A constraint sampling approach for multi-stage robust optimization

Phebe Vayanos, Daniel Kuhn, and Berç Rustem

Department of Computing, Imperial College London,
180 Queen's Gate, London SW7 2AZ, United Kingdom

November 25, 2010

Abstract

We propose a tractable approximation scheme for convex (not necessarily linear) multi-stage robust optimization problems. We approximate the adaptive decisions by finite linear combinations of prescribed basis functions and demonstrate how one can optimize over these decision rules at low computational cost through constraint randomization. We obtain a-priori probabilistic guarantees on the feasibility properties of the optimal decision rule by extending existing constraint sampling techniques from the single- to the multi-stage case. We demonstrate that for a suitable choice of basis functions, the approximation converges as the size of the basis and the number of sampled constraints tend to infinity. The approach yields an algorithm parameterized in the basis size, the probability of constraint violation and the confidence that this probability will not be exceeded. These three parameters serve to tune the trade-off between optimality and feasibility of the decision rules and the computational cost of the algorithm. We assess the convergence and scalability properties of our approach in the context of two inventory management problems.

Keywords multi-stage robust optimization, decision rules, scenario approximation, violation probability.

1 Introduction

Decision problems arising in engineering, finance, logistics etc. are usually dynamic and affected by uncertainty. Such problems are notoriously hard to solve (Shapiro and Nemirovski [2005]) and researchers from the fields of dynamic and stochastic programming (Birge and Louveaux [2000], Bertsekas [1995]) as well as robust optimization and control (Ben-Tal et al. [2009], Dullerud and Paganini [2005]) have spent substantial research efforts on devising efficient algorithmic solution procedures.

In this paper we focus on *robust* dynamic optimization problems with a worst-case objective and robust constraints that must hold for all possible realizations of the uncertain problem parameters. The label *dynamic* indicates the presence of adaptive decisions that must be modeled as functions of some (or all) of the uncertain parameters. One of the most recent methods to solve robust dynamic optimization problems of this type, pioneered by Ben-Tal et al. [2004], is to approximate these adaptive decisions by *decision rules*, that is, linear combinations of prescribed basis functions. If, for instance, the basis functions are chosen to be the Euclidean coordinate projections, we obtain the popular class of *linear decision rules* (Ben-Tal et al. [2004, 2009]). This decision rule approximation transforms the original dynamic problem to a *static* robust optimization problem whose decision variables are the coefficients of the linear combinations. The support of the uncertain parameters is referred to as *uncertainty set*.

While the original dynamic problem is typically intractable, the approximate static problem is sometimes equivalent to a linear, second-order conic or semidefinite program of moderate size and thus allows for an exact polynomial-time solution. This is the case, under mild structural assumptions, when the uncertainty set is characterized through conic inequalities and the employed decision rules are linear (Ben-Tal et al. [2004]), when the uncertainty set is rectangular and the decision rules are piecewise linear and separable (Ben-Tal et al. [2009], Georghiou et al. [2010]), or when the uncertainty set is ellipsoidal and the decision rules are quadratic (Ben-Tal et al. [2009]). Sometimes, however, the approximate static problem has no exact reformulation as a manifestly tractable conic program, and one has to resort to conservative approximations, yielding feasible yet sub-optimal solutions. This is the case, for instance, when the uncertainty set is semi-algebraic and the employed decision rules are polynomial (Bertsimas et al. [2010]), when the uncertainty set is polyhedral and the decision rules are piecewise linear (Goh and Sim [2010], Georghiou et al. [2010]), or when the uncertainty set is an intersection of concentric ellipsoids and the decision rules are quadratic (Ben-Tal et al. [2009]).

When the uncertainty set and/or the employed decision rules have no simple algebraic structure, there are usually no semi-analytical conic programming approaches of the type described above. In this case, the static robust optimization problem emerging from the decision rule approximation is typically intractable. However, it can be solved approximately via *constraint sampling*, see Calafiore and Campi [2005, 2006]. The basic idea of this approach is to replace the static problem, which typically has a continuum of constraints, by a *scenario counterpart* with a finite number of constraints, obtained through Monte Carlo sampling of the uncertain parameters. The scenario counterpart can be solved efficiently under standard assumptions even if the original static problem is intractable. Due to the outer approximation entailed by the sampling, the solution of the scenario counterpart may fail to satisfy some of the constraints of the original static problem. However, it has been shown to satisfy most of them provided the number of samples is sufficiently large (Calafiore and Campi [2005], Campi and Garatti [2008]). Allowing for a certain probability of constraint violation is a notion permeating the field of chance constrained optimization (see e.g. Section 8.3 in Prékopa [1995]), which is seen as an attractive alternative to robust optimization. In fact, the link between the constraint sampling methodology and chance constrained optimization is showcased by Campi and Garatti [2008], who determine an exact universal bound on the number of samples required to ensure that only a small portion of the constraints of the original static problem is violated.

