

ON CONDITIONAL RISK ASSESSMENTS IN SCENARIO OPTIMIZATION*

S. GARATTI[†] AND M.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 avail 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

AMS subject classifications. 90C15, 90C25, 62C20

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 $\delta_1, \dots, \delta_N$ from $(\Delta, \mathcal{D}, \mathbb{P})$, the scenario-based solution, [7, 14], is obtained by solving the following optimization problem

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

Problem (1.1) only involves finitely many constraints from Δ (in applications, Δ often has infinite cardinality) and the values $\delta_1, \dots, \delta_N$ 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, 47, 19, 36, 52, 28, 20, 15, 43, 18]. 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 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 of study for 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

*Submitted to the editors October 13th, 2021.

[†] Dipartimento di Elettronica, Informazione e Bioingegneria - Politecnico di Milano, piazza Leonardo da Vinci 32, 20133 Milano, Italia. E-mail: simone.garatti@polimi.it, web-site: <https://garatti.faculty.polimi.it/>

[‡] Dipartimento di Ingegneria dell’Informazione - Università di Brescia, via Branze 38, 25123 Brescia, Italia. E-mail: marco.campi@unibs.it, web-site: <https://marco-campi.unibs.it/>

functions $f(x)$ can be re-written in the form of (1.1) by an epigraphic reformulation; moreover, problems with an uncertain convex cost function can also be re-written 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 $\delta_1, \dots, \delta_N$, 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 $t_2(x)$, $t_3(x)$, ... if the tie still occurs. After breaking the tie, the solution is denoted by x_N^* . $\diamond$*

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., $t_1(x) = x_1$, $t_2(x) = x_2$, ...

Scenario programs have proven useful in multiple application domains, including control and systems design, [8, 17, 29, 46, 48, 2, 32, 38, 23, 44, 45], prediction, [10, 25, 26, 22, 27, 34, 24, 31], quantitative finance, [39, 40, 41, 33], and classification, [9, 11, 37, 21]. 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*

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

$V(x)$ 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, 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

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

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

¹Writing $\mathcal{X}_{\delta_{N+j}}$, 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 .

85 $\delta_1, \dots, \delta_N$), one obtains a random variable $V(x_N^*)$. The distribution of $V(x_N^*)$ is
 86 important because it characterizes the robustness level achieved by the solution to
 87 (1.1). One fundamental result proven in [12] shows that the distribution of $V(x_N^*)$
 88 is always (i.e., *independently of the type of constraints* $x \in \mathcal{X}_\delta$ and of $\mathbb{P}$) dominated
 89 by a Beta($d, N - d + 1$) distribution according to the formula

$$\begin{aligned}
 90 \quad (1.3) \quad & \mathbb{P}^N\{V(x_N^*) \leq \epsilon\} \geq 1 - \sum_{i=0}^{d-1} \binom{N}{i} \epsilon^i (1 - \epsilon)^{N-i}.^2
 \end{aligned}$$

91 Moreover, this result is tight because (1.3) holds with equality for a whole class of
 92 problems, named *fully-supported* in [12]. The fact that (1.3) is valid irrespective of
 93 $\mathbb{P}$ makes this result widely usable in applications, where assuming that one knows
 94 $\mathbb{P}$ is often not realistic.

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

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

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

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

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

108 ASSUMPTION 1.4 (non-degeneracy). *With probability 1, the scenario program*
 109 *that contains only the support constraints:*

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

112 *has the same solution x_N^* (possibly singled out by the same tie-break rule as for the*
 113 *initial program (1.1)) as program (1.1).* $\diamond$

114 This condition rules out situations where the boundary of various constraints
 115 group together anomalously so that if one of them is removed in isolation, then
 116 the solution does not change (and therefore this constraint is not of support) but
 117 a simultaneous removal of all the constraints that are not of support gives a new
 118 solution.³ Since the solution can be reconstructed from the support constraints,
 119 we can think that they “represent” the solution and the cardinality of the support

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

³Assumption 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 non-degeneracy 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 non-degeneracy condition here. In this paper, we have preferred to stay in the mainstream of Assumption 1.4 to avoid additional mathematical cluttering.

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., [51, 50, 47, 19, 15]. When this happens it comes spontaneous to ask whether a better result than (1.3) applies, and particularly whether in (1.3) d can be substituted by the actual number of support constraints, say h , and still obtain a valid bound *conditionally on seeing h support constraints*. In formal terms, this is written as

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

where s_N^* is a random variable that returns the number of support constraints. One important observation that motivates this study is that equation (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 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

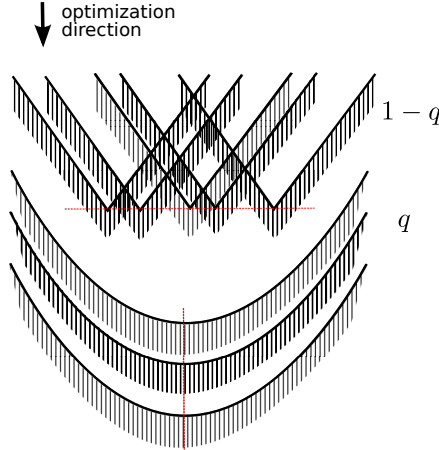


FIGURE 1. V-shaped and U-shaped constraints for Example 1.5.

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

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

153 This contradicts (1.4) because the Beta distribution on the right-hand side has a
 154 support that covers all $[0, 1]$, see Figure 2. Moreover, letting $q \rightarrow 0$, (1.5) shows that

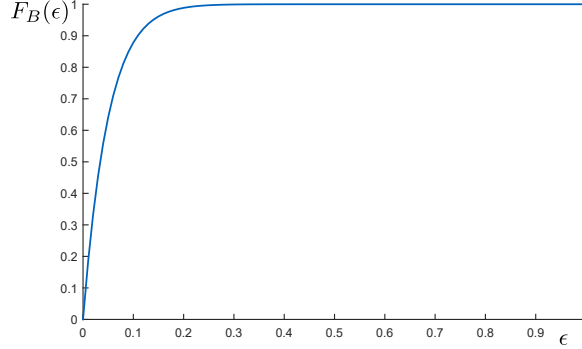


FIGURE 2. Cumulative distribution function of a $\text{Beta}(h, N - h + 1)$ distribution with $N = 20$ and $h = 1$.

154 $V(x_N^*)$ has conditional probability 0 of being less than a value ϵ that can be made
 155 close to 1 at will, an arbitrarily bad result.

156 Before closing, an additional fact is worth mentioning to shed light on how
 157 this example relates to the wait-and-judge theory of paper [15]. A simple com-
 158 putation shows that $s_N^* = 1$ is attained with probability $q^N + Nq^{N-1}(1 - q)$, so
 159 that when $q \simeq 0$ (corresponding to high violation conditional on $s_N^* = 1$) the
 160 probability that $s_N^* = 1$ tends to zero extremely fast with N . By inspecting [15],
 161 one sees that this fact is general: the main result of [15] proves that simulta-
 162 neously having small values of s_N^* and high violation occurs with very low prob-
 163 ability, that is, $\mathbb{P}^N\{\text{high } V(x_N^*) \wedge \text{small } s_N^*\} = \text{low value}$. On the other hand,
 164 $\mathbb{P}^N\{\text{high } V(x_N^*) \wedge \text{small } s_N^*\} = \mathbb{P}^N\{\text{high } V(x_N^*) \mid \text{small } s_N^*\} \cdot \mathbb{P}^N\{\text{small } s_N^*\}$, which
 165 shows that, if the probability of high violation conditional on seeing a small value of
 166 s_N^* is high, then it must be that observing that small value of s_N^* has low probab-
 167 ility. $\diamond$

