SIAM J. OPTIM.
Vol. 0, No. 0, pp. 1–26

ON CONDITIONAL RISK ASSESSMENTS IN SCENARIO OPTIMIZATION*

SIMONE GARATTI† AND MARCO C. CAMPI‡

Abstract. Scenario optimization is a data-driven technique in which one optimizes an objective function subject to a set of constraints, each given by a data point. In this article, we show that probabilistic claims on the violation of out-of-sample constraints (*risk*) conditional on the *complexity* of the solution (number of elements in the data set by which the solution can be reconstructed) are impossible if one does not use of extra information in addition to the data. While this article establishes this fundamental limitation, it also proves that a "mild" prior suffices to draw strong conditional conclusions. Precisely, a prior on the distribution of the complexity (which has support in a finite dimensional space) allows one to effectively bound the conditional distribution of the risk. Besides its intrinsic epistemological value, this result is useful for the conditional quantification of the risk of constraints violation in various application endeavors.

Key words. data-driven optimization, risk quantification, conditional risk, scenario approach, stochastic optimization, Bayesian statistics

MSC codes. 90C15, 90C25, 62C20

DOI. 10.1137/21M1451385

1. Introduction.

We consider optimization problems with uncertain convex constraints where $x \in \mathcal{X} \subseteq \mathbb{R}^d$ is the optimization variable, $c^T x$ is a cost to be minimized, and δ is a random outcome from a probability space $(\Delta, \mathcal{D}, \mathbb{P})$ that parameterizes the family of constraints $x \in \mathcal{X}_\delta$ (depending on the application domain, these constraints formalize a condition of correct classification in machine learning problems, saturation effects in control problems, etc.; see below for references pointing to various fields). All involved sets $\mathcal{X}$ and $\mathcal{X}_\delta$, $\delta \in \Delta$, are assumed to be *convex*. Given N ($N \geq d$) independent draws $\delta_1, \ldots, \delta_N$ from $(\Delta, \mathcal{D}, \mathbb{P})$, the scenario-based solution [7, 14] is obtained by solving the following optimization problem:

$$\begin{aligned} &\min_{x \in \mathcal{X}} c^T x \\ &\text{subject to } x \in \bigcap_{i=1,\ldots,N} \mathcal{X}_{\delta_i}. \end{aligned} \tag{1.1}$$

Problem (1.1) only involves finitely many constraints from Δ (in applications, Δ often has infinite cardinality) and the values $\delta_1, \ldots, \delta_N$ are observations called "scenarios." Hence, the *scenario optimization problem* (1.1) makes a selection of x that is optimal for a set of available observations.

Scenario problems of the form (1.1) have attracted much attention over the past decade; see, e.g., [1, 13, 15, 18, 19, 20, 28, 36, 43, 47, 52]. A central idea underlying these contributions is that enforcing the satisfaction of N constraints as is done in (1.1) provides robustness against most of the other constraints, those that correspond

* Received by the editors October 19, 2021; accepted for publication (in revised form) October 6, 2022; published electronically DATE
https://doi.org/10.1137/21M1451385

† Dipartimento di Elettronica, Informazione e Bioingegneria - Politecnico di Milano, piazza L. da Vinci 32, 20133 Milan, Italy (simone.garatti@polimi.it).

‡ Dipartimento di Ingegneria dell'Informazione - Università di Brescia, via Branze 38, 25123 Brescia, Italy (marco.campi@unibs.it).

to unseen realizations of $\delta \in \Delta$. Importantly, the robustness guarantees established in the aforementioned contributions do not depend on $\mathbb{P}$ so that the scenario approach is purely inductive and the user only needs to know the scenarios, while the underlying generative scheme for δ need not be explicitly described. Some of these theoretical results are reviewed below because they set the stage for study in the present contribution.

Notice that linearity of the cost function $c^T x$ in (1.1) introduces no loss of generality within a convex setup because problems with more general convex cost functions $f(x)$ can be rewritten in the form of (1.1) by an epigraphic reformulation; moreover, problems with an uncertain convex cost function can also be rewritten in the form of (1.1) and the interested reader is referred to [17] for a detailed exposition of these reformulations.

Throughout, we assume existence and uniqueness of the solution to program (1.1).

Assumption 1.1 (existence and uniqueness). For any $\delta_1, \ldots, \delta_N$, program (1.1) admits a solution. If more than one solution exists, it is assumed that a solution is singled out by a convex rule, that is, the tie is broken by minimizing an additional convex function $t_1(x)$, and, possibly, other convex functions $t_2(x)$, $t_3(x), \ldots$, if the tie still occurs. After breaking the tie, the solution is denoted by x_N^*.$\star$

An example of a tie-break function is the norm of x, $t_1(x) = \|x\|$. Another example is the lexicographic rule, which consists in minimizing the components of x in succession, i.e., $t_1(x) = x_1$, $t_2(x) = x_2, \ldots,$

Scenario programs have proven useful in multiple application domains, including control and systems design [2, 8, 17, 22, 29, 32, 38, 44, 45, 46, 48], prediction [10, 23, 24, 25, 26, 27, 31, 34], quantitative finance [33, 39, 40, 41], and classification [9, 11, 21, 37]. See also [14] for a book-length presentation of the scenario approach. The reader interested in the interpretation of (1.1) in specific domains is referred to these references, while in this paper we concentrate on a theoretically oriented study on conditional risk assessments.

1.1. Revision of previous results. Introduce the following definition of probability of violation.

Definition 1.2 (probability of violation). *Given an $x \in \mathcal{X}$, the probability of violation* (*also called violation for short*) *of x is defined as*

$$V(x) = \mathbb{P}\{\delta \in \Delta : x \notin \mathcal{X}_\delta\}. \qquad \star$$

$V(x)$ quantifies the probability with which a new, randomly drawn, constraint is not satisfied by x, and it quantifies the level of robustness of x against constraint violation. Depending on the application, $V(x)$ is the probability of not meeting the control specifications, of providing an incorrect prediction, or of not obtaining the expected reward in an investment, and the reader is referred to the literature referenced above for a contextualization in specific setups. Correspondingly, in some literature on the scenario approach $V(x)$ is also called the "risk." For future use, we also note that $V(x)$ can be given an interpretation in terms of repeated experiments as follows. Consider the infinite product probability space $(\Delta^\infty, \mathcal{D}^\infty, \mathbb{P}^\infty)$. By the law of large numbers, one has[1]

$$\lim_{M\to\infty} \frac{1}{M} \sum_{j=1}^{M} \mathbf{1}_{x \notin \mathcal{X}_{\delta_{N+j}}} = V(x), \quad \mathbb{P}^\infty\text{-almost surely.} \tag{1.2}$$

[1] Writing $\mathcal{X}_{\delta_{N+j}}$, instead of $\mathcal{X}_{\delta_j}$, is relevant to the subsequent use of this formula with x_N^* (which depends on the first N scenarios) in place of x.

Hence, $V(x)$ is the long-term average of times in which x does not satisfy a sequence of independent constraints. This is relevant to situations in which a decision is applied many times, for example, a digital filter is applied to many signals or an investment is kept for many subsequent days, etc.

When the variable x that appears in $V(x)$ is replaced by the random vector x_N^* (this is the solution to problem (1.1) and is random because it depends on $\delta_1, \ldots, \delta_N$), one obtains a random variable $V(x_N^*)$. The distribution of $V(x_N^*)$ is important because it characterizes the robustness level achieved by the solution to (1.1). One fundamental result proven in [12] shows that the distribution of $V(x_N^*)$ is always (i.e., *independently of the type of constraints* $x \in \mathcal{X}_\delta$ *and of* $\mathbb{P}$) dominated by a Beta$(d, N-d+1)$ distribution according to the formula[2]

$$\mathbb{P}^N\{V(x_N^*) \leq \epsilon\} \geq 1 - \sum_{i=0}^{d-1} \binom{N}{i} \epsilon^i (1-\epsilon)^{N-i}. \tag{1.3}$$

Moreover, this result is tight because (1.3) holds with equality for a whole class of problems, named *fully supported* in [12]. The fact that (1.3) is valid irrespective of $\mathbb{P}$ makes this result widely usable in applications, where assuming that one knows $\mathbb{P}$ is often not realistic.

An inspection of how the result (1.3) has been proven in [12] shows the central role played by the following concept of support constraint.

DEFINITION 1.3 (support constraint). *A constraint* $x \in \mathcal{X}_{\delta_{\bar{i}}}$ *of the scenario program* (1.1) *is called a support constraint if the scenario program obtained by removing this constraint, namely,*

$$\begin{aligned} &\min_{x \in \mathcal{X}} c^T x \\ &\text{subject to } x \in \bigcap_{i=1,\ldots,\bar{i}-1,\bar{i}+1,\ldots,N} \mathcal{X}_{\delta_i}. \end{aligned}$$

has a solution (possibly singled out by the same tie-break rule as for the initial program (1.1)*) different from* x_N^*. ⋆

In words, a support constraint is a constraint that is strictly needed to obtain the solution. In the following, we make a mild assumption of nondegeneracy (borrowed from [15, Assumption 2]), which requires that the support constraints are also sufficient to determine the solution.

Assumption 1.4 (nondegeneracy). With probability 1, the scenario program that contains only the support constraints

$$\begin{aligned} &\min_{x \in \mathcal{X}} c^T x \\ &\text{subject to } x \in \bigcap_{\substack{\text{support} \\ \text{constraints}}} \mathcal{X}_{\delta_{\bar{i}}} \end{aligned}$$

has the same solution x_N^* (possibly singled out by the same tie-break rule as for the initial program (1.1)) as program (1.1). ⋆

This condition rules out situations where the boundaries of various constraints group together anomalously so that if one of them is removed in isolation, then the

[2]In this formula, $\mathbb{P}^N$ is the probability according to which $\delta_1, \ldots, \delta_N$ is drawn and it is a product probability owing to independence of the scenarios.

solution does not change (and therefore this constraint is not of support) but a simultaneous removal of all the constraints that are not of support gives a new solution.[3] Since the solution can be reconstructed from the support constraints, we can think that they "represent" the solution, and the cardinality of the support constraint set can be interpreted as the *complexity* of the solution [30]. In [7], it is shown that the number of support constraints never exceeds the number of optimization variables d. This result is key in the analysis of [12] to bound the distribution of $V(x_N^*)$ as shown in (1.3). On the other hand, it is not rare that when one a posteriori evaluates the number of support constraints after that x_N^* has been computed, fewer support constraints are found than there are optimization variables, that is, the complexity is smaller than d. This is especially true for optimization problems in high dimensions where the gap between the number of support constraints and d is often large; see, e.g., [15, 19, 47, 50, 51]. When this happens it comes spontaneous to ask whether a better result than (1.3) applies, and particularly whether in (1.3) d one can substitute d with the actual number of support constraints h to obtain a valid bound *conditionally on seeing h support constraints*. In formal terms, this is written as

$$\mathbb{P}^N\{V(x_N^*) \leq \epsilon | s_N^* = h\} \geq 1 - \sum_{i=0}^{h-1} \binom{N}{i} \epsilon^i (1-\epsilon)^{N-i}, \tag{1.4}$$

where s_N^* is a random variable that returns the number of support constraints. One important observation that motivates this study is that (1.4) is incorrect and in fact an extreme, negative, result holds: no meaningful conditional results can be established at the level of generality at which this discussion is made here. The interpretation is that *talking after the actual complexity of the solution has been seen is too late a stage to make any meaningful claims*. We feel it is advisable to make this assertion compelling by way of an example.

Example 1.5. Let $x \in \mathbb{R}^2$, $c^T x = x_2$, and assume that $\mathcal{X}_\delta$ is either V-shaped or U-shaped as depicted in Figure 1.

Precisely, with probability $1-q$, set $\mathcal{X}_\delta$ is V-shaped with a vertex uniformly drawn from a horizontal segment, while, with probability q, $\mathcal{X}_\delta$ is U-shaped with a vertex uniformly distributed on a vertical segment; V-shaped constraints are all above U-shaped constraints. With N constraints, $s_N^* = 1$ happens if and only if either all constraints are U-shaped (in which case the support constraint is the highest among the U-shaped constraints) or all but one are U-shaped (in which case the support constraint is the only V-shaped constraint). In both cases, all V-shaped constraints (with the exception of the support V-shaped constraint in the second case) are violated, so that $V(x_N^*) \geq 1-q$ (the probability of V-shaped constraints). Thus, for any $\epsilon < 1-q$, we have

$$\mathbb{P}^N\{V(x_N^*) \leq \epsilon | s_N^* = 1\} = 0. \tag{1.5}$$

This contradicts (1.4) because the Beta distribution on the right-hand side has a support that covers all $[0,1]$; see Figure 2.