Combining decision rule and constraint sampling techniques provides a flexible modeling framework for robust dynamic optimization problems with a general nonlinear dependence on the uncertain parameters. The sampling approach offers almost complete freedom in the choice of the basis functions for the decision rules. This allows the modeler to exploit any knowledge about the structure of the optimal solution when designing the decision rule approximation. The sampling approach is particularly attractive for problems that cannot be solved using semi-analytical schemes. However, its benefits also extend to certain problems for which semi-analytical approaches are adequate. This is the case, for instance, when the semi-analytical techniques lift the problem to a higher complexity class. Then, constraint sampling techniques prove attractive as they circumvent the added complexity, thereby yielding problems that can potentially be solved with less computational effort. While the benefits of combining decision rule and sampling techniques have been explored by Calafiore and Nilim [2004], Bertsimas and Caramanis [2007], Skaf and Boyd [2009] and Lobel and Perakis [2010], several issues remain to be addressed. First, tractability results have only been provided for the case of polynomial decision rules. Furthermore, the interplay between the design parameters of the approximation has not been studied systematically. Finally, the asymptotic properties of the approximation have not been investigated rigorously.

The goal of this paper is to present a unifying methodology for solving multi-stage convex robust optimization problems under generic nonlinear decision rules that yield arbitrarily tight approximations. In particular, we introduce an axiomatic characterization of classes of decision rules that result in a tractable scenario counterpart and guarantee asymptotic consistency. We model the degree of flexibility of the decision rules through a user specified complexity parameter that, for a fixed feasibility probability, controls the trade-off between sub-optimality and computational complexity.

The main contributions of this paper are summarized below:

1. We consider general decision rule approximations for multi-stage robust optimization problems. The decision rules are modeled as finite linear combinations of a prescribed set of basis functions such as algebraic or trigonometric monomials, sigmoidal or Gaussian radial functions, etc. The flexibility to choose a generic set of basis functions is attractive since it enables the modeler to tailor the decision rule approximation to the problem instances if particular structural properties are known.

2. The synthesis of decision rule and constraint sampling techniques results in computationally tractable approximations of multi-stage problems. We illustrate how the underlying design parameters can be used, in a systematic fashion, to control the trade-off between sub-optimality and computational complexity of the approximation.

3. We provide a rigorous convergence proof for the decision rule approximation which applies, under mild technical assumptions, if the number of basis functions is driven to infinity. We also demonstrate almost sure convergence of the constraint randomization approach as the number of samples

tends to infinity.

The paper is organized as follows. The remainder of this section introduces the notation, while Section 2 presents the mathematical problem formulation. Section 3 is split into two parts concerned with the decision rule and constraint sampling approximations, respectively, and Section 4 provides a complexity analysis. The convergence properties of the approximation scheme are investigated in Section 5. Finally, Section 6 presents two problems from inventory management which are amenable to our approximation scheme, and Section 7 reports on numerical results.

Notation Throughout this paper, vectors (matrices) are denoted by boldface lowercase (uppercase) letters. For $\boldsymbol{x} \in \mathbb{R}^n$, we denote the closed n-ball of radius r centered at $\boldsymbol{x}$ by $B_r(\boldsymbol{x})$, the Euclidean norm of $\boldsymbol{x}$ by $|\boldsymbol{x}|$ and the i^{th} component of $\boldsymbol{x}$ by x_i. Also, for $\boldsymbol{\alpha} \in \mathbb{N}_0^n$, we let $\boldsymbol{x}^{\boldsymbol{\alpha}} := \prod_{i=1}^n (x_i)^{\alpha_i}$. For any matrix $\boldsymbol{X} \in \mathbb{R}^{n \times m}$ with columns $\boldsymbol{x}_i \in \mathbb{R}^n$, $i = 1, \ldots, m$, we denote by $\operatorname{vec}(\boldsymbol{X}) := (\boldsymbol{x}_1, \boldsymbol{x}_2, \ldots, \boldsymbol{x}_m) \in \mathbb{R}^{nm}$ the vector concatenation of its columns.