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

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

184 On the other hand, one should not underestimate the (practical) importance of
 185 unconditional claims. Unconditional claims capture how well a scenario program
 186 performs on average over potential data sets, and independently of how data sets
 187 are generated. Hence, if we interpret the scenario program as an algorithm that
 188 maps data sets into decisions, an unconditional judgment quantifies how good the
 189 algorithm is, and this may drive us in ranking algorithms (this may also be relevant
 190

to selecting hyper-parameters). Therefore, conditional claims are in use to judge solutions, while unconditional claims allows 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 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. $\diamond$

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 of “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 $D \rightarrow \mathbb{P}_\vartheta(D)$ is a probability distribution;
- ii. $\forall D \in \mathcal{D}$, the map $\vartheta \rightarrow \mathbb{P}_\vartheta(D)$ 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 $(\Delta^N \times \Theta, \mathcal{D}^N \otimes \mathcal{Q}, \mathbb{P})$ where, for any $E \in \mathcal{D}^N \otimes \mathcal{Q}$, $\mathbb{P}(E)$ is defined by $\mathbb{P}(E) = \int_\Theta \mathbb{P}_\vartheta^N(E_{/\vartheta}) \pi(d\vartheta)$, 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 $\delta_1, \dots, \delta_N$ is an i.i.d. 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 $(\Delta^N \times \Theta, \mathcal{D}^N \otimes \mathcal{Q}, \mathbb{P})$. 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,

$$F_V(\epsilon | s_N^* = h) = \mathbb{P}\{V_\vartheta(x_N^*) \leq \epsilon | s_N^* = h\}.$$
⁴

⁴ It is important to note that, while distribution F_V is with respect to $\mathbb{P}$ (which involves

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 the next 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 $\{0, 1, \dots, d\}$ 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 $\dots$ 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 the same $s_N^* = h$. These results are formalized and precisely stated in the next section.

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

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

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

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

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

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

both $\mathbb{P}_\vartheta$ and π), $V_\vartheta(x)$ only involves $\mathbb{P}_\vartheta$. This is because we are interested in quantifying the robustness 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 $P\{V_\vartheta(x_N^*) \leq 0.1\} = 1$, 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 $V(x) = \int_{\Theta} \mathbb{P}_\vartheta\{\delta \in \Delta : x \notin \mathcal{X}_\delta\} \pi(d\vartheta)$.

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

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:

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

Next, define $p_k := \mathbb{P}_\vartheta^N\{s_N^* = k\}$, $k = 0, 1, \dots, d$, and let: $\mathbf{p} = (p_0, p_1, \dots, p_d) \in S$ where S is the simplex in $\mathbb{R}^{d+1}$ (i.e., $\sum_{k=0}^d p_k = 1$, $p_k \geq 0$, $k = 0, 1, \dots, d$). $\mathbf{p}$ 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 $\mathbf{p}$ induced by π , equation (2.3) now gives the following theorem.

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

$$(2.4) \quad F_V(\epsilon|s_N^* = h) \geq 1 - \frac{\int_{\{p_h \leq b_h(\epsilon)\}} p_h \pi'_h(dp_h) + b_h(\epsilon) \pi'_h\{p_h > b_h(\epsilon)\}}{\int_{[0,1]} p_h \pi'_h(dp_h)},$$

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

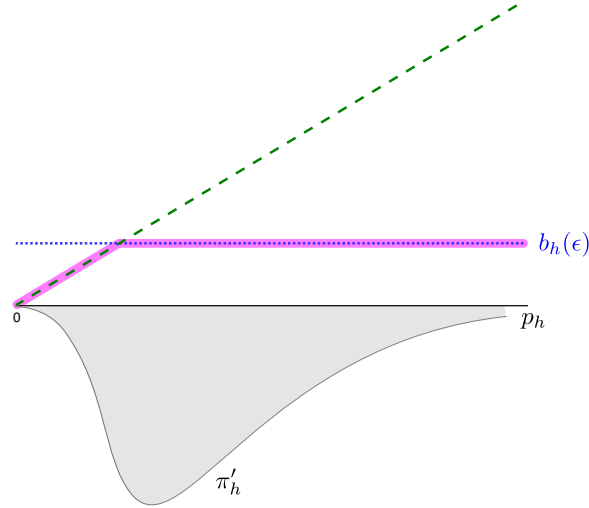


FIGURE 3. Visual interpretation of the second term in the right-hand side of (2.4): the numerator is obtained by integrating the pink solid line with respect to the probability distribution π'_h while the denominator, obtained by integration of p_h (green dashed line) with respect to π'_h , is the expected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

the right-hand side of (2.4) close to 1 for ϵ marginally bigger than h/N whenever 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 somehow articulated manner. First, in the next Section 2.1 we provide a quick approach to upper-bound $b_h(\epsilon)$ and also present an example (Example 2.5) that helps gain insight in equation (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 it has been previously discussed at the end of Section 1.2, the prior that is in use in this paper is π' (refer to equation (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 an exposure to the same (or a similar) problem in the past. This provides us with a ground for constructing the empirical frequency 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 Dirichelet 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 Dirichelet distribution of Section 4.* $\diamond$

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*

$$(2.5) \quad b_h(\epsilon) \leq \min \left\{ \binom{N}{h} (1 - \epsilon)^{N-h}, 1 \right\}.^8$$

$\diamond$

Proof of Theorem 2.3. Suppose that there are h support constraints and these are the first h constraints: $x \in \mathcal{X}_{\delta_1}, \dots, x \in \mathcal{X}_{\delta_h}$. Then, all other $N - h$ constraints must be satisfied by the solution x_N^* . If $V_\vartheta(x_N^*) > \epsilon$, this means that $\delta_{h+1}, \dots, \delta_N$ 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, $\mathbb{P}_\vartheta^N\{V_\vartheta(x_N^*) > \epsilon \wedge s_N^* =$

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

⁷The 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 to undergo a closer scrutiny; however, we shall not be engaged in discussing our philosophical positions in this publication, which is only geared towards establishing precise, and compelling, mathematical results in the context of decision making.

⁸As 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.

326 $h \wedge$ the support constraints are the first h constraints $\} \leq (1 - \epsilon)^{N-h}$. A similar
 327 argument applies to all other choices of h support constraints. Hence, summing
 328 over all possible choices of the h support constraints from a total of N constraints
 329 (which gives $\binom{N}{h}$ choices) one obtains:

$$\begin{aligned}
 330 \quad & \mathbb{P}_{\vartheta}^N \{V_{\vartheta}(x_N^*) > \epsilon \wedge s_N^* = h\} \\
 331 \quad &= \sum_{i=1}^{\binom{N}{h}} \mathbb{P}_{\vartheta}^N \left\{ V_{\vartheta}(x_N^*) > \epsilon \wedge s_N^* = h \wedge \begin{array}{l} \text{the support} \\ \text{constraints are the } i\text{-th group of } h \text{ constraints} \end{array} \right\} \\
 332 \quad & \\
 333 \quad &\leq \binom{N}{h} (1 - \epsilon)^{N-h},
 \end{aligned}$$

334 which holds true independently of ϑ . This establishes result (2.5). $\square$

335
 336 It is instructive to pause a second and reflect upon the meaning of equation
 337 (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,

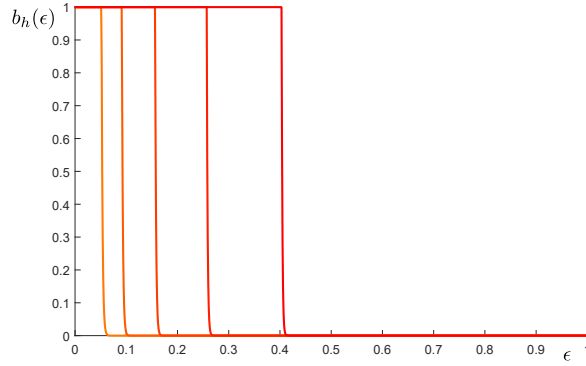


FIGURE 4. The upper bound to $b_h(\epsilon)$ in (2.5) for $N = 500$ and $h = 5, 10, 20, 40, 80$. As h grows, the function moves to the right.