[3] Assumption 1.4 is often satisfied when δ itself does not accumulate (for example, when it has a density). Moreover, section 5.1 of [30] also introduces a nondegeneracy condition for the sole sample of scenarios at hand (as opposed to Assumption 1.4 that is required to hold with probability 1), and the theory therein might possibly be used to lessen our nondegeneracy condition here. In this paper, we have preferred to stay in the mainstream of Assumption 1.4 to avoid additional mathematical cluttering.

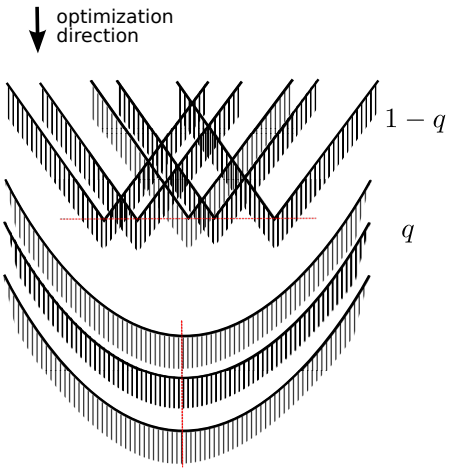


FIG. 1. *V-shaped and U-shaped constraints for Example* 1.5.

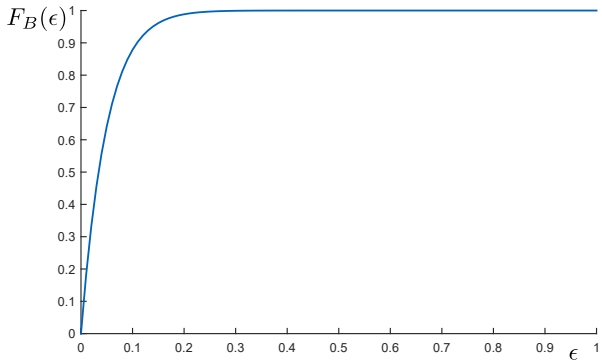


FIG. 2. *Cumulative distribution function of a* $\mathrm{Beta}(h, N-h+1)$ *distribution with* $N=20$ *and* $h=1$.

Moreover, letting $q \to 0$, (1.5) shows that $V(x_N^*)$ has conditional probability 0 of being less than a value ϵ that can be made close to 1 at will, an arbitrarily bad result.

Before closing, an additional fact is worth mentioning to shed light on how this example relates to the wait-and-judge theory of paper [15]. A simple computation shows that $s_N^* = 1$ is attained with probability $q^N + Nq^{N-1}(1-q)$, so that when $q \simeq 0$ (corresponding to high violation conditional on $s_N^* = 1$) the probability that $s_N^* = 1$ tends to zero extremely fast with N. By inspecting [15], one sees that this fact is general: the main result of [15] proves that simultaneously having small values of s_N^* and high violation occurs with very low probability, that is, $\mathbb{P}^N\{\text{high } V(x_N^*) \wedge \text{small } s_N^*\} = \text{low value}$. On the other hand, $\mathbb{P}^N\{\text{high } V(x_N^*) \wedge \text{small } s_N^*\} = \mathbb{P}^N\{\text{high } V(x_N^*) \mid \text{small } s_N^*\} \cdot \mathbb{P}^N\{\text{small } s_N^*\}$, which shows that if the probability of high violation conditional on seeing a small value of s_N^* is high, then it must be that observing that small value of s_N^* has low probability. $\star$

Remark 1.6 (a comparison between unconditional and conditional bounds). We feel we owe the reader an additional word to better clarify the difference between unconditional results, as per (1.3), and conditional results, and position them in relation to how they can be used in practice.

A conditional claim refers to the situation at hand: "Since I have seen this complexity, I am in a position to specifically judge the risk in the following way…." This clearly matters because it points to the specific situation as it has unfolded after the data points have been gathered. In this context, it is of the greatest importance that Example 1.5 rules out any possibility of making sensible claims on the conditional risk that hold without extra assumptions. This sets intrinsic limits separating what can be done from what cannot be done: it's not a matter of how the theory has been developed; it is an inherent limitation that exists per se, beyond the language in which it is formulated. In the remainder of this article, we aim at showing that suitable extra assumptions can be assigned in the form of a prior, thus embracing a Bayesian framework. Along this path, we shall see that mild priors suffice to draw strong conclusions and that these conclusions are little sensitive to the prior.

On the other hand, one should not underestimate the (practical) importance of unconditional claims. Unconditional claims capture how well a scenario program performs on average over potential data sets, and independently of how data sets are generated. Hence, if we interpret the scenario program as an algorithm that maps data sets into decisions, an unconditional judgment quantifies how good the algorithm is, and this may drive us in ranking algorithms (this may also be relevant to selecting hyperparameters). Therefore, conditional claims are in use to judge solutions, while unconditional claims allow one to evaluate the way solutions are obtained. As a case in point, consider the problem of classifying linearly separable points in $\mathbb{R}^{d-1}$ and suppose we aim at using a support vector machine (SVM) to build our classifiers. Since the SVM algorithm is convex, with d optimization variables if points are in $\mathbb{R}^{d-1}$, the Beta result in (1.3) can be applied and, depending on how small d is compared to the data set size N, we can draw conclusions on how well we expect this algorithm to perform. The reader is referred to [16] for details on SVM, as well as other methods used in machine learning problems.

While we are a bit afraid that the last point we want to make in this remark might slightly, and hopefully temporarily, mystify the reader, in the hope that a deeper order will come out of a possible initial hesitation we also feel it advisable to precisely position the results from paper [15] within our present discussion. In [15], it is shown that the risk can be evaluated from the complexity at a certain level of confidence. The essential difference between [15] and the present article is that the level of confidence in [15] is not conditional,it is in total probability. Therefore, the interpretation of the confidence in [15] is exactly the same as that in (1.3), while the difference that puts aside the result of [15] from (1.3) is that the risk evaluation in [15] is based on a statistic of the data (the complexity). As paper [15] proves, evaluating the unconditional confidence for risk evaluations that depend on the complexity is possible distribution-free; nonetheless, Example 1.5 shows that it is not possible to use [15] to derive conditional judgments valid for the complexity at hand without resorting to extra assumptions. $\star$

1.2. The result of this paper. Example 1.5 shows that no useful conditional results may exist in the general framework of [12]. In this paper, we move to a new setting where the information carried by data is complemented with prior knowledge. The ultimate goal we aim at is showing that strong conditional results can be established under mild priors.

To set the mathematical stage, consider an additional probability space $(\Theta, \mathcal{Q}, \pi)$ and let $\mathbb{P}_\vartheta$, $\vartheta \in \Theta$, be a transition probability function (see, e.g., Definition 1 in

Appendix 6 of [3]; the same notion is also known under the name "Markov kernel"—see, e.g., [4]) on $\Theta \times \mathcal{D}$ (recall that $\mathcal{D}$ is the σ-algebra over Δ), that is,

i. $\forall \vartheta \in \Theta$, the map $D \to \mathbb{P}_\vartheta(D)$ is a probability distribution;
ii. $\forall D \in \mathcal{D}$, the map $\vartheta \to \mathbb{P}_\vartheta(D)$ is $\mathcal{Q}$-measurable.

The interpretation is that, for any given $\vartheta \in \Theta$, $\mathbb{P}_\vartheta$ operates as $\mathbb{P}$ in the previous section, that is, it defines a mechanism by which constraints are generated. Our uncertainty about the mechanism for constraint generation is then modeled by assuming that ϑ distributes according to π, where π is *our prior over the constraint generation mechanism*.

Consider now the probability space $(\Delta^N \times \Theta, \mathcal{D}^N \otimes \mathcal{Q}, \mathrm{P})$, where, for any $E \in \mathcal{D}^N \otimes \mathcal{Q}$, $\mathrm{P}(E)$ is defined by $\mathrm{P}(E) = \int_\Theta \mathbb{P}_\vartheta^N(E_{/\vartheta})\pi(\mathrm{d}\vartheta)$, where $E_{/\vartheta}$ is the set in Δ^N given by E with the coordinate in Θ kept fixed at value ϑ. This is the space that hosts a ϑ along with a sample of N independent constraints obtained from a problem where δ distributes according to $\mathbb{P}_\vartheta$. In this context, x_N^* is the solution to (1.1) when $\delta_1, \ldots, \delta_N$ is an independent and identically distributed sample from $(\Delta, \mathcal{D}, \mathbb{P}_\vartheta)$ and ϑ is a realization from $(\Theta, \mathcal{Q}, \pi)$, and s_N^* is the corresponding complexity. x_N^* and s_N^* are random quantities defined over $(\Delta^N \times \Theta, \mathcal{D}^N \otimes \mathcal{Q}, \mathrm{P})$. The definition of violation for a given ϑ is $V_\vartheta(x) = \mathbb{P}_\vartheta\{\delta \in \Delta : x \notin \mathcal{X}_\delta\}$, and, for each ϑ, it admits the same long- term average interpretation as $V(x)$ in (1.2). Our objective is to evaluate the distribution of $V_\vartheta(x_N^*)$ conditionally on seeing h support constraints,[4]

$$F_V(\epsilon | s_N^* = h) = \mathrm{P}\{V_\vartheta(x_N^*) \leq \epsilon | s_N^* = h\}.$$

While this conditional distribution can in principle be calculated once π is given, the actual computation can be very difficult. Moreover, what is of paramount importance in our perspective is that, in practice, knowledge of π can be very hard to obtain (note that π sets a prior distribution on the constraint generation mechanism and, upon reflection, the reader will see that coming to a suitable formulation of a π is a formidable task in virtually any meaningful application; see also section 5). The good news is that in section 2 we shall show that tight and useful bounds for $F_V(\epsilon | s_N^* = h)$ can be drawn based on limited knowledge of π.[5] Precisely, let π' be the prior induced by π on the distributions of s_N^*. Note that π' is a much simpler object than π: in fact, s_N^* takes value in $\{0, 1, \ldots, d\}$ and, for a given ϑ, the distribution of s_N^* is a $(d+1)$-dimensional vector whose components assign the probability with which it happens that $s_N^* = 0$ or $s_N^* = 1$ or $\ldots$ or $s_N^* = d$. Since the sum of these probabilities is 1, the vector belongs to the simplex in $d+1$ dimensions and π' is just a finite-dimensional distribution over this simplex. One first useful result we shall show is that $F_V(\epsilon | s_N^* = h)$ can be effectively bounded based on π' only. Further, we show that the conditional distribution $F_V(\epsilon | s_N^* = h)$ is little sensitive to π' so that various individuals carrying different a priori beliefs draw similar conclusions after observing

[4]It is important to note that while distribution F_V is with respect to P (which involves both $\mathbb{P}_\vartheta$ and π), $V_\vartheta(x)$ only involves $\mathbb{P}_\vartheta$. This is because we are interested in quantifying the roustness level with respect to the optimization problem at hand as given by ϑ. To understand the importance of this point, suppose, for example, that $s_N^* = h$ holds with probability 1 and that $F_V(0.1 | s_N^* = h) = 1$. Then, it holds that $\mathrm{P}\{V_\vartheta(x_N^*) \leq 0.1\} = 1$, which gives that, for almost all ϑ's, $\mathbb{P}_\vartheta\{\delta \in \Delta : x_N^* \notin \mathcal{X}_\delta\} \leq 0.1$ holds $\mathbb{P}_\vartheta^N$-almost surely. Hence, independently of the optimization problem at hand as given by ϑ, one concludes that the long-term average of times in which x_N^* does not satisfy a sequence of independent constraints does not exceed 0.1. Such a strong conclusion might not have been drawn if the definition of violation would have mixed various values of ϑ, e.g., by the rule $V(x) = \int_\Theta \mathbb{P}_\vartheta\{\delta \in \Delta : x \notin \mathcal{X}_\delta\}\pi(\mathrm{d}\vartheta)$.

[5]In this connection, this study can be seen as being in the vein of robust Bayesian methods; see, e.g., [5, 6, 35, 49].

the same $s_N^* = h$. These results are formalized and precisely stated in the next section.