Uncertainty is modeled by the probability space $(\mathbb{R}^k, \mathcal{F}, \mathbb{P})$, which consists of the sample space $\mathbb{R}^k$, the Borel σ-algebra $\mathcal{F} := \mathcal{B}(\mathbb{R}^k)$ and the probability measure $\mathbb{P}$, whose support we denote by Ξ. The elements of the sample space are denoted by $\boldsymbol{\xi}$ and are assumed to possess a temporal structure in that they are representable as $\boldsymbol{\xi} := (\boldsymbol{\xi}_1, \boldsymbol{\xi}_2, \ldots, \boldsymbol{\xi}_T)$. The random vectors $\boldsymbol{\xi}_t \in \mathbb{R}^{k_t}$, $t \in \mathbb{T} := \{1, 2, \ldots, T\}$, have marginal supports $\Xi_t \subseteq \mathbb{R}^{k_t}$, where $\sum_{t \in \mathbb{T}} k_t = k$. For convenience, we define combined random vectors $\boldsymbol{\xi}^t := (\boldsymbol{\xi}_1, \ldots, \boldsymbol{\xi}_t) \in \mathbb{R}^{k^t}$, $t \in \mathbb{T}$, with marginal supports $\Xi^t \subseteq \mathbb{R}^{k^t}$, where $k^t := \sum_{\tau=1}^t k_\tau$. Furthermore, for each $t \in \mathbb{T}$, we introduce the information set

$$\mathcal{F}^t := \left\{ Z^t \times \left(\bigtimes_{\tau=t+1}^{T} \mathbb{R}^{k_\tau} \right) : Z^t \in \mathcal{B}\left(\mathbb{R}^{k^t}\right) \right\}$$

which corresponds to the σ-algebra generated by $\boldsymbol{\xi}^t$, the history of the random vector $\boldsymbol{\xi}$ up to time t. Finally, for each $t \in \mathbb{T}$ we denote by $\mathcal{C}(\Xi^t)$ the space of continuous real-valued functions on Ξ^t.

2 Problem description

We consider a multi-stage decision problem under uncertainty over the finite planning horizon $\mathbb{T}$. The aim is to find a sequence of decisions $\boldsymbol{x} := (\boldsymbol{x}_1, \boldsymbol{x}_2, \ldots, \boldsymbol{x}_T)$ that minimize a cost function $f_0(\boldsymbol{x}, \boldsymbol{\xi})$ in the worst-case realization of $\boldsymbol{\xi} \in \Xi$. These decisions are constrained by a set of inequalities which are required to be obeyed robustly, that is, for any realization of the uncertainties $\boldsymbol{\xi} \in \Xi$. The decision $\boldsymbol{x}_t \in \mathbb{R}^{n_t}$ is selected at time t after the history of realizations $\boldsymbol{\xi}^t$ has been observed but before the future outcomes $\{\boldsymbol{\xi}_f\}_{f>t}$ have been revealed. This motivates us to represent $\boldsymbol{x}_t$ as an $\mathcal{F}^t$-measurable function or *decision rule* of $\boldsymbol{\xi}$. The requirement that $\boldsymbol{x}_t$ be constant in $\{\boldsymbol{\xi}_f\}_{f>t}$, which is implied by the $\mathcal{F}^t$-measurability, reflects the *non-anticipative* nature of the decision process and ensures the *causality* of the dynamic decision problem. Modeling $\boldsymbol{x}_t$ as an $\mathcal{F}^t$-measurable function of $\boldsymbol{\xi}$ is essential to keep the decision model realistic and flexible but makes it computationally challenging. Robust optimization problems of the type described here can be formulated compactly as

$$\begin{aligned} \inf_{\boldsymbol{x}\in\mathcal{N}} \quad & \sup_{\boldsymbol{\xi}\in\Xi} \; f_0\left(\boldsymbol{x}\left(\boldsymbol{\xi}\right), \boldsymbol{\xi}\right) \\ \text{s.t.} \quad & f_i\left(\boldsymbol{x}\left(\boldsymbol{\xi}\right), \boldsymbol{\xi}\right) \leq 0 \quad \forall \boldsymbol{\xi} \in \Xi, \; i = 1, \ldots, I, \end{aligned} \tag{$\mathcal{R}$}$$

where $f_i : \mathbb{R}^n \times \mathbb{R}^k \to \mathbb{R}$, $i = 0, 1, \ldots, I$, are convex in $\boldsymbol{x}$ and continuous in $(\boldsymbol{x}, \boldsymbol{\xi})$, while

$$\mathcal{N} := \bigtimes_{t\in\mathbb{T}} L^{\infty}_{n_t}\left(\mathbb{R}^k, \mathcal{F}^t, \mathbb{P}\right)$$

denotes the space of all non-anticipative essentially bounded decision rules from $\mathbb{R}^k$ to $\mathbb{R}^n$, $n := \sum_{t\in\mathbb{T}} n_t$. Problem $\mathcal{R}$ provides an attractive alternative to its stochastic counterpart with an expectation objective since it yields solutions that are better suited to the needs of more risk averse decision-makers.