338
 339 it can be noted that, beyond $\bar{\epsilon}_h$, $b_h(\epsilon)$ rapidly saturates down to 0. Hence, by using
 340 the relation $\mathbb{P}_{\vartheta}^N \{V_{\vartheta}(x_N^*) > \epsilon \wedge s_N^* = h\} = \mathbb{P}_{\vartheta}^N \{V_{\vartheta}(x_N^*) > \epsilon \mid s_N^* = h\} \cdot \mathbb{P}_{\vartheta}^N \{s_N^* = h\}$,
 341 one draws the conclusion that, for a value of ϵ even moderately bigger than $\bar{\epsilon}_h$, either
 342 $V_{\vartheta}(x_N^*) > \epsilon$ occurs very rarely conditionally on seeing $s_N^* = h$, or seeing $s_N^* = h$
 343 must itself be a rare event. Referring back to Example 1.5, suppose that $q = 0.1$
 344 and $N = 20$. For $h = 1$, the value of $\bar{\epsilon}_1$ can be calculated to be 0.18. In Example
 345 1.5, the event $s_N^* = 1$ had probability equal to $0.1^{20} + 20 \cdot 0.1^{19} \cdot 0.9 = 1.81 \cdot 10^{-18}$,
 346 which is extremely low, and this was the circumstance that made it possible to have
 347 always a large violation, of at least 0.9, when $s_N^* = 1$.

348 **REMARK 2.4** (sensitivity to the prior). Referring again to Figure 4, pick a
 349 value of h and suppose for a moment that the transition from 1 to 0 of function
 350 $b_h(\epsilon)$ happens instantaneously for a single value $\bar{\epsilon}$ (in other words, $b_h(\epsilon)$ is a step
 351 function). Then, the right-hand side of equation (2.4) gives value 0 when $\epsilon < \bar{\epsilon}$
 352 and value 1 when $\epsilon > \bar{\epsilon}$, independently of the prior π_h' . Since the transition of
 353 $b_h(\epsilon)$ occurs abruptly but not instantaneously, this result holds only approximately.
 354 On the other hand, after the transition, $b_h(\epsilon)$ does go down to very low values
 355 quite rapidly. Hence, a significant departure from the above description can only
 356 occur when the prior π_h' all nests at very low values, below that of $b_h(\epsilon)$, as shown
 357 again by an inspection of equation (2.4): for the sake of the argument, say that all

358 π'_h concentrates below the value of $b_h(\epsilon)$ for some value of ϵ beyond the transition
 359 point, then one easily sees that the right-hand side of equation (2.4) still keeps the
 360 value 0 for that ϵ . $\diamond$

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

365 **EXAMPLE 2.5.** Consider again Example 1.5 with $N = 20$. Suppose that q is
 366 uncertain with only two possible values: $q = 0.1$ or $q = 1$. In the former, $p_1 =$
 367 $\mathbb{P}^N\{s_N^* = 1\} = 0.1^{20} + 20 \cdot 0.1^{19} \cdot 0.9 = 1.81 \cdot 10^{-18}$, while in the latter $p_1 = 1$.
 368 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

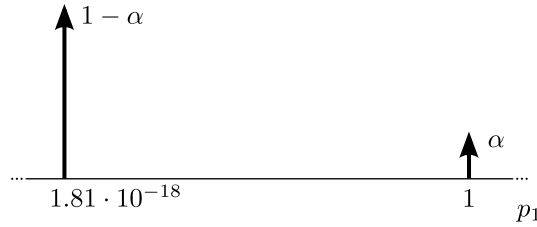


FIGURE 5. The prior π'_1 : $p_1 = 1.81 \cdot 10^{-18}$ with probability $1 - \alpha$ while $p_1 = 1$ with probability α .

369 of this very low value of α is that we bear a strong a priori belief in favor of the
 370 case $q = 0.1$ (in which case seeing 1 support constraint is extremely rare) while
 371 we do not completely exclude the case $q = 1$ (in which case one systematically
 372 sees 1 support constraint). Using formula (2.5) with $h = 1$ and $N = 20$ gives
 373 $b_1(\epsilon) \leq \min\{N(1 - \epsilon)^{N-1}, 1\} = \min\{20(1 - \epsilon)^{19}, 1\}$, which substituted in equation
 374 (2.4) yields
 375

$$\begin{aligned}
 376 \quad F_V(\epsilon | s_{20}^* = 1) &\geq 1 - \frac{\int_{\{p_1 \leq \min\{20(1-\epsilon)^{19}, 1\}\}} p_1 \pi'_1(dp_1)}{1.81 \cdot 10^{-18} \cdot (1 - 10^{-7}) + 1 \cdot 10^{-7}} \\
 377 \quad &\quad - \frac{\min\{20(1 - \epsilon)^{19}, 1\} \cdot \pi'_1\{p_1 > \min\{20(1 - \epsilon)^{19}, 1\}\}}{1.81 \cdot 10^{-18} \cdot (1 - 10^{-7}) + 1 \cdot 10^{-7}} \\
 378 \quad (2.6) \quad &= \begin{cases} 1 - \frac{1.81 \cdot 10^{-18} \cdot (1 - 10^{-7}) + \min\{20(1-\epsilon)^{19}, 1\} \cdot 10^{-7}}{1.81 \cdot 10^{-18} \cdot (1 - 10^{-7}) + 1 \cdot 10^{-7}}, & \text{if } \epsilon \leq 0.9005 \\ 1 - \frac{\min\{20(1-\epsilon)^{19}, 1\}}{1.81 \cdot 10^{-18} \cdot (1 - 10^{-7}) + 1 \cdot 10^{-7}}, & \text{if } \epsilon > 0.9005. \end{cases}
 \end{aligned}$$

379 The right-hand side of (2.6) is profiled in Figure 6 against the result which is ob-
 380 tained for the case in which $\alpha = 1$. There is very little difference between the two
 381 curves despite that the prior is substantially different in the two cases ($\alpha = 10^{-7}$
 382 against $\alpha = 1$). The reason is that for $\alpha = 10^{-7}$ one has an a priori strong belief
 383 in favor of $q = 0.1$ but, after seeing $s_{20}^* = 1$ (which is extremely rare when $q = 0.1$,
 384 with probability $1.81 \cdot 10^{-18}$) the prior is reversed into a posterior belief that favors
 385 $q = 1$, so leveling the two cases $\alpha = 10^{-7}$ and $\alpha = 1$.

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

$$390 \quad (2.7) \quad F_V(\epsilon | s_{20}^* = 1) \geq \begin{cases} 0, & \text{if } \epsilon \leq 0.9005 \\ 1 - \frac{\min\{20(1-\epsilon)^{19}, 1\}}{1.81 \cdot 10^{-18}}, & \text{if } \epsilon > 0.9005. \end{cases}$$

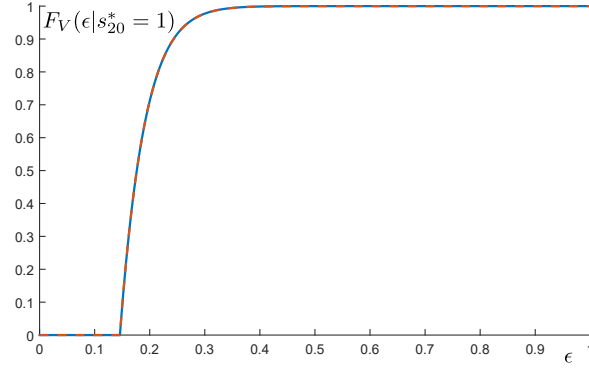


FIGURE 6. Bounds for $F_V(\epsilon | s_{20}^* = 1)$ for $\alpha = 10^{-7}$ (blue solid line) and $\alpha = 1$ (dashed red line). The two curves are almost indistinguishable.