2. Conditional distribution of the violation. Assume that $\mathrm{P}\{s_N^* = h\} > 0$; if not, over $\{s_N^* = h\}$ the conditional distribution $F_V(\epsilon|s_N^* = h)$ can be defined arbitrarily. Write

$$\begin{aligned} F_V(\epsilon|s_N^* = h) & \\ &= \mathrm{P}\{V_\vartheta(x_N^*) \le \epsilon | s_N^* = h\} \\ &= 1 - \mathrm{P}\{V_\vartheta(x_N^*) > \epsilon | s_N^* = h\} \\ &= 1 - \frac{\mathrm{P}\{V_\vartheta(x_N^*) > \epsilon \wedge s_N^* = h\}}{\mathrm{P}\{s_N^* = h\}} \\ &= 1 - \frac{\int_\Theta \mathbb{P}_\vartheta^N\{V_\vartheta(x_N^*) > \epsilon \wedge s_N^* = h\}\pi(\mathrm{d}\vartheta)}{\int_\Theta \mathbb{P}_\vartheta^N\{s_N^* = h\}\pi(\mathrm{d}\vartheta)}, \end{aligned} \tag{2.1}$$

We shall obtain tight evaluations for (2.1) by bounding the integrand at the numerator of (2.1) depending on whether $\mathbb{P}_\vartheta^N\{s_N^* = h\}$ is very small or not. Clearly, it holds that

$$\mathbb{P}_\vartheta^N\{V_\vartheta(x_N^*) > \epsilon \wedge s_N^* = h\} \le \mathbb{P}_\vartheta^N\{s_N^* = h\}, \tag{2.2}$$

which is used when $\mathbb{P}_\vartheta^N\{s_N^* = h\}$ is very small. Moreover, letting for brevity

$$b_h(\epsilon) := \sup_{\vartheta\in\Theta} \mathbb{P}_\vartheta^N\{V_\vartheta(x_N^*) > \epsilon \wedge s_N^* = h\},$$

we can bound $\mathbb{P}_\vartheta^N\{V_\vartheta(x_N^*) > \epsilon \wedge s_N^* = h\}$ in (2.1) with $b_h(\epsilon)$ when $\mathbb{P}_\vartheta^N\{s_N^* = h\}$ is not very small. Substituting (2.2) and $b_h(\epsilon)$ in (2.1) gives

$$\begin{aligned} & F_V(\epsilon|s_N^* = h) \\ & \quad \ge 1 - \frac{\int_{\{\vartheta:\, \mathbb{P}_\vartheta^N\{s_N^*=h\} \le b_h(\epsilon)\}} \mathbb{P}_\vartheta^N\{s_N^* = h\}\pi(\mathrm{d}\vartheta) + \int_{\{\vartheta:\, \mathbb{P}_\vartheta^N\{s_N^*=h\} > b_h(\epsilon)\}} b_h(\epsilon)\pi(\mathrm{d}\vartheta)}{\int_\Theta \mathbb{P}_\vartheta^N\{s_N^* = h\}\pi(\mathrm{d}\vartheta)} \\ & \quad = 1 - \frac{\int_{\{\vartheta:\, \mathbb{P}_\vartheta^N\{s_N^*=h\} \le b_h(\epsilon)\}} \mathbb{P}_\vartheta^N\{s_N^* = h\}\pi(\mathrm{d}\vartheta) + b_h(\epsilon)\pi(\vartheta:\, \mathbb{P}_\vartheta^N\{s_N^* = h\} > b_h(\epsilon))}{\int_\Theta \mathbb{P}_\vartheta^N\{s_N^* = h\}\pi(\mathrm{d}\vartheta)}. \end{aligned} \tag{2.3}$$

Next, define $p_k := \mathbb{P}_\vartheta^N\{s_N^* = k\}$, $k = 0, 1, \ldots, d$, and let $\boldsymbol{p} = (p_0, p_1, \ldots, p_d) \in S$, where S is the simplex in $\mathbb{R}^{d+1}$ (i.e., $\sum_{k=0}^d p_k = 1$, $p_k \ge 0$, $k = 0, 1, \ldots, d$). $\boldsymbol{p}$ is a random variable since it depends on ϑ (we do not indicate explicitly the dependence on ϑ for notational convenience). Letting π' be the probability distribution of $\boldsymbol{p}$ induced by π, (2.3) now gives the following theorem.

THEOREM 2.1. *Under Assumption* 1.1*, it holds that*

$$F_V(\epsilon|s_N^* = h) \ge 1 - \frac{\int_{\{p_h \le b_h(\epsilon)\}} p_h \pi'_h(\mathrm{d}p_h) + b_h(\epsilon)\pi'_h\{p_h > b_h(\epsilon)\}}{\int_{[0,1]} p_h \pi'_h(\mathrm{d}p_h)}, \tag{2.4}$$

where π'_h is the marginal of π' on the component p_h. ⋆

Equation (2.4) is the fundamental relation we shall use to evaluate $F_V(\epsilon|s_N^* = h)$. Notice that $F_V(\epsilon|s_N^* = h)$ is close to 1 when the second term in the right-hand side is close to zero. Figure 3 provides a visual interpretation of this second term.

As we shall see, $b_h(\epsilon)$ goes rapidly to zero for values of ϵ above h/N, thus making the right-hand side of (2.4) close to 1 for ϵ marginally bigger than h/N whenever

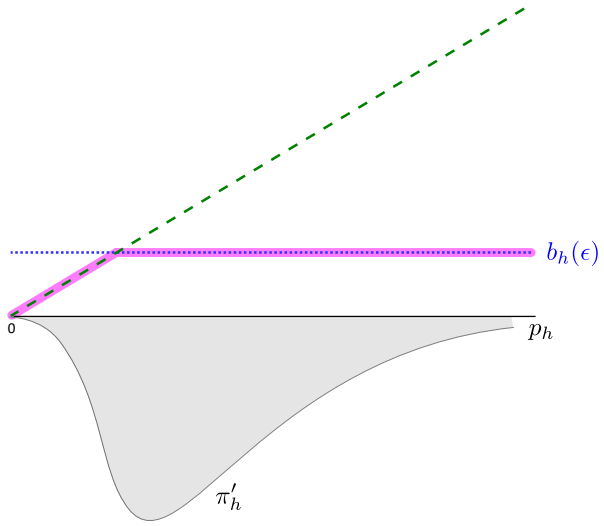


FIG. 3. *Visual interpretation of the second term in the right-hand side of* (2.4)*: the numerator is obtained by integrating the pink solid line with respect to the probability distribution* π'_h*, while the denominator, obtained by integration of* p_h *(green dashed line) with respect to* π'_h*, is the expected value of* p_h*.*

the prior π'_h does not concentrate about the zero value. Note also that, when π'_h concentrates in 1, the right-hand side of (2.4) becomes exactly $1 - b_h(\epsilon)$.

To proceed, we need to find a suitable expression for $b_h(\epsilon)$. While finding an exact expression is hard, Theorem 2.1 retains its validity if $b_h(\epsilon)$ is overestimated (i.e., $b_h(\epsilon)$ is replaced by a quantity that is equal to or bigger than $b_h(\epsilon)$—inspect the easy proof of Theorem 2.1 to draw this conclusion). This is the approach that is pursued in the following, in a somewhat articulated manner. First, in section 2.1 we provide a quick approach to upper bound $b_h(\epsilon)$ and present an example (Example 2.5) that helps gain insight into (2.4). As an alternative, section 3 presents a refinement of the evaluation of $b_h(\epsilon)$ given in section 2.1.[6]

Remark 2.2 (about the prior π'). As discussed at the end of section 1.2, the prior that is in use in this paper is π' (refer to (2.4)), which assigns a probability for the distribution of the complexity s^*_N. As previously noticed, π' is a relatively simple mathematical object as compared to a prior (called π in this paper) over the generation mechanism of the constraints. Nonetheless, by adopting a practical viewpoint one may wonder where this prior comes from in a given applied problem. While this question does contain subtle and, as we believe, unsettled philosophical issues, we do not want to go in this article to this foundational level.[7] Rather, we want to remark that the prior π' embodies our belief, however obtained, on how complexity distributes. At times, we have had exposure to the same (or a similar) problem in the past. This provides us with grounds for constructing the empirical frequency

[6]We advise the reader that section 3 is technically complex and can be skipped at first reading without loss of continuity.

[7]The justification of using a prior is a common aspect to all Bayesian statistics, and the reader is referred to any advanced textbook for a discussion of this issue. On our end, we maintain that the use and concept of prior demands closer scrutiny; however, we shall not be engaged in discussing our philosophical positions in this publication, which is only geared toward establishing precise, and compelling, mathematical results in the context of decision making.

with which various values of the complexity have been encountered. However, guided also by the idea that the empirical frequency is an imprecise descriptor because it is subject to stochastic variability, we may set out to build a distribution π' that spreads around the empirical frequency that has been found. The Dirichlet distribution described in section 4 is a perfect instrument to capture this situation and the reader is referred to that section for more discussion. In other cases, we cannot boast any real previous experience, in which situation a "principle of indifference" has been postulated by some as a proper way to set a prior: according to this principle, each case is deemed equally probable, which corresponds to a flat prior in our context. We notice that also a flat prior is a special case of the Dirichlet distribution of section 4. $\star$

2.1. An easy formula for $b_h(\epsilon)$. In this section, we establish the validity of a first bound for $b_h(\epsilon)$ that is easy to obtain.

THEOREM 2.3. *Under Assumption* 1.1 *and assuming that Assumption* 1.4 *holds true for any probability* $\mathbb{P}_\vartheta$, $\vartheta \in \Theta$, *it holds that* [8]

$$b_h(\epsilon) \leq \min\left\{ \binom{N}{h}(1-\epsilon)^{N-h}, 1 \right\}. \tag{2.5}$$

***Proof of Theorem* 2.3.** Suppose that there are h support constraints and these are the first h constraints: $x \in \mathcal{X}_{\delta_1}, \ldots, x \in \mathcal{X}_{\delta_h}$. Then, all other $N-h$ constraints must be satisfied by the solution x^*_N. If $V_\vartheta(x^*_N) > \epsilon$, this means that $\delta_{h+1}, \ldots, \delta_N$ must belong to an event whose probability is no more than $1-\epsilon$, and this happens with a probability that is no more than $(1-\epsilon)^{N-h}$ since the constraints are independent of each other. Hence, $\mathbb{P}^N_\vartheta\{V_\vartheta(x^*_N) > \epsilon \wedge s^*_N = h \wedge$ the support constraints are the first h constraints$\} \leq (1-\epsilon)^{N-h}$. A similar argument applies to all other choices of h support constraints. Hence, summing over all possible choices of the h support constraints from a total of N constraints (which gives $\binom{N}{h}$ choices) one obtains

$$\begin{aligned}
&\mathbb{P}^N_\vartheta\{V_\vartheta(x^*_N) > \epsilon \wedge s^*_N = h\} \\
&\quad = \sum_{i=1}^{\binom{N}{h}} \mathbb{P}^N_\vartheta\Big\{ V_\vartheta(x^*_N) > \epsilon \wedge s^*_N = h \wedge \text{the support} \\
&\qquad\qquad \text{constraints are the } i\text{th group of } h \text{ constraints} \Big\} \\
&\quad \leq \binom{N}{h}(1-\epsilon)^{N-h},
\end{aligned}$$

which holds true independently of ϑ. This establishes result (2.5). □

It is instructive to pause a second and reflect upon the meaning of (2.5). The function on the right-hand side of (2.5) is represented in Figure 4 for various values of h.

Letting $\bar{\epsilon}_h$ be the value of ϵ for which the function equals 0.5, it can be noted that, beyond $\bar{\epsilon}_h$, $b_h(\epsilon)$ rapidly saturates down to 0. Hence, by using the relation $\mathbb{P}^N_\vartheta\{V_\vartheta(x^*_N) > \epsilon \wedge s^*_N = h\} = \mathbb{P}^N_\vartheta\{V_\vartheta(x^*_N) > \epsilon \mid s^*_N = h\} \cdot \mathbb{P}^N_\vartheta\{s^*_N = h\}$, one draws the conclusion that, for a value of ϵ even moderately bigger than $\bar{\epsilon}_h$, either $V_\vartheta(x^*_N) > \epsilon$

[8] As already said, a tighter bound is established in section 3, which, however, requires a much more complicated theory than the one developed in this section. Bound (2.5) is introduced and used here because it allows for an easy understanding of the fundamental elements of the theory and, moreover, it provides satisfactory evaluations in many cases.

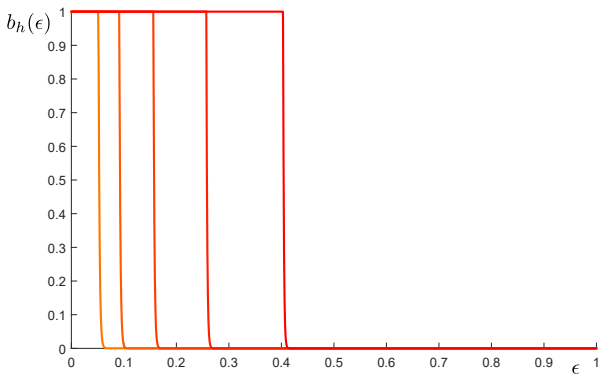


FIG. 4. *The upper bound to* $b_h(\epsilon)$ *in* (2.5) *for* $N = 500$ *and* $h = 5, 10, 20, 40, 80$. *As* h *grows, the function moves to the right.*

occurs very rarely conditionally on seeing $s_N^* = h$, or seeing $s_N^* = h$ must itself be a rare event. Referring back to Example 1.5, suppose that $q = 0.1$ and $N = 20$. For $h = 1$, the value of $\bar{\epsilon}_1$ can be calculated to be 0.18. In Example 1.5, the event $s_N^* = 1$ had probability equal to $0.1^{20} + 20 \cdot 0.1^{19} \cdot 0.9 = 1.81 \cdot 10^{-18}$, which is extremely low, and this was the circumstance that made it possible to have always a large violation, of at least 0.9, when $s_N^* = 1$.

