SIAM J. OPTIM.
Vol. 25, No. 4, pp. 2061–2080

SCENARIO MIN-MAX OPTIMIZATION AND THE RISK OF EMPIRICAL COSTS*

A. CARÈ†, S. GARATTI‡, AND M. C. CAMPI§

Abstract. We consider convex optimization problems in the presence of stochastic uncertainty. The min-max sample-based solution is the solution obtained by minimizing the max of the cost functions corresponding to a finite sample of the uncertainty parameter. The empirical costs are instead the cost values that the solution incurs for the various parameter realizations that have been sampled. Our goal is to evaluate the risks associated with the empirical costs, where the risk associated with a cost is the probability that the cost is exceeded when a new realization of the uncertainty parameter is seen. This task is accomplished without resorting to uncertainty realizations other than those used in optimization. The theoretical result proved in this paper is that these risks form a random vector whose probability distribution is an ordered Dirichlet distribution, irrespective of the probability measure of the stochastic uncertainty parameter. This result provides a distribution-free characterization of the risks associated with the empirical costs that can be used in a variety of application problems.

Key words. stochastic optimization, sample-based techniques, scenario approach, data-driven optimization

AMS subject classifications. 90C25, 90C15, 68W20

DOI. 10.1137/130928546

1. Introduction. In this paper, we consider min-max sample-based uncertain convex optimization problems. The uncertainty parameter is modeled as a random element δ that takes value in a set Δ according to a probability distribution $\mathbb{P}$, the optimization variable x takes value in a convex set $\mathcal{X} \subseteq \mathbb{R}^d$, and the cost function $f(x, \delta)$ is convex and continuous in x for all values of δ. We are provided with a sample of N independent realizations, or "scenarios," $\delta^{(1)}, \delta^{(2)}, \ldots, \delta^{(N)}$ of δ distributed according to $\mathbb{P}$. $\delta^{(1)}, \delta^{(2)}, \ldots, \delta^{(N)}$ is the only available information on the random element δ that is used to select a value of x. The min-max sample-based approach consists in solving the optimization problem

$$\min_{x \in \mathcal{X} \subseteq \mathbb{R}^d} \max_{i=1,\ldots,N} f(x, \delta^{(i)}), \tag{1}$$

which is called the "min-max scenario program" and whose solution is denoted by x^*.[1] Problem (1) arises in diverse applications. For example, given a random variable y, consider the problem of linearly regressing y against variables $u_1, \ldots, u_d$ based on a sample of N independent observations $\delta^{(i)} = (u_1^{(i)}, \ldots, u_d^{(i)}, y^{(i)})$, $i = 1, \ldots, N$.

*Received by the editors July 10, 2013; accepted for publication (in revised form) July 6, 2015; published electronically October 20, 2015. This work was supported by the Ministero dell'Istruzione, dell'Università e della Ricerca (MIUR).

http://www.siam.org/journals/siopt/25-4/92854.html

†Department of Electrical and Electronic Engineering, University of Melbourne, Parkville, VIC 3052, Australia (algo.care@unimelb.edu.au).

‡Dipartimento di Elettronica, Informazione e Bioingegneria, Politecnico di Milano, piazza Leonardo da Vinci 32, 20133 Milano, Italia (simone.garatti@polimi.it, http://home.deib.polimi.it/sgaratti/).

§Dipartimento di Ingegneria dell'Informazione, Università di Brescia, via Branze 38, 25123 Brescia, Italia (marco.campi@ing.unibs.it, http://www.ing.unibs.it/campi/).

[1] A more explicit notation for the solution would be x_N^*, emphasizing that the solution depends on N scenarios. Since N is kept fixed throughout, the simpler notation x^* has been preferred.

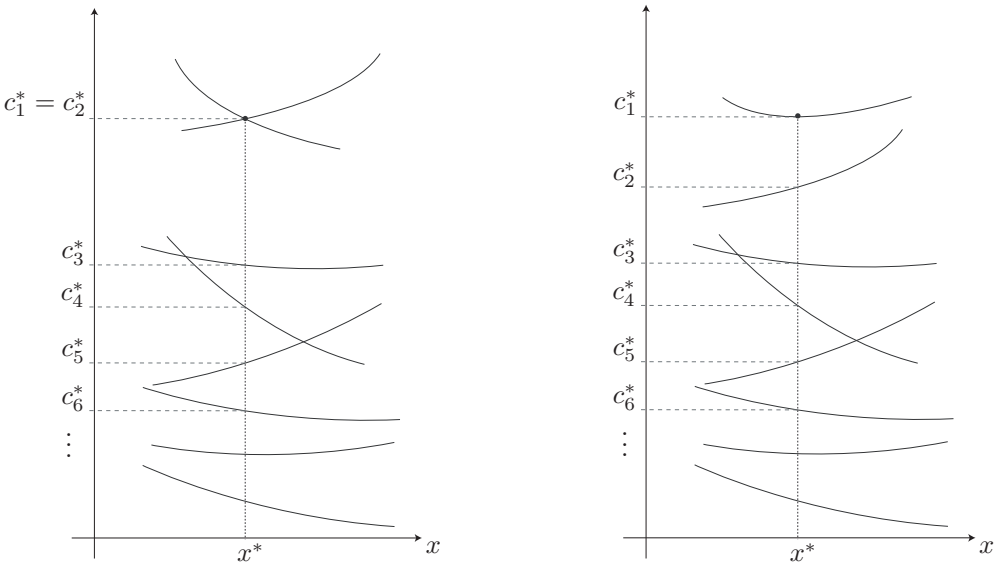


FIG. 1. *Two instances of the optimization problem in* (1). *Each function corresponds to a scenario* $\delta^{(i)}$.

The coefficients $x_1, \ldots, x_d$ in the regression model can be obtained according to the L_∞ criterion of best fit, which corresponds to solving (1) with $f(x, \delta^{(i)}) = |y^{(i)} - \sum_{j=1}^d x_j u_j^{(i)}|$; see, e.g., [27, 10, 22]. Optimization problems of the type (1) also arise in simulation-based control, e.g., when the controller parameters are decided based on N realizations of the disturbance so as to minimize the worst-case output variance [13, 9]. Yet another example of application is value at risk (VaR) portfolio optimization, where the portfolio is optimized based on a record of past asset returns [40]. The link between VaR portfolio optimization and (1) is discussed in [44, 50, 45].

When the computed x^* is applied to the real world, a new realization of the uncertainty parameter δ independent of $\delta^{(1)}, \delta^{(2)}, \ldots, \delta^{(N)}$ is experienced, and one issue that arises quite naturally is the assessment of the performance achieved by x^* for a new δ. This analysis is conducted in this paper by relying on the $\delta^{(i)}$'s without resorting to new scenarios. With this objective in mind, we introduce the following definitions.

DEFINITION 1 (empirical cost). *Consider the cost values* $f(x^*, \delta^{(i)})$, $i = 1, \ldots, N$, *achieved by the solution* x^* *of* (1) *for the seen scenarios* $\delta^{(i)}$*'s, and sort them in decreasing order:* $f(x^*, \delta^{(i_1)}) \geq f(x^*, \delta^{(i_2)}) \geq \cdots \geq f(x^*, \delta^{(i_N)})$. *The* k*th empirical cost is defined as*

$$c_k^* := f(x^*, \delta^{(i_k)}).$$

See Figure 1 for an illustration of the concept of empirical cost.

DEFINITION 2 (risk). *For any given* $x \in \mathbb{R}^d$ *and* $c \in \mathbb{R}$, *the risk associated with* (x, c) *is* $R(x, c) = \mathbb{P}\{\delta \in \Delta : f(x, \delta) > c\}$. *The risk of the empirical cost* c_k^* *is defined as*

$$R_k = R(x^*, c_k^*).$$

The risk R_k is defined as the composition of $R(x,c)$ with (x^*, c_k^*), so that R_k is a random variable that depends on $\delta^{(1)}, \delta^{(2)}, \ldots, \delta^{(N)}$ through x^* and c_k^*. The interpretation of R_k is that it is the conditional probability given $\delta^{(1)}, \delta^{(2)}, \ldots, \delta^{(N)}$ that a new realization of δ from $(\Delta, \mathbb{P})$ incurs a cost $f(x^*, \delta)$ greater than c_k^*, and R_k can be equivalently written as $R_k = \mathbb{P}^{N+1}\{f(x^*, \delta) > c_k^* | \delta^{(1)}, \delta^{(2)}, \ldots, \delta^{(N)}\}$, where $\mathbb{P}^{N+1} = \mathbb{P} \times \mathbb{P} \times \cdots \times \mathbb{P}$ is the probability distribution of $\delta^{(1)}, \delta^{(2)}, \ldots, \delta^{(N)}, \delta$, which is a product probability due to independence. As a shorthand notation, in what follows we shall also write $R_k = \mathbb{P}_\delta\{f(x^*, \delta) > c_k^*\}$, where $\mathbb{P}_\delta$ indicates that the probability is computed with respect to δ, while x^* and c_k^* are kept fixed.

To make the concepts of empirical cost and risk more concrete, refer to the linear regression example. Here, $c_1^* = \max_{i=1,\ldots,N} |y^{(i)} - \sum_{j=1}^d x_j^* u_j^{(i)}|$ is the largest vertical distance between the observations and the regression hyperplane, that is, twice c_1^* is the vertical thickness of a layer that contains all observations. Cost c_k^* is instead the kth largest vertical distance between the observations and the regression hyperplane. Thus, a layer whose vertical thickness is $2c_k^*$ contains all observations but $k-1$ of them. For given observations, the risk R_k is the probability that a new observation falls outside this layer. Knowledge of R_k is important in prediction problems.

To assess the performance achieved by x^* for a new random δ, theoretical bounds on R_k are established in this paper.[2] This goal is pursued without resorting to new realizations of the uncertainty parameter, that is, only the realizations $\delta^{(1)}, \delta^{(2)}, \ldots, \delta^{(N)}$ used in optimization are available. This set-up is of interest any time the realizations represent a costly and limited resource, as is the case in *data-driven* optimization problems where the scenarios are observations; see, e.g., [7, 6, 51]. This is different from assessing the value of R_k with new realizations of the uncertainty parameter by a Monte Carlo procedure; see, e.g., [39, 15, 34, 4, 5].