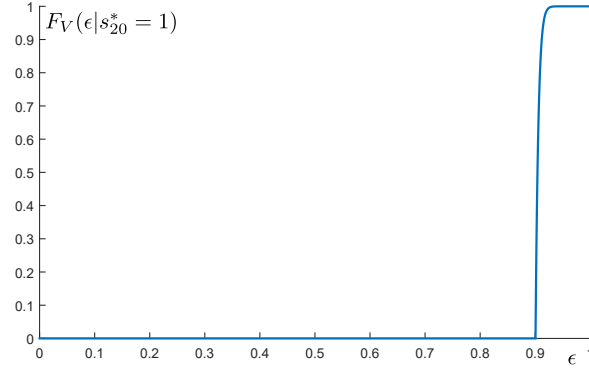


FIGURE 7. Bound for $F_V(\epsilon | s_{20}^* = 1)$ for $\alpha = 0$.

391 The right-hand side of (2.7) is profiled in Figure 7. Note that the curve keeps at
 392 zero value till 0.9005, which means that the theory does not provide any meaningful
 393 lower bound on the probability of the event $V_{\vartheta}(x_N^*) \leq 0.9005$ conditionally on seeing
 394 $s_{20}^* = 1$. This is in agreement with previous remarks in Example 1.5 where indeed
 395 seeing $s_{20}^* = 1$ implied that $V_{\vartheta}(x_N^*) \geq 0.9$ when $q = 0.1$. It is also interesting to note
 396 that beyond $\epsilon = 0.9005$ the conditional distribution $F_V(\epsilon | s_{20}^* = 1)$ saturates quickly
 397 to 1. $\diamond$

398 **2.2. A Bayesian interpretation.** While discussing the subject matter of this
 399 article with some eminent Bayesian statisticians, the present authors were prompted
 400 to more explicitly highlight the essentially Bayesian character of its content. This
 401 short section provides an explanation by showing how the various quantities involved
 402 can be rewritten in terms of the posterior distribution of ϑ given the data points.

403 Referring back to equation (2.1), the sets $\{V_{\vartheta}(x_N^*) > \epsilon\}$ and $\{s_N^* = h\}$ are
 404 events on $\Delta^N \times \Theta$ (see Figure 8 for a representation). The integrals in equation

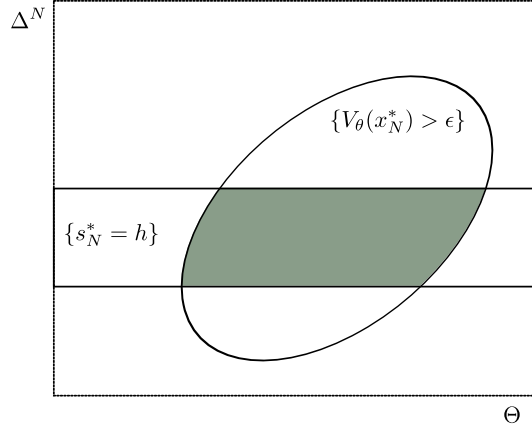


FIGURE 8. Representation of events $\{V_\vartheta(x_N^*) > \epsilon\}$ and $\{s_N^* = h\}$ on $\Delta^N \times \Theta$. The shaded region is $\{V_\vartheta(x_N^*) > \epsilon \wedge s_N^* = h\}$.

(2.1) can also be written as

$$\begin{aligned}
 & \int_{\Theta} \mathbb{P}_{\vartheta}^N \{V_{\vartheta}(x_N^*) > \epsilon \wedge s_N^* = h\} \pi(d\vartheta) \\
 &= \int_{\Delta^N} \pi_{/\delta_1, \dots, \delta_N} \{V_{\vartheta}(x_N^*) > \epsilon \wedge s_N^* = h\} P_{\Delta^N}(d\delta_1, \dots, d\delta_N) \\
 & \int_{\Theta} \mathbb{P}_{\vartheta}^N \{s_N^* = h\} \pi(d\vartheta) \\
 &= \int_{\Delta^N} \pi_{/\delta_1, \dots, \delta_N} \{s_N^* = h\} P_{\Delta^N}(d\delta_1, \dots, d\delta_N),
 \end{aligned}$$

where $\pi_{/\delta_1, \dots, \delta_N}$ 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 $\pi_{/\delta_1, \dots, \delta_N}$ 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 equations (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 evaluation of $b_h(\epsilon)$ and in this section we obtain an upper bound for $b_h(\epsilon)$ that is provably close to the best possible. Quantitatively, the final result outperforms, even significantly, the evaluations made with equation (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*

426 true for any probability $\mathbb{P}_\vartheta$, $\vartheta \in \Theta$, it holds that ⁹
 (3.1)

$$427 \quad b_h(\epsilon) \leq \begin{cases} \sum_{i=0}^{d-1} \binom{N}{i} \epsilon^i (1-\epsilon)^{N-i}, & \text{if } h = d \\ \sum_{i=0}^{d-1} \binom{N}{i} \left(\frac{d-1}{h}\epsilon\right)^i \left(1 - \frac{d-1-h}{h} \frac{\epsilon}{1-\epsilon}\right)^{N-i} \\ \quad + \sum_{i=d-h}^{N-h} \binom{N-h}{i} \left(\frac{d-1-h}{h} \frac{\epsilon}{1-\epsilon}\right)^i \left(1 - \frac{d-1-h}{h} \frac{\epsilon}{1-\epsilon}\right)^{N-h-i}, & \text{if } h \leq d-1 \\ \quad \times \frac{\binom{N}{h}}{\binom{d-1}{h}} \epsilon^h (1-\epsilon)^{N-h} \frac{(d-1)^{(d-1)}}{(d-1-h)^{(d-h-1)} h^h}, & \text{and } \epsilon \leq \frac{h}{d-1} \\ \frac{\binom{N}{h}}{\binom{d-1}{h}} (1-\epsilon)^{N-d+1}, & \text{if } h \leq d-1 \\ & \text{and } \epsilon > \frac{h}{d-1}. \end{cases}$$

428

◇

429 Before proving the theorem, we make some remarks about the result. Figure 9
 430 profiles bound (3.1) for $N = 500$, $d = 150$, and $h = 10, 20, 40$ against the bound in
 431 (2.5), which shows that the new bound marks a significant improvement. Moreover,
 432 it is a notable fact that there is little margin of further improvement over bound
 433 (3.1) provided that Θ is rich enough. This claim can be justified by making reference
 434 to the theory developed in paper [12]: suppose that Θ contains problems with h
 435 support constraints with probability 1,¹⁰ then the theory in [12] proves that for
 436 these problems it holds that

$$437 \quad (3.2) \quad \mathbb{P}^N \{V(x_N^*) > \epsilon \wedge s_N^* = h\} = \sum_{i=0}^{h-1} \binom{N}{i} \epsilon^i (1-\epsilon)^{N-i},$$

438 so that the right-hand side of (3.2) represents an insurmountable bound for $b_h(\epsilon)$.
 439 The tightness of (3.1) can be appreciated in Figure 9 where the right-hand side of
 440 (3.2) is also represented. Note that the value of ϵ for which the right-hand side of
 441 (3.2) equals 0.5 is approximately h/N , see, e.g., [42], so that bound (3.1) rapidly
 442 saturates to 0 for values of ϵ larger than h/N .