Remark 2.4 (sensitivity to the prior). Referring again to Figure 4, pick a value of h and suppose for a moment that the transition from 1 to 0 of function $b_h(\epsilon)$ happens instantaneously for a single value $\bar{\epsilon}$ (in other words, $b_h(\epsilon)$ is a step function). Then, the right-hand side of (2.4) gives value 0 when $\epsilon < \bar{\epsilon}$ and value 1 when $\epsilon > \bar{\epsilon}$, independently of the prior π'_h. Since the transition of $b_h(\epsilon)$ occurs abruptly but not instantaneously, this result holds only approximately. On the other hand, after the transition, $b_h(\epsilon)$ does go down to very low values quite rapidly. Hence, a significant departure from the above description can occur only when the prior π'_h all nests at very low values, below that of $b_h(\epsilon)$, as shown again by an inspection of (2.4): for the sake of the argument, say that all π'_h concentrates below the value of $b_h(\epsilon)$ for some value of ϵ beyond the transition point; then one easily sees that the right-hand side of (2.4) still keeps the value 0 for that ϵ. ⋆

The next example provides a case in point of Remark 2.4 and, still in the context of Example 1.5, shows that the conditional assessments bear very little sensitivity on the prior, until the prior expresses a very strong belief that seeing $s_N^* = 1$ is a rare event.

Example 2.5. Consider again Example 1.5 with $N = 20$. Suppose that q is uncertain with only two possible values: $q = 0.1$ or $q = 1$. In the former, $p_1 = \mathbb{P}^N\{s_N^* = 1\} = 0.1^{20} + 20 \cdot 0.1^{19} \cdot 0.9 = 1.81 \cdot 10^{-18}$, while in the latter $p_1 = 1$. The marginal distribution π'_1 for p_1 is shown in Figure 5, where parameter α defines our prior trust in the case $p_1 = 1$. Suppose that $\alpha = 10^{-7}$.

The interpretation of this very low value of α is that we bear a strong a priori belief in favor of the case $q = 0.1$ (in which case seeing 1 support constraint is extremely rare) while we do not completely exclude the case $q = 1$ (in which case one systematically sees 1 support constraint). Using formula (2.5) with $h = 1$ and $N = 20$ gives $b_1(\epsilon) \leq \min\{N(1-\epsilon)^{N-1}, 1\} = \min\{20(1-\epsilon)^{19}, 1\}$, which substituted in (2.4) yields

FIG. 5. *The prior* π_1'*:* $p_1 = 1.81 \cdot 10^{-18}$ *with probability* $1-\alpha$ *while* $p_1 = 1$ *with probability* α*.*

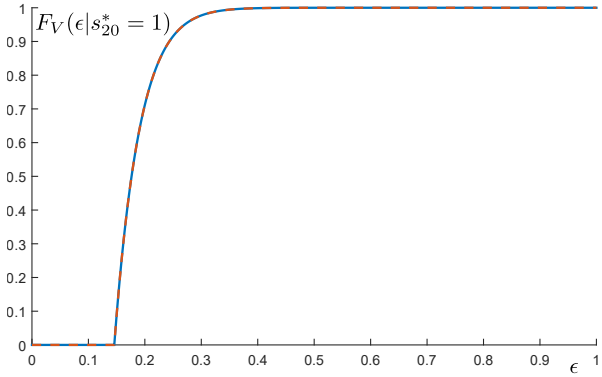


FIG. 6. *Bounds for* $F_V(\epsilon|s_{20}^* = 1)$ *for* $\alpha = 10^{-7}$ (*blue solid line*) *and* $\alpha = 1$ (*dashed red line*)*. The two curves are almost indistinguishable.*

$$
\begin{aligned}
F_V(\epsilon|s_{20}^* = 1) \geq 1 &- \frac{\int_{\{p_1 \leq \min\{20(1-\epsilon)^{19},1\}\}} p_1 \pi_1'(\mathrm{d}p_1)}{1.81 \cdot 10^{-18} \cdot (1 - 10^{-7}) + 1 \cdot 10^{-7}} \\
&- \frac{\min\{20(1-\epsilon)^{19}, 1\} \cdot \pi_1'\{p_1 > \min\{20(1-\epsilon)^{19}, 1\}\}}{1.81 \cdot 10^{-18} \cdot (1 - 10^{-7}) + 1 \cdot 10^{-7}} \\
&= \begin{cases} 1 - \frac{1.81 \cdot 10^{-18} \cdot (1-10^{-7}) + \min\{20(1-\epsilon)^{19},1\} \cdot 10^{-7}}{1.81 \cdot 10^{-18} \cdot (1-10^{-7}) + 1 \cdot 10^{-7}} & \text{if } \epsilon \leq 0.9005 \\ 1 - \frac{\min\{20(1-\epsilon)^{19},1\}}{1.81 \cdot 10^{-18} \cdot (1-10^{-7}) + 1 \cdot 10^{-7}} & \text{if } \epsilon > 0.9005. \end{cases}
\end{aligned} \tag{2.6}
$$

The right-hand side of (2.6) is profiled in Figure 6 against the result which is obtained for the case in which $\alpha = 1$.

There is very little difference between the two curves although the prior is substantially different in the two cases ($\alpha = 10^{-7}$ against $\alpha = 1$). The reason is that for $\alpha = 10^{-7}$ one has an a priori strong belief in favor of $q = 0.1$ but, after seeing $s_{20}^* = 1$ (which is extremely rare when $q = 0.1$, with probability $1.81 \cdot 10^{-18}$) the prior is reversed into a posterior belief that favors $q = 1$, so leveling the two cases $\alpha = 10^{-7}$ and $\alpha = 1$.

Interestingly enough, the above described mechanism maintains its validity till very low values of α. On the other hand, it is also instructive to go to the extreme case where $\alpha = 0$. If so, seeing $s_{20}^* = 1$ is interpreted as that a rare event has occurred. Correspondingly, (2.4) now gives

$$
F_V(\epsilon|s_{20}^* = 1) \geq \begin{cases} 0 & \text{if } \epsilon \leq 0.9005 \\ 1 - \frac{\min\{20(1-\epsilon)^{19},1\}}{1.81 \cdot 10^{-18}} & \text{if } \epsilon > 0.9005. \end{cases} \tag{2.7}
$$

The right-hand side of (2.7) is profiled in Figure 7.

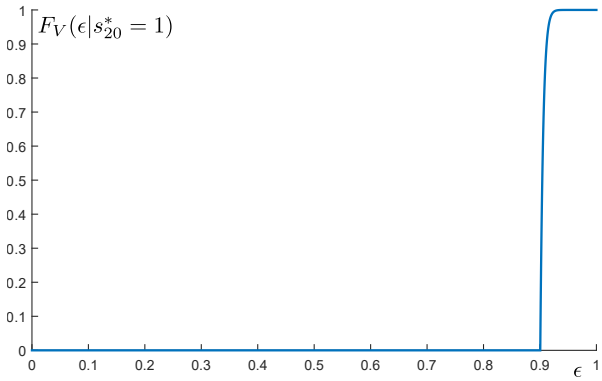


FIG. 7. *Bound for* $F_V(\epsilon|s^*_{20}=1)$ *for* $\alpha=0$.

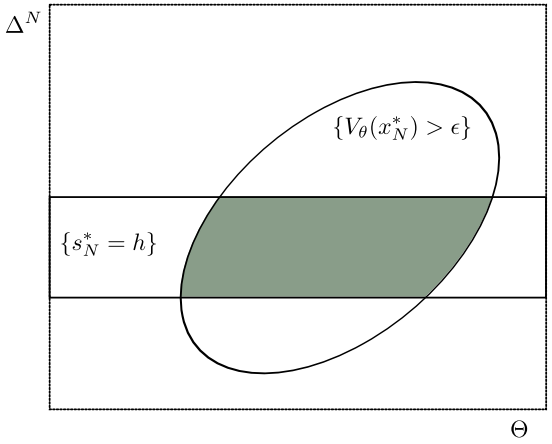


FIG. 8. *Representation of events* $\{V_\vartheta(x^*_N)>\epsilon\}$ *and* $\{s^*_N=h\}$ *on* $\Delta^N\times\Theta$. *The shaded region is* $\{V_\vartheta(x^*_N)>\epsilon\wedge s^*_N=h\}$.

Note that the curve stays at zero value till 0.9005, which means that the theory does not provide any meaningful lower bound on the probability of the event $V_\vartheta(x^*_N)\leq 0.9005$ conditionally on seeing $s^*_{20}=1$. This is in agreement with previous remarks in Example 1.5, where indeed seeing $s^*_{20}=1$ implied that $V_\vartheta(x^*_N)\geq 0.9$ when $q=0.1$. It is also interesting to note that beyond $\epsilon=0.9005$ the conditional distribution $F_V(\epsilon|s^*_{20}=1)$ saturates quickly to 1. $\star$

2.2. A Bayesian interpretation. While discussing the subject matter of this article with some eminent Bayesian statisticians, the present authors were prompted to more explicitly highlight the essentially Bayesian character of its content. This short section provides an explanation by showing how the various quantities involved can be rewritten in terms of the posterior distribution of ϑ given the data points.

Referring back to (2.1), the sets $\{V_\vartheta(x^*_N)>\epsilon\}$ and $\{s^*_N=h\}$ are events on $\Delta^N\times\Theta$ (see Figure 8 for a representation).

The integrals in (2.1) can also be written as

$$\int_\Theta \mathbb{P}^N_\vartheta\{V_\vartheta(x^*_N)>\epsilon\wedge s^*_N=h\}\pi(\mathrm{d}\vartheta)$$

$$= \int_{\Delta^N} \pi_{/\delta_1,\ldots,\delta_N}\{V_\vartheta(x_N^*) > \epsilon \wedge s_N^* = h\} \mathrm{P}_{\Delta^N}(\mathrm{d}\delta_1,\ldots,\mathrm{d}\delta_N),$$

$$\int_\Theta \mathbb{P}_\vartheta^N\{s_N^* = h\}\pi(\mathrm{d}\vartheta) = \int_{\Delta^N} \pi_{/\delta_1,\ldots,\delta_N}\{s_N^* = h\} \mathrm{P}_{\Delta^N}(\mathrm{d}\delta_1,\ldots,\mathrm{d}\delta_N),$$

where $\pi_{/\delta_1,\ldots,\delta_N}$ is the a posteriori distribution on Θ and P_{Δ^N} is the marginal of P on Δ_N. It is important to note that computing the posterior $\pi_{/\delta_1,\ldots,\delta_N}$ is a frightening goal in real applications due to attendant overwhelming computations. Even more importantly, such computations apply to a complete model of all the probabilistic elements involved in the problem. One fundamental achievement of (2.4) and (2.5) is that informative bounds on $F_V(\epsilon|s_N^* = h)$ can be obtained from a limited prior information expressed by π'_h.

3. A tight evaluation of $b_h(\epsilon)$. The fundamental formula (2.4) can be used with any valid upper bound of $b_h(\epsilon)$ and in this section we obtain a bound for $b_h(\epsilon)$ that is provably close to the best possible. Quantitatively, the final result outperforms, even significantly, the evaluations made with (2.5). This rather technical section leverages recent results on "wait-and-judge" scenario optimization established in [15].