In the literature, the problem of assessing the risk associated with empirical costs has been studied for c_1^*, and various results are available that cover both the asymptotic case when $N \to \infty$ (see, e.g., [50] and the references therein) and the finite sample case, which has been considered in a series of papers by the authors of this contribution [8, 9, 11]. Moreover, extensions to a nonconvex context [37, 1] and to a multistage set-up [52] are also available. The present work is in the vein of the so-called scenario approach of [8, 9, 11, 12, 23]. In [11], the sharpest possible characterization of the risk R_1 is provided. It is shown that R_1, which, we recall, is a random variable that depends on $\delta^{(1)}, \delta^{(2)}, \ldots, \delta^{(N)}$ through x^* and c_1^*, has a cumulative distribution function that is lower-bounded by a Beta distribution with parameters $(d+1, N-d)$. From this fact it follows that R_1 tends to zero with probability 1 as $N \to \infty$. In [2] it has been shown that the tail of the Beta distribution beyond the value $\rho := \frac{1}{N}(d + \ln\frac{1}{\beta} + \sqrt{2d\ln\frac{1}{\beta}})$ has a probability smaller than β, so that, based on the result in [11], for any finite N relation $R_1 \le \rho$ holds with confidence $1-\beta$. These results have opened new avenues to address stochastic optimization problems with a VaR risk measure. Indeed, relation $R_1 \le \rho$ means that the value attained by the scenario solution x^* exceeds c_1^* with probability no more than ρ, that is, the VaR at level $1-\rho$ is smaller than or equal to c_1^*. For more discussion, see [8, 11].

In this paper, we move an important step beyond the results in [11]. One first observation is that the set of all empirical costs $c_1^*, c_2^*, \ldots, c_N^*$ provides a much more

[2] In other words, we study the feasibility of (x^*, c_k^*) in a chance-constrained sense. See [46, 47, 48, 16] for general references on chance-constrained optimization and [28, 19, 49, 18, 29, 43, 42, 30, 38, 54, 36] for recent advances.

complete characterization of the goodness of the solution x^* than c_1^* only. Suppose, for instance, that the gaps between the costs c_k^* are large. Then, intuitively, it is expected that a new δ will obtain a cost $f(x^*, \delta)$ significantly smaller than c_1^* with high probability. On the other hand, when most values c_k^* concentrate near c_1^*, it is expected that $f(x^*, \delta)$ takes a value close to c_1^* with high probability. This idea is not new. A similar approach is found in [31, 32], where the empirical costs are used in a financial decision optimization context. What this paper offers is a precise theory to put such reasonings on a solid quantitative ground. Precisely, we compute the *joint probability distribution* of all risks $R_1, R_2, \ldots, R_N$ and show that this joint probability distribution is lower-bounded by an *ordered Dirichlet distribution*. This result represents a rigorous tool to support decisions in many real applications even for small sample sizes. In particular, since the ordered Dirichlet distribution is thin-tailed, the risks can be bounded with high confidence. Based on these findings, we further show that the cumulative distribution function of the cost $f(x^*, \delta)$ belongs to a probability box with high confidence, and this result provides an easy-to-inspect characterization of the quality of the sample-based solution x^*. All the results of this paper hold independently of the probability $\mathbb{P}$, i.e., they are *distribution-free*, so that they are well-suited for *data-driven optimization*, where knowledge on the probability $\mathbb{P}$ is missing. The significance of the found results is highlighted by an application example on the equalization of a communication channel.

1.1. Structure of the paper. Section 2 provides the main results of the paper. The practical use of the results is discussed in section 3. Section 4 presents a numerical example, while the proofs are in section 5.

2. The risk of empirical costs: Theoretical results. The following assumption on the solution of problem (1) is in force throughout the paper.

Assumption 1 (existence and uniqueness). For every value of N and $\delta^{(1)}, \delta^{(2)}, \ldots, \delta^{(N)}$, the optimal solution x^* to (1) exists and is unique.

Although problem (1) is always feasible, existence of the solution may be lost when the cost value improves as x drifts away toward infinity in some directions. This behavior can be prevented by confining optimization to a compact domain $\mathcal{X}$. In this way, existence in Assumption 1 is secured. In addition, uniqueness can be enforced by introducing suitable tie-break rules as discussed in section 2.1 of [11].

For future use, it is convenient to rewrite problem (1) in epigraphic form as follows:

$$\begin{aligned} \text{EPI}_N : \quad & \min_{x \in \mathcal{X} \subseteq \mathbb{R}^d, c \in \mathbb{R}} c \\ & \text{subject to } f(x, \delta^{(i)}) \leq c, \quad i = 1, \ldots, N. \end{aligned} \tag{2}$$

The following Definitions 3 and 4 are taken from [11].

DEFINITION 3 (support scenario). *The scenario $\delta^{(i)}$, $i \in \{1, \ldots, N\}$, is called a* support scenario *for problem* (2) *if its removal changes the solution (x^*, c_1^*) of* (2).

Loosely speaking, support scenarios are those preventing the solution from "falling" to a lower position. It can be proven[3] that the number of support scenarios of problem (2) is at most $d+1$. Figure 1 shows two cases with $d = 1$ where the number of support scenarios are two and one, respectively.

[3] A proof of this result stated in a slightly different but equivalent way can be found in [33]. The proof of [33] is based on Helly's theorem. A self-contained proof of the result stated in the terminology of the present paper is given in [8].

DEFINITION 4 (fully supported problem). *Problem* (2) *is* fully supported *if, for every* $N \geq d+1$*, with probability one with respect to the random sample* $\delta^{(1)}, \delta^{(2)}, \ldots, \delta^{(N)}$*, the number of its support scenarios is* $d+1$.

By the very definition of support scenario, all the support scenarios attain the same cost, and when the problem is fully supported we have that $c_1^* = c_2^* = \cdots = c_{d+1}^*$. The fact that the empirical costs from the $(d+1)$th on are not distinct, instead, is regarded as a situation of degeneracy.

DEFINITION 5 (nondegenerate problems). *Problem* (2) *is* nondegenerate *if, for every* $N \geq d+1$*, with probability one with respect to the random sample* $\delta^{(1)}, \delta^{(2)}, \ldots, \delta^{(N)}$*, it holds that*

$$c_{d+1}^* \neq c_{d+2}^* \neq \cdots \neq c_N^*. \tag{3}$$

A sufficient condition for nondegeneracy is that, for any given $x \in \mathbb{R}^d$ and $c \in \mathbb{R}$, it holds that $\mathbb{P}\{\delta \in \Delta : f(x,\delta) = c\} = 0$. In other words, for any x, the probability distribution of $f(x,\delta)$ has no concentrated mass. Though less general than (3), this condition is easier to check and it may be helpful in some situations.

Theorem 1 below characterizes the joint probability distribution of the risks $R_{d+1}, R_{d+2}, \ldots, R_N$ for nondegenerate problems; Theorem 2 extends the result to when the nondegeneracy condition is removed.

Before the theorems are stated, we recall that the *ordered Dirichlet distribution* with parameters

$$(d+1, \underbrace{1, 1, \ldots, 1}_{N-d})$$

is the probability distribution whose density function is

$$p(\nu_{d+1}, \nu_{d+2}, \ldots, \nu_N) = \frac{N!}{d!}\nu_{d+1}^d \mathbb{1}\{0 \leq \nu_{d+1} \leq \nu_{d+2} \leq \cdots \leq \nu_N \leq 1\},$$

where $\mathbb{1}\{\cdot\}$ denotes the indicator function; see, e.g., [53, p. 182]. Its cumulative distribution function is

$$F_{d,N}(\epsilon_{d+1}, \ldots, \epsilon_N) = \frac{N!}{d!}\int_0^{\epsilon_{d+1}} \nu_{d+1}^d \int_0^{\epsilon_{d+2}} \int_0^{\epsilon_{d+3}} \cdots \int_0^{\epsilon_N} \mathbb{1}\{0 \leq \nu_{d+1} \leq \cdots \leq \nu_N \leq 1\} \mathrm{d}\nu_N \cdots \mathrm{d}\nu_{d+3} \mathrm{d}\nu_{d+2} \mathrm{d}\nu_{d+1}.$$

Section 3.4 provides additional information on Dirichlet distributions. In the theorems, $\mathbb{P}^N = \mathbb{P} \times \mathbb{P} \times \cdots \times \mathbb{P}$ is the probability distribution of $\delta^{(1)}, \delta^{(2)}, \ldots, \delta^{(N)}$; it is a product probability since the scenarios are independent.

THEOREM 1. *If problem* (2) *is nondegenerate, then the joint probability distribution function of* $R_{d+1}, \ldots, R_N$ *is given by* $F_{d,N}(\epsilon_{d+1}, \epsilon_{d+2}, \ldots, \epsilon_N)$*, i.e.,*

$$\mathbb{P}^N\{R_{d+1} \leq \epsilon_{d+1}, R_{d+2} \leq \epsilon_{d+2}, \ldots, R_N \leq \epsilon_N\} = F_{d,N}(\epsilon_{d+1}, \epsilon_{d+2}, \ldots, \epsilon_N). \tag{4}$$

Proof. See section 5.1. ∎

Theorem 1 gives the joint probability distribution function of the risks from the $(d+1)$th on and shows that this distribution does not depend on $\mathbb{P}$ (*distribution-free* result). Since $c_i^* \geq c_{d+1}^*$, $i = 1, \ldots, d$, we have that $R_i \leq R_{d+1}$, $i = 1, \ldots, d$, and the bound ϵ_{d+1} automatically applies also to R_i, $i = 1, \ldots, d$. Arguably, this is the most general distribution-free result possible, since the distribution of R_i, $i =$

$1, \dots, d$, is problem dependent. When the problem is fully supported, it holds that $c_1^* = c_2^* = \cdots = c_{d+1}^*$ so that $R_1 = R_2 = \cdots = R_{d+1}$, and Theorem 1 exactly characterizes the joint distribution of all risks R_i, $i = 1, \dots, N$. Thus, we see that the class of fully supported problems admits "universal" risks $R_1, R_2, \dots, R_N$, in the sense that their joint probability distribution function is the same irrespective of the problem at hand.