Remark 1 (Role of the distribution $\mathbb{P}$) *We note at this point that the probability distribution $\mathbb{P}$ only influences problem $\mathcal{R}$ through its support Ξ. Indeed, the worst-case objective and the robust constraints are independent of probabilities. Although the distribution $\mathbb{P}$ does not play a role per se in problem $\mathcal{R}$, it will, however, assume a bivalent position when the constraint sampling approximation is introduced. Firstly, it constitutes the distribution used for sampling. Secondly, it provides the basis for establishing probabilistic guarantees on the solution quality. We remark that our approach remains applicable even if $\mathbb{P}$ is unknown. In this case, we merely require that a finite (yet "sufficiently large") number of samples from $\mathbb{P}$, e.g. in the form of past observations, be available.*

In what follows, we adopt an approach which, although aimed at approximating the robust problem $\mathcal{R}$, mitigates the sometimes criticized over-conservatism of robust optimization. In particular, we take the view of a decision-maker who may accept to be unprotected against events that have a low probability of occurrence. This idea is well known from chance constrained programming and is often conceptually preferred to the "hard" robust paradigm since the choice in the immunization level may be tailored to the risk tolerance of the decision-maker. Indeed, sacrificing some robustness can sometimes result in a significant gain in optimality.

As discussed in the introduction, multi-stage problems of the form $\mathcal{R}$ are very difficult to solve. In what follows, we discuss two successive parametric approximations that enable us to find near-optimal solutions for $\mathcal{R}$, while controlling the quality and complexity of the approximation.

3 Tractable reformulation

In this section, we construct a computationally tractable approximation for problem $\mathcal{R}$ based on a flexible decision rule and constraint sampling approach.

3.1 Decision rule approximation

The decision variables in the original robust problem $\mathcal{R}$ range over a space of non-anticipative measurable functions from $\mathbb{R}^k$ to $\mathbb{R}^n$. As a first step towards obtaining a computationally tractable model, we restrict the space of all measurable decision rules to those representable as finite linear combinations of certain prescribed basis functions. We first describe how to construct vectors of basis functions of a given complexity. We further provide conditions ensuring that the flexibility of the decision rules increases with their complexity and that the size of the basis vectors is polynomially bounded in the dimension k of the random vectors $\boldsymbol{\xi}$. We then use the basis vectors to construct approximations for problem $\mathcal{R}$ which accommodate only a finite number of decision variables.

The construction of the basis vectors proceeds in two steps:

1. For each $t \in \mathbb{T}$, select a sequence of continuous functions $b_{t,m} : \mathbb{R}^{k^t} \to \mathbb{R}$, $m \in \mathbb{N}$.

2. Select a complexity parameter $d \in \mathbb{N}_0$. For any fixed d, choose an increasing function $s_d : \mathbb{N}_0 \to \mathbb{N}$ and define the basis vector $\boldsymbol{b}_t^d : \mathbb{R}^{k^t} \to \mathbb{R}^{s_d(k^t)}$ as $\boldsymbol{b}_t^d := (b_{t,1}, b_{t,2}, \ldots, b_{t,s_d(k^t)})$ for all $t \in \mathbb{T}$. We will refer to $s_d(k^t)$ as the basis size at time t.

Throughout the rest of the paper, we assume the following conditions to hold true.

(C1) The linear hull of $\{b_{t,m}\}_{m\in\mathbb{N}}$ is dense in $\mathcal{C}(\Xi^t)$ with respect to the supremum norm.

(C2) For any fixed $d \in \mathbb{N}_0$, the function s_d is bounded above by a polynomial.

Once the basis vectors have been constructed as described above, we restrict the functional decisions $\boldsymbol{x}_t$ in $\mathcal{R}$ to be representable as linear combinations of the components of $\boldsymbol{b}_t^d$ for each $t \in \mathbb{T}$. Thus, we focus on decision rules of the form

$$\boldsymbol{x}_t(\boldsymbol{\xi}) = \boldsymbol{X}_t \, \boldsymbol{b}_t^d\left(\boldsymbol{\xi}^t\right), \tag{3.1}$$

where the matrix $\boldsymbol{X}_t \in \mathbb{R}^{n_t \times s_d(k^t)}$ contains the coefficients of the linear combinations, which become the new decision variables. With the restriction (3.1), the epigraph formulation of problem $\mathcal{R}$ reduces to

$$\begin{aligned} \min_{\theta\in\mathbb{R},\, \boldsymbol{y}\in\mathbb{R}^{n_y}} \quad & \theta \\ \text{s.t.} \quad & \tilde{f}_0(\boldsymbol{y},\boldsymbol{\xi}) - \theta \le 0, \quad \forall \boldsymbol{\xi} \in \Xi \\ & \tilde{f}_i(\boldsymbol{y},\boldsymbol{\xi}) \le 0, \quad i = 1,\ldots,I, \quad \forall \boldsymbol{\xi} \in \Xi, \end{aligned}$$