THEOREM 3.1. *Under Assumption* 1.1 *and assuming that Assumption* 1.4 *holds true for any probability* $\mathbb{P}_\vartheta$, $\vartheta \in \Theta$, *it holds that* [9]

(3.1)

$$b_h(\epsilon) \leq \begin{cases} \sum_{i=0}^{d-1} \binom{N}{i} \epsilon^i (1-\epsilon)^{N-i} & \text{if } h = d \\ \sum_{i=0}^{d-1} \binom{N}{i} (\frac{d-1}{h}\epsilon)^i (1-\frac{d-1}{h}\epsilon)^{N-i} & \\ \quad + \sum_{i=d-h}^{N-h} \binom{N-h}{i} (\frac{d-1-h}{h}\frac{\epsilon}{1-\epsilon})^i (1-\frac{d-1-h}{h}\frac{\epsilon}{1-\epsilon})^{N-h-i} & \text{if } h \leq d-1 \\ \quad \times \frac{\binom{N}{h}}{\binom{d-1}{h}} \epsilon^h (1-\epsilon)^{N-h} \frac{(d-1)^{(d-1)}}{(d-1-h)^{(d-h-1)} h^h} & \text{and } \epsilon \leq \frac{h}{d-1} \\ \frac{\binom{N}{h}}{\binom{d-1}{h}} (1-\epsilon)^{N-d+1} & \text{if } h \leq d-1 \\ & \text{and } \epsilon > \frac{h}{d-1}. \end{cases}$$

[9] Despite its apparent complexity, this bound can be computed relatively easily. Indeed, it can be rewritten as

$$b_h(\epsilon) \leq \begin{cases} 1 - \mathrm{I}_\epsilon(d, N-d+1) & \text{if } h = d, \\ 1 - \mathrm{I}_{\frac{d-1}{h}\epsilon}(d, N-d+1) & \\ \quad + \mathrm{I}_{\frac{d-1-h}{h}\frac{\epsilon}{1-\epsilon}}(d-h, N-d+1) \times \mathrm{e}^{\sum_{k=0}^{h-1}[\log(N-k)-\log(d-1-k)]} & \text{if } h < d \\ \quad \times \mathrm{e}^{h\log(\epsilon)+(N-h)\log(1-\epsilon)+(d-1)\log(d-1)-(d-1-h)\log(d-1-h)-h\log(h)} & \text{and } \epsilon \leq \frac{h}{d-1}, \\ \mathrm{e}^{\sum_{k=0}^{h-1}[\log(N-k)-\log(d-1-k)]+(N-d-1)\log(1-\epsilon)} & \text{if } h < d \\ & \text{and } \epsilon > \frac{h}{d-1}, \end{cases}$$

where $\mathrm{I}_x(a,b) := \int_0^x \frac{\Gamma(a+b)}{\Gamma(a)\Gamma(b)} t^{a-1}(1-t)^{b-1}\mathrm{d}t$ is the well-known regularized incomplete Beta function, which, for integers values of a and b, equals $\sum_{i=a}^{b+a-1} \binom{b+a-1}{i} x^i (1-x)^{b+a-1-i}$ (we leave to the reader the simple verification that the two bounds coincide when the latter formula is used). The value of $\mathrm{I}_x(a,b)$ can be efficiently computed in various scientific computing environments (for example, in MATLAB, $\mathrm{I}_x(a,b)$ is implemented by the function betainc(x,a,b)), while the exponentials and logarithms serve the purpose of making the computation of the other terms numerically stable.

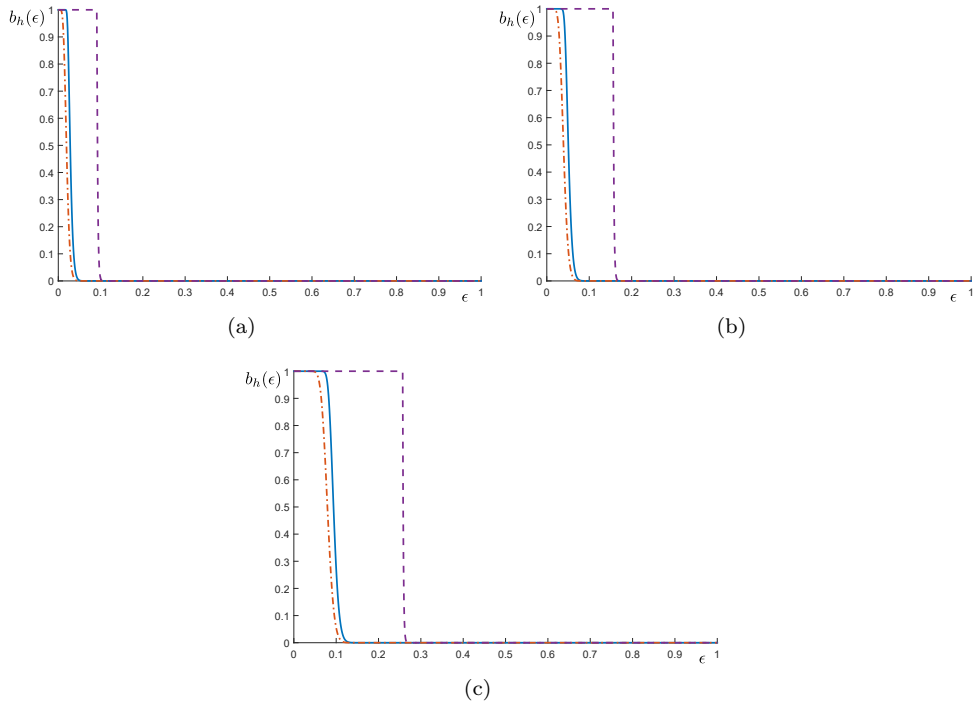


FIG. 9. *Bound* (3.1) (*solid blue line*) *versus bound* (2.5) (*dashed purple line*) *and the unsurmountable lower bound* $\sum_{i=0}^{h-1}\binom{N}{i}\epsilon^i(1-\epsilon)^{N-i}$ (*dashed-dotted red line*). $N=500$, $d=150$, *and* (a) $h=10$; (b) $h=20$; (c) $h=40$.

Before proving the theorem, we make some remarks about the result. Figure 9 profiles bound (3.1) for $N=500$, $d=150$, and $h=10,20,40$ against the bound in (2.5), which shows that the new bound marks a significant improvement.

Moreover, it is a notable fact that there is little margin of further improvement over bound (3.1) provided that Θ is rich enough. This claim can be justified by making reference to the theory developed in paper [12]: suppose that Θ contains problems with h support constraints with probability 1;[10] then the theory in [12] proves that for these problems it holds that

$$\mathbb{P}^N\{V(x_N^*)>\epsilon \wedge s_N^*=h\}=\sum_{i=0}^{h-1}\binom{N}{i}\epsilon^i(1-\epsilon)^{N-i}, \tag{3.2}$$

so that the right-hand side of (3.2) represents an insurmountable bound for $b_h(\epsilon)$. The tightness of (3.1) can be appreciated in Figure 9, where the right-hand side of (3.2) is also represented. Note that the value of ϵ for which the right-hand side of (3.2) equals 0.5 is approximately h/N (see, e.g., [42]) so that bound (3.1) rapidly saturates to 0 for values of ϵ larger than h/N.

***Proof of Theorem* 3.1.** Let $\bar{\epsilon}(k)$, $k\in\{0,1,\dots,d\}$, be equal to ϵ when $k=h$ and equal to 1 for $k\neq h$. Since $\mathbb{P}_\vartheta^N\{V_\vartheta(x_N^*)>1\}=0$, we have that

[10] In [12] a problem in dimension h (i.e., $x\in\mathbb{R}^h$) that has h support constraints with probability 1 is called *fully supported*. It is not difficult to embed a fully supported problem with h support constraints in a problem in dimension $d>h$ by adding $d-h$ dummy variables.

$$\mathbb{P}_\vartheta^N\{V_\vartheta(x_N^*) > \epsilon \wedge s_N^* = h\} = \sum_{k=0}^{d} \mathbb{P}_\vartheta^N\{V_\vartheta(x_N^*) > \bar{\epsilon}(k) \wedge s_N^* = k\}$$
$$= \mathbb{P}_\vartheta^N\{V_\vartheta(x_N^*) > \bar{\epsilon}(s_N^*)\}.$$

A direct application of Theorem 1 of [15] now gives that, for any ϑ,

$$\mathbb{P}_\vartheta^N\{V_\vartheta(x_N^*) > \bar{\epsilon}(s_N^*)\} \leq \gamma_h^*,$$

where

$$\gamma_h^* := \inf_{\xi(\cdot)\in \mathsf{C}^d[0,1]} \xi(1) \tag{3.3a}$$

$$\text{subject to } \frac{1}{h!}\frac{\mathrm{d}^h}{\mathrm{d}t^h}\xi(t) \geq \binom{N}{h} t^{N-h} \cdot \mathbf{1}_{[0,1-\epsilon)}, \quad t \in [0,1], \tag{3.3b}$$

$$\frac{1}{k!}\frac{\mathrm{d}^k}{\mathrm{d}t^k}\xi(t) \geq 0, \quad t \in [0,1],\ k = 0,1,\ldots,d,\ k \neq h, \tag{3.3c}$$

and $\mathsf{C}^d[0,1]$ is the set of continuous functions on $[0,1]$ with continuous derivative up to order d. To prove (3.1) we shall next show that γ_h^* is indeed smaller than or equal to the right-hand side of (3.1). To this aim, we exhibit a function $\bar{\xi}(t)$ that is feasible for (3.3) for which $\bar{\xi}(1)$ equals the right-hand side of (3.1).

Function $\bar{\xi}(t)$ is given by

$$\bar{\xi}(t) = t^N \cdot \mathbf{1}_{[0,\tau)} + \sum_{i=0}^{d-1} \binom{N}{i} (t-\tau)^i \tau^{N-i} \cdot \mathbf{1}_{[\tau,1]} + dA(t-\tau)^{d-1} \cdot \mathbf{1}_{[\tau,1]},$$

where $\mathbf{1}_{[0,\tau)}$ is the indicator function of the interval $[0,\tau)$ and $\mathbf{1}_{[\tau,1]}$ the indicator function of the interval $[\tau,1]$,

$$\tau = \begin{cases} 1-\epsilon & \text{if } h = d \\ \max\{0, 1 - \frac{d-1}{h}\epsilon\} & \text{if } h \leq d-1, \end{cases} \tag{3.4}$$

and

$$A = \begin{cases} 0 & \text{if } h = d \\ \frac{\binom{N}{h}(1-\epsilon)^{N-h}}{d\binom{d-1}{h}(1-\epsilon-\tau)^{d-h-1}} \cdot \sum_{i=d-h}^{N-h} \binom{N-h}{i} \left(1 - \frac{\tau}{1-\epsilon}\right)^i \left(\frac{\tau}{1-\epsilon}\right)^{N-h-i} & \text{if } h \leq d-1. \end{cases} \tag{3.5}$$

We first show the validity of (3.3b) and (3.3c).

For $k \leq d-1$, differentiating $\bar{\xi}(t)$ gives

$$\frac{1}{k!}\frac{\mathrm{d}^k}{\mathrm{d}t^k}\bar{\xi}(t) = \binom{N}{k} t^{N-k} \cdot \mathbf{1}_{[0,\tau)} + \sum_{i=k}^{d-1} \binom{N}{i}\binom{i}{k} (t-\tau)^{i-k} \tau^{N-i} \cdot \mathbf{1}_{[\tau,1]}$$
$$+ d\binom{d-1}{k} A(t-\tau)^{d-k-1} \cdot \mathbf{1}_{[\tau,1]},$$

which are all continuous for $k \leq d-2$, as revealed by a direct inspection. However, for $k = d-1$, $\frac{1}{k!}\frac{\mathrm{d}^k}{\mathrm{d}t^k}\bar{\xi}(t)$ has a jump in $t = \tau$ of height dA and $\frac{1}{d!}\frac{\mathrm{d}^d}{\mathrm{d}t^d}\bar{\xi}(t)$ becomes meaningful as a generalized function only:

$$\frac{1}{d!}\frac{\mathrm{d}^d}{\mathrm{d}t^d}\bar{\xi}(t) = \binom{N}{d} t^{N-d} \cdot \mathbf{1}_{[0,\tau)} + A\delta(t-\tau),$$

where δ is the Dirac delta function. It is thus a fact that $\bar{\xi}(t)$ is not in $\mathsf{C}^d[0,1]$; however, as we shall show later in the proof, this difficulty can be circumvented by a small modification of $\bar{\xi}(t)$. For now we concentrate on showing that $\bar{\xi}(t)$ satisfies the constraints (3.3b) and (3.3c). Constraint (3.3c) is clearly satisfied because $\frac{1}{k!}\frac{\mathrm{d}^k}{\mathrm{d}t^k}\bar{\xi}(t)$ is a sum of positive terms. Then, consider constraint (3.3b).

For $h=d$, we have that $\tau=1-\epsilon$ and $A=0$ (see (3.4) and (3.5)). Thus,

$$\frac{1}{d!}\frac{\mathrm{d}^d}{\mathrm{d}t^d}\bar{\xi}(t)=\binom{N}{d}t^{N-d}\cdot\mathbf{1}_{[0,1-\epsilon)},$$

which satisfies (3.3b).

