. Note that we can achieve similar models by relaxing a chance-constraint problem with known probability distribution to one over a family of distributions; this leads to the notion of *distributionally robust chance-constraints*. In fact, authors have shown a one-to-one correspondence between uncertainty sets for certain linear optimization problems and conservative approximations to distributionally robust linear chance-constraints [1]. In this way, we see that data-driven optimization bridges RO and SO methods and offers a framework for interpolating between intuitive but computationally expensive SO problems and less intuitive but often more efficient RO problems.

Working with a finite sample in both forms of DDO motivates the question of generalization error and asymptotic performance. Following the stochastic optimization

approach, the SAA method relaxes problems of the form

$$\min_{x\in\mathcal{X}} \mathbb{E}_F[c(x,\xi)] \tag{6}$$

with cost function $c(x,\xi)$, decision variable x and uncertainty ξ by approximating them with problems of the form

$$\min_{x\in\mathcal{X}} \frac{1}{n}\sum_{i=1}^{n} c(x,\xi_i). \tag{7}$$

From the empirical estimate, we can use concentration inequalities [13] and hypothesis testing [3] bound the generalization error. In the data-driven robust optimization literature, a common approach is to use a data-driven constructions of distributionally robust chance constraints followed by a concentration inequality bound.

3. Chance-constrained programming. We now consider chance-constraints in more detail from SO, RO and data-driven perspectives. A chance-constrained problem can can be formulated as

$$\begin{aligned} \underset{x\in\mathcal{X}}{\text{minimize}} \quad & f_0(x,\xi) \\ \text{subject to} \quad & \mathbb{P}\left[f(x,\xi)\leq 0\right] \geq \alpha \end{aligned} \tag{8}$$

for

- control variable $x \in \mathcal{X} \subseteq \mathbb{R}^d$ and random parameter vector $\xi \in \Xi \subseteq \mathbb{R}^s$
- objective function $f_0 : \mathcal{X} \times \Xi \to \mathbb{R}$,
- constraint $f(x,\xi)$ consisting of m component quantities $f_1(x,\xi),\ldots,f_m(x,\xi)$ with $f_i : \mathcal{X}\times\Xi \to \mathbb{R}^p$
- distribution function $F_{f(x,\xi)}$ governing the random variable $f(x,\xi)$ (so the probability in the constraint is with respect to F).

Chance constraints represent a middle ground between computationally efficient but non-robust expectation constraints of the form

$$f(x,\mathbb{E}\left[\xi\right]) \leq 0$$

and possibly expensive, non closed-form, or overly-conservative but robust worst-case constraints of the form

$$f(x,\xi)\leq 0, \quad \forall \xi \in \Xi.$$

Thus, chance-constraints capture some level of robustness to the possible outcomes and are tractable under certain distributional assumptions. Additional considerations that are important when solving such problems include: (1) the distinction between individual and joint constraints; (2) constraint separability; and (3) convexity.

3.1. Individual vs. joint constraints. When $m > 1$, the component constraints may be enforced either individually or jointly. That is, the m constraints can be enforced *individually* as

$$\left(\mathbb{P}\left[f_1(x,\xi)\leq 0\right]\geq\alpha_1\right),\ldots,\left(\mathbb{P}\left[f_m(x,\xi)\leq 0\right]\geq\alpha_m\right), \quad \alpha_i\in\mathbb{R}$$

or *jointly* [11] as

$$\mathbb{P}\left[\{f_1(x,\xi)\leq 0\}\cap\ldots\cap\{f_m(x,\xi)\leq 0\}\right]\geq\alpha, \quad \alpha\in\mathbb{R}.$$

Joint constraints are considerably more difficult to solve because we need

- the analytic joint distribution $F(Y_1, \ldots, Y_m)$ for $Y_i = f_i(x, \xi)$; or
- an approximate joint distribution and thus simulate from $F(Y_1, \ldots, Y_m)$ using Monte Carlo [16]

unless the random variables $f_i(x, \xi)$ are independent, in which case we can formulate the constraints individually. Note that we could decompose a joint constraint into m individual constraints and apply a Bonferroni correction to bound the joint probability but that this approach is only a conservative approximation [5].

3.2. Separability. Separability refers to properties of the constraint components. If $f_i(x, \xi)$ is *separable*, then the constraint $\mathbb{P}[f_i(x, \xi) \leq 0]$ may be formulated as

$$\mathbb{P}[h_i(\xi) \leq g_i(x)] = F_{h_i(\xi)}(g_i(x))$$

for $f_i(x, \xi) = -g_i(x) + h_i(\xi)$ with distribution function $F_{h_i(\xi)}$ and appropriately defined functions g_i, h_i. Thus, the probabilistic constraint now takes an algebraic functional form based on the distribution F, and the feasible set can be analyzed in terms of the (left) quantile function, which for a random variable Z we recall is

$$F_Z^-(p) = \inf\{z \in \mathbb{R} : F_Z(z) \geq p\}. \tag{9}$$

In the separable case, assuming knowledge of distribution $F_{h_i(\xi)}$, we can describe a constraint component deterministically in terms of (9), e.g.,

$$g_i(x) \geq F_{h_i(\xi)}^-(\alpha).$$

3.3. Convexity. In both separable and non-separable cases, it is possible to formulate the chance-constrained problem as a convex problem given that f_i satisfies appropriate conditions [17].

In the separable case, because the formulation of the original constraint is equivalent to a formulation in terms of $F_\xi(g_i(x))$, the continuity and differentiability properties of F_ξ and f_i are important to the analysis of the constraint. An important property for F and the f_is to possess is α-concavity (in both arguments x and ξ). α-concavity is a generalization of convexity applied to event probabilities, and it leads to the following result:

THEOREM 3.1 (Convexity of separable CC set). *[17] Let the mapping $f : \mathbb{R}^n \to \mathbb{R}^m$ be such that each component f_i is a concave function. Furthermore, assume that the random vector ξ has independent components and the one-dimensional marginal distribution functions $F_{\xi_i}, i = 1, \ldots, m$, are α_i-concave. Let $\sum_{i=1}^k 1/\alpha_i > 0$. Then*

$$X_0 = \{x \in \mathbb{R}^n : \mathbb{P}_F[f(x) \geq \xi] \geq p\}$$

is convex.

In the non-separable case, the problem is convex provided that $\mathbb{P}_F[f(x, \xi) \geq 0]$ is α-concave. However, checking the optimality conditions requires computing a subdifferential of the probability function and characteristic function of the event $\{f_i(x, \xi) \geq 0\}$ which may be discontinuous.

3.4. Distributionally robust chance-constraints. Seeking a robust method of measuring risk, we would like to allow for ambiguity in the knowledge of the distribution governing the chance-constraint. Building toward this, we first consider risk metrics and then consider relaxations for framing chance-constraints in a distributionally robust and computationally tractable way.

3.4.1. Risk measures. We first consider *Value at Risk* (VaR) and its related notion *Conditional Value at Risk* (CVaR).

DEFINITION 3.2 (VaR). *[17] Let $\alpha \in (0,1)$ be a given confidence level and Z_x be a random variable characterizing the "loss" in a particular system under decision x. Then*

$$\begin{aligned}\mathrm{VaR}_\alpha[Z_x] \coloneqq F^-_{Z_x}(1-\alpha) &= \inf\{t : F_{Z_x}(t) \geq 1-\alpha\} \\ &= \inf\{t : \mathbb{P}[Z_x \leq t] \geq 1-\alpha\} \\ &= \inf\{t : \mathbb{P}[Z_x > t] \leq \alpha\}\end{aligned}$$

VaR is an appealing measure in that it relates the size of a possible loss to the probability of a loss equally or more extreme, i.e., losses exceeding $\mathrm{VaR}_\alpha[Z]$ occur with probability at most α, but it is not a convex in Z_x. On the other hand, CVaR is convex in Z_x(see Appendix A.1), and it is often used as a relaxation for VaR-type constraints.

DEFINITION 3.3 (CVaR). *[17] Under the same scenario as VaR, define*

$$\mathrm{CVaR}_\alpha[Z_x] \coloneqq \inf_{t\in\mathbb{R}}\{t + \alpha^{-1}\mathbb{E}[[Z_x - t]_+]\}$$

and if F_{Z_x} is smooth, then we have

$$\mathrm{CVaR}_\alpha[Z_x] = \alpha^{-1}\int_{1-\alpha}^{1} \mathrm{VaR}_{1-s}[Z_x]\,ds$$

Thus we see that CVaR quantifies the expected loss conditional on knowing that losses exceed the α quantile. In the language of probability distribution functions, the quantile function (a maximal monotone relationship) can be thought of as the subdifferential of CVaR [15], and there is a nice connection between a random variable's distribution function and the convex conjugate (details in Appendix A.5). Moreover, since CVaR is convex, it offers computational tractability, though other convex approximations of the original chance-constraint (8) exist [4, 13] but are not discussed further.

From Definition 3.2 of VaR, we see that it is possible to formulate chance-constraints in terms of VaR restrictions

$$\mathrm{VaR}_\alpha[Z_x] \leq 0 \iff \inf\{t : \mathbb{P}[Z_x \leq t] \geq 1-\alpha\} \leq 0 \iff \mathbb{P}[Z_x \leq 0] \geq 1-\alpha. \tag{10}$$

The first equivalence holds by definition. To see the second equivalence, $\implies$ holds immediately since if in the first case $t < 0$ gives $\mathbb{P}_Z[Z_x \leq t] \geq 1-\alpha$, surely $\mathbb{P}_Z[Z_x \leq 0] \geq 1-\alpha$. The reverse direction holds since we require the greatest lower bound of the $1-\alpha$ quantile which is bounded above by 0.

In the case of (8), we can formulate the constraint with $Z_x \coloneqq f(x,\xi)$ to make the relationship with the distribution function explicit

$$\mathrm{VaR}_{1-\alpha}[Z_x] \leq 0 \overset{(10)}{\iff} \mathbb{P}[Z_x \overset{(8)}{\leq} 0] \geq \alpha \iff \mathbb{P}[Z_x \geq 0] \leq 1-\alpha \iff F_{Z_x}(0) \geq \alpha \tag{11}$$

and can use probability tools as a method of solution. However, F_{Z_x} may not be an α-concave function, so $1 - F_{Z_x} = \mathbb{P}[Z_x \geq 0]$ may not be convex, which would be problematic for implementing an efficient convex approximation in light of Theorem 3.1.

This is clear to see by rewriting the VaR constraint from (11) with indicators as

$$\mathbb{P}\left[Z_x \geq 0\right] \leq 1-\alpha \iff \mathbb{E}\left[\mathbb{1}\{Z_x \geq 0\}\right] \leq 1-\alpha \tag{12}$$

and observing that the step function is non-convex and discontinuous at 0, meaning that we do not have an obvious tractable solution method.

3.4.2. Tractability. Since VaR constraints involve possibly non-convex distribution functions, we instead seek a convex constraint that guarantees VaR feasibility when satisfied. Such a constraint is called a conservative convex relaxation. To build one, we can upper bound the left-hand side of the VaR constraint (10) by a convex function, and further, we can ask for the tightest bound. For a convex formulation, note that we still require convexity in x for $f(x,\xi) = Z_x$.

Following [13] and [17], consider

$$\mathrm{VaR}_\alpha\left[Z_x\right] \leq 0 \iff \mathbb{P}[Z_x \leq 0] \geq 1-\alpha \iff \mathbb{P}[Z_x > 0] \leq \alpha \iff \mathbb{E}[\mathbb{1}\{Z_x > 0\}] \leq \alpha \tag{13}$$

To build a conservative relaxation, we now seek to approximate the step function with a convex, non-negative, non-decreasing function $\psi : \mathbb{R} \to \mathbb{R}$ such that $\psi(z) \geq \mathbb{1}\{z > 0\}$ for every $z \in \mathbb{R}$ (in particular, we also require $\psi(0) = 1$ at the edge case $z = 0$). An important observation is that ψ is scale invariant in its input since $\psi(tz) \geq \mathbb{1}\{tz > 0\} = \mathbb{1}\{z > 0\}$ for any $t > 0$. We will use this property in searching for the tightest bound.

Using the scale invariance with any parameter $t > 0$, we have

$$\mathbb{E}[\psi(t^{-1}Z_x)] \geq \mathbb{E}[\mathbb{1}\{Z_x > 0\}]\,\forall t > 0 \implies \inf_{t>0}\left\{\mathbb{E}[\psi(t^{-1}Z_x)]\right\} \geq \mathbb{E}[\mathbb{1}\{Z_x > 0\}]. \tag{14}$$

Note that we have preserved convexity in x of Z_x since this is an increasing convex function of a convex function of x.