443

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

$$446 \quad \mathbb{P}_\vartheta^N \{V_\vartheta(x_N^*) > \epsilon \wedge s_N^* = h\} = \sum_{k=0}^d \mathbb{P}_\vartheta^N \{V_\vartheta(x_N^*) > \bar{\epsilon}(k) \wedge s_N^* = k\}$$

$$447 \quad = \mathbb{P}_\vartheta^N \{V_\vartheta(x_N^*) > \bar{\epsilon}(s_N^*)\}.$$

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

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

where $I_x(a, b) := \int_0^x \frac{\Gamma(a+b)}{\Gamma(a)\Gamma(b)} t^{a-1} (1-t)^{b-1} dt$ 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 $I_x(a, b)$ can be efficiently computed in various scientific computing environments (for example, in MATLAB, $I_x(a, b)$ 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.

¹⁰In [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.

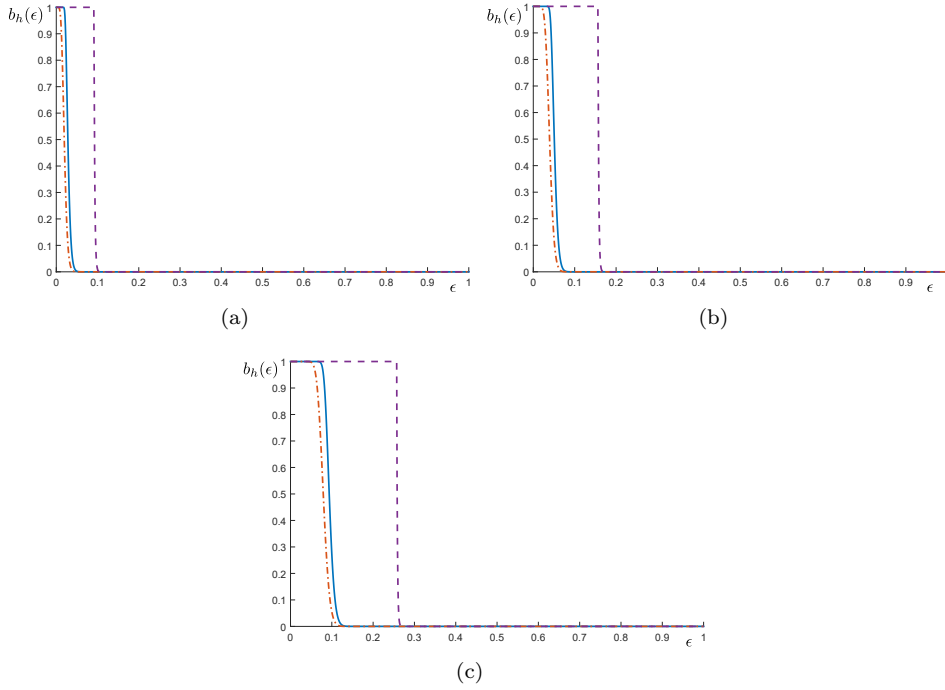


FIGURE 9. Bound (3.1) (solid blue line) vs. bound (2.5) (dashed purple line) and the unsurmountable lower bound $\sum_{i=0}^{h-1} \binom{N}{i} \epsilon^i (1-\epsilon)^{N-i}$ (dashed-dotted red line). $N = 500$, $d = 150$ and (a) $h = 10$; (b) $h = 20$; (c) $h = 40$.

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

449
$$\mathbb{P}_{\vartheta}^N \{V_{\vartheta}(x_N^*) > \bar{\epsilon}(s_N^*)\} \leq \gamma_h^*,$$

450 where

451 (3.3a)
$$\gamma_h^* := \inf_{\xi(\cdot) \in C^d[0,1]} \xi(1)$$

452 (3.3b)
$$\text{subject to } \frac{1}{h!} \frac{d^h}{dt^h} \xi(t) \geq \binom{N}{h} t^{N-h} \cdot \mathbf{1}_{[0,1-\epsilon)}, \quad t \in [0, 1]$$

453 (3.3c)
$$\frac{1}{k!} \frac{d^k}{dt^k} \xi(t) \geq 0, \quad t \in [0, 1], \quad k = 0, 1, \dots, d, \quad k \neq h$$

and $C^d[0, 1]$ is the set of continuous function 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

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

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

456 (3.4)
$$\tau = \begin{cases} 1 - \epsilon, & \text{if } h = d \\ \max\{0, 1 - \frac{d-1}{h}\epsilon\}, & \text{if } h \leq d-1, \end{cases}$$

457 and
 (3.5)

$$458 \quad A = \begin{cases} 0, & \text{if } h = d \\ \frac{\binom{N}{h}(1-\epsilon)^{N-h}}{d\binom{d-1}{h}(1-\epsilon-\tau)^{d-h-1}} \cdot \sum_{i=d-h}^{N-h} \binom{N-h}{i} \left(1 - \frac{\tau}{1-\epsilon}\right)^i \left(\frac{\tau}{1-\epsilon}\right)^{N-h-i}, & \text{if } h \leq d-1. \end{cases}$$

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

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

$$461 \quad \frac{1}{k!} \frac{d^k}{dt^k} \bar{\xi}(t) = \binom{N}{k} t^{N-k} \cdot \mathbf{1}_{[0,\tau)} + \sum_{i=k}^{d-1} \binom{N}{i} \binom{i}{k} (t-\tau)^{i-k} \tau^{N-i} \cdot \mathbf{1}_{[\tau,1]} \\ 462 \quad + d \binom{d-1}{k} A (t-\tau)^{d-k-1} \cdot \mathbf{1}_{[\tau,1]},$$

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

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

463 where δ is the Dirac delta function. It is thus a fact that $\bar{\xi}(t)$ is not in $C^d[0,1]$,
 464 however, as we shall show later in the proof, this difficulty can be circumvented by
 465 a small modification of $\bar{\xi}(t)$. For now we concentrate on showing that $\bar{\xi}(t)$ satisfies
 466 the constraints (3.3b) and (3.3c). Constraint (3.3c) is clearly satisfied because
 467 $\frac{1}{k!} \frac{d^k}{dt^k} \bar{\xi}(t)$ 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,

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

468 which satisfies (3.3b).

469 For $h \leq d-1$, (3.3b) is clearly satisfied for $t \in [0, \tau)$ because on this interval
 470 $\frac{1}{h!} \frac{d^h}{dt^h} \bar{\xi}(t)$ coincides with $\binom{N}{h} t^{N-h}$. For $t \in [1-\epsilon, 1]$, (3.3b) is also satisfied because
 471 $\frac{1}{h!} \frac{d^h}{dt^h} \bar{\xi}(t) \geq 0$. Over the interval $[\tau, 1-\epsilon)$, substituting the expression for A given
 472 in (3.5) in $\frac{1}{h!} \frac{d^h}{dt^h} \bar{\xi}(t)$ gives

$$473 \quad \frac{1}{h!} \frac{d^h}{dt^h} \bar{\xi}(t) = \sum_{i=h}^{d-1} \binom{i}{h} \binom{N}{i} (t-\tau)^{i-h} \tau^{N-i} \\ 474 \quad + \frac{\binom{N}{h}(1-\epsilon)^{N-h}}{(1-\epsilon-\tau)^{d-h-1}} \sum_{i=d-h}^{N-h} \binom{N-h}{i} \left(1 - \frac{\tau}{1-\epsilon}\right)^i \left(\frac{\tau}{1-\epsilon}\right)^{N-h-i} (t-\tau)^{d-h-1}. \\ 475 \quad (3.6)$$