It is well known (see, e.g., [53]) that the marginals of an ordered Dirichlet distribution are Beta distributions. From this, one can infer that the probability distribution function of R_k, $k = d+1, \dots, N$, is a Beta distribution with parameters $(k, N-k+1)$, that is,

$$\mathbb{P}^N\{R_k \le \epsilon\} = 1 - \sum_{i=0}^{k-1} \binom{N}{i} \epsilon^i (1-\epsilon)^{N-i}. \tag{5}$$

For $k = d+1$, and recalling that $R_1 \le R_{d+1}$, we obtain

$$\mathbb{P}^N\{R_1 \le \epsilon\} \ge \mathbb{P}^N\{R_{d+1} \le \epsilon\} = 1 - \sum_{i=0}^{d} \binom{N}{i} \epsilon^i (1-\epsilon)^{N-i}. \tag{6}$$

This bound on $\mathbb{P}^N\{R_1 \le \epsilon\}$ is bound (6) in Theorem 1 of [11], which is recovered as a byproduct of the general theory of this paper.[4] Actually, (6) proves more than the result in [11] since the right-hand side of (6) is recognized to be the exact probability distribution function of the risk of c_{d+1}^*.

The nondegeneracy assumption in Theorem 1 cannot be removed while preserving the equality in (4). In fact, for the sake of the argument, suppose, e.g., that the probability distribution $\mathbb{P}$ is concentrated on a unique scenario $\bar{\delta}$. Then, all of the costs $c_1^*, c_2^*, \dots, c_N^*$ are equal and have zero risk. Although Theorem 1 does not hold for degenerate problems, it can be shown that the distribution of the risks $R_{d+1}, R_{d+2}, \dots, R_N$ is always lower-bounded by the ordered Dirichlet distribution, as is formally stated in the next theorem.

THEOREM 2. *For any problem* (2)*, the joint probability distribution function of* $R_{d+1}, \dots, R_N$ *is lower-bounded by* $F_{d,N}(\epsilon_{d+1}, \dots, \epsilon_N)$*, i.e.,*

$$\mathbb{P}^N\{R_{d+1} \le \epsilon_{d+1}, R_{d+2} \le \epsilon_{d+2}, \dots, R_N \le \epsilon_N\} \ge F_{d,N}(\epsilon_{d+1}, \epsilon_{d+2}, \dots, \epsilon_N). \tag{7}$$

Proof. See section 5.2. ∎

Since $R_1 \le \cdots \le R_{d+1}$, the following corollary that covers all the R_i, $i = 1, \dots, N$, follows.

COROLLARY 1. *For any problem* (2)*, it holds that*

$$\begin{aligned}\mathbb{P}^N\{&R_1 \le \epsilon_{d+1}, \dots, R_{d+1} \le \epsilon_{d+1}, R_{d+2} \le \epsilon_{d+2}, \dots, R_N \le \epsilon_N\} \\ &\ge F_{d,N}(\epsilon_{d+1}, \epsilon_{d+2}, \dots, \epsilon_N).\end{aligned}$$

3. Practical use of the results and discussion.

3.1. Connection with order statistics. Consider N independent realizations of a continuous random variable Y with cumulative distribution function F, and sort them in decreasing order to obtain

$$Y_1 \ge Y_2 \ge \cdots \ge Y_N.$$

[4] To help the reader match the terminology of this paper to that of [11], we notice that in [11] the risk R_1 is named the *violation probability* of the solution (x^*, c_1^*) of (2).

These variables are called order statistics, and it is well known that the vector $(1 - F(Y_1), 1 - F(Y_2), \dots, 1 - F(Y_N))$ is a random element whose joint probability distribution is an ordered Dirichlet [53, 26]. Order statistics are recovered as a particular case of the theory developed in this paper by letting $\delta = Y$ and $f(x, \delta) = \delta$, i.e., when we are in a purely descriptive set-up where no optimization is performed. The surprising fact expressed by Theorem 1 is that joint probability distribution of the risks is still an ordered Dirichlet when an optimization variable is present, a framework that is way more complex than that of order statistics. As a matter of fact, order statistics are ordered values from the real line, while the empirical costs lie on a random line passing through x^*, which is selected by solving an optimization problem. From the mathematical side, this fact implies that our results cannot be traced back to order statistics and their derivation demands a genuinely new approach.

3.2. Postexperiment analysis and experiment design. The results presented in section 2 can be applied in various ways. Two examples are in order.

Postexperiment analysis. The user has available a sample $\delta^{(1)}, \delta^{(2)}, \dots, \delta^{(N)}$ and solves problem (1) obtaining x^* and the corresponding empirical costs c_k^*, $k = 1, \dots, N$. He/she then chooses a confidence parameter value, e.g., $\beta = 10^{-7}$, and determines values for $\epsilon_{d+1}, \epsilon_{d+2}, \dots, \epsilon_N$ such that $F_{d,N}(\epsilon_{d+1}, \epsilon_{d+2}, \dots, \epsilon_N) \geq 1 - \beta$. By applying Theorem 2, the user can claim that $\mathbb{P}_\delta\{f(x^*, \delta) > c_k^*\} \leq \epsilon_k$ holds true simultaneously for all $k = d + 1, \dots, N$ with high confidence $1 - \beta$.

Experiment design. The user chooses a confidence parameter value, e.g., $\beta = 10^{-7}$. Then he/she fixes desired upper bounds on the risks of the empirical costs, that is, $0 \leq \epsilon_{d+1} \leq \epsilon_{d+2} \leq \cdots \leq \epsilon_N \leq 1$ (selecting $\epsilon_k = 1$ for some k corresponds to having no constraints on the risk of c_k^*). Then, he/she computes the minimum number N of scenarios that guarantees that $F_{d,N}(\epsilon_{d+1}, \epsilon_{d+2}, \dots, \epsilon_N) \geq 1 - \beta$, and samples N scenarios to be used in problem (1). Theorem 2 can be applied to give the same guarantees as in the postexperiment analysis.

3.3. Bounding the cumulative distribution function of $f(x^*, \delta)$. By the definition of risk, $R(x^*, c) = \mathbb{P}_\delta\{f(x^*, \delta) > c\} = 1 - \mathbb{P}_\delta\{f(x^*, \delta) \leq c\}$. Thus, if $\epsilon_{d+1}, \epsilon_{d+2}, \dots, \epsilon_N$ are chosen such that $F_{d,N}(\epsilon_{d+1}, \epsilon_{d+2}, \dots, \epsilon_N) \geq 1 - \beta$, then, by letting $\epsilon_k = \epsilon_{d+1}$ for $k \leq d$, from Corollary 1 we have with confidence $1 - \beta$ that

$$\mathbb{P}_\delta\{f(x^*, \delta) \leq c_k^*\} \geq 1 - \epsilon_k \quad \text{for all } k = 1, \dots, N. \tag{8}$$

Observing that $\mathbb{P}_\delta\{f(x^*, \delta) \leq c\}$ is increasing with c, (8) implies that

$$\mathbb{P}_\delta\{f(x^*, \delta) \leq c\} \geq L(c), \tag{9}$$

where

$$L(c) = \begin{cases} 1 - \epsilon_1 & \text{if } c \geq c_1^*, \\ 1 - \epsilon_k & \text{if } c_k^* \leq c < c_{k-1}^* \ (k = 2, \dots, N), \\ 0 & \text{if } c < c_N^*. \end{cases}$$

The right-hand side of (9), $L(c)$, is a step function that, with confidence $1 - \beta$, lower bounds the cumulative distribution function of the cost $f(x^*, \delta)$ (first order stochastic dominance[5]). The lower step function in Figure 2 gives an example of this construction.

[5] An interesting stream of research in stochastic optimization introduces stochastic dominance as a constraint; see, e.g., [20, 21, 17, 35]. The difference between these papers and the present contribution is that the distribution on the right-hand side of (9) is here a posteriori computed, and it is not a priori enforced as a constraint.

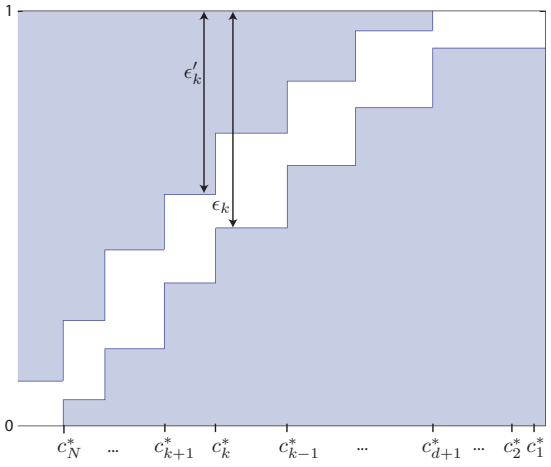


FIG. 2. *A "probability box" for the cumulative distribution function of $f(x^*, \delta)$. With confidence $1 - \beta - \beta'$, the whole graph of $\mathbb{P}_\delta\{f(x^*, \delta) \leq c\}$ lies in the white area bounded by the two step functions.*

Result (9) can be refined when the problem is nondegenerate and Theorem 1 applies. Since the joint distribution of $R_{d+1}, \ldots, R_N$ is exactly known in this case, in addition to the ϵ_k's computed above, values $\epsilon'_{d+1}, \epsilon'_{d+2}, \ldots, \epsilon'_N$ can be obtained such that, with confidence $1 - \beta'$, it holds that

$$\mathbb{P}_\delta\{f(x^*, \delta) \leq c^*_k\} \leq 1 - \epsilon'_k \quad \text{for all } k = d+1, \ldots, N. \tag{10}$$

By the monotonicity of $\mathbb{P}_\delta\{f(x^*, \delta) \leq c\}$, we conclude that, with confidence $1-\beta-\beta'$, $\mathbb{P}_\delta\{f(x^*, \delta) \leq c\}$ can be bounded from below and from above as follows:

$$U(c) \geq \mathbb{P}_\delta\{f(x^*, \delta) \leq c\} \geq L(c), \tag{11}$$

where

$$U(c) = \begin{cases} 1 & \text{if } c > c^*_{d+1}, \\ 1 - \epsilon'_k & \text{if } c^*_{k+1} < c \leq c^*_k \ (k = d+1, \ldots, N-1), \\ 1 - \epsilon'_N & \text{if } c \leq c^*_N. \end{cases}$$

Relation (11) defines, with confidence $1 - \beta - \beta'$ with respect to the variability of $\delta^{(1)}, \delta^{(2)}, \ldots, \delta^{(N)}$, a "probability box"[6] for the conditional cumulative distribution function of $f(x^*, \delta)$ given x^*; see again Figure 2. See section 4 for the construction of the probability box in a concrete example.