Now we turn to optimizing over the form of ψ. The approximation is tightest when ψ is smallest, and from this perspective, the piecewise linear (hinge-loss) function $\psi(z) = [1+\gamma z]_+$ for $\gamma > 0$ is tightest. There may be other cases where analytically it would be useful to bound with a smooth function like $\exp(z)$, but our end case is simulation, so we only consider the piecewise linear function. Because of the scale invariance of ψ and the requirement that $\psi(0) = 1$, we can take $\gamma = 1$ without loss of generality and thus define $\psi(z) \coloneqq [1+z]_+ = \max\{1+z, 0\}$.

As formulated, $\inf_{t>0}\left\{\mathbb{E}[\psi(t^{-1}Z_x)]\right\}$ may not be convex in both t and x. To ensure convexity in both arguments we can reformulate it as the perspective function which takes $(x,t) \mapsto t\psi(x/t)$ and is jointly convex by multiplying both sides of the constraint by t [12, 6]. Accordingly, we have

$$\inf_{t>0}\left\{t\,\mathbb{E}[\psi(t^{-1}Z_x)]\right\} \leq \alpha t \implies \mathbb{E}[\mathbb{1}\{Z_x > 0\}] \leq \alpha. \tag{15}$$

Rearranging the inequality on the left, substituting $\psi(z)$, replacing $t' = -t$, and rescaling by α, we have

$$\begin{aligned}
\inf_{t>0}\left\{t\,\mathbb{E}[\psi(t^{-1}Z_x)] - \alpha t\right\} &= \inf_{t>0}\left\{t\,\mathbb{E}[[1+t^{-1}Z_x)]_+] - \alpha t\right\} \\
&= \inf_{t>0}\left\{\mathbb{E}[[t+Z_x]_+] - \alpha t\right\} \\
&= \inf_{t'<0}\left\{\mathbb{E}[[Z_x - t']_+] + \alpha t'\right\} = \inf_{t'\in\mathbb{R}}\left\{\alpha^{-1}\mathbb{E}[[Z_x - t']_+] + t'\right\} \\
&= \mathrm{CVaR}_\alpha\left[Z_x\right] \qquad (16)
\end{aligned}$$

where we observe that it is unnecessary to constrain $t' < 0$ (see Appendix A.4 for more detail). Finally, we see that $\text{CVaR}_\alpha[Z_x] \leq 0$ is a conservative, convex approximation for the constraint $\text{VaR}_\alpha[Z_x] \leq 0$.

3.4.3. Computation. Even though CVaR is convex, the true expectation may be difficult to compute and work with analytically. If we have sample observations of $f(x,\xi)$ on hand, then we might use a sample average approximation for the expectation and seek a bound on the distance between the approximation and the truth. Asymptotically, with more samples, the SAA approaches the true CVaR, and we can use Hoeffding's inequality to bound the rate of convergence (with assumptions on the support of the samples). This offers a data-driven approach that we briefly discuss.

Assuming that we are unable to compute the expectation directly, we resort to empirical approximation of the expectation. Hoeffding's inequality provides a probabilistic statement about the size of the difference between the true expectation under distribution F of $f(x,\xi)$

$$T = \mathbb{E}_F\left[[f(x,\xi)+t]_+\right]$$

and an empirical estimate

$$\hat{T} = \frac{1}{N}\sum_{i=1}^{N}[f(x,\xi_i)+t]_+$$

based on observed sample data $\xi,\ i = 1,\ldots,N$ (given that $f(x,\xi)$ is bounded above and below). We state Hoeffding's inequality and then apply it to f under the assumption that $|f(x,\xi)|$ is bounded by $\Gamma/2$.

THEOREM 3.4 (Hoeffding). *[10] Let $X_1,\ldots,X_n$ be independent, bounded random variables such that $X_i \in [a_i,b_i]\ \forall i = 1,\ldots,n$. Then we have*

$$\mathbb{P}\left(\frac{1}{n}\sum_{i=1}^{n}X_i - \mathbb{E}\left[\frac{1}{n}\sum_{i=1}^{n}X_i\right] \geq \delta\right) \leq \exp\left(\frac{-2n^2\delta^2}{\sum_{i=1}^{n}(b_i-a_i)^2}\right)$$

From Hoeffding's inequality, we see that we want $\mathbb{P}(\hat{T}-T \geq \delta)$ to be small for some small threshold δ. Before applying Hoeffding's inequality, define Ω_0 as the set of events where the error between the true expectation and empirical approximation is small (bounded by δ) and Ω_1 as the set of events where the true expectation exceeds the approximation by more than δ. That is, for sample space Ω,

$$\Omega_0 = \{\omega \in \Omega : \hat{T}(\omega) - T(\omega) \leq \delta\} \text{ and}$$
$$\Omega_1 = \Omega\backslash\Omega_0 = \{\omega \in \Omega : \hat{T}(\omega) - T(\omega) < \delta\}$$

where ω is the argument of the random variable ξ above. Now applying Hoeffding's inequality to Ω_1 gives

$$\mathbb{P}(\Omega_1) \leq \exp\left(\frac{-2N\delta^2}{\Gamma^2}\right) \iff 1-\mathbb{P}(\Omega_1) = \mathbb{P}(\Omega_0) \geq 1-\exp\left(\frac{-2N\delta^2}{\Gamma^2}\right)$$

and provides a lower bound on the probability of feasibility if the approximate constraint is satisfied.

Explicitly, for $\omega \in \Omega_0$, we have $T \leq \hat{T} + \delta$, and for $t > 0$ we conclude by substitution

$$\hat{T} \leq T + \delta \leq t(1-\alpha) \tag{17}$$

$$\implies \frac{\frac{1}{N}\sum_{i=1}^{N}[f(x,\xi_i)+t]_+ + \delta}{t} \leq \inf_{t>0}\left[\frac{\mathbb{E}_F\left[f(x,\xi)+t\right]_+}{t}\right] \leq 1-\alpha \tag{18}$$

$$\implies \mathbb{P}_F(f(x,\xi) \geq 0) \leq 1-\alpha \tag{19}$$

with probability greater than or equal to $1 - \exp\left(\frac{-2N\delta^2}{\Gamma^2}\right)$ depending on parameter δ, sample size N, and random variable bound Γ. In other words, an optimal solution x^* to the approximate problem is feasible (and thus provides an upper bound) for the original problem (8) with probability exceeding $1 - \exp\left(\frac{-2N\delta^2}{\Gamma^2}\right)$ if we can bound the empirical expectation $\hat{T}+\delta$ by $t(1-\alpha)$. It is also possible to utilize information about the variance of the distribution to find a tighter relationship through Bernstein-type inequalities [18].

3.4.4. Summary. We saw that a convex risk-measure relaxation can ensure computational tractability of a chance-constrained approach. Furthermore, it has the added benefit that in the distributionally robust case, we can derive confidence bounds on the performance of an empirically constructed chance-constraint that are independent of explicit distributional choices (aside from specified support assumptions). In this way, we see that the CVaR approximation and concentration inequality bound provide a form of distributionally robust chance-constrained programming with a practical numerical implementation.

4. Numerical studies. In this section, we introduce the DRCC framework, apply it to two test problems (portfolio optimization and a continuously indexed chance-constrained problem), and discuss the method in the context of SO, RO, and data-driven approaches.

4.1. General DRCC formulation. We now turn to considering a general distributionally robust chance-constrained problem of the following form:

$$\begin{aligned} \text{minimize} \quad & f_0(x) \\ \text{subject to} \quad & \sup_{F \in \mathcal{D}} \mathbb{P}_F\left[f(x,\xi) \leq 0\right] \geq \alpha \\ & \text{supp}(\mathcal{D}) \subseteq [a,b] \end{aligned} \tag{20}$$

where $x \in \mathbb{R}^d$ is the control variable, $f : \mathbb{R}^{d+p} \to \mathbb{R}$ constrains the feasible set x, is indexed by random parameter vector $\xi \in \mathbb{R}^p$, and is convex in x. We also assume that the distribution F of random variable $f(x,\xi)$ is contained in a family of distributions $\mathcal{D}$ with known support bounded by $[a,b]$. Finally, α is a problem parameter specifying the acceptable error threshold.

The distributionally robust chance-constrained formulation generalizes the standard chance-constrained problems by allowing for uncertainty in the probability distribution governing the constraints. In the below sections, we will see how it provides a flexible framework for interpreting problems with RO, SO, and data-driven elements.

4.2. Portfolio optimization. We consider single-period, multi-objective portfolio optimization problems of the form

$$\begin{array}{ll} \underset{x\in\mathbb{R}^n_+}{\text{maximize}} & \mathbb{E}\left[W_1\right] = \mathbb{E}\left[x^T\xi\right] = \mu^T x \\ \underset{x\in\mathbb{R}^n_+}{\text{minimize}} & R(x) \\ \text{subject to} & \mathbb{1}^T x = W_0 \\ & x_i \geq 0,\, i = 1,\ldots,n \end{array} \tag{21}$$

where x is the nonnegative decision variable specifying the initial wealth W_0 invested in each of the n assets, $R(x)$ is a measure of risk, and ξ is a vector of random returns (mean μ and covariance Σ) for the single-period. We consider three variants for addressing portfolio risk

1. Markowitz: $R(x) = x^T\Sigma x$
2. Chance-constrained (VaR constraint): $R(x) = \mathbb{P}(x^T\xi \leq \rho) \leq \epsilon$ for specified return threshold ρ and certainty parameter ϵ
3. Distributionally robust chance-constrained: $R(x) = \mathbb{P}_{F\in\mathcal{D}}(\xi^T x \leq \rho) \leq \epsilon$ for specified return threshold ρ, certainty parameter ϵ, and distributional set $\mathcal{D}$

To formulate the risk objective as a tractable problem, we can translate the $R(x)$ objectives into constraints of the form $R(x) \leq \eta$ for some particular η. Similarly, we use $\mathbb{P}(x^T\xi \leq \rho) \leq \epsilon$ in the probabilistic formulation of the chance-constraint. For the distributionally robust version, we assume that we have data generated from the bounded family of distributions in $\mathcal{D}$ and use a relaxation. The optimization problems we consider are

$$\begin{array}{ll} \underset{x\in\mathbb{R}^n_+}{\text{maximize}} & \mu^T x \\ \text{subject to} & x^T\Sigma x \leq \eta \\ & \mathbb{1}^T x = W_0 \\ & x_i \geq 0,\, i = 1,\ldots,n \end{array} \tag{22}$$

$$\begin{array}{ll} \underset{x\in\mathbb{R}^n_+}{\text{maximize}} & \mu^T x \\ \text{subject to} & \mathbb{P}(x^T\xi \leq \rho) \leq \epsilon \\ & \mathbb{1}^T x = W_0 \\ & x_i \geq 0,\, i = 1,\ldots,n \end{array} \tag{23}$$

and

$$\begin{array}{ll} \underset{x\in\mathbb{R}^n_+}{\text{maximize}} & \mu^T x \\ \text{subject to} & \mathbb{P}_{F\in\mathcal{D}}(\xi^T x \leq \rho) \leq \epsilon \\ & \mathbb{1}^T x = W_0 \\ & x_i \geq 0,\, i = 1,\ldots,n \end{array} \tag{24}$$

Interpreting the above problems, the Markowitz problem may be seen as a robust optimization problem with trivial uncertainty set. With minor reformulations, we can also view the chance-constrained problem as a form of robust optimization. For the

chance-constraint, suppose that we know $\xi \in \mathcal{U} \coloneqq \{\mu + \Sigma^{1/2}u : \|u\|_2 \leq \Phi^{-1}(\epsilon)\}$. Then we have the constraint $\mu^T x \leq \rho \; \forall \xi \in \mathcal{U}$ is equivalent to the constraint $\mu^T x + \Phi^{-1}(\epsilon)\sqrt{^T\Sigma x} \leq \rho$ [8]. This can be seen as a robust optimization formulation since we have specified an explicit uncertainty set. Finally, the DRCC problem can be thought of as a RO problem by specifying an uncertainty set in a similar way to the CC problem, a SO problem with the standard random variable interpretation, or as a data driven problem (which we will see in its relaxed, computationally tractable form).

4.2.1. Normal distribution. We first study the familiar setup where $\xi \sim N(\mu, \Sigma)$. In this section, we generate data from n hypothetical assets over m days according to the following:

1. hyperparameters
   - true covariance matrix $\Sigma_0 \sim IW(2n, \sqrt{n}I)$ (inverse-Wishart)
   - true mean $\mu_0 \sim N(0, I)$
2. data
   - n-dimensional asset return vector $\xi \sim N(\mu_0, \Sigma_0)$
   - m observations $\xi_1, \ldots, \xi_m \sim N(\mu_0, \Sigma_0)$
3. estimators
   - mean $\hat{\mu} = \frac{1}{m}\sum_{i=1}^m \xi_i$
   - covariance $\hat{\Sigma} = \frac{1}{m-1}\sum_{i=1}^m (\xi_i - \hat{\mu})(\xi_i - \hat{\mu})^T$

After simulating our dataset, we solve the two optimization problems for $\hat{x}_{markowitz}$ and $\hat{x}_{chance}$ with Julia's `JuMP` optimization toolbox.