476 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

477 letting $j = i - h$ in the first sum, (3.6) can also be written as

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

482 where the second equality follows by a simple re-arrangement of the terms in the
 483 second sum. Since the above expression is in use for $t \in [\tau, 1-\epsilon]$, it holds that

$$484 \quad \left(\frac{1-\epsilon-\tau}{t-\tau} \right)^{i-(d-h-1)} \geq 1. \text{ This yields the inequality}$$

$$\begin{aligned}
 485 \quad \frac{1}{h!} \frac{d^h}{dt^h} \bar{\xi}(t) &\geq \binom{N}{h} \left[\sum_{j=0}^{d-h-1} \binom{N-h}{j} (t-\tau)^j \tau^{N-h-j} + \sum_{i=d-h}^{N-h} \binom{N-h}{i} (t-\tau)^i \tau^{N-h-i} \right] \\
 486 \quad &= \binom{N}{h} t^{N-h} \quad (\text{where we have used the binomial theorem}). \\
 487
 \end{aligned}$$

488 This shows that (3.3b) is satisfied.

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

$$\begin{aligned}
 492 \quad f_n(t) &= \binom{N}{d} t^{N-d} \cdot \mathbf{1}_{[0, \tau)} + A \left[n^2(t-\tau+2/n) \mathbf{1}_{[\tau-\frac{2}{n}, \tau-\frac{1}{n})} - n^2(t-\tau) \mathbf{1}_{[\tau-\frac{1}{n}, \tau]} \right] \\
 493 \quad &\quad + \binom{N}{d} \tau^{N-d} (1-n(t-\tau)) \cdot \mathbf{1}_{[\tau, \tau+\frac{1}{n}]},
 \end{aligned}$$

which provide continuous approximations of $\frac{1}{d!} \frac{d^d}{dt^d} \bar{\xi}(t)$ (see Figure 10). For each n ,

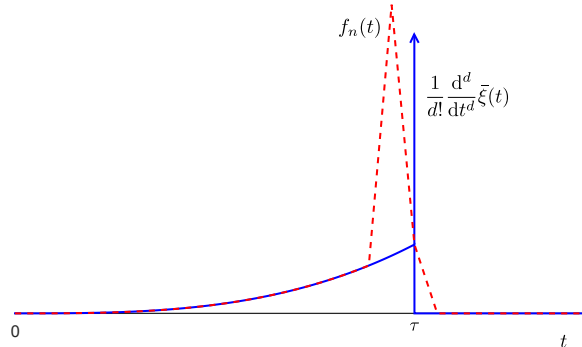


FIGURE 10. $\frac{1}{d!} \frac{d^d}{dt^d} \bar{\xi}(t)$ (solid blue line) vs. $f_n(t)$ (red dashed line).

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

$$\bar{\xi}_n(t) = \int_0^t \int_0^{s_1} \cdots \int_0^{s_{d-1}} f_n(s_d) ds_d \cdots ds_1$$

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

497 To conclude the proof one has to show that $\bar{\xi}(1)$ equals the right-hand side of
 498 (3.1), which is a cumbersome, but straightforward, computation that is left to the
 499 reader. $\square$

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

Suppose therefore that

$$\pi' = \text{Dir}(\alpha_0, \alpha_1, \dots, \alpha_d),$$

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

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

507 ($\Gamma(\cdot)$ denotes the Gamma function) and further p_0 remains determined by relation
 508 $p_0 = 1 - \sum_{k=1}^d p_k$. Coefficients $\alpha_0 > 0, \alpha_1 > 0, \dots, \alpha_d > 0$ are free parameters
 that can be selected by the user. Figure 11 shows the density of some Dirichlet

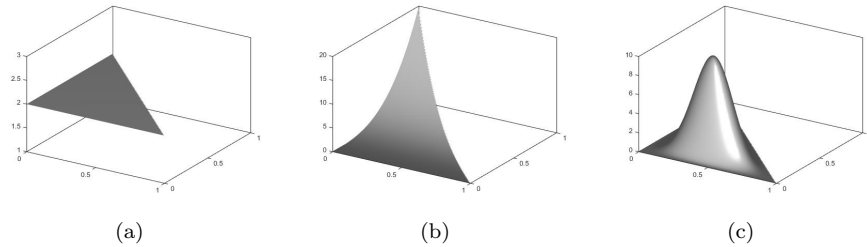


FIGURE 11. For $d = 2$, density of p_1, p_2 for a Dirichlet distribution with: (a) $\alpha_0 = \alpha_1 = \alpha_2 = 1$, (b) $\alpha_0 = \alpha_1 = 1, \alpha_2 = 4$, (c) $\alpha_0 = \alpha_1 = \alpha_2 = 10$.

509 distributions obtained by specific choices of the free parameters for $d = 2$.
 510

511 As is the case for panel (a) of Figure 11, the choice $\alpha_0 = \alpha_1 = \dots = \alpha_d = 1$ gives
 512 for any d a flat (uniform) distribution, which corresponds to adopting a “principle
 513 of indifference” (see Remark 2.2). Instead, for $\alpha_k > 1, \forall k$, the Dirichlet distribution

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

514 becomes uni-modal with mean $(\alpha_0/\sum_k \alpha_k, \alpha_1/\sum_k \alpha_k, \dots, \alpha_d/\sum_k \alpha_k)$. Moreover,
 515 the bigger $\sum_k \alpha_k$, the more concentrated the distribution around its mean. This
 516 may be used to accommodate the case in which one wants to build a prior around
 517 an empirical frequency for the complexity (see again Remark 2.2). The reader
 518 interested in shaping a Dirichlet distribution beyond these simple rules can consult
 519 any statistical textbook for further information.

We next show how a Dirichlet prior can be used in equation (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:

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

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 $\text{Beta}(a, b)$
 (which will be used in subsequent derivations) are that its mean is $a/(a+b)$ while
 its cumulative distribution function

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

520 is the regularized incomplete Beta function $I_x(a, b)$, which can be easily evaluated,
 521 e.g., via the `betainc` command in MATLAB (we have already encountered $I_x(a, b)$
 522 in Section 3).

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

$$525 \quad (4.1) \quad \int_{[0,1]} p_h \pi'_h(dp_h) = \frac{\alpha_h}{\sum_{k=0}^d \alpha_k}.$$

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

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

and

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

(4.3)

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 $\pi' = \text{Dir}(\alpha_0, \alpha_1, \dots, \alpha_d)$, then equation (2.4) becomes*

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

(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 the following Section 5.

EXAMPLE 4.2. *Let $N = 50$, $d = 10$ and $\alpha_k = 1$ for all $k = 0, \dots, 10$ 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*

$$(4.5) \quad F_V(\epsilon | s_{50}^* = 4) \geq 1 - [I_{b_4(\epsilon)}(2, 10) + b_4(\epsilon) \cdot 11 (1 - I_{b_4(\epsilon)}(1, 10))],$$

which, using equation (3.1) for $b_h(\epsilon)$, is the distribution represented in Figure 12(b) (solid blue line). As we can see, seeing 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,

$$(4.6) \quad F_V(\epsilon | s_{50}^* = 4) \geq 1 - [I_{b_4(\epsilon)}(6, 10) + b_4(\epsilon) \cdot 3 (1 - I_{b_4(\epsilon)}(5, 10))],$$

¹²As 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.