where we identify $\boldsymbol{y}$ with $(\mathrm{vec}(\boldsymbol{X}_1),\ldots,\mathrm{vec}(\boldsymbol{X}_T))$, set $n_y := \sum_{t\in\mathbb{T}} n_t s_d(k^t)$ and define

$$\tilde{f}_i(\boldsymbol{y},\boldsymbol{\xi}) := f_i\left(\left(\boldsymbol{X}_1\, \boldsymbol{b}_1^d\left(\boldsymbol{\xi}^1\right),\ldots,\boldsymbol{X}_T\, \boldsymbol{b}_T^d\left(\boldsymbol{\xi}^T\right)\right),\boldsymbol{\xi}\right), \quad i = 0,\ldots,I.$$

The functions $\tilde{f}_i : \mathbb{R}^{n_y} \times \mathbb{R}^k \to \mathbb{R}$ are convex in $\boldsymbol{y}$ and continuous in $(\boldsymbol{y},\boldsymbol{\xi})$. For notational convenience, we set $\boldsymbol{w} := (\theta, \boldsymbol{y})$ and introduce the following equivalent robust optimization problem with basis functions

$$\begin{aligned} \min_{\boldsymbol{w} \in \mathbb{R}^{n_w}} \quad & w_1 \\ \text{s.t.} \quad & f(\boldsymbol{w}, \boldsymbol{\xi}) \leq 0 \quad \forall \boldsymbol{\xi} \in \Xi, \end{aligned} \tag{$\mathcal{RB}$}$$

where $n_w := 1 + n_y$ and $f : \mathbb{R}^{n_w} \times \mathbb{R}^k \to \mathbb{R}$ is defined through

$$f(\boldsymbol{w}, \boldsymbol{\xi}) := \max \left\{ \max_{1 \leq i \leq I} \tilde{f}_i(\boldsymbol{y}, \boldsymbol{\xi}), \tilde{f}_0(\boldsymbol{y}, \boldsymbol{\xi}) - \theta \right\}.$$

Several observations are now in order. The function f is convex in $\boldsymbol{w}$ and continuous in $(\boldsymbol{w}, \boldsymbol{\xi})$ as it is obtained by taking the maximum of $I + 1$ functions with these properties. The dimension n_w of the decision vector in $\mathcal{RB}$ is linear in the number n of functional decision variables in the original problem and polynomial in the number k of uncertainties.

Thus far, we have derived an approximation for problem $\mathcal{R}$ with a finite number of decision variables that remains polynomially bounded in n and k. In order to illustrate the flexibility of this approach, we now provide several examples of decision rules and associated basis vectors. These examples highlight the advantages of non-affine decision rules, which can result in arbitrarily tight approximations for the original problem $\mathcal{R}$ (see Section 5.1). An illustration of their approximation capabilities is provided in Figure 1.

Example 1 (Algebraic polynomials) *A convenient choice for $\boldsymbol{b}_t^d$ is the vector of monomials in the $\mathbb{R}$-vector space of multivariate polynomials in $\boldsymbol{\xi}^t$ with degree at most d, that is,*

$$\boldsymbol{b}_t^d(\boldsymbol{\xi}^t) := \left\{ \left(\boldsymbol{\zeta}^t(\boldsymbol{\xi}^t) \right)^{\boldsymbol{\alpha}} : \boldsymbol{\alpha} \in \mathbb{N}_0^{k^t}, \|\boldsymbol{\alpha}\|_1 \leq d \right\}.$$

The affine scaling function $\boldsymbol{\zeta}^t : \mathbb{R}^{k^t} \to \mathbb{R}^{k^t}$ is defined through

$$\boldsymbol{\zeta}^t(\boldsymbol{\xi}^t) := \left(2\frac{\xi_1 - l_1}{u_1 - l_1} - 1, \ldots, 2\frac{\xi_{k^t} - l_{k^t}}{u_{k^t} - l_{k^t}} - 1 \right), \tag{3.2}$$

where $[l_s, u_s] \subseteq \mathbb{R}$ denotes an interval covering the marginal support of ξ_s. The scaling ensures that $\boldsymbol{b}_t^d(\Xi^t) \subseteq [-1, 1]^{s_d(k^t)}$, which in turn implies that the decision rules are uniformly bounded on Ξ^t. This helps to avoid numerical instabilities due to poorly scaled optimization problems. By the Stone-Weierstraß theorem, the algebraic polynomials (of arbitrary finite degree) are dense in $\mathcal{C}(\Xi^t)$. Therefore, condition (C1) is satisfied. The dimension of $\boldsymbol{b}_t^d$ is given by $s_d(k^t) = \binom{k^t+d}{d} \in \mathcal{O}((k^t)^d)$, which is polynomial in k^t, as required by condition (C2). The complexity parameter d corresponds to the degree of the polynomial decision rules.