Note that we can choose parameters η and (ρ, ϵ) so that problems (1) and (2) give equivalent solutions. The chance-constraint in the second optimization problem can be reformulated as a SOCP constraint

$$\begin{aligned}
\mathbb{P}\left(\xi^T x \leq \rho\right) \leq \epsilon &\iff \mathbb{P}\left([\xi^T x - \mu^T x]/\sqrt{x^T\Sigma x} \leq [\rho - \mu^T x/[\sqrt{x^T\Sigma x}]\right) \leq \epsilon \\
&\iff \mathbb{P}\left(Z \leq [\rho - \mu^T x]/\sqrt{x^T\Sigma x}\right) \leq \epsilon \\
&\iff \Phi\left([\rho - \mu^T x]/\sqrt{x^T\Sigma x}\right) \leq \epsilon \\
&\iff [\rho - \mu^T x]/\sqrt{x^T\Sigma x} \leq \Phi^{-1}(\epsilon) = z_\epsilon \\
&\iff x^T\Sigma x \leq \left(\frac{\mu^T x - \rho}{z_\epsilon}\right)^2.
\end{aligned}$$

Therefore, choosing

$$\eta = \left(\frac{\mu^T \hat{x}_{chance} - \rho}{z_\epsilon}\right)^2$$

makes the first two problems equivalent. Note that we need to solve the chance-constrained problem first, though. An example is presented below.

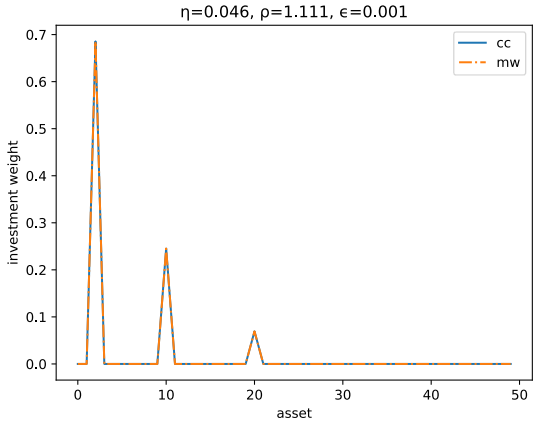


FIG. 1. *Equivalence of Markowitz (Sharpe) and chance-constrained formulations*

For the distributionally robust chance-constraint

$$\mathbb{P}_{F\in\mathcal{D}}(\xi^T x \le \rho) \le \epsilon \tag{25}$$

we assume the true distribution is contained in some family of bounded distributions $\mathcal{D}$ (note that in particular, this is valid for $F = \arg\ \sup_{F'}\{\mathbb{P}_{F'}(\xi^T x \le \rho) : F' \in \mathcal{D}\}$). From (19), we follow the pattern

$$\frac{1}{N}\sum_{i=1}^{N}[f(x,\xi_i)+t]_+ + \delta \le t\epsilon \implies \mathbb{P}(f(x,\xi)\ge 0) \le \epsilon$$

and formulate (25) as

$$\frac{1}{N}\sum_{i=1}^{N}[\rho - \xi_i^T x + t]_+ + \delta \le \epsilon \implies \mathbb{P}(\xi^T x \le \rho) \le \epsilon$$

with feasibility according to Hoeffding so that the distributionally robust portfolio optimization problem is formulated as

$$\begin{aligned} \underset{x\in\mathbb{R}^n_+, t\in\mathbb{R}}{\text{maximize}} \quad & \mu^T x \\ \text{subject to} \quad & \frac{1}{m}\sum_{i=1}^{m}\left[\rho - \xi_i^T x + t\right]_+ + \delta \le t\epsilon \\ & \mathbb{1}^T x = W_0 \\ & x_i \ge 0,\ i = 1,\dots,n \end{aligned} \tag{26}$$

Simulations. We construct hypothetical datasets consisting of $n = 50$ assets across $d = 75$ days to generate a particular problem instance and estimate the first two moments from the data. Next, we compute the unconstrained optimal decision (corresponding to 100% weight in the asset with maximum expected return) and require ρ to be a fixed fraction (we choose 1/3) of the "unconstrained" return for the chance-constrained distributionally robust chance-constrained problems. Similarly,

we choose ϵ to be 1%. We also compute the η that corresponds to equating the Markowitz problem with the chance-constrained problem; such an η is important because the Markowitz optimum will provide a basis of comparison when we explore non-Normal distributions. Next we compute the optimal decision weighting for the assets under each model. Finally, we simulate $1,000$ observations from the Normal distribution seeded by the true parameter values which generated the original dataset.

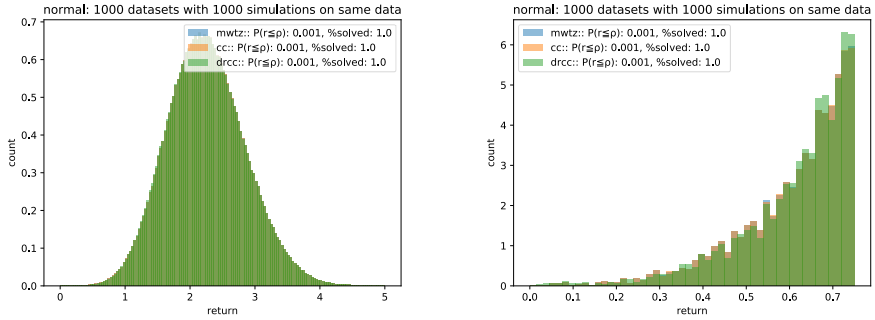


FIG. 2. *Normal simulation study: we observe similar results amongst the three methods. Each method solves all of the simulated problems, and there is little difference in the overall return distribution (Left) and the tail losses (Right).*

We observe that each of the three methods produces similar decisions when the underlying distribution is known to be normal. In particular, both Markowitz and the chance-constrained procedures produce the same decisions (as expected), and the DRCC approach doesn't differ much. Furthermore, each method outperforms the goal of achieving 1/3 of the unconstrained return with probability greater than 99%, and the lower and upper tails of the return distributions exhibit similar properties among the models. The ellipsoidal uncertainty set posited by the chance-constrained model fits the true distribution well, and we see strong performance of the model. Furthermore, it is worth mentioning that under the assumption of normality, VaR is a coherent risk measure (i.e., it regains subadditivity). We see this since for any two assets Z_1 and Z_2 distributed jointly normal with standard deviation $\sigma_{Z_1}, \sigma_{Z_2}$ and correlation ρ, their sum $Z = Z_1 + Z_2$ has standard deviation

$$\sigma_Z = \sigma_{Z_1+Z_2} = \sqrt{\sigma_{Z_1}^2 + \sigma_{Z_2}^2 + 2\rho\sigma_{Z_1}\sigma_{Z_2}} \leq \sigma_{Z_1} + \sigma_{Z_2}.$$

Combined with the homogeneity property (of choosing a quantile), we see that VaR is coherent under normality, which lends further explanation to the similarity of the CC and DRCC cases.

4.2.2. Beta distribution. Next we consider Beta distributions and simulate n assets with distribution Beta$(1 + n - i, 1 + i)$ for $i = 1, \ldots, n$. These distributions range from highly concentrated at 1 to highly concentrated at 0 as i increases (with concentrations depending on n). In this set of simulations, it would be easy for each method to determine $i = 1$ as the optimum when provided with 75 samples. To increase the uncertainty, we only sample 3 data points from 5 assets (it still would be relatively easy to distinguish in the case of 75 assets and 3 data points).

Simulations. The simulations below indicate that each model selects a similar portfolio (in nearly every case, a single asset that the model deems the best per-

former). Each model produces decisions that exceed the required fraction of maximum return, but the DRCC does a slightly better job at capturing upside while not sacrificing performance on the lower tail. This indicates that in a few instances, CC and Markowitz selected suboptimal portfolios but that DRCC was more robust to the lack of data. Compared to the Normal case, in about 1% of the simulations, the simulated dataset results in infeasible problems under the CC/DRCC models indicating that the ellipsoidal uncertainty set is not entirely tuned to this sort of distribution, as expected.

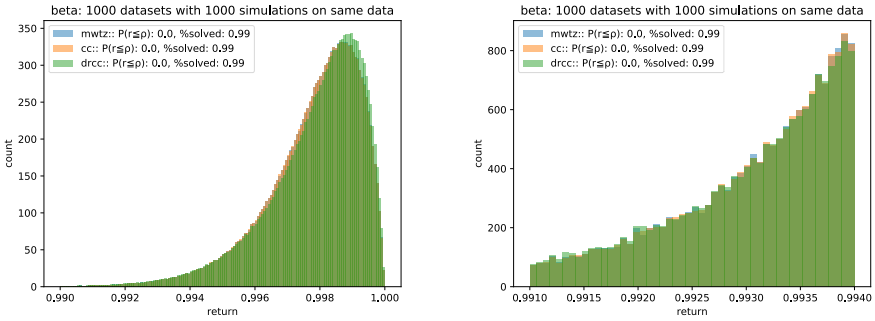


FIG. 3. *Beta simulation study: in these simulations, we observe that the overall return distribution is slightly more favorable in the DRCC case; the tail errors are similar amongst the three methods.*

4.2.3. Gaussian mixture distribution. Next, we compare the performance of the three optimization methods on a Gaussian mixture. The only difference from the above study is the distribution; as before, we seed the true covariances from an Inverse Wishart and the true means from a standard Gaussian. Next, we randomly draw mixture proportions by simulating three uniform RVs and normalizing by their sum. This consists of one "experiment", and for each experiment, we simulate $1,000$ observations and test the original decision computed from the experiment's original dataset.

Simulations. For some datasets, we observe that any of the three methods may be infeasible for our chosen parameters. In particular, in any experiment, the Markowitz calculation depends on the chance-constrained calculation for η, so if the chance constrained is infeasible, we skip the problem and track the percentage of datasets which can be solved. We compute $\mathbb{P}[\xi^T x \geq \rho]$ for each experiment and weight the final probability by the number of feasible Comparisons are below for $1,000$ datasets where each decision method is applied to the experimental dataset but tested with $1,000$ separately drawn observations.

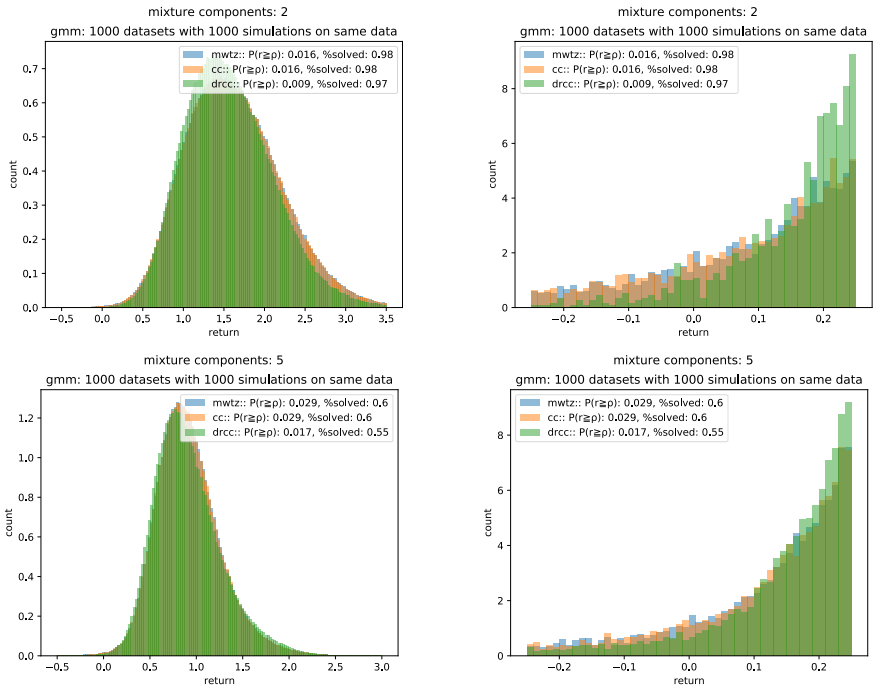


FIG. 4. *GMM simulation study, top: 2 mixture components, bottom: 5 mixture components. DRCC method performs better in the tails but gives up returns in the middle of the distribution.*

In the above figures, we observe that the DRCC performs as expected and reduces the observed frequency of choosing a portfolio with returns below a specified threshold. For models with less uncertainty (2 components), we observe more pronounced risk-mitigation on the low-return spectrum relative to models with more uncertainty (5 components). On the other hand, for the higher uncertainty model, the DRCC exhibits relatively better upside. It is also interesting to note that the CC and Markowitz models don't differ much in either of the GMM models, indicating a shortcoming of the CC model for generalizing to non-Normality. Also note that the averaged probability of observing a return lower than an experiment's ρ is nearly twice as high for the CC and Markowitz models than the DRCC.