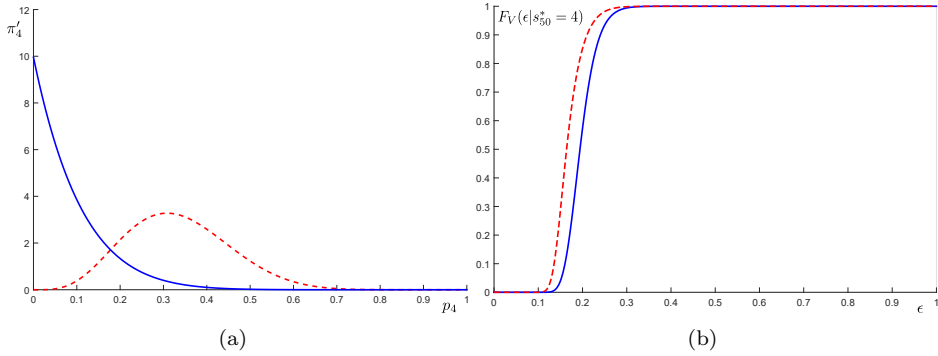


FIGURE 12. Left: π'_4 . Right: $F_V(\epsilon|s_{50}^* = 4)$ (solid blue line: $\alpha_k = 1$, $k = 0, \dots, 10$; dashed red line: $\alpha_k = 1$ for $k \neq 4$, $\alpha_4 = 5$).

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 shows that the posterior bears little sensitivity on the prior. $\diamond$

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., $\alpha_k = 1$ for all 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

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

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 $\mathbb{P}\{s_{2500}^* = 10\} = 1/11$ (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 for all k and d is left free, equation (4.4) becomes

$$(4.7) \quad F_V(\epsilon|s_{2500}^* = 10) \geq 1 - [\mathbf{I}_{b_h(\epsilon)}(2, d) + b_h(\epsilon) \cdot (d + 1) (1 - \mathbf{I}_{b_h(\epsilon)}(1, d))],$$

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

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

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 $\mathbb{P}\{s_{2500}^* = 10\} = 1/(1 + 10^3)$. 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 that for $d = 10^3$ we have a strong belief in that the probability of violation is below the value 0.015. $\diamond$

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, among which we select the following 25 non-categorical 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 $\delta_1, \delta_2, \dots, \delta_{500}$ (random sorting is used to cancel any possible ordering introduced in the data set at the time it has been uploaded in the repository). $\delta_1, \delta_2, \dots, \delta_{500}$ 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 the 500 scenarios:

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

where $v = [v_1, v_2, \dots, v_{25}]^T$ are the coordinates of the box vertex with left-most components, $\ell = [\ell_1, \ell_2, \dots, \ell_{25}]^T$ are the lengths of the box edges, and inequalities are meant component-wise. 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 $\text{Dir}(1, 1, \dots, 1)$.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 belonged to the box.¹⁶ For estimating $F_V(\epsilon | s_{500}^* = 13)$, we only kept the

¹³<http://archive.ics.uci.edu/ml/datasets/Corel+Image+Features>

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

¹⁵Notice that a probability $\mathbb{P}_\theta$ 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.

¹⁶This 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.

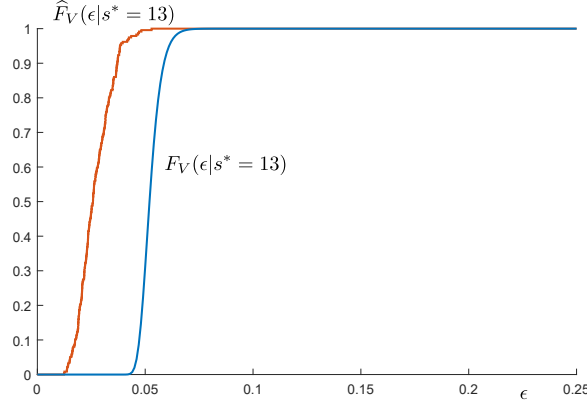


FIGURE 13. The bound to $F_V(\epsilon | s_{500}^* = 13)$ (continuous blue curve) vs. $\hat{F}_V(\epsilon | s_{500}^* = 13)$ (stairwise red curve).

636 runs that gave $s_{500}^* = 13$ and computed $\hat{F}_V(\epsilon | s_N^* = 13) = (\text{no. of cases with } \hat{v} \leq$
 637 $\epsilon \text{ and } s_{500}^* = 13) / (\text{no. of cases with } s_{500}^* = 13)$. The obtained $\hat{F}_V(\epsilon | s_N^* = 13)$ is
 638 depicted in Figure 13 as a stairwise red curve. Empirical evidence is an agreement
 639 with the expectation given by the theorem.

640 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

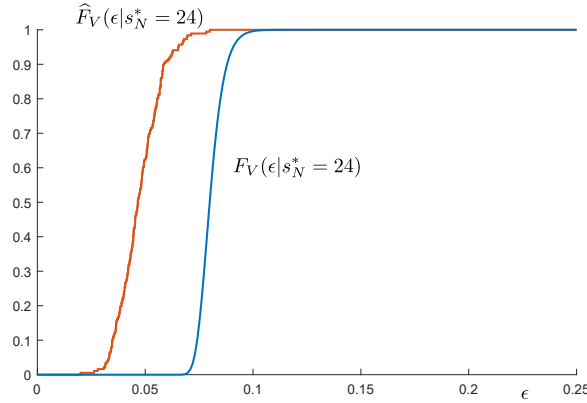


FIGURE 14. The bound to $F_V(\epsilon | s_N^* = 24)$ (continuous blue curve) vs. $\hat{F}_V(\epsilon | s_N^* = 24)$ (stairwise red curve).

641 depicts curves similar to Figure 13 for $s_{500}^* = 24$.

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

653 **Acknowledgments.** We would like to gratefully thank Professors E. Regazz-
 654 ini, L. Ladelli and A. Guglielmi for fruitful discussions.