3.4. Computational issues for the ordered Dirichlet distribution.

Ordered Dirichlet versus Dirichlet distributions. Equation (4) states that the random vector $(R_{d+1}, R_{d+2}, \ldots, R_N)$ is distributed according to an *ordered Dirichlet distribution*. By the transformation

$$D_{d+1} = R_{d+2} - R_{d+1}, D_{d+2} = R_{d+3} - R_{d+2}, \ldots, D_{N-1} = R_N - R_{N-1}, D_N = 1 - R_N,$$

[6] For the definition of "probability box" and a discussion of its usefulness in statistics, see, e.g., [3].

one obtains vector $(D_{d+1}, D_{d+2}, \ldots, D_N)$, which is distributed according to the so-called Dirichlet distribution, [53]. Hence, the evaluation of an ordered Dirichlet distribution function can be converted into the problem of evaluating a Dirichlet distribution function. The reader is referred to [24, 25] and references therein for studies on computational issues for Dirichlet distributions.

Marginal distributions. We have already observed that for nondegenerate problems the probability distribution function of each R_k is a Beta with parameters $(k, N-k+1)$ for $k = d+1, \ldots, N$; see (5). Notably, the right-hand side of (5) can be easily evaluated by means of common tools, like the betainc function in MATLAB, or pbeta in R. As is clear, a lower bound for the joint probability distribution function of $R_{d+1}, R_{d+2}, \ldots, R_N$ is given by the sum of the marginals, so that one obtains

$$\begin{aligned} &\mathbb{P}^N\{R_{d+1} \le \epsilon_{d+1}, \ldots, R_N \le \epsilon_N\} \\ &\ge 1 - \sum_{k=d+1}^{N} \mathbb{P}^N\{R_k > \epsilon_k\} \\ &= 1 - \sum_{k=d+1}^{N} \sum_{i=0}^{k-1} \binom{N}{i} \epsilon_k^i (1-\epsilon_k)^{N-i}. \end{aligned} \tag{12}$$

See section 4 for an example of the use of formula (12).

An explicit expression for N. Based on (12), and following similar calculations as in the proof of (12) in [23], it can be shown that, for a given $\beta \in (0,1)$, if

$$N \ge \max_{k=d+1,\ldots,N} N^{(k)},$$

where

$$N^{(k)} = \left\lfloor \frac{2}{\epsilon_k}\left(k + \ln\frac{1}{\beta}\right) + \frac{4}{\epsilon_k}\ln\left(\frac{2}{\epsilon_k}\left(k + \ln\frac{1}{\beta}\right)\right)\right\rfloor + 1$$

($\lfloor\cdot\rfloor$ denotes integer part), then $\mathbb{P}^N\{R_{d+1} \le \epsilon_{d+1}, \ldots, R_N \le \epsilon_N\} \ge 1-\beta$, i.e., conditions $R_k \le \epsilon_k$, $k = d+1, \ldots, N$, hold simultaneously with high confidence $1-\beta$. This bound has a logarithmic dependence of N on β, a fact that shows that a very high confidence can be enforced without increasing N too much.

4. A numerical example.

The example in this section is inspired by the equalizer design problem in [41].

4.1. Problem formulation.

In a digital communication system, a signal $u(t)$, $t = \ldots, -2, -1, 0, 1, 2, \ldots$, is sent from a transmitter to a receiver through a communication channel C. In general, the signal $\tilde{u}(t)$ at the receiver end is different from the transmitted signal owing to the distortion introduced by the channel. We assume that the channel acts approximately as a linear filter so that its behavior is characterized by its frequency response $C(\omega)$, which is a complex-valued function of $\omega \in [-\pi, \pi]$ linking the Fourier transform $U(\omega)$ of $u(t)$ to the Fourier transform $\tilde{U}(\omega)$ of $\tilde{u}(t)$ according to the equation $\tilde{U}(\omega) = C(\omega)U(\omega)$. If the distortion introduced by the channel is unacceptably high, a device E, called the equalizer, can be added at the receiver end to improve the quality of the received signal.

The equalizer E is a filter whose frequency response is denoted by $E(\omega)$. We consider a d-tap finite impulse response equalizer:

$$E(\omega) = \sum_{k=0}^{d-1} x_k e^{-ik\omega}, \tag{13}$$

where i is the imaginary unit and $x_0, x_1, \ldots, x_{d-1}$ are real parameters through which the frequency response of E can be shaped. Overall, the frequency response of the channel-equalizer cascade is $C(\omega)E(\omega)$, and the aim is to design the equalizer E so as to make $C(\omega)E(\omega)$ as similar as possible to a desired frequency response that incorporates the idea that the equalized channel should introduce little distortion. In line with [41], the desired frequency response we consider is $e^{-iD\omega}$, the frequency response of a pure delay of D time steps. As for the cost function, a grid $\omega_k = \frac{k}{100}\pi$, $k = 0, \pm 1, \ldots, \pm 100$ of $[-\pi, \pi]$ is considered, and we take

$$f(x) = \frac{1}{201} \sum_{k=-100}^{100} |C(\omega_k)E(\omega_k) - e^{-iD\omega_k}| + \max_{k=-100,\ldots,100} |C(\omega_k)E(\omega_k) - e^{-iD\omega_k}|. \tag{14}$$

The first term is the average deviation and takes care of the global behavior over the whole range of frequencies, while the second term penalizes the presence of large deviations localized at given frequencies caused by resonant peaks in $C(\omega)E(\omega)$. Resonant peaks are undesirable because they generate annoying whistling noise in audio communications.

The cost function in (14) assumes that $C(\omega)$ is known. In real-world applications, the frequency response of the channel is often not completely known because of imperfections in the procedure used to estimate $C(\omega)$, or due to intrinsic variability of the environment, as is the case, for example, in mobile communication. Hence, in what follows we consider a channel function $C(\omega, \delta)$ in place of $C(\omega)$ in (14), where δ is a parameter describing uncertainty, and the cost function is correspondingly written as $f(x, \delta)$.

4.2. The scenario approach. In this simulation example, the scenarios are generated according to the model

$$C(\omega, \delta) = \frac{1}{e^{i2\omega} + \delta_1 e^{i\omega} + \delta_2},$$

where $\delta = (\delta_1, \delta_2)$ is uniformly distributed over $[-0.4, 0.4] \times [0.5, 0.8]$. We take $N = 3000$, $d = 10$, and $D = 8$. According to the scenario approach, the equalizer E^* is obtained by solving

$$\min_{x \in \mathbb{R}^{10}} \max_{i=1,\ldots,3000} f(x, \delta^{(i)}). \tag{15}$$

The solution we found is $x^* = (7.08 \cdot 10^{-2}, 1.00 \cdot 10^{-3}, -6.64 \cdot 10^{-2}, 1.42 \cdot 10^{-3}, 4.71 \cdot 10^{-2}, 3.73 \cdot 10^{-4}, 8.37 \cdot 10^{-1}, 2 \cdot 10^{-3}, 5.09 \cdot 10^{-1}, -3.46 \cdot 10^{-4})$, and the empirical costs c_k^* were also evaluated, and they are plotted in Figure 3. The empirical costs c_k^* are monotonically decreasing and do not accumulate, a fact that is not surprising since δ has a density. In a real application, Theorem 2 is always applicable since it holds without any assumption. However, observing that the c_k^*'s do not accumulate, possibly used in conjunction with prior knowledge on the application domain, may justify that one assumes that the problem is nondegenerate and Theorem 1 is applied.

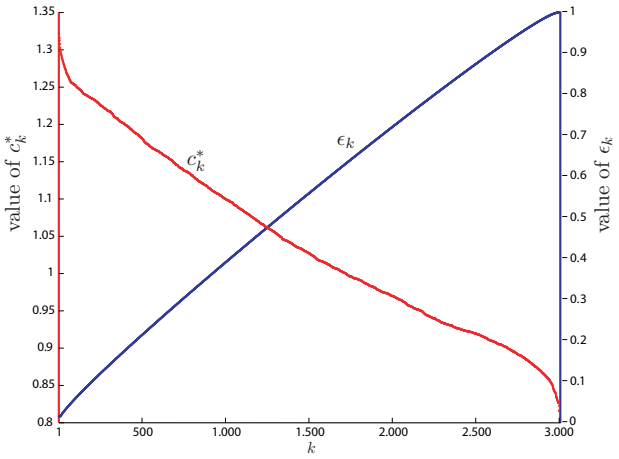


FIG. 3. *Functions c_k^* and ϵ_k, $k = 1, \ldots, 3000$. As k increases, c_k^* goes down from* 1.322 *to* 0.816*, and the bound ϵ_k on the risk to exceed c_k^* increases from* 0.0159 *to* $1 - 10^{-14}$.