4.3. Tuned portfolio optimization. In a separate testing paradigm, we consider a "tuning" a model for a given dataset. In this structure, we tune parameters for each of the three models by performing a grid-search to determine the best parameter settings under the criterion of maximizing the minimum return on the original dataset.

4.3.1. Normal distribution. The first data model we consider is a Normal model where hyperparameters are chosen as described above. We first generate a single "historical" dataset and then grid-search our parameters over this data. Once we have determined optimal parameter values, we then test each method's decision on 1,000 new observations generated from the same true Normal distribution which generated our dataset. Results for a single, tuned problem instance and 100 tuned problem instances are included below.

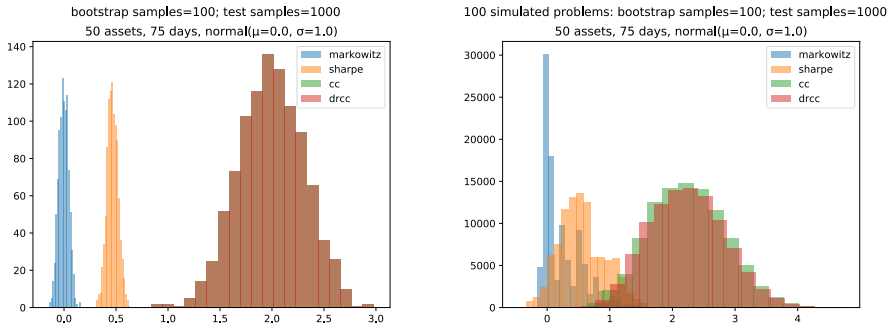


FIG. 5. *Left: single problem instance; Right: 100 tuned problem instances. We observe similar performance in the CC/DRCC methods for the simulated problems which outperforms the other methods. Both the CC/DRCC have more available tuning parameters, so this makes sense.*

From the above figure, we observe that the traditional solutions (Markowitz, Sharpe) tend to produce better variance results (and a more diversified portfolio) but a poor lower bound on the minimum return. The DRCC solution is often similar to the CC solution, and over 100 problem instances, it appears slightly more robust than the CC solution both in terms of parameter and hyperparameter uncertainty. It's interesting to note that though the Sharpe/Markowitz formulation can theoretically perform as well as the chance-constrained problem by the explicit mapping given above, the results for a coarse grid-search show that this equivalence may be isolated and difficult to find. This lends favor for the chance-constrained and its distributional robust formulations for computational gains.

Next we consider a more difficult correlation structure where an asset's variance is increased proportionally by its mean. Specifically, the diagonal of Σ_0 is scaled up by the drawn μ_0 for those asset j such that $\mu_0^{(j)} > 1$. Under this simulation framework, we observe skewed distributions, and the CC/DRCC again outperform the traditional approaches, giving support for their robustness across various distributional forms.

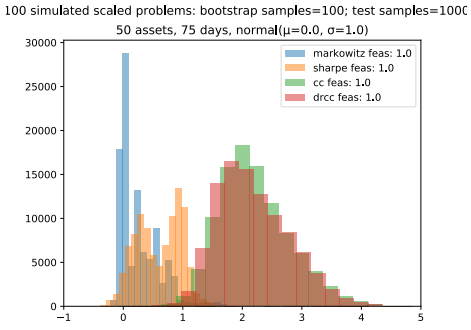


FIG. 6. *scaled problem instance*

4.3.2. Beta distribution. In these simulations, the problem again is defined by simulating returns for n assets over d days, but now each asset is drawn from an independent Beta distribution. Repeating the testing and parameter optimization as above, we again see that DRCC/CC outperform the traditional methods.

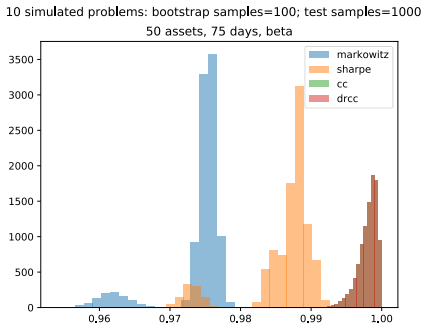


FIG. 7. *independent Beta*

4.4. Continuously indexed chance constraint. We now consider a slight variant on (20) with *continuously-indexed* constraints

$$\begin{aligned} \underset{x\in C}{\text{minimize}} \quad & f_0(x) \\ \text{subject to} \quad & \sup_{F\in\mathcal{D}} \mathbb{P}_F\left[\exists u : y(x,u,\xi) \le 0\right] \le \epsilon \\ & u \in [a,b] \\ & \text{supp}(\mathcal{D}) \subseteq [l,r] \end{aligned} \tag{27}$$

where x is the control variable in some convex, bounded domain C, $y(x,u,\xi)$ is now a semi-infinite constraint indexed by continuous variable u and random parameter vector ξ that is linear in x, u, and ξ. We still assume that the distribution F of random variable $y(x,u,\xi)$ is contained in a family of distributions $\mathcal{D}$ with known support bounded by $[l,r]$ and that ϵ is a problem parameter specifying the acceptable error threshold.

As a toy model of (27), we consider a specific problem where the decision lives in compact domain $[-2,2]$

$$\begin{aligned} \underset{x\in[-2,2]}{\text{minimize}} \quad & f_0(x) = x \\ \text{subject to} \quad & \sup_{F\in\mathcal{D}} \mathbb{P}_F\left[\exists u : y(x,u,\xi) = 1+\xi+x\sin(u) \le 0\right] \le \epsilon \\ & u \in [0,2\pi] \\ & \text{supp}(\mathcal{D}') \subseteq [0,1] \end{aligned} \tag{28}$$

In (28), we consider $\mathcal{D}'$ to be the family of distributions for ξ. Requiring $\text{supp}(F') = [0,1]$, $F \in \mathcal{D}'$ is equivalent to the constraint $\text{supp}(F) \subseteq [-1,4]$, $F \in \mathcal{D}$

$$\begin{cases} \max_{\xi,x}(1+\xi+x) = 2+\max(x) = 4 \\ \max_{\xi,x}(1+\xi-x) = 2+\max(-x) = 4 \\ \min_{\xi,x}(1+\xi+x) = 1+\min(x) = -1. \\ \min_{\xi,x}(1+\xi-x) = 1+\min(-x) = -1. \end{cases}$$

Thus, $\xi \in [0,1]$ yields $y(x,u,\xi) \in [-1,4]$..

4.4.1. Analytic solution. Ignoring the boundary cases of the inequality, the first constraint yields a "good" set:

$$\Omega_0 = \{x, u, \xi : 1 + \xi + x\sin(u) \geq 0\} = \{x, u, \xi : -1 - \xi - x\sin(u) \leq 0\}$$

and a "bad" set:

$$\Omega_1 = \{x, u, \xi : 1 + \xi + x\sin(u) \leq 0\} = \{x, u, \xi : -1 - \xi - x\sin(u) \geq 0\}.$$

Framed probabilistically, we bound the occurrences of the "bad" case and consider

$$\mathbb{P}_F\left[\exists u \in [0, 2\pi] : 1 + \xi + x\sin(u) \leq 0\right].$$

The existence of a u such that the constraint is violated can be accounted for by minimizing $y(x, u, \xi) = 1 + \xi + x\sin(u)$ with respect to u to seek the worst case in the semi-infinite constraint. Since $\sin(u)$ achieves both -1 and 1 on $[0, 2\pi]$, $\min_u x\sin(u)$ will depend on the sign of x. For $x > 0$, $\min_u y(x, u, \xi) = -x$ with $\sin(u) = -1$, and for $x < 0$, $\min_u y(x, u, \xi) = -x$ with $\sin(u) = 1$. This gives

$$\min_u \sin(u) = -|x|$$

and $\min_u y(x, u, \xi) = 1 + \xi - |x|$. Now we have the constraint

$$\mathbb{P}_F\left[1 + \xi - |x| < 0\right] \leq \epsilon$$

which can be inverted to find x if we know (or can approximate) F and F^{-1}

$$F^{-1}\left(\mathbb{P}_F\left[\xi < |x| - 1\right]\right) \leq F^{-1}(\epsilon) \implies |x| \leq 1 + F^{-1}(\epsilon).$$

Again there are two branches

$$\begin{cases} x \leq 1 + F^{-1}(\epsilon) \\ x \geq -(1 + F^{-1}(\epsilon)) \end{cases}$$

and taking the minimum of the two (based on objective function $f(x) = x$) yields

$$x \geq -(1 + F^{-1}(\epsilon)) \implies x_* = -(1 + F^{-1}(\epsilon)). \tag{29}$$

However, we are not entirely done since our solution depends on F, the distribution of $y(x, u, \xi)$. We address this case in the following section.

4.4.2. Adverse distribution (analytic problem). If for every $F \in \mathcal{D}$ we have

$$\mathbb{P}_F\left[1 + \xi - |x| < 0\right] = \mathbb{P}_F\left[\xi < |x| + 1\right] \leq \epsilon$$

then we have

$$\sup_{F \in \mathcal{D}} \left\{\mathbb{P}_F\left[\xi < |x| - 1\right]\right\} \leq \epsilon.$$

We can then characterize F by its left quantile function $F^-(t) = \inf\{z : F(z) \geq t\}$ so that

$$\sup_{F \in \mathcal{D}} \left\{\mathbb{P}_F\left[\xi < |x| - 1\right]\right\} = \sup_{F \in \mathcal{D}} \left\{F(|x| - 1)\right\} \leq \epsilon \tag{30}$$

$$\implies F(|x|-1) \le \epsilon \quad \forall F \in \mathcal{D}$$
$$\implies |x| \le F^-(\epsilon) + 1 \quad \forall F \in \mathcal{D}$$
$$\implies |x| \le \inf_{F \in \mathcal{D}} \{F^-(\epsilon)\} + 1. \tag{31}$$

If we are able to characterize F by F^- [15] Appendix A.4, then we have equivalence between (30) and (31) since we can reverse the argument. Then we have $|x| \le 1$ if $F^-(\epsilon) = 0$ for any ϵ for the worst case F, which means F is a point mass at -1. Translated to the distribution $\mathcal{D}'$, this equates to the condition that ξ is a point mass at zero. Thus, a robust feasible solution must be feasible with $1-\epsilon$ probability even under the most adverse case of a point mass distribution.

4.4.3. Approximate solution. For some compact $\mathbb{K}$, we can apply the generating function bound from (16) to $y'(x,u,\xi) - M$. This means that satisfying the $\inf\left[t^{-1}\mathbb{E}_F(\cdot)\right]$ constraint will satisfy the original chance constraint. To apply the CVaR-type approximation, we rewrite the constraint in terms of the random variable $y'(x,u,\xi) = -(1+\xi+x\sin(u))$ as

$$\begin{aligned}
\mathbb{P}_F\left[\exists u \in [0,2\pi] : 1+\xi+x\sin(u) \le 0\right] &= \mathbb{P}_F\left[\exists u \in [0,2\pi] : -(1+\xi+x\sin(u)) \ge 0\right] \\
&= \mathbb{P}_F\left[\max_{u\in[0,2\pi]} \{-(1+\xi+x\sin(u))\} \ge 0\right] \\
&= \mathbb{P}_F\left[-\min_{u\in[0,2\pi]} \{1+\xi+x\sin(u)\} \ge 0\right] \\
&= \mathbb{P}_F\left[-(1+\xi+\min_{u\in[0,2\pi]} \{x\sin(u)\}) \ge 0\right]
\end{aligned}$$

so that

$$\mathbb{P}_F\left[-(1+\xi+\min_{u\in[0,2\pi]} \{x\sin(u)\}) \ge 0\right] \le \tag{32}$$
$$\inf_{t>0}\left[\frac{1}{t}\mathbb{E}_F\left[[-(1+\xi+\min_{u\in[0,2\pi]} \{x\sin(u)\}) + t]_+\right]\right] \le \epsilon. \tag{33}$$

Supposing that we have prior observations on $y'(x,u,\xi)$, we can bound the difference between the true expectation and the sample average approximation (SAA) using McDiarmid's inequality. Thus we have

$$T = \inf_{t>0}\left[\frac{1}{t}\mathbb{E}_F\left[[-(1+\xi+\min_{u\in[0,2\pi]} \{x\sin(u)\}) + t]_+\right]\right]$$
$$\hat{T} = \frac{1}{t}\left[\frac{1}{N}\sum_{i=1}^{N}[-(1+\xi_i+\min_{u\in[0,2\pi]} \{x\sin(u)\}) + t]_+ + \delta\right]$$
$$T \le \hat{T} \quad \text{w.p. } q \ge 1 - \exp\left(\frac{-2N\delta^2}{\Gamma^2}\right)$$

where $1+\xi_1+x\sin(u), \ldots, 1+\xi_N+x\sin(u) \underset{iid}{\sim} F$ with known support $[-1,4]$.