Example 2 (Trigonometric polynomials) *The vector of monomials in the $\mathbb{R}$-vector space of multi-*

variate trigonometric polynomials in $\boldsymbol{\xi}^t$ *with degree at most* d *can be written as*

$$\boldsymbol{b}_t^d\left(\boldsymbol{\xi}^t\right) := \left\{\cos \boldsymbol{c}^\top \boldsymbol{\zeta}^t\left(\boldsymbol{\xi}^t\right) : \boldsymbol{c} \in \mathbb{Z}^{k^t},\ \|\boldsymbol{c}\|_1 \leq d\right\} \bigcup \left\{\sin \boldsymbol{s}^\top \boldsymbol{\zeta}^t\left(\boldsymbol{\xi}^t\right) : \boldsymbol{s} \in \mathbb{Z}^{k^t},\ \|\boldsymbol{s}\|_1 \leq d,\ \boldsymbol{s} \neq \boldsymbol{0}\right\},$$

where the scaling function $\boldsymbol{\zeta}^t$ *is defined as in* (3.2)*. Here, the scaling is needed due to the periodicity of the trigonometric polynomials. By the Stone-Weierstraß theorem, the multivariate trigonometric polynomials (of arbitrary finite degree) are dense in* $\mathcal{C}(\Xi^t)$*.*

The total number of vectors $\boldsymbol{c} \in \mathbb{N}_0^{k^t}$ *with exactly* $i \leq \min(d, k^t)$ *non-zero elements satisfying* $\|\boldsymbol{c}\|_1 = j$*,* $i \leq j \leq d$*, is* $\binom{k^t}{i}\binom{j-1}{j-i}$*. For each vector* $\boldsymbol{c} \in \mathbb{Z}^{k^t}$ *with* i *non-zero elements, there are* 2^i *sign combinations which give rise to distinct cosine-type basis functions. A similar argument can be made for the sine-type basis functions. Thus, the basis size at time* t *is given by*

$$s_d(k^t) = 1 + \sum_{i=1}^{\min(d,k^t)} \sum_{j=i}^{d} 2^{i+1}\binom{k^t}{i}\binom{j-1}{j-i} \in \mathcal{O}\left((k^t)^d\right),$$

which is polynomially bounded in k^t *for any fixed* d*. The complexity parameter* d *corresponds to the degree of the trigonometric polynomial decision rules.*

Example 3 (Sigmoidal basis functions) *Let* $\sigma : \mathbb{R} \to \mathbb{R}$ *be any continuous sigmoidal function such as* $\sigma(t) = 1/(1 + \exp(-t))$ *or* $\sigma(t) = \max\{0, \min\{(t+1)/2, 1\}\}$*, and let* $\boldsymbol{\gamma} : \mathbb{Z}^{k^t+1} \to \mathbb{Q}^{k^t+1}$ *be a bijection (see e.g Yu-Ting [1980] for an example of such a bijection). We can now define a basis vector* $\boldsymbol{b}_t^d$ *of complexity* d *through*

$$\boldsymbol{b}_t^d\left(\boldsymbol{\xi}^t\right) := \left\{\sigma\left(\boldsymbol{c}^\top \boldsymbol{\zeta}^t\left(\boldsymbol{\xi}^t\right) + g\right) :\ (\boldsymbol{c}, g) = \boldsymbol{\gamma}\left(\boldsymbol{m}\right),\ \boldsymbol{m} \in \mathbb{Z}^{k^t+1},\ \|\boldsymbol{m}\|_1 \leq d\right\},$$

where the scaling function $\boldsymbol{\zeta}^t$ *is defined as in* (3.2)*. The scaling ensures that the basis functions corresponding to low values of* d *have significant variability within* Ξ^t*. By Cybenko's theorem [Cybenko, 1989, Theorem 1], the linear hull of all sigmoidal basis functions (for arbitrary values of* d*) is dense in* $\mathcal{C}(\Xi^t)$*.*

Following a similar reasoning as in Example 2, it can be shown that

$$s_d\left(k^t\right) = 1 + \sum_{i=1}^{\min(d,k^t+1)} \sum_{j=i}^{d} 2^i\binom{k^t+1}{i}\binom{j-1}{j-i} \in \mathcal{O}\left((k^t)^d\right).$$

The basis size is therefore polynomially bounded in k^t *for any fixed* d*. The complexity parameter* d *corresponds to the maximum Manhattan norm of the vector* $\boldsymbol{m}$ *encoding the rationals.*

Example 4 (Gaussian radial basis functions) *Following a similar approach as in Example 3, we can construct a basis vector* $\boldsymbol{b}_t^d$ *of complexity* d *by setting*