REFERENCES

- [1] T. Alamo, R. Tempo, and E.F. Camacho. A randomized strategy for probabilistic solutions of uncertain feasibility and optimization problems. *IEEE Transactions on Automatic Control*, 54(11):2545–2559, 2009.
- [2] T. Alamo, R. Tempo, A. Luque, and D. R. Ramirez. Randomized methods for design of uncertain systems: sample complexity and sequential algorithms. *Automatica*, 51:160–172, 2015.
- [3] J.R. Barra. *Mathematical Basis of Statistics*. Academic Press, New York, NY, USA, 1981.
- [4] H. Bauer. *Probability theory*. De Gruyter, Berlin, 1996.
- [5] J.O. Berger. The robust Bayesian viewpoint. In J. Kadane, editor, *Robustness of Bayesian analysis*. North-Holland, Amsterdam, 1984.
- [6] J.O. Berger. Robust Bayesian analysis: sensitivity to the prior. *J. Statist. Plan. Inform.*, 25:303–328, 1990.
- [7] G.C. Calafiore and M.C. Campi. Uncertain convex programs: randomized solutions and confidence levels. *Mathematical Programming*, 102(1):25–46, 2005.
- [8] G.C. Calafiore and M.C. Campi. The scenario approach to robust control design. *IEEE Transactions on Automatic Control*, 51(5):742–753, 2006.
- [9] M.C. Campi. Classification with guaranteed probability of error. *Machine Learning*, 80:63–84, 2010.
- [10] M.C. Campi, G. Calafiore, and S. Garatti. Interval predictor models: identification and reliability. *Automatica*, 45(2):382–392, 2009.
- [11] M.C. Campi and A. Carè. Random convex programs with L_1 -regularization: sparsity and generalization. *SIAM Journal on Control and Optimization*, 51(5):3532–3557, 2013.
- [12] M.C. Campi and S. Garatti. The exact feasibility of randomized solutions of uncertain convex programs. *SIAM Journal on Optimization*, 19(3):1211–1230, 2008.
- [13] M.C. Campi and S. Garatti. A sampling-and-discarding approach to chance-constrained optimization: feasibility and optimality. *Journal of Optimization Theory and Applications*, 148(2):257–280, 2011.
- [14] M.C. Campi and S. Garatti. *Introduction to the Scenario Approach*. MOS-SIAM Series on Optimization, 2018.
- [15] M.C. Campi and S. Garatti. Wait-and-judge scenario optimization. *Mathematical Programming*, 167(1):155–189, 2018.
- [16] M.C. Campi and S. Garatti. A theory of the risk for optimization with relaxation and its application to support vector machines. *Journal of Machine Learning Research*, 22(288):1–38, 2021.
- [17] M.C. Campi, S. Garatti, and M. Prandini. The scenario approach for systems and control design. *Annual Reviews in Control*, 33(2):149 – 157, 2009.
- [18] M.C. Campi, S. Garatti, and F.A. Ramponi. A general scenario theory for non-convex optimization and decision making. *IEEE Transactions on Automatic Control*, 63(12):4067–4078, 2018.
- [19] A. Carè, S. Garatti, and M.C. Campi. FAST - Fast Algorithm for the Scenario Technique. *Operations Research*, 62(3):662–671, 2014.
- [20] A. Carè, S. Garatti, and M.C. Campi. Scenario min-max optimization and the risk of empirical costs. *SIAM Journal on Optimization*, 25(4):2061–2080, 2015.
- [21] A. Carè, F.A. Ramponi, and M.C. Campi. A new classification algorithm with guaranteed sensitivity and specificity for medical applications. *IEEE Control Systems Letters*, 2(3):393–398, 2018.
- [22] L. G. Crespo, S. P. Kenny, and D. P. Giesy. Interval predictor models with a linear parameter dependency. *Journal of Verification, Validation and Uncertainty Quantification*, 1(2):1–10, 2016.
- [23] L.G. Crespo, B.K. Colbert, S.P. Kenny, and D.P. Giesy. On the quantification of aleatory and epistemic uncertainty using sliced-normal distributions. *Systems and Control Letters*, 134, 2019.
- [24] L.G. Crespo, D. Giesy, S. Kenny, and J. Deride. A scenario optimization approach to system identification with reliability guarantees. In *Proceedings of the 2019 American Control Conference (ACC)*, pages 2100–2106, Philadelphia, USA, 2019.
- [25] L.G. Crespo, D.P. Giesy, and S.P. Kenny. Interval predictor models with a formal characterization of uncertainty and reliability. In *Proceedings of the 53rd IEEE Conference on Decision and Control (CDC)*, pages 5991–5996, Los Angeles, CA, USA, 2014.
- [26] L.G. Crespo, S.P. Kenny, and D.P. Giesy. Random predictor models for rigorous uncertainty quantification. *International Journal for Uncertainty Quantification*, 5(5):469–489, 2015.
- [27] L.G. Crespo, S.P. Kenny, D.P. Giesy, R.B. Norman, and S.R. Blattnig. Application of interval predictor models to space radiation shielding. In *Proceedings of the 18th AIAA Non-Deterministic Approaches Conference*, San Diego, CA, USA, 2016.
- [28] P. Mohajerin Esfahani, T. Sutter, and J. Lygeros. Performance bounds for the scenario

- approach and an extension to a class of non-convex programs. *IEEE Transactions on Automatic Control*, 60(1):46–58, 2015.
- [29] S. Garatti and M.C. Campi. Modulating robustness in control design: principles and algorithms. *IEEE Control Systems Magazine*, 33(2):36–51, 2013.
- [30] S. Garatti and M.C. Campi. Risk and complexity in scenario optimization. *Mathematical Programming*, 191(1):243–279, 2022.
- [31] S. Garatti, M.C. Campi, and A. Caré. On a class of interval predictor models with universal reliability. *Automatica*, 110, 2019. Article 108542.
- [32] S. Grammatico, X. Zhang, K. Margellos, P.J. Goulart, and J. Lygeros. A scenario approach for non-convex control design. *IEEE Transactions on Automatic Control*, 61(2):334–345, 2016.
- [33] L.J. Hong, Z. Hu, and G. Liu. Monte Carlo methods for value-at-risk and conditional value-at-risk: a review. *ACM Transactions on Modeling and Computer Simulation*, 24(4):22:1–22:37, 2014.
- [34] M.J. Lacerda and L. G. Crespo. Interval predictor models for data with measurement uncertainty. In *Proceedings of the 2017 American Control Conference*, pages 1487–1492, Seattle, WA, USA, 2017.
- [35] M. Lavine. An approach to robust Bayesian analysis for multidimensional parameter spaces. *J. Amer. Statist. Assoc.*, 86:400–403, 1991.
- [36] K. Margellos, P.J. Goulart, and J. Lygeros. On the road between robust optimization and the scenario approach for chance constrained optimization problems. *IEEE Transactions on Automatic Control*, 59(8):2258–2263, 2014.
- [37] K. Margellos, M. Prandini, and J. Lygeros. On the connection between compression learning and scenario based single-stage and cascading optimization problems. *IEEE Transactions on Automatic Control*, 60(10):2716–2721, 2015.
- [38] H.A. 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 Transactions on Control Systems Technology*, 27(4):1397–1410, 2019.
- [39] B.K. Pagnoncelli, S. Ahmed, and A. Shapiro. Sample average approximation method for chance constrained programming: theory and applications. *Journal of Optimization Theory and Applications*, 142(2):399–416, 2009.
- [40] 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. *Journal of Optimization Theory and Applications*, 155(2):707–722, 2012.
- [41] B.K. Pagnoncelli and S. Vanduffel. A provisioning problem with stochastic payments. *European Journal of Operational Research*, 221(2):445–453, 2012.
- [42] M.E. Payton, L.J. Young, and J.H. Young. Bounds for the difference between median and mean of beta and negative binomial distributions. *Metrika*, 36:347–354, 1989.
- [43] F.A. Ramponi. Consistency of the scenario approach. *SIAM Journal on Optimization*, 28:135–162, 2018.
- [44] R. Rocchetta, L.G. Crespo, and S.P. Kenny. Solution of the benchmark control problem by scenario optimization. In *Proceedings of the ASME 2019 Dynamic Systems and Control Conference (DSCC 2019)*, Park City, UT, USA, 2019.
- [45] R. Rocchetta, L.G. Crespo, and S.P. Kenny. A scenario optimization approach to reliability-based design. *Reliability Engineering and System Safety*, 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(12):3009–3018, 2014.
- [47] G. Schildbach, L. Fagiano, and M. Morari. Randomized solutions to convex programs with multiple chance constraints. *SIAM Journal on Optimization*, 23(4):2479–2501, 2013.
- [48] T. Sutter, P. Mohajerin Esfahani, and J. Lygeros. Approximation of constrained average cost markov control processes. In *Proceedings of the 53rd IEEE Conference on Decision and Control*, pages 6597–6602, Los Angeles, CA, USA, 2014.
- [49] L. Wasserman. Recent methodological advances in robust Bayesian inference. In J.M. Bernardo, J.O. Berger, A.P. David, and A.F.M. Smith, editors, *Bayesian Statistics 4*, pages 483–502. Oxford University Press, 1990.
- [50] J.S. Welsh and H. Kong. Robust experiment design through randomisation with chance constraints. In *Proceedings of the 18th IFAC World Congress*, Milan, Italy, 2011.
- [51] J.S. Welsh and C.R. 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.J. Goulart, and J. Lygeros. On the sample size of random convex programs with structured dependence on the uncertainty. *Automatica*, 60:182–188, 2015.