4.4.4. Solver implementation. The only remaining approximation is to relax the constraint to hold at finitely many (M) mesh points between 0 and 2π in order to model the semi-infinite constraint. We transform the generating function constraint:

$$\frac{1}{t}\left[\frac{1}{N}\sum_{i=1}^{N}[-(1+\xi_i+\min_{u\in[0,2\pi]} \{x\sin(u)\}) + t]_+ + \delta\right] \le \epsilon, \quad (\text{and } t \ge 0)$$

into the discretized constraint:

$$\sum_{i=1}^{N}[-(1+\xi_i+\min_{j=1,\dots,M}\{x\sin(u_j)\})+t]_+ \le N(t\epsilon-\delta), \quad (\text{and } t\ge 0).$$

4.4.5. Approximation tightness. This toy model is meant to exhibit the same types of approximation error as more complicated, chance-PDE constrained models where it might be necessary to numerically approximate the PDE. To recap, the approximations present in our model are

1. CVaR bound of VaR
2. Sample-average approximation of CVaR
3. Discretization of semi-infinite constraint

We investigate the contributions of each in the context of a uniform distribution on the randomness ξ.

CVaR bound of VaR. The CVaR of $U\sim \text{unif}(a,b)$ is

$$\begin{aligned}\text{CVaR}_\alpha[U] &= \frac{1}{\alpha}\int_{1-\alpha}^{1}\text{VaR}_{1-s}[U]\,ds = \frac{1}{\alpha}\int_{1-\alpha}^{1}s(b-a)+a\,ds\\ &= \frac{1}{2}\alpha(a-b)+b\end{aligned}$$

Note that in our problem, both a and b depend on $\min_u x\sin(u)$. For the first approximation, suppose that $x\in[-2,2]$ is fixed and again that $\xi\sim\text{unif}(0,1)$. Then we have that

$$y'(x,u,\xi)\sim\text{unif}(-(1+\xi_{hi}-|x|,-(1+\xi_{lo}-|x|)=\text{unif}(-2+|x|,-1+|x|)$$

Thus the CVaR constraint on $a=-2+|x|$ and $b=-1+|x|$ gives

$$\frac{1}{2}\alpha(a-b)+b=-\frac{1}{2}\alpha-1+|x|\le 0,$$

and we see that $x^*=-(1+\alpha/2)$ by setting $\alpha:=\epsilon$.

Sample-average approximation of CVaR. Performing the simulation for various α, the computed approximate solution $\hat{x}$ closely follows the CVaR bound but is far from the analytic bound.

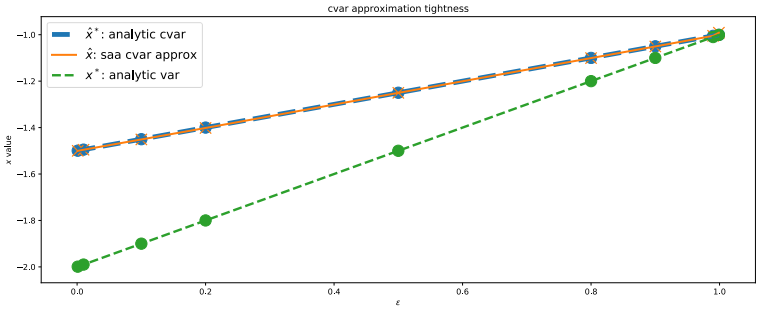


FIG. 8. *Effect of analytic vs approximate CVaR. The blue line shows the solution $\hat{x}^*$ of the CVaR problem using analytic CVaR; the orange line shows the solution $\hat{x}$ of the CVaR problem using a sample-average CVaR approximation; the green line shows the solution x^* of the VaR problem using analytic VaR. We observe that the CVaR/VaR relaxation is loose and that the sample-average approximation is reasonably tight.*

Discretization of semi-infinite constraint. Based on the looseness of the CVaR approximation to VaR, we expect the discretization effect to be relatively small. We verify this hypothesis computationally and observe only second-order differences between the optimal computed values with the discretization vs the analytic $\min_u x sin(u)$.

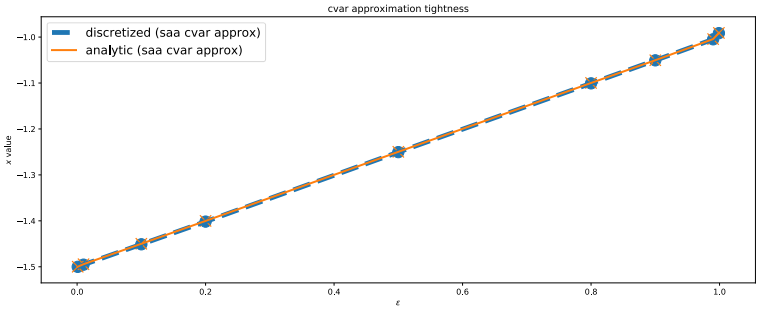


FIG. 9. *Effect of analytic vs discretization. The blue line shows the solution of the CVaR problem using a sample-average CVaR approximation and the discretized solution of* $\min_u x \sin(u)$*; the orange line shows the solution of the CVaR problem using a sample-average CVaR approximation and the analytic solution of* $\min_u x \sin(u)$*. We observe that the discretization approximation is reasonably tight.*

4.4.6. Adverse distribution (relaxed problem). To find the most adverse distribution for the relaxed problem, we consider solving the relaxed optimization problem analytically in the Bernoulli case. Under the assumption that $\xi \sim \text{Bern}(p)$, the original problem

$$\begin{aligned} \underset{x,t}{\text{minimize}} \quad & x \\ \text{subject to} \quad & \frac{1}{N}\sum_{i=1}^{n}[-(1+\xi-|x|)+t]_+ + \delta - t\epsilon \le 0 \\ & -t \le 0 \\ & |x| \le 2 \end{aligned} \tag{34}$$

can be rewritten with the constraint

$$\begin{aligned} \tfrac{1}{N}\sum_{i=1}^{n}[-(1+\xi-|x|)+t]_+ + \delta - t\epsilon &= \tfrac{1}{N}\left[\sum_{i\in\mathcal{I}_1}^{n}[t+|x|-2]_+ + \sum_{i\in\mathcal{I}_0}^{n}[t+|x|-1]_+\right] + \delta - t\epsilon \\ &= \hat{p}_1[t+|x|-2]_+ + \hat{p}_0[t+|x|-1]_+ + \delta - t\epsilon \end{aligned} \tag{35}$$

where $\mathcal{I}_1 = \{i : \xi_i = 1\}$, $\mathcal{I}_0 = \{i : \xi_i = 0\}$, and $\hat{p}_1 = |\mathcal{I}_1|/N$, $\hat{p}_0 = |\mathcal{I}_0|/N$. Our goal is to find the "worst" p for fixed δ, ϵ in two senses: it leads to decisions $\hat{x}$ which are

1. overly conservative
2. overly aggressive

This amounts to the two ways which a solution can be "fooled by the data".

We can reformulate the problem as the following LP

$$
\begin{aligned}
&\underset{x,t}{\text{minimize}} && x \\
&\text{subject to} && \hat{p}_1(t+|x|-2)+\hat{p}_0(t+|x|-1)+\delta-\epsilon t \le 0 && (c_1)\\
&&& \hat{p}_1(t+|x|-2)+\delta-\epsilon t \le 0 && (c_2)\\
&&& \hat{p}_0(t+|x|-1)+\delta-\epsilon t \le 0 && (c_3)\\
&&& \delta-\epsilon t \le 0 && (c_4)\\
&&& -t \le 0 \\
&&& |x| \le 2
\end{aligned}
\tag{36}
$$

using the relationship

$$\max\{x,0\} \le c \iff c \ge x,\ c \ge 0$$

Now we have an LP, so the solution will be a vertex. Since we have two decision variables, we can plot the feasible set for various settings of δ and ϵ. For example, with $\delta = 0.01, \epsilon = 0.1$ and choosing $p = 0.99$, we plot the feasible region of (34) and (36). From the feasible regions, we can read off the solution $\hat{x}$.

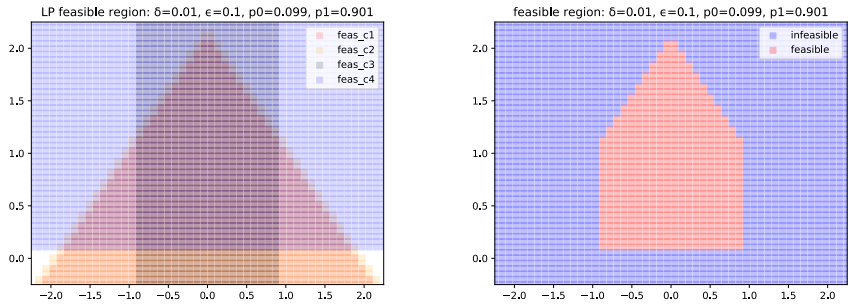


FIG. 10. *The intersection of the feasible regions defined by the four linear constraints in (36) (left figure) is equivalent to the feasible region of the original constraint in (34) (right figure)*

For the worst case p, we consider the two variants. Recall also that the approximation algorithm requires $\hat{p} \coloneqq \hat{p}_1$, the empirical success rate to produce a solution $\hat{x}$ and the analytic solution depends on the cdf, i.e., $x^* = -(1 + F^{-1}(\epsilon))$. Thus

1. choosing $p > 1 - \epsilon$ gives analytic solution $x^* = -2$, but the approximation algorithm isn't sensitive enough to pick up this and chooses the conservative approximation $\hat{x} = 1 - \delta/(1 - \hat{p})$ (i.e., where (c_3) intersects (c_1) in (36)) if p is close to ϵ
2. choosing $p < 1 - \epsilon$ restricts the analytic solution to $-(1 - 0)$, but choosing p close enough to ϵ will result in overreaches when the empirical $\hat{p} > 1 - \epsilon$ (again based on the binding (c_3) and (c_1) intersection)

Thus we see that choosing $p \approx 1 - \epsilon$ pushes the approximation algorithm toward over-conservativeness or over-aggressiveness.

4.4.7. Simulation Results. We simulate the toy problem under various parameter and distributional assumptions to gauge the performance of the relaxation relative to its analytic bounds. First, fixing $F \sim \text{unif}(0, 1)$, we test the model parameters N, δ, ϵ, and Γ and observe that the parameters behave as expected:

- increasing N yields a higher predicted and observed feasibility rate
- decreasing δ yields a more aggressive approximation ($x_a - x_c$ diminishes)
- decreasing ϵ leads to more conservative approximation
- the model appears to be relatively robust to Γ, indicating that using a conservative Γ does not significantly affect performance

Next, we fix the models parameters and test various distributions over $[0, 1]$. The results indicate that the convex relaxation is a reasonable approximation in terms of identifying decision x that is both feasible and aggressive:

- point-masses: feasibility is observed in all cases and aggressiveness increases with point-masses approaching 1
- restricted intervals: similar to point-masses, feasibility is observed in all cases and aggressive increases with intervals approaching the right side of $[0, 1]$
- negative vs positive skew: similar to the above two cases, feasibility is observed in all cases and aggressiveness increases with means approaching the right side of $[0, 1]$. It is interesting to note that we observe less approximation variability with negative-skewed distributions
- unit-interval:
  - unimodal Beta: similar trends to point-mass and restricted intervals. Mass on left gave worse approximations; however the "right" mix of good mass and bad mass gave infeasibility (still rare occurrence)
  - bimodal Beta: approximation performed worse when bimodality increased
  - Bernoulli: feasibility was observed in each case, but the approximation was sensitive to the uncertainty parameter p. In fact, the Bernoulli model was the case which performed worst in terms of relative approximation performance. The analytic solution still performed better than the analytic solution of the worst-case point mass, but the relative approximation is worse. This indicates a susceptibility of the approximation algorithm and an important computational question in exploring the worst case distribution for the algorithm which can differ from the analytic worst-case distribution.

5. Conclusions. In this overview, we have discussed robust, stochastic, and data-driven optimization and highlighted their relationships through two numerical studies. The portfolio optimization study indicated that using the DRCC formulation and subsequent relaxation into a computationally tractable problem offered performance improvements over the standard Markowitz and RO/SO chance-constrained approaches. It also motivated the intuitive formulation of risk minimization problems as SO problems but provided a practical approximation technique for distributions which would be difficult to approximate under the SO approach. Thus, it showed how RO and data-driven approaches can be used to reinterpret SO problems as robust models that can be solved efficiently.