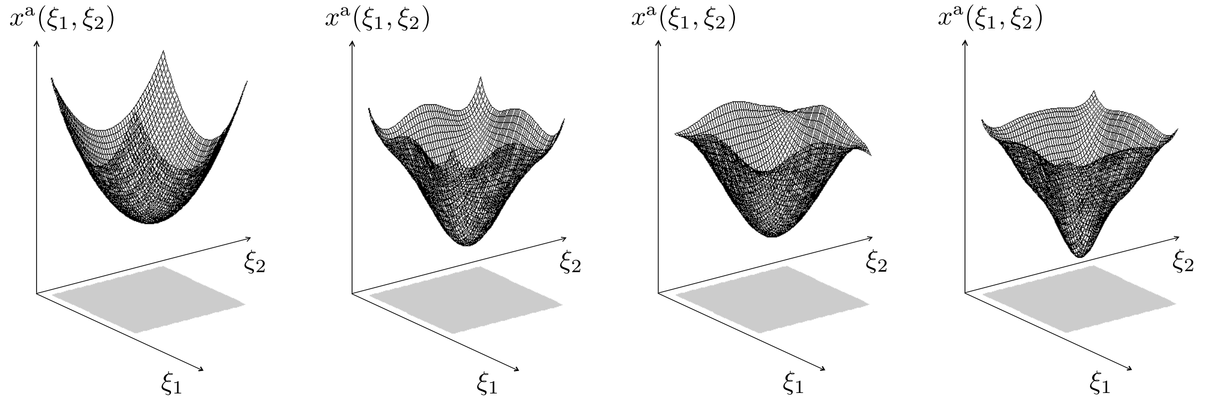
$$\boldsymbol{b}_t^d\left(\boldsymbol{\xi}^t\right) := \left\{\exp\left(-g^2|\boldsymbol{\zeta}^t\left(\boldsymbol{\xi}^t\right) - \boldsymbol{c}|^2\right) :\ (\boldsymbol{c}, g) = \boldsymbol{\gamma}\left(\boldsymbol{m}\right),\ \boldsymbol{m} \in \mathbb{Z}^{k^t+1},\ \|\boldsymbol{m}\|_1 \leq d\right\},$$

where the scaling function $\boldsymbol{\zeta}^t$ is as in (3.2) *and the bijection $\boldsymbol{\gamma}$ is as in Example 3. The linear hull of all Gaussian radial basis functions (for arbitrary values of d) is dense in $\mathcal{C}(\Xi^t)$, see e.g. Proposition B.1 in Girosi and Poggio [1990].*

Following a similar approach as in Example 3, we obtain

$$s_d\left(k^t\right) = 1 + \sum_{i=1}^{\min(d,k^t+1)} \sum_{j=i}^{d} 2^i \binom{k^t+1}{i}\binom{j-1}{j-i} \in \mathcal{O}\left((k^t)^d\right).$$

Thus, the basis size is polynomially bounded in k^t for any fixed d. As in Example 3, the complexity parameter d corresponds to the maximum Manhattan norm of the vector $\boldsymbol{m}$ which encodes the rationals.



(a) Polynomial (b) Trigonometric (c) Sigmoidal (d) Gaussian radial

Figure 1: Approximation of $x(\xi_1, \xi_2) := \max(\xi_1, \xi_2) + \max(-\xi_1, -\xi_2) + \max(\xi_1, -\xi_2) + \max(-\xi_1, \xi_2)$ on $[-3, 3] \times [-3, 3]$ by (a) polynomial (b) trigonometric polynomial (c) sigmoidal and (d) Gaussian radial basis functions. In all cases, the complexity parameter is set to $d = 3$.

We note that the number of decision variables in problem $\mathcal{RB}$ can always be reduced by limiting the memory of the decision rules, that is, by representing $\boldsymbol{x}_t$ as a combination of functions that only depend on $(\boldsymbol{\xi}_{t-m}, \ldots, \boldsymbol{\xi}_t)$ for some $m \in \mathbb{N}$. Furthermore, a different basis vector may be used for each decision variable, thereby leading to a great deal of modeling flexibility.

So far, we described a methodology for reducing the number of decision variables of problem $\mathcal{R}$. However, we note that when Ξ has infinite cardinality (as is the case if $\mathbb{P}$ is continuous), the resulting problem $\mathcal{RB}$ falls in the category of semi-infinite programs, which are extremely hard to solve in general, see e.g. Blondel and Tsitsiklis [2000]. In particular, it is known that checking feasibility of a generic semi-infinite constraint $f(\boldsymbol{w}, \boldsymbol{\xi}) \leq 0 \ \forall \boldsymbol{\xi} \in \Xi$ for a fixed $\boldsymbol{w} \in \mathbb{R}^{n_w}$ is NP-hard even if Ξ is a simplex and f is indefinite quadratic in $\boldsymbol{\xi}$. In the next section, we therefore apply a constraint sampling approach to problem $\mathcal{RB}$ to achieve computational tractability.