Values $\epsilon_{11}, \ldots, \epsilon_{3000}$ are chosen as follows: β was set to 10^{-7} and, for each $k = 11, 12, \ldots, 3000$, ϵ_k is obtained by solving the equation

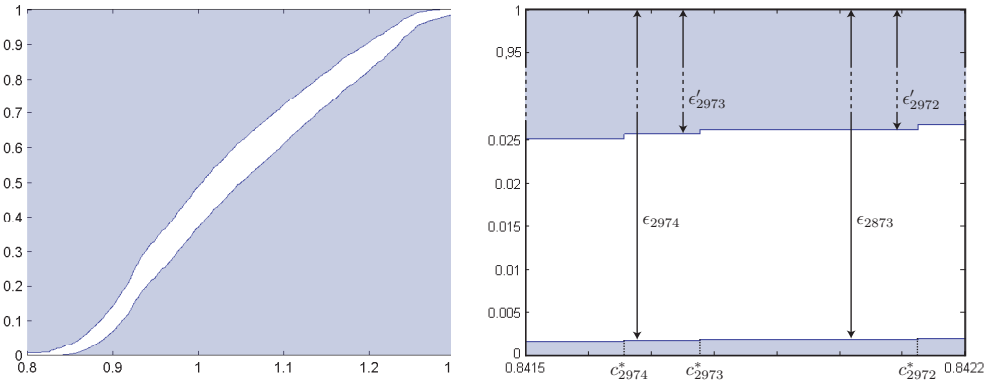
$$\sum_{i=0}^{k-1} \binom{N}{i} \epsilon_k^i (1 - \epsilon_k)^{N-i} = \frac{\beta}{2990},$$

which, assuming nondegeneracy, corresponds to choosing the ϵ_k's so that the marginal probability $\mathbb{P}^N\{R_k > \epsilon_k\}$ is equal to $\frac{\beta}{2990}$ for all $k = 11, 12, \ldots, 3000$. The values of ϵ_k are also displayed in Figure 3. Recalling (12), this choice implies that $\mathbb{P}^N\{R_{11} \leq \epsilon_{11}, \ldots, R_{3000} \leq \epsilon_{3000}\} \geq 1 - 10^{-7}$. Thus, we can, e.g., claim that the risk that the equalizer E^* incurs a cost greater than $c_{11}^* = 1.298$ is no more than $\epsilon_{11} = 1.59\%$, i.e., cost 1.298 is guaranteed for 98.41% of the channel frequency responses $C(\omega, \delta)$. Likewise, cost $c_{12}^* = 1.297$ is guaranteed for the $1 - \epsilon_{12} = 98.35\%$ of the channel frequency responses, and so on for any value of k. These claims are true simultaneously for all k with high confidence $1 - \beta = 1 - 10^{-7}$. We can also construct a probability box for the cumulative distribution function of $f(x^*, \delta)$ as explained in section 3.3. Values $\epsilon'_{11}, \ldots, \epsilon'_{3000}$ are chosen such that the marginal probability $\mathbb{P}^N\{R_k < \epsilon'_k\}$ is equal to $\frac{\beta}{2990}$ for all $k = 11, 12, \ldots, 3000$, that is,

$$1 - \sum_{i=0}^{k-1} \binom{N}{i} (\epsilon'_k)^i (1 - \epsilon'_k)^{N-i} = \frac{\beta}{2990}.$$

Hence, $\epsilon'_k \leq R_k \leq \epsilon_k$ holds for each $k = 11, \ldots, 3000$, with confidence at least $1 - 2 \cdot 10^{-7}$. Figure 4 represents the probability box found in this example.

In conclusion, the user has at his disposal an evaluation of the probability of the various costs when the design x^* is applied. A graph like that visualized in Figure 4 is interpreted that, if a vertical line is drawn from any cost value, this line crosses the white region over a segment that contains the probability with which that cost is incurred. This result is established without using any knowledge of the distribution $\mathbb{P}$.



(a) (b)

FIG. 4. (a) *With confidence* $1-2\cdot 10^{-7}$ *the cumulative distribution function of the cost function* $f(x^*,\delta)$ *lies in the white strip.* (b) *Zoomed-in detail of Figure* 4(a).

5. Proofs.

5.1. Proof of Theorem 1. For any fixed $(x,\underline{c},\bar{c})\in\mathbb{R}^{d+2}$, let $D(x,\underline{c},\bar{c})=\mathbb{P}\{\delta\in\Delta : \underline{c}<f(x,\delta)\le\bar{c}\}$ and, for any integer k such that $d+1\le k\le N$, let

$$D_k = D(x^*, c^*_{k+1}, c^*_k), \tag{16}$$

where c^*_{N+1} is defined to be equal to $-\infty$. Similarly to the R_k's, the D_k's are random variables, since they depend on the sample $(\delta^{(1)},\dots,\delta^{(N)})$ through $x^*, c^*_{d+1},\dots,c^*_N$. The interpretation of D_k is that it is the conditional probability with respect to x^*, c^*_{k+1}, c^*_k that a new realization of δ incurs a cost between levels c^*_k and c^*_{k+1}. The variables D_k's and R_k's are related by the following simple linear transformations:

$$\begin{aligned}
&D_{d+1} = R_{d+2} - R_{d+1} && && R_{d+1} = 1 - \sum_{i=d+1}^{N} D_k\\
&D_{d+2} = R_{d+3} - R_{d+2} && \text{or, equivalently,} && R_{d+2} = 1 - \sum_{i=d+2}^{N} D_k\\
&\quad\vdots && && \quad\vdots\\
&D_{N-1} = R_N - R_{N-1} && && R_{N-1} = 1 - \sum_{i=N-1}^{N} D_k\\
&D_N = 1 - R_N && && R_N = 1 - D_N.
\end{aligned} \tag{17}$$

Thanks to (17), the joint probability distribution function of the R_k's can be easily derived from the joint probability distribution function of the D_k's and vice versa. Hence, we proceed by computing the joint probability distribution function of the D_k's first. In order to do so, we consider $\mathbb{E}[D_{d+1}^{k_{d+1}}\cdots D_N^{k_N}]$, the multivariate moment of $D_{d+1},\dots,D_N$, and evaluate it for each possible assignment of nonnegative integers $k_{d+1},\dots,k_N$. The joint distribution function of $D_{d+1},\dots,D_N$ can then be deduced from the resulting moment problem.

To ease the notation, define $M_d = N$, $M_{d+1} = N + k_{d+1}$, $M_{d+2} = N + k_{d+1} + k_{d+2}$, etc., until $M_N = N + \sum_{i=d+1}^{N} k_i$. By (16), the product $D_{d+1}^{k_{d+1}} D_{d+2}^{k_{d+2}}\cdots D_N^{k_N}$

gives the conditional probability with respect to $x^*, c^*_{d+1}, \ldots, c^*_N$, i.e., with respect to $(\delta^{(1)}, \ldots, \delta^{(N)})$, that $M_N - N$ new independent realizations of the uncertainty parameter, say, $\delta^{(N+1)}, \ldots, \delta^{(M_N)}$, are such that the first k_{d+1} (i.e., $\delta^{(N+1)}, \ldots, \delta^{(M_{d+1})}$) incur a cost between c^*_{d+1} and c^*_{d+2}, the next k_{d+2} (i.e., $\delta^{(M_{d+1}+1)}, \ldots, \delta^{(M_{d+2})}$) incur a cost between c^*_{d+2} and c^*_{d+3}, and so on till the last k_N incurring a cost below c^*_N (recall that $c^*_{N+1} = -\infty$). Therefore, the product $D^{k_{d+1}}_{d+1} D^{k_{d+2}}_{d+2} \ldots D^{k_N}_N$ can be expressed as

(18)
$$\prod_{i=d+1}^{N} D_i^{k_i} = \mathbb{P}^{M_N - N}_{\boldsymbol{\delta}^{M_N}_{N+1}} \{c^*_{i+1} < f(x^*, \delta^{(j)}) \le c^*_i,\ i = d+1, \ldots, N,\ j = M_{i-1}+1, \ldots, M_i\},$$

where $\mathbb{P}^{M_N - N} = \mathbb{P} \times \cdots \times \mathbb{P}$ denotes the product probability measure of $\delta^{(N+1)}, \ldots, \delta^{(M_N)}$, and $\boldsymbol{\delta}^{M_N}_{N+1}$ is shorthand for $\delta^{(N+1)}, \ldots, \delta^{(M_N)}$. Expressing probability as the integral of an indicator function and using the notation $\Delta^{M_N}_{N+1} = \Delta \times \Delta \times \cdots \times \Delta$ to indicate the domain for $\boldsymbol{\delta}^{M_N}_{N+1}$, (18) can be rewritten as

$$\prod_{i=d+1}^{N} D_i^{k_i} = \int_{\Delta^{M_N}_{N+1}} \mathbb{1}\{c^*_{i+1} < f(x^*, \delta^{(j)}) \le c^*_i,\ i = d+1, \ldots, N,\ j = M_{i-1}+1, \ldots, M_i\} \mathbb{P}^{M_N - N}\{\mathrm{d}\boldsymbol{\delta}^{M_N}_{N+1}\}.$$

As $\delta^{(1)}, \ldots, \delta^{(N)}$ vary, $\prod_{i=d+1}^{N} D_i^{k_i}$ takes on various values and we are interested in computing its expected value, i.e.,

$$\begin{aligned}
&\mathbb{E}\left[\prod_{i=d+1}^{N} D_i^{k_i}\right] \\
&= \int_{\Delta_1^N} \prod_{i=d+1}^{N} D_i^{k_i} \mathbb{P}^N\{\mathrm{d}\boldsymbol{\delta}^N_1\} \\
&= \int_{\Delta_1^N} \int_{\Delta^{M_N}_{N+1}} \mathbb{1}\{c^*_{i+1} < f(x^*, \delta^{(j)}) \le c^*_i,\ j = M_{i-1}+1, \ldots, M_i,\ i = d+1, \ldots, N\} \mathbb{P}^{M_N - N}\{\mathrm{d}\boldsymbol{\delta}^{M_N}_{N+1}\} \mathbb{P}^N\{\mathrm{d}\boldsymbol{\delta}^N_1\},
\end{aligned}$$

which, by Fubini's theorem, can be restated as

$$\int_{\Delta_1^{M_N}} \mathbb{1}\{c^*_{i+1} < f(x^*, \delta^{(j)}) \le c^*_i,\ i = d+1, \ldots, N,\ j = M_{i-1}+1, \ldots, M_i\} \mathbb{P}^{M_N}\{\mathrm{d}\boldsymbol{\delta}^{M_N}_1\}.$$

Thus, the moment $\mathbb{E}[D^{k_{d+1}}_{d+1} \cdots D^{k_N}_N]$ is interpreted as the total probability with respect to all variables $\delta^{(1)}, \ldots, \delta^{(N)}, \delta^{(N+1)}, \ldots, \delta^{(M_N)}$ that $\delta^{(N+1)}, \ldots, \delta^{(M_{d+1})}$ incur a cost between c^*_{d+1} and c^*_{d+2}, $\delta^{(M_{d+1}+1)}, \ldots, \delta^{(M_{d+2})}$ incur a cost between c^*_{d+2} and c^*_{d+3}, and so on.