Next, we considered a continuously indexed chance-constrained problem and applied the empirical distributionally robust chance-constraint framework to approximate the solution. For a toy model, we showed that the most adverse distribution of the analytic model might not coincide with the worse-case distribution of the relaxed model. We showed that this susceptibility can be attributable to the approximation induced by the DRCC convex relaxation; it also may be influenced by the semi-infinite relaxation, but we did not explore this.

Appendix A. Chance-constrained optimization.

A.1. VaR and CVaR. Let t_i be such that $\mathrm{CVaR}_\alpha[Y_i] = t_i + \frac{1}{1-\alpha}\mathbb{E}[[Y_i - a_i]_+]$. Then $f(y) = [y-a]_+$ is convex, so

$$\begin{aligned}
&\mathrm{CVaR}_\alpha[\theta Y_1 + (1-\theta)Y_2] \\
&\quad \le \theta t_1 + (1-\theta)t_2 + \frac{1}{1-\alpha}\mathbb{E}[[\theta Y_1 + (1-\theta)Y_2 - \theta t_1 + (1-\theta)t_2]_+] \\
&\quad \le \theta t_1 + (1-\theta)t_2 + \frac{\theta}{1-\alpha}\mathbb{E}[[Y_1 - t_1]_+] + \frac{1-\theta}{1-\alpha}\mathbb{E}[[Y_2 - t_2]_+] \\
&\quad = \theta\,\mathrm{CVaR}_\alpha[Y_1] + (1-\theta)\,\mathrm{CVaR}_\alpha[Y_2]
\end{aligned}$$

A.2. Chance-constraint. We derive the chance-constraint from the original constraint in (1) by choosing

$$g_1(x,\omega) = \begin{cases} \alpha - 1, \text{ if } f_1(x,\omega) \le 0 \\ \alpha, \text{ else} \end{cases}$$

and observing that

$$\begin{aligned}
G_1(x,\omega) &= \int_\Omega g_1(x,\omega)\ F(d\omega) \\
&= \int_\Omega [(\alpha-1)\mathbb{1}\{f_1(x,\omega) \le 0\} + \alpha\mathbb{1}\{f_1(x,\omega) > 0\}]\ F(d\omega) \\
&= \alpha - \mathbb{P}[f_1(x,\omega) \le 0]
\end{aligned}$$

so that $G_1(x,\omega) \le 0 \iff \alpha - \mathbb{P}[f_1(x,\omega) \le 0] \le 0$, giving an α level of reliability.

A.3. Generating function bound.

CVaR constraint. To show that we can remove the constraint $t < 0$ in the final step of (16), we consider the function $\phi(t) \coloneqq t + \alpha^{-1}\mathbb{E}[[Z_x - t]_+]$. Computing $\phi'(t) = 1 + \alpha^{-1}[F_{Z_x}(t) - 1]$ for cdf F_{Z_x} continuous at t (otherwise, the left and right derivatives at t are given with corresponding left and right limits of $F_{Z_x}(\cdot)$. This means that the minimum of $\phi(t)$ occurs in an interval $[t_l, t_r]$ such that $t_l = \inf\{s : F_{Z_x}(s) \ge 1\alpha\}$ and $t_r = \sup\{s : F_{Z_x}(s) \le 1\alpha\}$. But, these are the left and right quantiles, respectively, and $\mathrm{VaR}_\alpha[Z_x]$ is defined as the left quantile, so the minimum occurs at t_l.

This means that

$$\mathrm{CVaR}_\alpha[Z_x] = \mathrm{VaR}_\alpha[Z_x] + \alpha^{-1}\mathbb{E}[[Z_x - t_l]_+],$$

but noting that the last term is nonnegative, we have

$$\inf_{t\in\mathbb{R}}\{t + \alpha^{-1}\mathbb{E}[[Z_x - t_l]_+]\} \le 0 \implies t_l \le 0$$

so $t < 0$ is not needed [17].

Interpreting CVaR. Next, we consider working directly with the CVaR constraint. We follow [6]. Optimizing over t explicitly (assuming continuity) as

$$\frac{\partial}{\partial t}\left[t + \alpha^{-1}\mathbb{E}[[Z_x - t_l]_+]\right] = 1 - \frac{1}{\alpha}\mathbb{E}[\mathbb{1}\{Z_x \ge t\}] = 1 - \frac{1}{\alpha}\mathbb{P}[Z_x \ge t] = 0,$$

if we set t^* such that $\alpha = \mathbb{P}[Z_x \geq t^*]$, then

$$0 = 1 - \frac{1}{\alpha}\mathbb{P}[Z_x \geq t^*] \implies \mathrm{CVaR}_\alpha[Z_x] = \frac{1}{\alpha}\mathbb{E}[[Z_x - t^*]_+] + t^* = \frac{1}{\alpha}\mathbb{E}[[Z_x - t^*]_+] + \mathrm{VaR}_\alpha[Z_x]$$

as we saw above. This gives the interpretation that CVaR upper bounds VaR and penalizes excess over the VaR t^*.

A.4. Generalized distribution function. [15] point out that "the correspondence between distribution functions and quantile functions is one-to-one, with [distribution function] F_X recoverable from [quantile function] Q_X by"

$$F_X(x) = \begin{cases} \max\{p : Q_X(x) \leq x\} & x \in (\inf Q_X, \sup Q_X], \\ 1 & x > \sup Q_x, \\ 0 & x \leq \inf Q_x. \end{cases}$$

A.5. CVaR relation to convex conjugate. For a random variable X with distribution function F, consider the *superexpectation* function defined as [15]

$$\mathbb{E}_X[x] \coloneqq \mathbb{E}[\max\{x, X\}]. \tag{37}$$

Rearranging gives

$$\begin{aligned} \mathbb{E}_X[x] \coloneqq \mathbb{E}[\max\{x, X\}] &= \int_{-\infty}^{\infty} \max\{x, \hat{x}\}\, dF_X(\hat{x}) \\ &= x \int_{-\infty}^{x} dF(\hat{x}) + \int_x^{\infty} \hat{x}\, dF(\hat{x}) \\ &= x\, F(x) + \int_x^{\infty} (\hat{x} - x)\, dF(\hat{x}) + x \int_x^{\infty} dF(\hat{x}) \\ &= x\, F(x) + \int_{-\infty}^{\infty} \max\{\hat{x} - x, 0\}\, dF(\hat{x}) + x\,(1 - F(x)) \\ &= x + \mathbb{E}[\max\{X - x, 0\}] \end{aligned}$$

To find the Fenchel conjugate of $\mathbb{E}_X[x]$, consider

$$\begin{aligned} (\mathbb{E}_X[y^*])^* &= \sup_x \{y^* x - \mathbb{E}_X[x]\} \\ &= \sup_x \{y^* x - (x + \mathbb{E}[\max\{X - x, 0\}])\} \\ &= -\inf_x \{-y^* x + (x + \mathbb{E}[\max\{X - x, 0\}])\} \end{aligned}$$

so that multiplying by $1/(y^* - 1)$ gives

$$\frac{1}{y^* - 1}(\mathbb{E}_X[y^*])^* = \inf_x \left\{ x + \frac{1}{1 - y^*}\mathbb{E}[\max\{X - x, 0\}] \right\} = \mathrm{CVaR}_{y^*}[X] \tag{38}$$

and following [15] we obtain the relationship

$$(\mathbb{E}_X[y^*])^* = \begin{cases} -(1 - y^*)\,\mathrm{CVaR}_{y^*}[X] & y^* \in (0, 1), \\ -\mathbb{E}[X] & y^* = 0, \\ 0 & y^* = 1, \\ \infty & y^* \notin [0, 1] \end{cases} \tag{39}$$

Appendix B. Distributionally robust chance-constrained optimization.

B.1. Experiments.

Setup. We define the following setup for a simulation $S_e^{(k)}$ where $e = 1, 2, \ldots, E$ indexes the experiment number and $k = 1, 2, \ldots, K$ indexes the simulation number within experiment e. Each simulation $S_e^{(k)}$ computes or generates:

- Data: $\xi_1, \xi_2, \ldots, \xi_N$ drawn $i.i.d$ from known distribution F used in computing the approximate optimum
- Optimum points: x_a (analytic optimum) and x_c (computed approximate optimum); note that within an experiment e, $x_a^{(1)} = \ldots = x_a^{(k)} = \ldots = x_a^{(K)}$
- Feasibility certificate: $f \in \{0, 1\}$ where $f = 1$ if $x_c > x_a$ (feasible), or $f = 0$ otherwise; average feasibility for experiment e is computed as $K^{-1} \sum_{k=1}^{K} f^{(k)}$

given specified experiment parameters:

- Buffer: $\delta > 0$, the approximate optimum's desired aggressiveness
- Error: ϵ, the permissible probability that the constraint is infeasible
- Discretization index: M, the number of mesh points discretizing the interval
- Support bound: γ, the known, bounded range of F's support

and optimization parameter:

- Auxiliary variable: t, which is required for the generating function bound.

Metrics. To compare the results of different experiments, we focus on the following metrics:

- Predicted feasibility: the feasibility bound predicted by $1 - \exp\left(-\frac{2N\delta^2}{\Gamma^2}\right)$
- Observed feasibility: the empirical estimate of feasibility proportion computed as $\frac{1}{K} \sum_{k=1}^{K} f^{(k)}$
- Absolute approximation aggressiveness: x_c; the more negative, the more "aggressive"
- Relative approximation aggressiveness: x_c/x_a, and its related measures of dispersion (empirical mean and variance estimates)
- Dispersion of x_c: characterized by its empirical mean and variance

B.1.1. Model testing.

- When testing the effect of sample size N on feasibility and optimum, we observe the expected effect where feasibility increases with sample size. We also observe a corresponding decrease in both relative/absolute approximation aggressiveness.

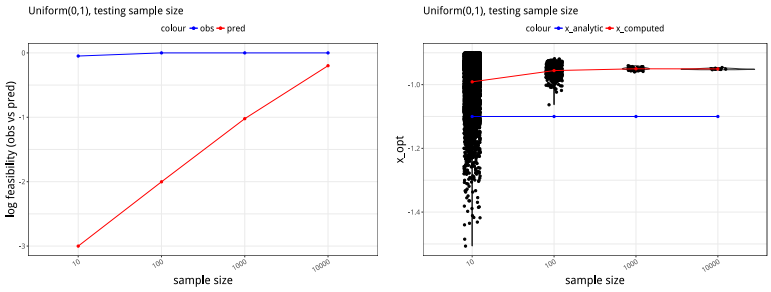


FIG. 11. *Varying sample size with model parameters:* $\xi \sim unif(0, 1)$, $\epsilon = 0.1$, $\delta = 0.01$, $M = 100$, $\Gamma = 2$, *and* $K = 5\times$ (1/ *"predicted feasibility"*). *Left: observed feasibility. Right: computed optimums (black points) and computed averages (red line) versus analytic optimum (blue).*

- When testing the effect of δ on feasibility and optimum, we note that the approximation becomes more aggressive as δ shrinks, which provides support for the intuition that δ behaves as a "buffer" parameter.

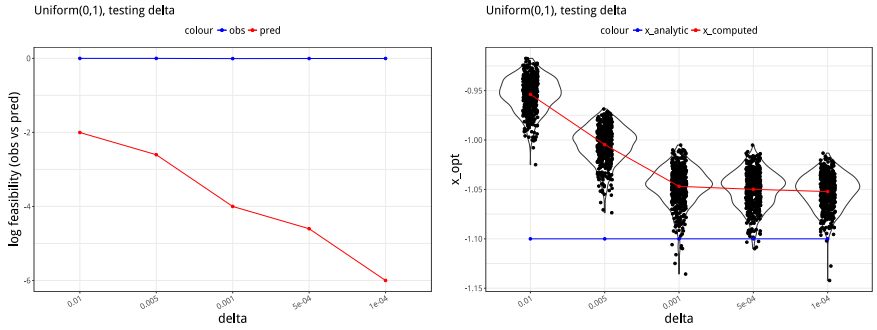


FIG. 12. *500 simulations varying δ with model parameters: $\xi \sim unif(0,1)$, $\epsilon = 0.1$, $N = 100$, $M = 100$, and $\Gamma = 2$. Left: observed feasibility. Right: computed optimums (black points) and computed averages (red line) versus analytic optimum (blue).*

- We test ϵ in two separate studies for fixed δ: (1) δ is small relative to ϵ; and (2) when ϵ is large relative to δ. Note that feasibility is obtained in every simulation instance, and we have suppressed the output.
  In the first case, we chose $\epsilon = c\delta$ for $c \in (50, 900)$). We observe that ϵ behaves as expected: smaller values of ϵ result in more conservative *absolute* approximations. We also observe that smaller ϵs lead to more aggressive *relative* approximations. The reason for this is related to t, the ancillary optimization parameter in the problem formulation. In the simulations, we find that $x_c - t = x_a$, and t takes larger values for larger ϵ values, so t plays a significant role in the approximation behavior.