For $h\le d-1$, (3.3b) is clearly satisfied for $t\in[0,\tau)$ because on this interval $\frac{1}{h!}\frac{\mathrm{d}^h}{\mathrm{d}t^h}\bar{\xi}(t)$ coincides with $\binom{N}{h}t^{N-h}$. For $t\in[1-\epsilon,1]$, (3.3b) is also satisfied because $\frac{1}{h!}\frac{\mathrm{d}^h}{\mathrm{d}t^h}\bar{\xi}(t)\ge 0$. Over the interval $[\tau,1-\epsilon)$, substituting the expression for A given in (3.5) in $\frac{1}{h!}\frac{\mathrm{d}^h}{\mathrm{d}t^h}\bar{\xi}(t)$ gives

$$\begin{aligned}\frac{1}{h!}\frac{\mathrm{d}^h}{\mathrm{d}t^h}\bar{\xi}(t)&=\sum_{i=h}^{d-1}\binom{i}{h}\binom{N}{i}(t-\tau)^{i-h}\tau^{N-i}\\&\quad+\frac{\binom{N}{h}(1-\epsilon)^{N-h}}{(1-\epsilon-\tau)^{d-h-1}}\sum_{i=d-h}^{N-h}\binom{N-h}{i}\left(1-\frac{\tau}{1-\epsilon}\right)^{i}\left(\frac{\tau}{1-\epsilon}\right)^{N-h-i}(t-\tau)^{d-h-1}.\end{aligned}\tag{3.6}$$

Noting that $\binom{i}{h}\binom{N}{i}=\binom{N-h}{i-h}\binom{N}{h}$ and that $(1-\epsilon)^{N-h}=(1-\epsilon)^i(1-\epsilon)^{N-h-i}$, and letting $j=i-h$ in the first sum, (3.6) can also be written as

$$\begin{aligned}\frac{1}{h!}\frac{\mathrm{d}^h}{\mathrm{d}t^h}\bar{\xi}(t)&=\binom{N}{h}\left[\sum_{j=0}^{d-1-h}\binom{N-h}{j}(t-\tau)^j\tau^{N-h-j}\right.\\&\qquad\left.+\sum_{i=d-h}^{N-h}\binom{N-h}{i}(1-\epsilon-\tau)^i\tau^{N-h-i}\cdot\left(\frac{t-\tau}{1-\epsilon-\tau}\right)^{d-h-1}\right]\\&=\binom{N}{h}\left[\sum_{j=0}^{d-h-1}\binom{N-h}{j}(t-\tau)^j\tau^{N-h-j}\right.\\&\qquad\left.+\sum_{i=d-h}^{N-h}\binom{N-h}{i}(t-\tau)^i\tau^{N-h-i}\cdot\left(\frac{1-\epsilon-\tau}{t-\tau}\right)^{i-(d-h-1)}\right],\end{aligned}$$

where the second equality follows by a simple rearrangement of the terms in the second sum. Since the above expression is in use for $t\in[\tau,1-\epsilon)$, it holds that $(\frac{1-\epsilon-\tau}{t-\tau})^{i-(d-h-1)}\ge 1$. This yields the inequality

$$\begin{aligned}&\frac{1}{h!}\frac{\mathrm{d}^h}{\mathrm{d}t^h}\bar{\xi}(t)\\&\quad\ge\binom{N}{h}\left[\sum_{j=0}^{d-h-1}\binom{N-h}{j}(t-\tau)^j\tau^{N-h-j}+\sum_{i=d-h}^{N-h}\binom{N-h}{i}(t-\tau)^i\tau^{N-h-i}\right]\\&\quad=\binom{N}{h}t^{N-h}\quad\text{(where we have used the binomial theorem).}\end{aligned}$$

This shows that (3.3b) is satisfied.

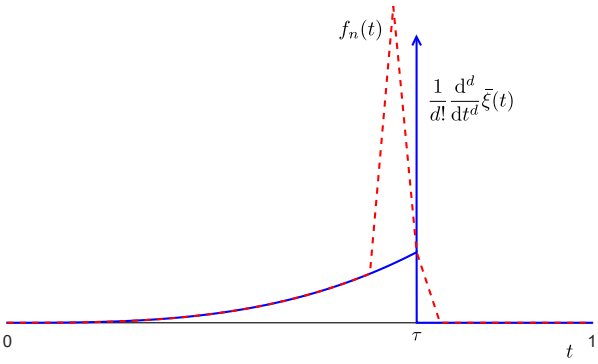


FIG. 10. $\frac{1}{d!}\frac{\mathrm{d}^d}{\mathrm{d}t^d}\bar{\xi}(t)$ (*solid blue line*) *versus* $f_n(t)$ (*red dashed line*).

We now come back to the difficulty that $\bar{\xi}(t)$ is not in $\mathsf{C}^d[0,1]$ and show that a slight modification of $\bar{\xi}(t)$ gives a function in $\mathsf{C}^d[0,1]$ while preserving satisfaction of the constraints (3.3b) and (3.3c). For $n=1,2,\ldots$, consider the functions

$$f_n(t) = \binom{N}{d} t^{N-d}\cdot \mathbf{1}_{[0,\tau)} + A\left[n^2(t-\tau+2/n)\mathbf{1}_{[\tau-\frac{2}{n},\tau-\frac{1}{n})} - n^2(t-\tau)\mathbf{1}_{[\tau-\frac{1}{n},\tau]}\right] + \binom{N}{d}\tau^{N-d}\Big(1-n(t-\tau)\Big)\cdot \mathbf{1}_{[\tau,\tau+\frac{1}{n}]},$$

which provide continuous approximations of $\frac{1}{d!}\frac{\mathrm{d}^d}{\mathrm{d}t^d}\bar{\xi}(t)$ (see Figure 10).

For each n, let $\bar{\xi}_n(t)$ be function $f_n(t)$ integrated d times, i.e.,

$$\bar{\xi}_n(t) = \int_0^t \int_0^{s_1} \cdots \int_0^{s_{d-1}} f_n(s_d)\,\mathrm{d}s_d \cdots \mathrm{d}s_1.$$

This way we clearly obtain a function $\bar{\xi}_n(t)$ that belongs to $\mathsf{C}^d[0,1]$; moreover, $\bar{\xi}_n(t)$ satisfies (3.3b) and (3.3c).[11] Hence, $\bar{\xi}_n(t)$ is feasible for (3.3), so that $\gamma_h^* \le \bar{\xi}_n(1)$, from which $\gamma_h^* \le \bar{\xi}(1)$ follows by taking the limit for $n\to+\infty$.

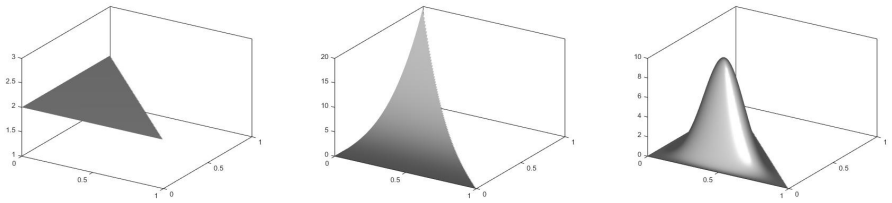
To conclude the proof one has to show that $\bar{\xi}(1)$ equals the right-hand side of (3.1), which is a cumbersome, but straightforward, computation that is left to the reader. ∎

4. Dirichlet priors. Equation (2.4) bounds $F_V(\epsilon|s_N^*=h)$ using π' only, which is the prior for $\boldsymbol{p}=(p_0,p_1,\ldots,p_d)$ over the simplex S. In this section we derive explicit expressions for $F_V(\epsilon|s_N^*=h)$ when π' is a Dirichlet distribution. This choice is motivated by the fact that the Dirichlet distribution is naturally supported over the simplex, while it also allows for enough flexibility to accommodate many situations of practical interest. Moreover, as we shall see, the Dirichlet distribution allows for explicit, closed-form, calculations.

Suppose therefore that

$$\pi' = \mathrm{Dir}(\alpha_0,\alpha_1,\ldots,\alpha_d),$$

[11] Indeed, $\frac{1}{d!}\frac{\mathrm{d}^d}{\mathrm{d}t^d}\bar{\xi}_n(t) = f_n(t)$ always satisfies its associated constraint both when $h\neq d$, because $\frac{1}{d!}\frac{\mathrm{d}^d}{\mathrm{d}t^d}\bar{\xi}_n(t)\ge 0$, and when $h=d$, since $\tau=1-\epsilon$ and $A=0$ in this case; instead, for $k<d$, the result follows by observing that $\frac{1}{k!}\frac{\mathrm{d}^k}{\mathrm{d}t^k}\bar{\xi}_n(t) \ge \frac{1}{k!}\frac{\mathrm{d}^k}{\mathrm{d}t^k}\bar{\xi}(t)$ because in $f_n(t)$ the mass that was concentrated in $t=\tau$ in $\frac{1}{d!}\frac{\mathrm{d}^d}{\mathrm{d}t^d}\bar{\xi}(t)$ has been moved to the left.



(a) (b) (c)

FIG. 11. *For $d = 2$, density of p_1, p_2 for a Dirichlet distribution with* (a) $\alpha_0 = \alpha_1 = \alpha_2 = 1$, (b) $\alpha_0 = \alpha_1 = 1, \alpha_2 = 4$, (c) $\alpha_0 = \alpha_1 = \alpha_2 = 10$.

where $\mathrm{Dir}(\alpha_0, \alpha_1, \ldots, \alpha_d)$ is a Dirichlet distribution, whose density over the support $\sum_{i=1}^{d} p_k \leq 1$, $p_k \geq 0$, $k = 1, \ldots, d$, is known to be

$$\frac{\Gamma\left(\sum_{k=0}^{d} \alpha_k\right)}{\prod_{k=0}^{d} \Gamma(\alpha_k)} \left(1 - \sum_{k=1}^{d} p_k\right)^{\alpha_0 - 1} \prod_{k=1}^{d} p_k^{\alpha_k - 1}$$

($\Gamma(\cdot)$ denotes the Gamma function) and further p_0 remains determined by relation $p_0 = 1 - \sum_{k=1}^{d} p_k$. Coefficients $\alpha_0 > 0$, $\alpha_1 > 0$, …, $\alpha_d > 0$ are free parameters that can be selected by the user.

Figure 11 shows the density of some Dirichlet distributions obtained by specific choices of the free parameters for $d = 2$.

As is the case for panel (a) of Figure 11, the choice $\alpha_0 = \alpha_1 = \cdots = \alpha_d = 1$ gives for any d a flat (uniform) distribution, which corresponds to adopting a "principle of indifference" (see Remark 2.2). Instead, for $\alpha_k > 1$, $\forall k$, the Dirichlet distribution becomes unimodal with mean $(\alpha_0/\sum_k \alpha_k, \alpha_1/\sum_k \alpha_k, \ldots, \alpha_d/\sum_k \alpha_k)$. Moreover, the bigger the $\sum_k \alpha_k$, the more concentrated the distribution around its mean. This may be used to accommodate the case in which one wants to build a prior around an empirical frequency for the complexity (see again Remark 2.2). The reader interested in shaping a Dirichlet distribution beyond these simple rules can consult any statistical textbook for further information.

We next show how a Dirichlet prior can be used in (2.4). To this end, it is useful to recall a well-known result on Dirichlet distributions, namely, that the marginal distribution π'_h is a Beta distribution,

$$\pi'_h = \mathrm{Beta}\left(\alpha_h, \sum_{k \neq h} \alpha_k\right),$$

which has density

$$\frac{\Gamma\left(\sum_{k=0}^{d} \alpha_k\right)}{\Gamma(\alpha_h)\Gamma(\sum_{k \neq h} \alpha_k)} p_h^{\alpha_h - 1} (1 - p_h)^{\sum_{k \neq h} \alpha_k - 1}$$

over the support $p_h \in [0, 1]$. The main properties of the Beta distribution $\mathrm{Beta}(a, b)$ (which will be used in subsequent derivations) are that its mean is $a/(a + b)$ while its cumulative distribution function

$$F_B(x) = \int_0^x \frac{\Gamma(a + b)}{\Gamma(a)\Gamma(b)} t^{a-1} (1 - t)^{b-1} \mathrm{d}t$$

is the regularized incomplete Beta function $\mathrm{I}_x(a,b)$, which can be easily evaluated, e.g., via the `betainc` command in MATLAB (we have already encountered $\mathrm{I}_x(a,b)$ in section 3).