Now, let $\bar{S} = \{j_1, \ldots, j_N\}$ be a generic subset of N indexes taken from $\{1, \ldots, M_N\}$ and let $z^*_{|\bar{S}} := (x^*_{|\bar{S}}, c^*_{1|\bar{S}})$ be the optimal solution to problem

$$\mathrm{EPI}_{|\bar{S}}: \min_{c \in \mathbb{R}, x \in \mathcal{X} \subseteq \mathbb{R}^d} c \quad \text{subject to } f(x, \delta^{(i)}) \le c, \quad i \in \bar{S}.$$

Moreover, for $k = 1, \ldots, N$, let $c^*_{k|\bar{S}} = \max\{c \in \mathbb{R} : c \leq f(x^*_{|\bar{S}}, \delta^{(i)})$ for a choice of k indexes i among $\bar{S}\}$, i.e., the $c^*_{k|\bar{S}}$ are the empirical costs associated with $x^*_{|\bar{S}}$, and let $c^*_{N+1|\bar{S}} = -\infty$. Finally, for each $i = d+1, \ldots, N$, let $S_i = \{j_1, \ldots, j_{k_i}\}$ be a subset of k_i indexes from $\{1, \ldots, M_N\} \setminus \bar{S}$ so that $S_m \cap S_n = \emptyset$ if $m \neq n$. Due to the independently and identically distributed nature of $(\delta^{(1)}, \delta^{(2)}, \ldots, \delta^{(M_N)})$, the total probability that the realizations with indexes in S_{d+1} incur a cost between $c^*_{d+1|\bar{S}}$ and $c^*_{d+2|\bar{S}}$, those with indexes in S_{d+2} incur a cost between $c^*_{d+2|\bar{S}}$ and $c^*_{d+3|\bar{S}}$, and so on till those with indexes in S_N incurring a cost below $c^*_{N|\bar{S}}$, does not depend in any way on the choice of $\bar{S}, S_{d+1}, \ldots, S_N$. Then

$$\int_{\Delta_1^{M_N}} \mathbb{1}\{c^*_{i+1} < f(x^*, \delta^{(j)}) \leq c^*_i,\ j = M_{i-1}+1, \ldots, M_i,\ i = d+1, \ldots, N\} \mathbb{P}^{M_N}\{\mathrm{d}\boldsymbol{\delta}_1^{M_N}\}$$
$$= \int_{\Delta_1^{M_N}} \mathbb{1}\{c^*_{i+1|\bar{S}} < f(x^*_{|\bar{S}}, \delta^{(j)}) \leq c^*_{i\,|\bar{S}},\ j \in S_i,\ i = d+1, \ldots, N\} \mathbb{P}^{M_N}\{\mathrm{d}\boldsymbol{\delta}_1^{M_N}\}$$
for all $(\bar{S}, S_{d+1}, \ldots, S_N) \in \mathcal{S}$,

where $\mathcal{S}$ is the set of all feasible (i.e., of suitable sizes) choices of $\bar{S}, S_{d+1}, \ldots, S_N$ from $\{1, \ldots, M_N\}$. Indicating with $|\mathcal{S}|$ the cardinality of $\mathcal{S}$, we then have

$$\begin{aligned}
&\mathbb{E}[D_{d+1}^{k_{d+1}} \cdots D_N^{k_N}] \\
&= \int_{\Delta_1^{M_N}} \mathbb{1}\{c^*_{i+1} < f(x^*, \delta^{(j)}) \leq c^*_i,\ j = M_{i-1}+1, \ldots, M_i, \\
&\qquad\qquad i = d+1, \ldots, N\} \mathbb{P}^{M_N}\{\mathrm{d}\boldsymbol{\delta}_1^{M_N}\} \\
&= \frac{1}{|\mathcal{S}|} \sum_{(\bar{S}, S_{d+1}, \ldots, S_N) \in \mathcal{S}} \int_{\Delta_1^{M_N}} \mathbb{1}\{c^*_{i+1|\bar{S}} < f(x^*_{|\bar{S}}, \delta^{(j)}) \leq c^*_{i\,|\bar{S}}, \\
&\qquad\qquad j \in S_i,\ i = d+1, \ldots, N\} \mathbb{P}^{M_N}\{\mathrm{d}\boldsymbol{\delta}_1^{M_N}\} \\
&= \frac{1}{|\mathcal{S}|} \int_{\Delta_1^{M_N}} \sum_{(\bar{S}, S_{d+1}, \ldots, S_N) \in \mathcal{S}} \mathbb{1}\{c^*_{i+1|\bar{S}} < f(x^*_{|\bar{S}}, \delta^{(j)}) \leq c^*_{i\,|\bar{S}}, \\
&\qquad\qquad j \in S_i,\ i = d+1, \ldots, N\} \mathbb{P}^{M_N}\{\mathrm{d}\boldsymbol{\delta}_1^{M_N}\}.
\end{aligned}$$

For a fixed sample $\boldsymbol{\delta}_1^{M_N}$ the integrand

$$\sum_{(\bar{S}, S_{d+1}, \ldots, S_N) \in \mathcal{S}} \mathbb{1}\{c^*_{i+1|\bar{S}} < f(x^*_{|\bar{S}}, \delta^{(j)}) \leq c^*_{i\,|\bar{S}},\ j \in S_i,\ i = d+1, \ldots, N\}$$

counts the number of partitions of the indexes of given uncertainty realizations $\delta^{(1)}$, $\delta^{(2)}, \ldots, \delta^{(M_N)}$ into sets $\bar{S}, S_{d+1}, \ldots, S_N$, such that the costs associated with the realizations with index in $S_{d+1}, \ldots, S_N$ satisfy

$$c^*_{i+1|\bar{S}} < f(x^*_{|\bar{S}}, \delta^{(j)}) \leq c^*_{i|\bar{S}}, \quad j \in S_i, \quad i = d+1, \ldots, N. \tag{19}$$

It is a fact that such a number is almost surely equal to 1, as is formally stated in the next proposition, whose proof is postponed to section 5.1.1.

PROPOSITION 1.

$$\sum_{(\bar{S}, S_{d+1}, \ldots, S_N) \in \mathcal{S}} \mathbb{1}\{c^*_{i+1|\bar{S}} < f(x^*_{|\bar{S}}, \delta^{(j)}) \leq c^*_{i|\bar{S}},\ j \in S_i,\ i = d+1, \ldots, N\} = 1 \ \textit{almost surely.}$$

Thanks to Proposition 1, we have

$$\mathbb{E}[D_{d+1}^{k_{d+1}} \cdots D_N^{k_N}] = \frac{1}{|\mathcal{S}|} = \frac{1}{\binom{M_N}{N, k_{d+1}, \ldots, k_N}}, \tag{20}$$

where the denominator in the last expression is the multinomial coefficient

$$\binom{M_N}{N, k_{d+1}, \ldots, k_N} = \prod_{i=0}^{N-d-1} \binom{M_{N-i}}{k_{N-i}} = \frac{M_N!}{N! k_{d+1}! \cdots k_N!}.$$

Note that (20) holds true for every value of $k_{d+1}, \ldots, k_N$ so that (20) provides all the multivariate moments of $D_{d+1}, \ldots, D_N$. Hence, the joint distribution function of $D_{d+1}, \ldots, D_N$ remains uniquely determined [14]. In particular, by integration one can check that the density of the Dirichlet distribution,

$$p_D(x_{d+1}, x_{d+2}, \ldots, x_N) = \frac{N!}{d!} \left(1 - \sum_{i=d+1}^{N} x_i\right)^d \mathbb{1}\left\{\sum_{i=d+1}^{N} x_i \le 1, \quad 0 \le x_i \le 1\right\}, \tag{21}$$

satisfies the moment problem posed by (20), so that the conclusion is drawn that (21) is the density of $D_{d+1}, \ldots, D_N$.

Go back now to (17). Using this transformation we obtain the joint density p_R of $R_{d+1}, \ldots, R_N$ as follows:

$$\begin{aligned} &p_R(r_{d+1}, r_{d+2}, \ldots, r_N) \\ &= p_D(r_{d+2} - r_{d+1}, r_{d+3} - r_{d+2}, \ldots, r_N - r_{N-1}, 1 - r_N) \\ &= \frac{N!}{d!} r_{d+1}^d \mathbb{1}\{0 \le r_{d+1} \le r_{d+2} \le \cdots \le r_N \le 1\}, \end{aligned} \tag{22}$$

and (4) follows by integrating (22).