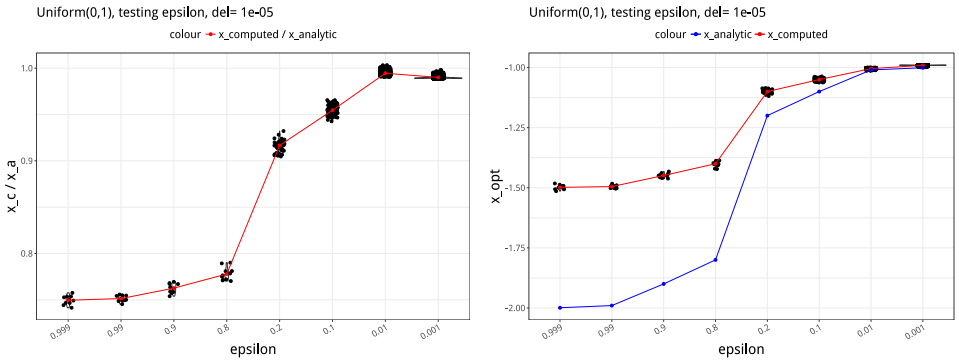


FIG. 13. *500 simulations varying ϵ with model parameters: $\xi \sim unif(0,1)$, $\delta = 1e{-}5$, $N = 100$, $M = 100$, and $\Gamma = 2$. Left: Relative approximation aggressiveness. Right: absolute approximation aggressiveness*

In the second case, we chose $\epsilon = c\delta$ for $c \in (1, 20)$. We observe that at some point, for large enough buffer, it becomes increasingly difficult to achieve the desired error rate. Clearly, for $\delta > \epsilon$, it is impossible to do so, as indicated by the sharp changes in the plots below.

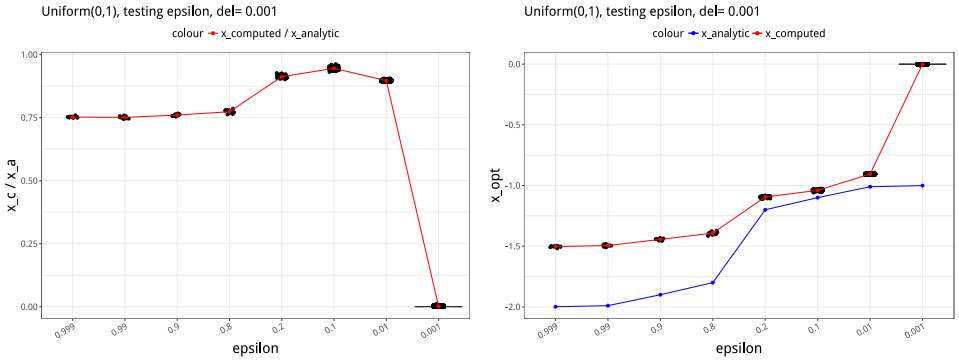
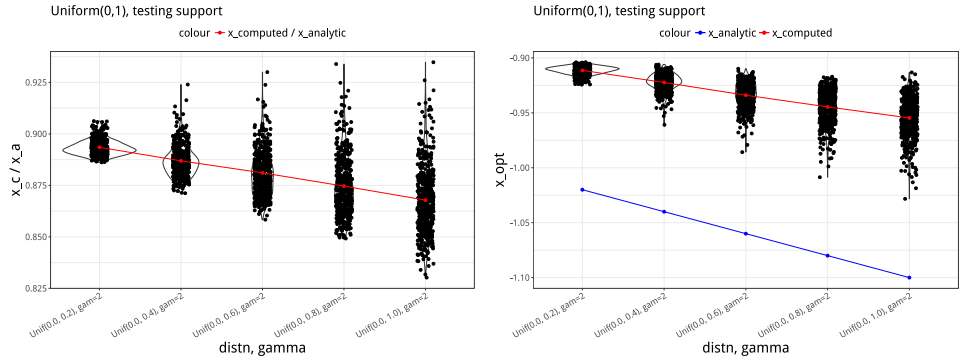
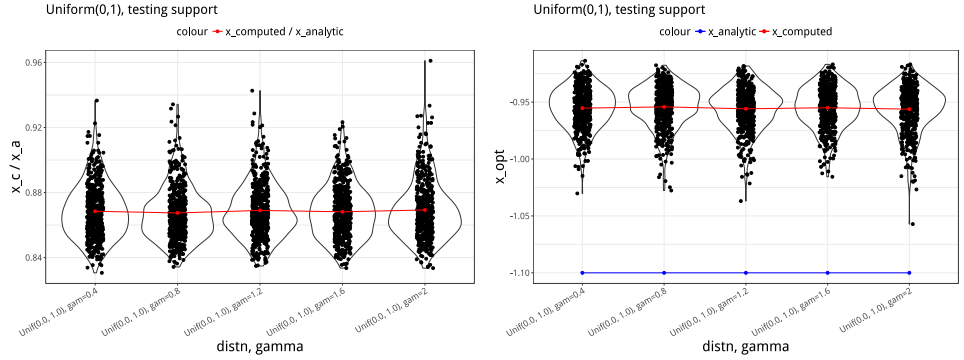


FIG. 14. *500 simulations varying ϵ with model parameters: $\xi \sim unif(0,1)$, $\delta = 1e{-}3$, $N = 100$, $M = 100$, and $\Gamma = 2$. Left: Relative approximation aggressiveness. Right: absolute approximation aggressiveness*

- Next we explore Γ in three ways. First we incrementally increase the support of F from $(0, 0.2)$ to $(0, 1)$ with a fixed Γ large enough to cover the whole range. Here we observe that for larger support, the approximation aggressiveness diminishes (and feasibility is observed in all cases).



Next we fix the distribution over $(0, 1)$ and vary γ from covering the whole support to covering only a small portion of the support. In these simulations, we observe negligible effect from choice of Γ.



Finally, we vary both the support of F and make the appropriate choice of γ given the known support. In this case, we observe similar results to the first case and note that the approximation aggressiveness appears insensitive to γ, at least in the current setting.

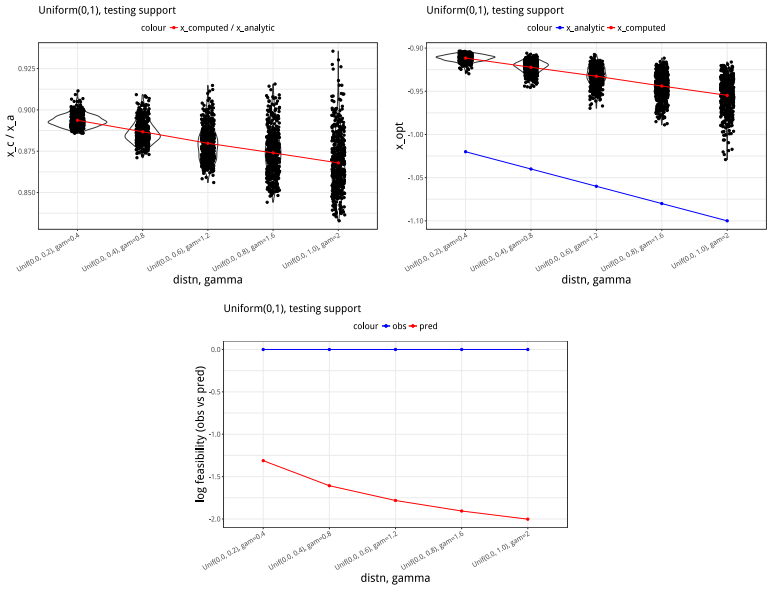


FIG. 15. *Feasibility is observed in every case which implies that the bound is not the limiting factor.*

B.1.2. Adverse distribution testing. In this set of simulations, we vary the form of the known distribution taking support on $[0, 1]$.

- *Point masses.* In this set of simulations, we consider the least amount of randomness in F, point masses on $[0, 1]$, with the goal of characterizing the worst-case and best-case expected performance. We run experiments with point masses equally spaced between 0 and 1 and observe that feasibility is observed in each experiment. This indicates that the approximation works. Furthermore, we note that the absolute and relative approximation becomes more aggressive as the point masses approach the upper bound of 1.

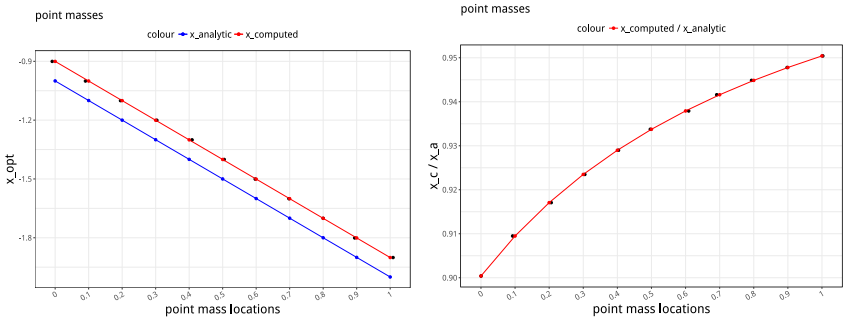


FIG. 16. *Varying point mass location with model parameters:* $\epsilon = 0.1$, $\delta = 0.01$, $M = 100$, $\Gamma = 2$, *and* $K = 1$. *Left: computed optimum versus analytic optimum. Right: relative approximation aggressiveness*

- *Uniform on intervals.* We consider symmetric $F \sim \text{unif}(a, b)$ on restricted intervals $(a, b) \in [0, 1]$ and run six simulations where (a, b) ranges from $[0.0, 0.2)$ to $(0.8, 1.0]$. The observed probability of feasibility is again 1 in each experiment. Furthermore, we recover the the same trend of better approximations as the intervals approach the upper bound of $(0.8, 1.0]$.

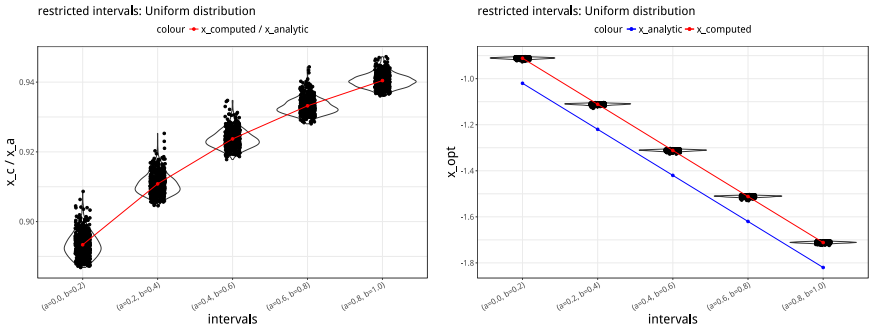


FIG. 17. *500 simulations varying restricted uniform intervals with model parameters:* $\epsilon = 0.1$, $\delta = 0.01$, $M = 100$, *and* $\Gamma = 2$. *Left: computed optimum versus analytic optimum. Right: relative approximation aggressiveness*

- *Negative skew on intervals.* We move to considering negatively skewed distributions on intervals of the form (a, b) and use the generalized Beta distribution with density $f_Y(y) = \frac{(y-a)^{\alpha-1}(b-y)^{\beta-1}}{(b-p)^{\alpha+\beta-1}\text{B}(\alpha,\beta)}$, which comes from transforming $X \sim Beta(\alpha, \beta)$ to $Y = (b - a)X + a$. In these simulations, we chose $\alpha = 5$ and $\beta = 1$. Feasibility is observed in every simulation and the aggressiveness trend is preserved.

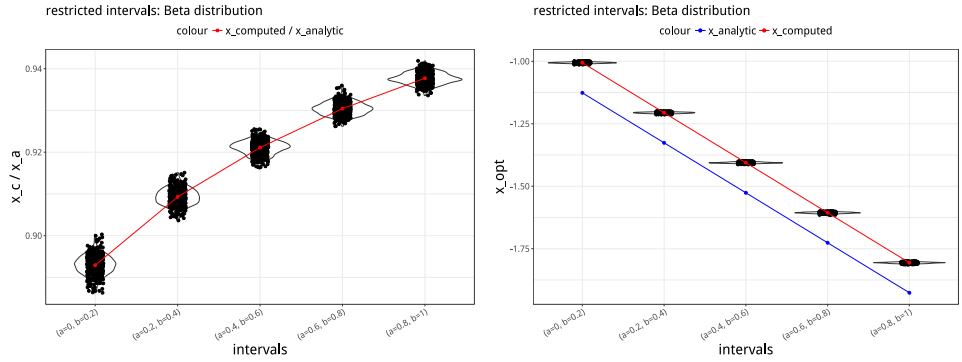


FIG. 18. *Negative skew restricted intervals*

- *Positive skew on intervals.* We move to considering positively skewed distributions on restricted intervals of the form (a, b) and follow the outline from the prior section but with $\alpha = 1$ and $\beta = 5$. Feasibility is observed in every simulation and the aggressiveness trend is preserved, but the variance of the computed optimums in the positive skew cases decreases relative to the negative skewed distributions.