Now, consider the right-hand side of (2.4). The denominator is easily recognized to be equal to the mean of a Beta$(\alpha_h, \sum_{k\neq h}\alpha_k)$:

$$\int_{[0,1]} p_h \pi'_h(\mathrm{d}p_h) = \frac{\alpha_h}{\sum_{k=0}^d \alpha_k}. \tag{4.1}$$

The terms in the numerator can instead be computed as follows:

$$\begin{aligned}\pi'_h\{p_h > b_h(\epsilon)\} &= \int_{\{p_h > b_h(\epsilon)\}} \frac{\Gamma\left(\sum_{k=0}^d \alpha_k\right)}{\Gamma(\alpha_h)\Gamma(\sum_{k\neq h}\alpha_k)} p_h^{\alpha_h - 1}(1-p_h)^{\sum_{k\neq h}\alpha_k - 1}\mathrm{d}p_h \\ &= 1 - \mathrm{I}_{b_h(\epsilon)}\left(\alpha_h, \sum_{k\neq h}\alpha_k\right) \qquad (4.2)\end{aligned}$$

and

$$\begin{aligned}&\int_{\{p_h \leq b_h(\epsilon)\}} p_h \pi'_h(\mathrm{d}p_h) \\ &= \int_{\{p_h \leq b_h(\epsilon)\}} p_h \cdot \frac{\Gamma\left(\sum_{k=0}^d \alpha_k\right)}{\Gamma(\alpha_h)\Gamma(\sum_{k\neq h}\alpha_k)} p_h^{\alpha_h - 1}(1-p_h)^{\sum_{k\neq h}\alpha_k - 1}\mathrm{d}p_h \\ &= \frac{\alpha_h}{\sum_{k=0}^d \alpha_k} \int_{\{p_h \leq b_h(\epsilon)\}} \frac{\Gamma\left(\sum_{k=0}^d \alpha_k + 1\right)}{\Gamma(\alpha_h + 1)\Gamma(\sum_{k\neq h}\alpha_k)} p_h^{\alpha_h}(1-p_h)^{\sum_{k\neq h}\alpha_k - 1}\mathrm{d}p_h \\ &\quad \text{(where we have used relation } [\Gamma(x+1) = x\Gamma(x)]) \\ &= \frac{\alpha_h}{\sum_{k=0}^d \alpha_k} \cdot \mathrm{I}_{b_h(\epsilon)}\left(\alpha_h + 1, \sum_{k\neq h}\alpha_k\right). \qquad (4.3)\end{aligned}$$

Using the three equations (4.1), (4.2), and (4.3) in (2.4) we obtain the following result.[12]

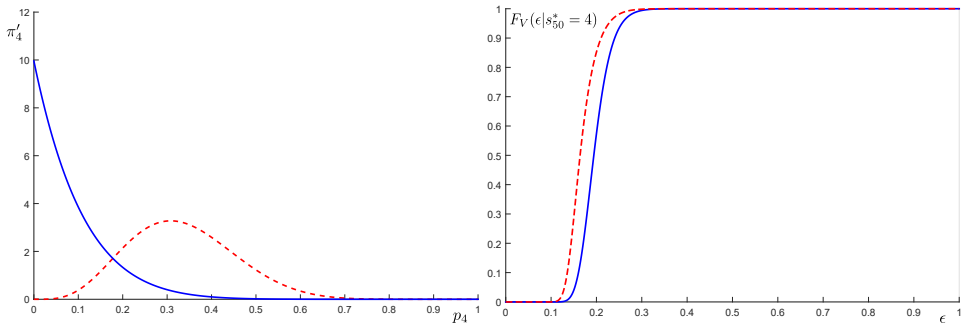
THEOREM 4.1. *Assume that* $\pi' = \mathrm{Dir}(\alpha_0, \alpha_1, \ldots, \alpha_d)$*; then* (2.4) *becomes*

$$\begin{aligned}&F_V(\epsilon | s_N^* = h) \\ &\quad \geq 1 - \left[\mathrm{I}_{b_h(\epsilon)}\left(\alpha_h + 1, \sum_{k\neq h}\alpha_k\right) + b_h(\epsilon) \cdot \frac{\sum_{k=0}^d \alpha_k}{\alpha_h}\left(1 - \mathrm{I}_{b_h(\epsilon)}\left(\alpha_h, \sum_{k\neq h}\alpha_k\right)\right)\right].\end{aligned} \tag{4.4}$$

The right-hand side of (4.4) has to be combined with the expression of $b_h(\epsilon)$ given in (2.5) or that given in (3.1) to obtain an explicit evaluation of $F_V(\epsilon | s_N^* = h)$.

We next use (4.4) with (3.1) in two simulated examples, which we believe allows for a better understanding of formula (4.4); instead, an example based on real data is presented in section 5.

[12] As is clear, similar computations as in (4.1), (4.2), and (4.3) can be performed for priors other than the Dirichlet distributions (for instance, the logistic normal distributions or even the truncated normal distributions); however, the computations may turn out to be more complex and require numerical evaluations. For example, even the mean as in the left-hand side of (4.1) does not admit an analytical expression for the two distributions mentioned above.



(a) (b)

FIG. 12. *Left:* π'_4. *Right:* $F_V(\epsilon|s^*_{50}=4)$ (*solid blue line:* $\alpha_k=1$, $k=0,\dots,10$*; dashed red line:* $\alpha_k=1$ *for* $k\neq 4$, $\alpha_4=5$).

Example 4.2. Let $N=50$, $d=10$, and $\alpha_k=1\ \forall k=0,\dots,10$ in the Dirichlet prior (the reason for taking small values of N in simulated examples is to prevent distributions from concentrating excessively, so that they visually convey the intended message). π'_4 is graphically represented in Figure 12(a) (solid blue line). Equation (4.4) with $h=4$ becomes in this case

$$F_V(\epsilon|s^*_{50}=4)\geq 1-\left[\mathrm{I}_{b_4(\epsilon)}(2,10)+b_4(\epsilon)\cdot 11\left(1-\mathrm{I}_{b_4(\epsilon)}(1,10)\right)\right], \tag{4.5}$$

which, using (3.1) for $b_h(\epsilon)$, is the distribution represented in Figure 12(b) (solid blue line). As we can see, having 4 support constraints in an optimization problem with 10 variables and 50 data points leads to a belief that the risk will be below 0.22–0.23 with high confidence.

Consider instead a Dirichlet prior with $\alpha_k=1$ for $k\neq 4$ and $\alpha_4=5$, which gives the π'_4 that can also be seen in Figure 12(a) (dashed red line). In this case,

$$F_V(\epsilon|s^*_{50}=4)\geq 1-\left[\mathrm{I}_{b_4(\epsilon)}(6,10)+b_4(\epsilon)\cdot 3\left(1-\mathrm{I}_{b_4(\epsilon)}(5,10)\right)\right], \tag{4.6}$$

which is represented in Figure 12(b) (dashed red line). It can be observed that in this second case $F_V(\epsilon|s^*_{50}=4)$ shifts to the left, so that one expects less violation than in the first case. This is not surprising since in this second case one has very little trust in low values of p_4. On the other hand, one can further observe that the difference in $F_V(\epsilon|s^*_{50}=4)$ between the two cases is minor as compared to the wide disagreement present in the priors, which experimentally shows that the posterior bears little sensitivity on the prior. ⋆

Example 4.3. In this second example, we discuss the role of the dimension d under the assumption that our prior π' is uniform over the simplex, i.e., $\alpha_k=1\ \forall k$.

For $d=10$, $N=2500$, $h=10$, and $\epsilon=0.01$, formula (3.1) gives $b_h(\epsilon)=2.1\cdot 10^{-4}$, so that

$$\sup_{\vartheta\in\Theta}\mathbb{P}^{2500}_{\vartheta}\{V_\vartheta(x^*_{2500})>0.01\wedge s^*_{2500}=10\}\leq 2.1\cdot 10^{-4}.$$

Since $\alpha_k=1$, $k=0,\dots,10$, we do not a priori reckon that any value of h is more likely than any other one so that $\mathrm{P}\{s^*_{2500}=10\}=1/11$ (one over the total number of cases). This is not a rare event and, hence, we tend to think that $V_\vartheta(x^*_{2500})>0.01$ will not

happen. This is confirmed by the theory. When α_k is set equal to 1 $\forall k$ and d is left free, (4.4) becomes

$$F_V(\epsilon|s^*_{2500}=10) \geq 1 - \left[\mathrm{I}_{b_h(\epsilon)}(2,d) + b_h(\epsilon)\cdot(d+1)\left(1-\mathrm{I}_{b_h(\epsilon)}(1,d)\right)\right], \tag{4.7}$$

which, for $d=10$ and $b_h(\epsilon)=2.1\cdot 10^{-4}$, returns the value $1-0.0023=0.9977$, implying high confidence that the probability of violation is below 0.01.

Suppose now that $d=10^3$, $N=2500$, $h=10$, and $\epsilon=0.01$, which gives $b_h(\epsilon)=2.8\cdot 10^{-3}$ when these values are used in formula (3.1). With a uniform prior, how confident are we that $V_\vartheta(x^*_{2500})>0.01$ will not happen? For one thing it holds that

$$\sup_{\vartheta\in\Theta} \mathbb{P}^{2500}_\vartheta\{V_\vartheta(x^*_{2500})>0.01 \wedge s^*_{2500}=10\} \leq 2.8\cdot 10^{-3}. \tag{4.8}$$

On the other hand, $\{s^*_{2500}=10\}$ is a rare event in itself in this case since s^*_{2500} takes one value among a large set of possibilities, in fact $1+10^3$ possibilities, so that $\mathrm{P}\{s^*_{2500}=10\}=1/(1+10^3)$. Using formula (4.7), we now find $F_V(0.01|s^*_{2500}=10)\geq 1-0.9383=0.0617$, a small value.

As an additional remark, one can verify that raising ϵ to value 0.015 in this example with $d=10^3$ leads to $F_V(0.015|s^*_{2500}=10)\geq 1-0.0014=0.9986$, which is interpreted as that for $d=10^3$ we have a strong belief that the probability of violation is below the value 0.015. $\star$

5. An example with real data. We consider the data set called Corel Image Features, which is publicly available at the UCI Machine Learning Repository.[13] This data set consists of 68040 records, each corresponding to a JPEG image taken from a Corel image collection. Each record is characterized by 89 attributes, from which we select the following 25 noncategorical ones:[14]

- 9 color moments, which are the mean, the standard deviation, and the skewness of the hue, the saturation, and the value of the color ($3\times 3=9$ attributes in total);
- 16 co-occurrence textures, which are the second angular moment, the contrast, the inverse difference moment, and the entropy of 4 co-occurrence matrices calculated along the principal directions (horizontal, vertical, and the two diagonals) of the image converted in gray-scale ($4\times 4=16$ attributes in total).

Each record in the data set is regarded as an instance of an uncertain element δ, which is a 25-dimensional vector with real components.

After randomly sorting the 68040 records we picked the first 500 records δ_1, δ_2, …, δ_{500} (random sorting is used to cancel any possible ordering introduced in the data set at the time it has been uploaded in the repository). $\delta_1,\delta_2,\ldots,\delta_{500}$ is regarded as an independent sample of images. Then, we considered the following scenario program that aims at constructing the "most compact" box in $\mathbb{R}^{25}$ containing all 500 scenarios:

$$\begin{aligned} &\min_{v\in\mathbb{R}^{25},\ell\in\mathbb{R}^{25}} \sum_{k=1}^{25}\ell_k \\ &\text{subject to } v\leq\delta_i\leq v+\ell, \quad i=1,\ldots,500, \end{aligned} \tag{5.1}$$

[13] http://archive.ics.uci.edu/ml/datasets/Corel+Image+Features.

[14] The reader is referred to the UCI Machine Learning Repository website for a more detailed description of the attributes.

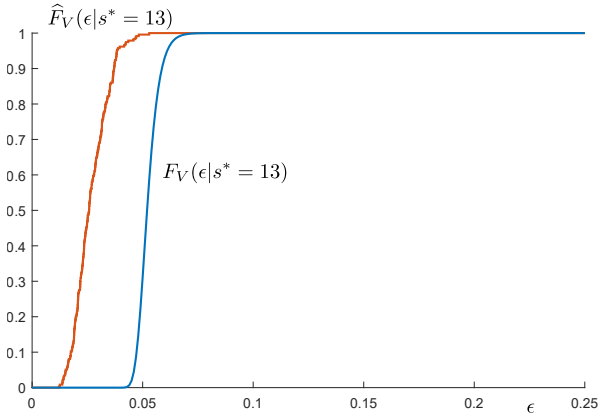


FIG. 13. *The bound to* $F_V(\epsilon|s^*_{500}=13)$ (*continuous blue curve*) *versus* $\widehat{F}_V(\epsilon|s^*_{500}=13)$ (*stairwise red curve*).

where $v=[v_1,v_2,\dots,v_{25}]^T$ are the coordinates of the box vertex with leftmost components, $\ell=[\ell_1,\ell_2,\dots,\ell_{25}]^T$ are the lengths of the box edges, and inequalities are meant componentwise. Note that $d=50$ in this case. The number of support constraints s^*_{500} is the amount of δ_i's that lie on the boundary of the box. The optimal box obtained after solving (5.1) can be thought of as a descriptor of the variability in the images and the probability of violation corresponds to the chance of observing a new image that lies outside the box (and, hence, some of its attributes are not correctly predicted).

We describe our prior π' as a uniform Dirichlet distribution $\mathrm{Dir}(1,1,\dots,1)$.[15] Upon solving the problem we found $s^*_{500}=13$. The continuous blue curve in Figure 13 is a plot of the bound to $F_V(\epsilon|s^*_{500}=13)$ that is obtained from Theorem 4.1 with formula (3.1) used to compute $b_h(\epsilon)$.