5.1.1. Proof of Proposition 1. Consider the optimization problem with all the uncertainty realizations $\delta^{(1)}, \ldots, \delta^{(N)}$, $\delta^{(N+1)}, \ldots, \delta^{(M_N)}$ in place,

$$\begin{aligned} \mathrm{EPI}_{M_N} : &\min_{c \in \mathbb{R}, x \in \mathcal{X} \subseteq \mathbb{R}^d} c \\ &\text{subject to } f(x, \delta^{(i)}) \le c, \ i = 1, \ldots, M_N, \end{aligned} \tag{23}$$

and let $(\tilde{x}, \tilde{c}_1)$ be the optimal solution. Moreover, let $\tilde{c}_k = \max\{c \in \mathbb{R} : f(\tilde{x}, \delta^{(i)}) \ge c$ for a choice of k indexes i among $\{1, \ldots, M_N\}\}$, $k = 1, \ldots, M_N$. Clearly, $\tilde{c}_k \le \tilde{c}_{k'}$ when $k > k'$. The nondegenerate assumption guarantees that the following strict ordering holds true almost surely:

$$\tilde{c}_{d+1} > \tilde{c}_{d+2} > \cdots > \tilde{c}_{M_N}.$$

In order for (19) to hold, observe that $\bar{S}$ must be such that $(\tilde{x}, \tilde{c}_1)$, the optimal solution to (23), coincides with $(x^*_{|\bar{S}}, c^*_{1|\bar{S}})$, the optimal solution computed with the uncertainty realizations in $\bar{S}$ only. Indeed, if this were not the case, there would be a $\delta^{(j)}$ in one of the sets $S_{d+1}, \ldots, S_N$ such that $f(x^*_{|\bar{S}}, \delta^{(j)}) > c^*_{1|\bar{S}}$. But then, by definition of $c^*_{d+1|\bar{S}}$, this would entail that $f(x^*_{|\bar{S}}, \delta^{(j)}) > c^*_{d+1|\bar{S}}$, which is in contrast with (19). Once the fact that $\tilde{x} = x^*_{|\bar{S}}$ has been established, the thesis easily follows. Indeed, $\bar{S}$ must

contain the indexes of the first $d+1$ functions $f(\tilde{x}, \delta^{(i)})$ counted starting from the top; S_{d+1} is the set of the indexes of the next k_{d+1} functions; the index of the function that comes immediately after must belong to $\bar{S}$; S_{d+2} is the set of the indexes of the next k_{d+2} functions, and so on. This construction determines the only possible partition of $\{1, \ldots, M_N\}$ in the subsets $\bar{S}, S_{d+1}, \ldots, S_N$.

5.2. Proof of Theorem 2. The reasoning is inspired to that used to prove the general bound (6) in [11], recalled in section 2, equation (6), of this paper. The idea consists in perturbing the sampled functions ("heating") so as to go back to the setting of Theorem 1 and then inferring the sought result via a limiting process ("cooling").

Heating. Given a real $\rho > 0$, let $H = [-\rho, \rho]$, and $\delta' = (\delta, h) \in \Delta'$, with $\Delta' = \Delta \times H$. Indicating with $\mathbb{U}$ the uniform measure on H, $\mathbb{P}' = \mathbb{P} \times \mathbb{U}$ defines a probability over Δ'. Moreover, for each $x \in \mathcal{X}$ and $\delta' = (\delta, h)$, let $f'(x, \delta') = f(x, \delta) + h$. The problem with N constraints obtained as realizations from $(\Delta', \mathbb{P}')$ is called the heated scenario problem:

$$\text{H-EPI}_N : \min_{c \in \mathbb{R}, x \in \mathcal{X} \subseteq \mathbb{R}^d} c \quad \text{subject to } f(x, \delta^{(i)}) + h^{(i)} \leq c, \quad i = 1, \ldots, N. \tag{24}$$

Since for any (x, c) we have that $\mathbb{P}'\{\delta' \in \Delta' : f'(x, \delta') = c\} = 0$, H-EPI$_N$ is nondegenerate and Theorem 1 applies. Hence, letting $(x'^*, c_1'^*)$ be the solution of H-EPI$_N$, $c_k'^*$, $k = 1, \ldots, N$, be the empirical costs, and $R'_k = \mathbb{P}'_{\delta'}\{f'(x'^*, \delta') > c_k'^*\}$, $k = 1, \ldots, N$, be the corresponding risks, the joint probability distribution function $\mathbb{P}'^N\{R'_{d+1} \leq \epsilon_{d+1}, \ldots, R'_N \leq \epsilon_N\}$ is computed according to (4), and it is given by $F_{d,N}(\epsilon_{d+1}, \epsilon_{d+2}, \ldots, \epsilon_N)$.

Convergence of the heated solution to the original solution by cooling. Fix $\delta^{(1)}, \delta^{(2)}, \ldots, \delta^{(N)}$, and compute the solution of EPI$_N$, (x^*, c_1^*), as well as the empirical costs $c_{d+1}^*, \ldots, c_N^*$. Let $\rho_n \downarrow 0$ be a sequence of reals monotonically decreasing to zero. For every n, pick N arbitrary numbers $h_n^{(1)}, \ldots, h_n^{(N)}$ from the interval $H_n = [-\rho_n, \rho_n]$, and let $(x'^*, c_1'^*)$ and $c_{d+1}'^*, \ldots, c_N'^*$ be the solution and the empirical costs of problem (24), where $h^{(i)} = h_n^{(i)}$. By mimicking [11], it is easy to show that the solution, as well as the heated costs of the heated problem, converges to the original solution, and to the empirical costs, of the original problem as $n \to \infty$. In formal terms,

$$\lim_{n \to \infty} \sup_{h_n^{(1)}, \ldots, h_n^{(N)} \in H_n} ||(x'^*, c_1'^*) - (x^*, c_1^*)|| = 0$$

and

$$\lim_{n \to \infty} \sup_{h_n^{(1)}, \ldots, h_n^{(N)} \in H_n} |c_k'^* - c_k^*| = 0, \ k = d+1, \ldots, N. \tag{25}$$

Derivation of (7). Fix a "bad" sample $\delta^{(1)}, \delta^{(2)}, \ldots, \delta^{(N)}$, i.e., a sample such that the condition $R_j > \epsilon_j$ is true for at least one $j \in \{d+1, \ldots, N\}$. As above, consider a sequence of heating parameters $\rho_n \downarrow 0$. In line with [11], it can be shown

that, thanks to (25), there exists a big enough $\bar{n}$ such that for all $n > \bar{n}$ and for every choice of $h_n^{(1)}, \ldots, h_n^{(N)}$, the heated sample $(\delta^{(1)}, h_n^{(1)}), \ldots, (\delta^{(N)}, h_n^{(N)})$ is such that $R'_j > \epsilon_j$, i.e., it is bad in the heated setting. Now, note that

$$\begin{aligned}
&(\mathbb{P} \times \mathbb{U})^N \{\exists j : R'_j > \epsilon_j\} \\
&\quad = \int_{\Delta^N} \int_{H_n^N} \mathbb{1}\{\exists j : R'_j > \epsilon_j\} \frac{\mathrm{d}\boldsymbol{h}_1^N}{(2\rho_n)^N} \, \mathbb{P}\{\mathrm{d}\boldsymbol{\delta}_1^N\} \\
&\quad \geq \int_{\Delta^N} \mathbb{1}\{\exists j : R_j > \epsilon_j\} \left[\int_{H_n^N} \mathbb{1}\{\exists j : R'_j > \epsilon_j\} \frac{\mathrm{d}\boldsymbol{h}_1^N}{(2\rho_n)^N} \right] \mathbb{P}\{\mathrm{d}\boldsymbol{\delta}_1^N\}.
\end{aligned}$$

The first indicator function limits the integration domain to samples in Δ^N that are bad. As previously noted, for every fixed bad sample in Δ^N the inner integral is equal to 1 for a sufficiently large n. Thus, by the dominated convergence theorem,

$$\begin{aligned}
&\lim_{n\to\infty} \int_{\Delta^N} \mathbb{1}\{\exists j : R_j > \epsilon_j\} \left[\int_{H_n^N} \mathbb{1}\{\exists j : R'_j > \epsilon_j\} \frac{\mathrm{d}\boldsymbol{h}_1^N}{(2\rho_n)^N} \right] \mathbb{P}\{\mathrm{d}\boldsymbol{\delta}_1^N\} \\
&\quad = \int_{\Delta^N} \mathbb{1}\{\exists j : R_j > \epsilon_j\} \mathbb{P}\{\mathrm{d}\boldsymbol{\delta}_1^N\}.
\end{aligned}$$

It follows that

$$\begin{aligned}
&\lim_{n\to\infty} (\mathbb{P} \times \mathbb{U})^N \{\exists j : R'_j > \epsilon_j\} \\
&\quad \geq \int_{\Delta^N} \mathbb{1}\{\exists j : R_j > \epsilon_j\} \mathbb{P}\{\mathrm{d}\boldsymbol{\delta}_1^N\} \\
&\quad = \mathbb{P}^N \{\exists j : R_j > \epsilon_j\} \\
&\quad = 1 - \mathbb{P}^N \{R_{d+1} \leq \epsilon_{d+1}, R_{d+2} \leq \epsilon_{d+2}, \ldots, R_N \leq \epsilon_N\},
\end{aligned}$$

from which

$$\begin{aligned}
&\mathbb{P}^N \{R_{d+1} \leq \epsilon_{d+1}, R_{d+2} \leq \epsilon_{d+2}, \ldots, R_N \leq \epsilon_N\} \\
&\quad \geq 1 - \lim_{n\to\infty} (\mathbb{P} \times \mathbb{U})^N \{\exists R'_j > \epsilon_j\} \\
&\quad = F_{d,N}(\epsilon_{d+1}, \epsilon_{d+2}, \ldots, \epsilon_N),
\end{aligned}$$

where the last equality follows from the discussion in the "heating" part of the proof. This establishes the validity of (7).

6. Summary and conclusions. In various application endeavors one relies on samples to optimize. In this paper, min-max sample-based optimization has been considered. After solving the optimization problem, one can evaluate the performance of the obtained solution corresponding to the samples that have been used in optimization. These performance values are called the empirical costs. Intuitively, the empirical costs carry useful information on the performance that can be expected when the solution is applied to a new situation, which is not in the set of initial samples. This idea is put on a solid mathematical ground in this paper. We have shown that precise limits to the probability of exceeding the empirical costs can be set. These results are tight in that they provide exact evaluations in situations precisely described in the paper, while they are also distribution-free, so that their application

does not require that prior knowledge on the underlying probability distribution of the samples is available to the user.