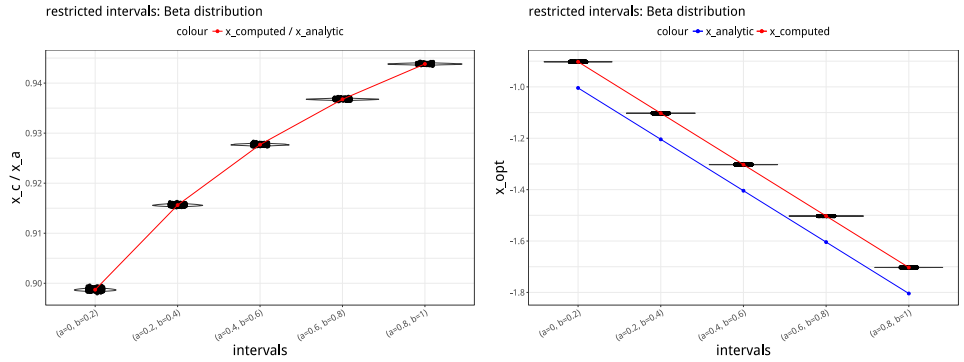


FIG. 19. *Left: relative efficiency of computed solution x_c to analytic solution x_a, Right: analytic and approximate solutions changing with Beta distribution confined to restricted subintervals of* $[0, 1]$

- *Unimodal Beta.* In this set of simulations, we consider Beta distributions on $(0, 1)$ whose means correspond to (or approach) the means and variances of those in the restricted interval tests. More precisely, we fix the variance at $(b - a)^2/12$ to correspond to the variance of the uniform over the partial intervals and vary the mean between $0.1, \ldots, 0.9$. Feasibility is observed in every simulation, and the aggressiveness trend persists, though at a diminished level as the means approach 1. We also see significantly higher variance in the approximation and lower aggressiveness on the whole.

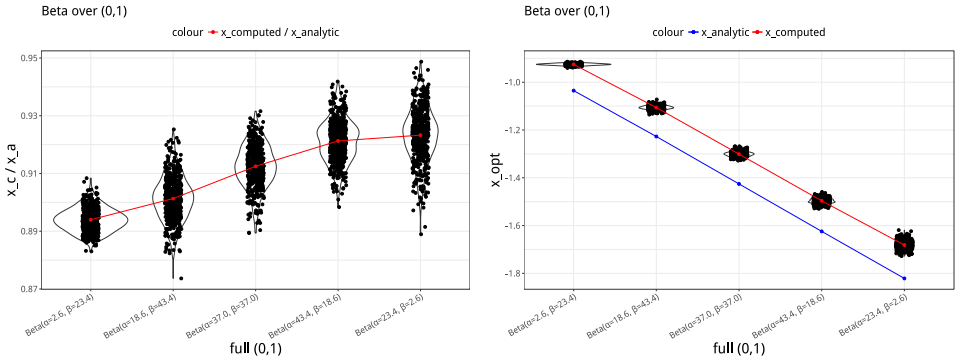


FIG. 20. *Left: relative efficiency of computed solution x_c to analytic solution x_a, Right: analytic and approximate solutions changing with Beta distribution*

- *Bimodal Beta.* In this set of simulations, we consider a set of increasingly severe bimodal Beta distributions on $(0, 1)$. We observe that as the separation becomes more distinct, the absolute approximation becomes worse.

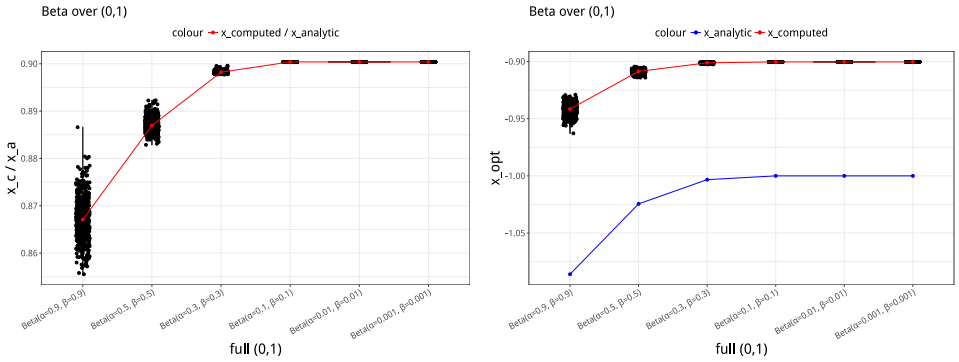


FIG. 21. *Left: relative efficiency of computed solution x_c to analytic solution x_a, Right: analytic and approximate solutions changing with severity of bimodal Beta distribution*

- *Bernoulli.* Finally, we consider bimodal (discrete) distributions of different height using Bernoulli variables and varying the parameter p.

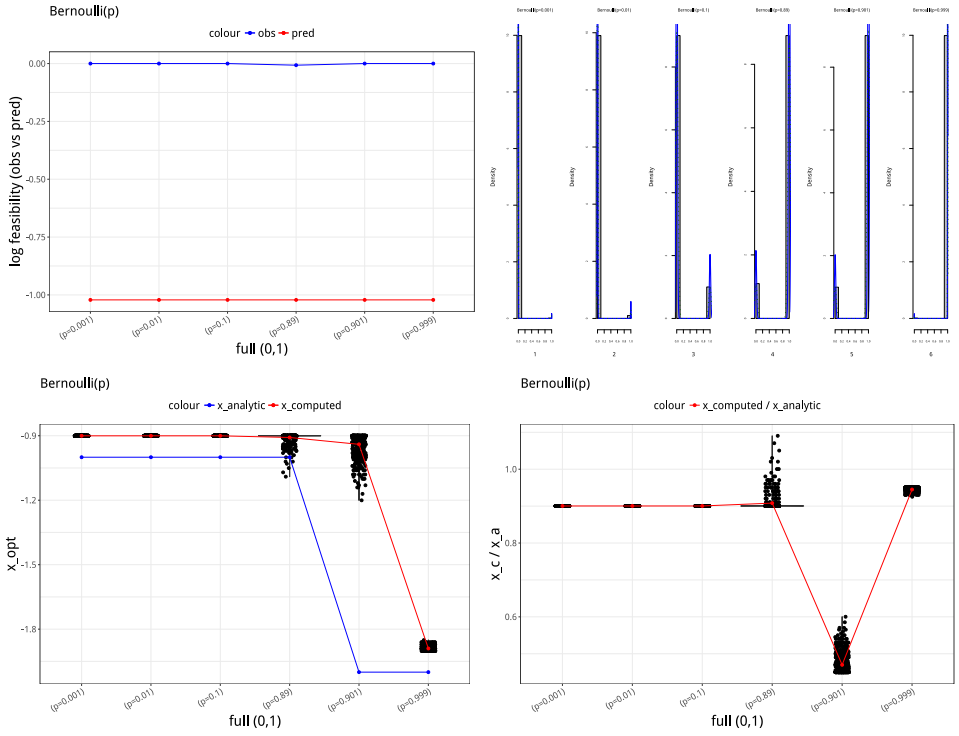


FIG. 22. *Top Left: observed feasibility of computed solution versus predicted feasibility, Top Right: histogram of samples, Bottom Left: relative efficiency of computed solution x_c to analytic solution x_a, Bottom Right: analytic and approximate solutions versus parameter p*

B.2. Discussion. The primary goal of the simulations was to identify the most adverse distribution over the unit interval for problem (8). In this case, the worst possible x_c would be -1. For our simulations, we fixed parameters $\delta = 0.01, \epsilon = 0.1, \Gamma = 2$ and performed two sets of tests over the unit interval. The first set segmented ran-

domness into piecewise components of $(0, 1)$, and the second considered randomness over the full $(0, 1)$ interval.

In the first setup, simulations provide evidence that a point-mass at zero is the most adverse of the distributions on segmented intervals in terms of absolute approximation. Similarly, if we know that randomness is confined to a specific subinterval, then the most adverse distribution is a positive skewed distribution over the interval that places most of its mass near 0, which provides additional support for the theoretical result. We note that in the cases where we introduce randomness away from the point-mass at zero, then the absolute approximation becomes less adverse.

In the second setup, we consider distributions over the entire unit interval. We observe that unimodal Beta distributions that place more mass toward zero are more adverse than other unimodal Beta distributions. However, compared to the point mass at zero, we also see that introducing mass away from zero produces slightly more aggressive absolute approximations, in line with what we observed in the previous restricted randomness case. Next, we observe that bimodal Betas that place mass primarily at either 0 or 1 give an absolute approximation as adverse as the point mass. We compare this with a Bernoulli distribution and observe the same phenomenon for a certain range of p.

Acknowledgments. Thanks Mihai for his patience and guidance which helped me internalize concepts, and thanks to my family for their support.

REFERENCES

[1] A. Ben-Tal, L. E. Ghaoui, and A. Nemirovski, *Robust Optimization*, Princeton University Press, 2009.

[2] D. Bertsimas, V. Gupta, and N. Kallus, *Data-driven robust optimization*, Mathematical Programming, (2017), https://doi.org/10.1007/s10107-017-1125-8.

[3] D. Bertsimas, V. Gupta, and N. Kallus, *Robust sample average approximation*, Mathematical Programming, (2017), https://doi.org/10.1007/s10107-017-1174-z.

[4] Y. Cao and V. Zavala, *A sigmoidal approximation for chance-constrained nonlinear problems*, Optimization Online, (2017), http://www.optimization-online.org/DB_FILE/2017/10/6236.pdf.

[5] W. Chen, M. Sim, S. Jie, and C.-P. Teo, *From cvar to uncertainty set: Implications in joint chance-constrained optimization*, Operations Research, Articles in Advance (2009), pp. 1–16, https://doi.org/10.1287/opre.1090.0712.

[6] J. Duchi, *Stanford ee364b notes: Optimization with uncertain data*, May 2015, https://doi.org/http://stanford.edu/class/ee364b/lectures/robust_notes.pdf.

[7] Y. Emoliev and R. J.-B. Wets, *Numerical Techniques for Stochastic Optimization*, Springer-Verlag, 1988.

[8] L. E. Ghaoui, *Robust optimization: Chance constraints*, Lecture Notes: EE227A, (2008), https://doi.org/https://people.eecs.berkeley.edu/~elghaoui/Teaching/EE227A/lecture24.pdf.

[9] B. L. Gorissen, I. Yankolu, and D. den Hertog, *A practical guide to robust optimization*, Omega, 53 (2015), pp. 124–137–996, https://doi.org/10.1016/j.omega.2014.12.006.

[10] W. Hoeffding, *Probability inequalities for sums of bounded random variables*, Journal of the American Statistical Association, 58 (1963), pp. 13–30, https://doi.org/10.1080/01621459.1963.10500830.

[11] C. Li, F. You, and D. Yue, *Northwestern University Open Text Book on Process Optimization*, 2017, https://optimization.mccormick.northwestern.edu/index.php/Main_Page.

[12] P. Maréchal, *On a functional operation generating convex functions, part 1: Duality*, Journal of Optimization Theory and Applications, 126 (2005), pp. 175–189, https://doi.org/10.1007/s10957-005-2667-0.

[13] A. Nemirovski and A. Shaprio, *Convex approximations of chance constraints*, SIAM Journal of Optimization, 17 (2006), pp. 969–996, https://doi.org/10.1137/050622328.

[14] A. G. Quaranta and A. Zaffaroni, *Robust optimization of conditional value at risk and portfolio selection*, Journal of Banking & Finance, 32 (2008), pp. 2046–2056, https://doi.org/https://www.sciencedirect.com/science/article/pii/S0378426607004281.

[15] R. T. Rockafeller and J. O. Royset, *Random variables, monotone relations, and convex analysis*, Springer, 148 (2014), p. 297, https://doi.org/http://sites.math.washington.edu/~rtr/papers/rtr226-Relations.pdf.

[16] A. Ruszczynski, *Probabilistic programming with discrete distributions and precedence constrained knapsack polyhedra*, SIAM Journal of Optimization, 17 (2002), pp. 969–996.

[17] A. Shapiro, D. Dentcheva, and A. Ruszczynski, *Lectures on Stochastic Programming*, Society for Industrial and Applied Mathematics and the Mathematical Programming Society, 2009.

[18] L. Wasserman, *Carnegie mellon 10/36-702 statistical machine learning: Concentration of measure*, http://www.stat.cmu.edu/~larry/=sml/Concentration.pdf (accessed 2018-07-22).