Since the bound rapidly saturates to 1 after the value $\epsilon=0.07$, this result induces a strong belief in the fact that the violation has to be below 7%.

We next computed an empirical estimate of $F_V(\epsilon|s^*_{500}=13)$ as follows. Problem (5.1) is solved 10000 times. In each run, 500 scenarios were randomly extracted from the available data set, the solution and the corresponding number of support constraints s^*_{500} were computed, and the violation of the solution was empirically estimated as the proportion $\hat{v}$ of the remaining $68040-500=67540$ data points that do not belong to the box.[16] For estimating $F_V(\epsilon|s^*_{500}=13)$, we kept only the runs that gave $s^*_{500}=13$ and computed $\widehat{F}_V(\epsilon|s^*_N=13)=$ (no. of cases with $\hat{v}\le\epsilon$ and $s^*_{500}=13$)/(no. of cases with $s^*_{500}=13$). The obtained $\widehat{F}_V(\epsilon|s^*_N=13)$ is depicted in Figure 13 as a stairwise red curve. Empirical evidence is in agreement with the expectation given by the theorem.

Similar results were obtained also for other values of s^*_{500} (in our simulations s^*_{500} ranged from 10 to 30 with higher frequencies for central values) and Figure 14 depicts curves similar to Figure 13 for $s^*_{500}=24$.

[15]Notice that a probability $\mathbb{P}_\vartheta$ in this problem is a distribution on $\mathbb{R}^{25}$ and assigning a complete prior π would consists in providing a distribution over the domain of 25-dimensional distributions, quite a formidable task.

[16]This procedure, which corresponds to multiple shuffles of the original data set, introduces a slight correlation between one case and the others; however, given the numerosity of the baseline data set, this correlation can be regarded as negligible.

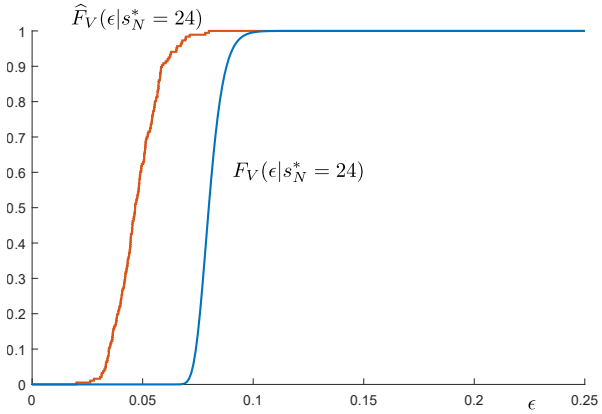


FIG. 14. *The bound to* $F_V(\epsilon|s_N^*=24)$ (*continuous blue curve*) *versus* $\widehat{F}_V(\epsilon|s_N^*=24)$ (*stairwise red curve*).

6. Conclusions. The number of support constraints of a scenario program is an observable that carries fundamental information to estimate the violation of the solution and it is natural that one asks for evaluations of the distribution of the violation conditional on the number of support constraints. In this article, we have shown that results for the conditional distribution are impossible in a distribution-free setup; however, modest prior information—which only refers to the distribution of the cardinality of the support constraint set—suffices to secure strong, practically useful, evaluations of the conditional distribution. These results open a door for the use of the theory in an optimization context where assuming a full prior is not realistic.

Acknowledgments. We would like to gratefully thank Professors E. Regazzini, L. Ladelli, and A. Guglielmi for fruitful discussions.

REFERENCES

[1] T. ALAMO, R. TEMPO, AND E. CAMACHO, *A randomized strategy 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, 51 (2015), pp. 160–172.
[3] J. BARRA, *Mathematical Basis of Statistics*, Academic Press, New York, 1981.
[4] H. BAUER, *Probability Theory*, De Gruyter, Berlin, 1996.
[5] J. BERGER, *The robust Bayesian viewpoint,* in Robustness of Bayesian analysis, J. Kadane, ed., North-Holland, Amsterdam, 1984, pp. 63–124
[6] J. BERGER, *Robust Bayesian analysis: Sensitivity to the prior*, J. Statist. Plann. Inference, 25 (1990), pp. 303–328.
[7] G. CALAFIORE AND M. CAMPI, *Uncertain convex programs: Randomized solutions and confidence levels*, Math. Program., 102 (2005), pp. 25–46, https://doi.org/10.1007/s10107-003-0499-y.
[8] G. CALAFIORE AND M. CAMPI, *The scenario approach to robust control design*, IEEE Trans. Automat. Control, 51 (2006), pp. 742–753.
[9] M. CAMPI, *Classification with guaranteed probability of error*, Mach. Learn., 80 (2010), pp. 63–84.
[10] M. CAMPI, G. CALAFIORE, AND S. GARATTI, *Interval predictor models: Identification and reliability*, Automatica, 45 (2009), pp. 382–392, https://doi.org/10.1016/j.automatica.2008.09.004.
[11] M. CAMPI AND A. CARÈ, *Random convex programs with* L_1*-regularization: Sparsity and generalization*, SIAM J. Control Optim., 51 (2013), pp. 3532–3557.

[12] M. Campi and S. Garatti, *The exact feasibility of randomized solutions of uncertain convex programs*, SIAM J. Optim., 19 (2008), pp. 1211–1230, https://doi.org/10.1137/07069821X.
[13] M. 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.
[14] M. Campi and S. Garatti, *Introduction to the Scenario Approach*, MOS-SIAM Ser. Optim. 26, SIAM, Philadelphia, 2018.
[15] M. Campi and S. Garatti, *Wait-and-judge scenario optimization*, Math. Program., 167 (2018), pp. 155–189.
[16] M. Campi and S. Garatti, *A theory of the risk for optimization with relaxation and its application to support vector machines*, J. Mach. Learn. Res., 22 (2021), pp. 1–38.
[17] M. Campi, S. Garatti, and M. Prandini, *The scenario approach for systems and control design*, Ann. Rev. Control, 33 (2009), pp. 149–157, https://doi.org/10.1016/j.arcontrol.2009.07.001.
[18] M. Campi, S. Garatti, and F. Ramponi, *A general scenario theory for non-convex optimization and decision making*, IEEE Trans. Automat. Control, 63 (2018), pp. 4067–4078.
[19] A. Carè, S. Garatti, and M. Campi, FAST—*Fast algorithm for the scenario technique*, Oper. Res., 62 (2014), pp. 662–671.
[20] A. Carè, S. Garatti, and M. Campi, *Scenario min-max optimization and the risk of empirical costs*, SIAM J. Optim., 25 (2015), pp. 2061–2080.
[21] A. Caré, F. Ramponi, and M. Campi, *A new classification algorithm with guaranteed sensitivity and specificity for medical applications*, IEEE Control Systems Lett., 2 (2018), pp. 393–398.
[22] L. Crespo, B. Colbert, S. Kenny, and D. Giesy, *On the quantification of aleatory and epistemic uncertainty using sliced-normal distributions*, Systems Control Lett., 134 (2019).
[23] L. Crespo, D. Giesy, and S. Kenny, *Interval predictor models with a formal characterization of uncertainty and reliability*, in Proceedings of the 53rd Conference on Decision and Control, IEEE, Los Angeles, CA, 2014, pp. 5991–5996.
[24] L. Crespo, D. Giesy, S. Kenny, and J. Deride, *A scenario optimization approach to system identification with reliability guarantees*, in Proceedings of the American Control Conference, Philadelphia, 2019, pp. 2100–2106.
[25] L. Crespo, S. Kenny, and D. Giesy, *Random predictor models for rigorous uncertainty quantification*, Int. J. Uncertain. Quantif., 5 (2015), pp. 469–489.
[26] L. Crespo, S. Kenny, D. Giesy, R. Norman, and S. Blattnig, *Application of interval predictor models to space radiation shielding*, in Proceedings of the 18th AIAA Non-Deterministic Approaches Conference, San Diego, CA, 2016.
[27] L. G. Crespo, S. P. Kenny, and D. P. Giesy, *Interval predictor models with a linear parameter dependency*, J. Verif. Valid. Uncertain. Quantif., 1 (2016), pp. 1–10.
[28] P. M. Esfahani, T. Sutter, and J. Lygeros, *Performance bounds for the scenario approach and an extension to a class of non-convex programs*, IEEE Trans. Automat. Control, 60 (2015), pp. 46–58, https://doi.org/10.1109/TAC.2014.2330702.
[29] S. Garatti and M. Campi, *Modulating robustness in control design: Principles and algorithms*, IEEE Control Syst. Mag., 33 (2013), pp. 36–51, https://doi.org/10.1109/MCS.2012.2234964.
[30] S. Garatti and M. Campi, *Risk and complexity in scenario optimization*, Math. Program., 191 (2022), pp. 243–279.
[31] S. Garatti, M. Campi, and A. Caré, *On a class of interval predictor models with universal reliability*, Automatica, 110 (2019), 108542.
[32] S. Grammatico, X. Zhang, K. Margellos, P. Goulart, and J. Lygeros, *A scenario approach for non-convex control design*, IEEE Trans. Automat. Control, 61 (2016), pp. 334–345.
[33] L. Hong, Z. Hu, and G. Liu, *Monte Carlo methods for value-at-risk and conditional value-at-risk: A review*, ACM Trans. Model. Comput. Sim., 24 (2014), pp. 22:1–22:37.
[34] M. Lacerda and L. G. Crespo, *Interval predictor models for data with measurement uncertainty*, in Proceedings of the 2017 American Control Conference, Seattle, WA, 2017, pp. 1487–1492.
[35] M. Lavine, *An approach to robust Bayesian analysis for multidimentional parameter spaces*, J. Amer. Statist. Assoc., 86 (1991), pp. 400–403.
[36] K. Margellos, P. Goulart, and J. Lygeros, *On the road between robust optimization and the scenario approach for chance constrained optimization problems*, IEEE Trans. Automat. Control, 59 (2014), pp. 2258–2263.
[37] K. Margellos, M. Prandini, and J. Lygeros, *On the connection between compression learning and scenario based single-stage and cascading optimization problems*, IEEE Trans. Automat. Control, 60 (2015), pp. 2716–2721.

[38] H. Nasir, A. Caré, and E. Weyer, *A scenario-based stochastic MPC approach for problems with normal and rare operations with an application to rivers*, IEEE Trans. Control Syst. Technol., 27 (2019), pp. 1397–1410.

[39] B. 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.

[40] B. Pagnoncelli, D. Reich, and M. 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.

[41] B. Pagnoncelli and S. Vanduffel, *A provisioning problem with stochastic payments*, European. J. Oper. Res., 221 (2012), pp. 445–453.

[42] M. Payton, L. Young, and J. Young, *Bounds for the difference between median and mean of beta and negative binomial distributions*, Metrika, 36 (1989), 347?354.

[43] F. Ramponi, *Consistency of the scenario approach*, SIAM J. Optim., 28 (2018), pp. 135–162.

[44] R. Rocchetta, L. Crespo, and S. Kenny, *Solution of the benchmark control problem by scenario optimization*, in Proceedings of the ASME 2019 Dynamic Systems and Control Conference Park City, UT, 2019.

[45] R. Rocchetta, L. Crespo, and S. Kenny, *A scenario optimization approach to reliability-based design*, Reliab. Eng. Syst. Saf., 196 (2020).

[46] G. Schildbach, L. Fagiano, C. Frei, and M. Morari, *The scenario approach for stochastic model predictive control with bounds on closed-loop constraint violations*, Automatica, 50 (2014), pp. 3009–3018.

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

[48] T. Sutter, P. M. Esfahani, and J. Lygeros, *Approximation of constrained average cost markov control processes*, in Proceedings of the 53rd Conference on Decision and Control, Los Angeles, CA, IEEE, 2014, pp. 6597–6602, https://doi.org/10.1109/CDC.2014.7040424.

[49] L. Wasserman, *Recent methodological advances in robust Bayesian inference*, in Bayesian Statistics 4, J. Bernardo, J. Berger, A. David, and A. Smith, eds., Oxford University Press, Oxford, UK, 1990, pp. 483–502.

[50] J. Welsh and H. Kong, *Robust experiment design through randomisation with chance constraints*, in Proceedings of the 18th IFAC World Congress, Milan, 2011.

[51] J. Welsh and C. Rojas, *A scenario based approach to robust experiment design*, in Proceedings of the 15th IFAC Symposium on System Identification, Saint-Malo, France, 2009.

[52] X. Zhang, S. Grammatico, G. Schildbach, P. Goulart, and J. Lygeros, *On the sample size of random convex programs with structured dependence on the uncertainty*, Automatica, 60 (2015), pp. 182–188.