One important feature of the methods developed in this paper is that all evaluations are carried out without resorting to new samples of the uncertainty parameter in addition to those that are used for optimization. This fact is key to the applicability of the methods to contexts where the samples are observations, so that they represent a costly and limited resource.

REFERENCES

[1] T. Alamo, R. Tempo, and E. F. Camacho, *Randomized strategies for probabilistic solutions of uncertain feasibility and optimization problems*, IEEE Trans. Automat. Control, 54 (2009), pp. 2545–2559.

[2] T. Alamo, R. Tempo, A. Luque, and D. R. Ramirez, *Randomized methods for design of uncertain systems: Sample complexity and sequential algorithms*, Automatica, 52 (2015), pp. 160–172.

[3] C. Baudrit and D. Dubois, *Practical representations of incomplete probabilistic knowledge*, Comput. Statist. Data Anal., 51 (2006), pp. 86–108.

[4] G. Bayraksan and D. P. Morton, *Assessing solution quality in stochastic programs*, Math. Program., 108 (2006), pp. 495–514.

[5] G. Bayraksan and D. P. Morton, *Assessing solution quality in stochastic programs via sampling*, in Tutorials in Operations Research, M. R. Oskoorouchi, ed., INFORMS, Hannover, MD, 2009, pp. 102–122.

[6] D. Bertsimas and A. Thiele, *A data-driven approach to newsvendor problems*, Optimization Online (2006).

[7] D. Bertsimas and A. Thiele, *Robust and data-driven optimization: modern decision-making under uncertainty*, in Tutorials on Operations Research, M. R. Oskoorouchi, ed., INFORMS, Hannover, MD, 2006.

[8] G. C. Calafiore and M. C. Campi, *Uncertain convex programs: Randomized solutions and confidence levels*, Math. Program., 102 (2005), pp. 25–46.

[9] G. C. Calafiore and M. C. Campi, *The scenario approach to robust control design*, IEEE Trans. Automat. Control, 51 (2006), pp. 742–753.

[10] M. C. Campi, G. Calafiore, and S. Garatti, *Interval predictor models: Identification and reliability*, Automatica, 45 (2009), pp. 382–392.

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

[12] M. C. Campi and S. Garatti, *A Sampling-and-discarding approach to chance-constrained optimization: Feasibility and optimality*, J. Optim. Theory Appl., 148 (2011), pp. 257–280.

[13] M. C. Campi, S. Garatti, and M. Prandini, *The scenario approach for systems and control design*, Ann. Rev. Control, 33 (2009), pp. 149–157.

[14] H. Cramér and H. Wold, *Some theorems on distribution functions*, J. Lond. Math. Soc., s1-11 (1936), pp. 290–294.

[15] T. Homem de Mello, *Variable-sample methods for stochastic optimization*, ACM Trans. Model. Comput. Simul., 13 (2003), pp. 108–133.

[16] D. Dentcheva, *Optimization models with probabilistic constraints*, in Probabilistic and Randomized Methods for Design Under Uncertainty, G. Calafiore and F. Dabbene, eds., Springer-Verlag, London, 2006.

[17] D. Dentcheva, R. Henrion, and A. Ruszczyǹski, *Stability and sensitivity of optimization problems with first order stochastic dominance constraints*, SIAM J. Optim., 18 (2007), pp. 322–337.

[18] D. Dentcheva, B. Lai, and A. Ruszczyǹski, *Dual methods for probabilistic optimization problems*, Math. Methods Oper. Res. (ZOR), 60 (2004), pp. 331–346.

[19] D. Dentcheva, A. Prèkopa, and A. Ruszczyǹski, *Concavity and efficient points of discrete distributions in probabilistic programming*, Math. Program., 89 (2000), pp. 55–77.

[20] D. Dentcheva and A. Ruszczyǹski, *Stochastic optimization with dominance constraints*, SIAM J. Optim., 14 (2003), pp. 548–566.

[21] D. Dentcheva and A. Ruszczyǹski, *Semi-infinite probabilistic optimization: First-order stochastic dominance constraints*, Optimization, 53 (2004), pp. 583–601.

[22] S. Garatti and M. C. Campi, *L-infinity layers and the probability of false prediction*, in Proceedings of the 15th IFAC Symposium on System Identification, Saint Malo, France, 2009.

[23] S. Garatti and M. C. Campi, *Modulating robustness in control design: Principles and algorithms*, IEEE Control Syst. Mag., 33 (2013), pp. 36–51.
[24] A. Gouda and T. Szántai, *New sampling techniques for calculation of Dirichlet probabilities*, CEJOR Cent. Eur. J. Oper. Res., 12 (2004), pp. 389–403.
[25] A. Gouda and T. Szántai, *On numerical calculation of probabilities according to Dirichlet distribution*, Ann. Oper. Res., 177 (2010), pp. 185–200.
[26] H. N. Nagaraja H. A. David, *Order Statistics*, 3rd ed., Wiley, New York, 2003.
[27] H. L. Harter, *Minimax methods*, in Encyclopedia of Statistical Sciences, Vol. 4, Wiley, New York, 1982, pp. 514–516.
[28] R. Henrion and W. Römisch, *Metric regularity and quantitative stability in stochastic programs with probabilistic constraints*, Math. Program., 84 (1999), pp. 55–88.
[29] R. Henrion and W. Römisch, *Hölder and Lipschitz stability of solution sets in programs with probabilistic constraints*, Math. Program., 100 (2004), pp. 589–611.
[30] R. Henrion and C. Strugarek, *Convexity of chance constraints with independent random variables*, Comput. Optim. Appl., 41 (2008), pp. 263–276.
[31] R. Hochreiter, *An evolutionary computation approach to scenario-based risk-return portfolio optimization for general risk measures*, in Applications of Evolutionary Computing, M. Giacobini, ed., Lecture Notes in Comput. Sci. 4448, Springer, New York, 2007, pp. 199–207.
[32] R. Hochreiter, *Evolutionary stochastic portfolio optimization*, in Natural Computing in Computational Finance, A. Brabazon and M. O'Neill, eds., Stud. Comput. Intell. 100, Springer, Berlin, 2008, pp. 67–87.
[33] V. L. Levin, *Application of E. Helly's theorem to convex programming, problems of best approximation and related questions*, Sb. Math., 8 (1969), pp. 235–247.
[34] J. T. Linderoth, A. Shapiro, and S. Wright, *The empirical behavior of sampling methods for stochastic programming*, Ann. Oper. Res., 142 (2006), pp. 215–241.
[35] J. Luedtke, *New formulations for optimization under stochastic dominance constraints*, SIAM J. Optim., 19 (2008), pp. 1433–1450.
[36] J. Luedtke, *A branch-and-cut decomposition algorithm for solving chance-constrained mathematical programs with finite support*, Math. Program., 146 (2014), pp. 219–244.
[37] J. Luedtke and S. Ahmed, *A sample approximation approach for optimization with probabilistic constraints*, SIAM J. Optim., 19 (2008), pp. 674–699.
[38] J. Luedtke, S. Ahmed, and G. L. Nemhauser, *An integer programming approach for linear programs with probabilistic constraints.*, Math. Program., 122 (2010), pp. 247–272.
[39] W. K. Mak, D. P. Morton, and R. K. Wood, *Monte Carlo bounding techniques for determining solution quality in stochastic programs*, Oper. Res. Lett., 24 (199), pp. 47–56.
[40] H. Markowitz, *Portfolio selection*, J. Finance, 7 (1952), pp. 77–91.
[41] A. Mutapcic, S. J. Kim, and S. P. Boyd, *Robust Chebyshev FIR equalization*, in Proceedings of the 50th IEEE Global Communication Conference (GLOBECOM '07), Washington, DC, 2007.
[42] A. Nemirovski and A. Shapiro, *Convex approximations of chance constrained programs*, SIAM J. Optim., 17 (2006), pp. 969–996.
[43] A. Nemirovski and A. Shapiro, *Scenario approximations of chance constraints*, in Probabilistic and Randomized Methods for Design Under Uncertainty, G. Calafiore and F. Dabbene, eds., Springer-Verlag, London, 2006.
[44] B. K. Pagnoncelli, S. Ahmed, and A. Shapiro, *Sample average approximation method for chance constrained programming: theory and applications*, J. Optim. Theory Appl., 142 (2009), pp. 399–416.
[45] B. K. Pagnoncelli, D. Reich, and M. C. Campi, *Risk-return trade-off with the scenario approach in practice: A case study in portfolio selection*, J. Optim. Theory Appl., 155 (2012), pp. 707–722.
[46] A. Prèkopa, *Contributions to the theory of stochastic programming*, Math. Program., 4 (1973), pp. 202–221.
[47] A. Prèkopa, *Stochastic Programming*, Kluwer, Boston, 1995.
[48] A. Prèkopa, *Probabilistic programming*, in Stochastic Programming, A. Ruszczyǹski and A. Shapiro, eds., Handbooks Oper. Res. Management Sci. 10, Elsevier, London, 2003.
[49] A. Ruszczyǹski, *Probabilistic programming with discrete distributions and precedence constrained knapsack polyhedra*, Math. Program., 93 (2002), pp. 195–215.
[50] A. Shapiro, D. Dentcheva, and A. Ruszczyński, *Lectures on Stochastic Programming: Modeling and Theory*, MOS-SIAM Ser. Optim., Philadelphia, 2009.
[51] A. Thiele, *Robust stochastic programming with uncertain probabilities*, IMA J. Manag. Math., 19 (2008), pp. 289–321.

[52] P. Vayanos, D. Kuhn, and B. Rustem, *A constraint sampling approach for multistage robust optimization*, Automatica, 48 (2012), pp. 459–471.
[53] S. S. Wilks, *Mathematical Statistics*, Wiley, New York, 1962.
[54] S. Zymler, D. Kuhn, and B. Rustem, *Distributionally robust joint chance constraints with second-order moment information*, Math. Program., 137 (2013), pp. 167–198.