3.2 Constraint sampling approximation

Calafiore and Campi [2005] suggest to solve single-stage robust optimization problems that would otherwise be intractable due to their semi-infinite nature by considering their "scenario" counterparts: a finite set of N constraints is chosen at random from the typically uncountable set of constraints, and the resulting tractable problem is solved. The main result of Calafiore and Campi [2005], which was later improved by Campi and Garatti [2008], is the observation that any solution of the sampled problem will also satisfy most of the constraints of the original problem (which were not sampled).

Here we apply this approach to the decision rule approximation $\mathcal{RB}$, which has in fact the structure of a single-stage robust optimization problem. Indeed, drawing N independent samples $\boldsymbol{\xi}^{(1)}, \boldsymbol{\xi}^{(2)}, \ldots, \boldsymbol{\xi}^{(N)}$ distributed according to $\mathbb{P}$, we can approximate $\mathcal{RB}$ by the following scenario problem with basis functions.

$$\begin{aligned} \min_{\boldsymbol{w} \in \mathbb{R}^{n_w}} \quad & w_1 \\ \text{s.t.} \quad & f(\boldsymbol{w}, \boldsymbol{\xi}^{(s)}) \leq 0, \quad s = 1, \ldots, N \end{aligned} \qquad (\mathcal{SB}_N)$$

Since the function f is convex in $\boldsymbol{w}$ for each $\boldsymbol{\xi} \in \Xi$, $\mathcal{SB}_N$ is again a convex optimization problem. We observe that since the samples $\boldsymbol{\xi}^{(1)}, \boldsymbol{\xi}^{(2)}, \ldots, \boldsymbol{\xi}^{(N)}$ are random variables, problem $\mathcal{SB}_N$ is itself random. However, the hope is that the variability of the solution set and the optimal value over different samples of size N is small if N is chosen sufficiently large (see Section 5.2).

Recall that problem $\mathcal{R}$ has infinitely many variables and constraints parameterized by $\boldsymbol{\xi} \in \Xi$. The decision rule approximation described in Section 3.1 transformed $\mathcal{R}$ to the semi-infinite problem $\mathcal{RB}$ with finitely many variables and infinitely many constraints. Constraint sampling, in turn, yields problem $\mathcal{SB}_N$, which has finitely many variables *and* constraints. We note that all three problems are convex. While $\mathcal{RB}$ provides an upper bound on $\mathcal{R}$ due to a restriction of the feasible set, the sampled problem $\mathcal{SB}_N$ constitutes a relaxation of problem $\mathcal{RB}$, and thus its solution provides a lower bound on $\mathcal{RB}$.

A fundamental question now arising is whether these approximations break the curse of dimensionality (see e.g. Section 13.1.3 in Ben-Tal et al. [2009]) so that scalability to multi-stage problems is actually possible. In the following section, we thus investigate the trade-off between the accuracy and the tractability of the two approximations.

4 Complexity analysis

As the constraints of the scenario problem $\mathcal{SB}_N$ are randomly extracted, its optimal solutions are random variables which depend on the set of extractions $\{\boldsymbol{\xi}^{(s)}\}_{s=1}^{N}$. These solutions typically fail to satisfy all constraints of the robust problem $\mathcal{RB}$. However, they can be shown to satisfy the constraints of $\mathcal{RB}$ with high probability. The following fundamental question is addressed, among others, by Calafiore and Campi [2005, 2006] and by Campi and Garatti [2008]: what confidence do we have that an optimal solution of $\mathcal{SB}_N$ will violate the constraints of $\mathcal{RB}$ with probability less than ϵ, where ϵ is a prescribed probability level as in chance-constrained programming? To the best of our knowledge, the most general answer to this question is provided by Campi and Garatti [2008] who establish, for any convex optimization problem and any probability distribution $\mathbb{P}$ with support Ξ, the so-called "exact feasibility" of the randomized solution. We repeat their results here, together with the definition of violation probability as introduced in Calafiore and Campi [2005].

Definition 1 (Calafiore and Campi [2005]) *The violation probability of a given* $\boldsymbol{w} \in \mathbb{R}^{n_w}$ *is defined as* $V(\boldsymbol{w}) := \mathbb{P}(\boldsymbol{\xi} \in \Xi : f(\boldsymbol{w}, \boldsymbol{\xi}) > 0)$.

Theorem 1 (Campi and Garatti [2008]) *For any given probability level* $\epsilon \in (0,1)$ *and confidence level* $\beta \in (0,1)$*, let*

$$N(\epsilon, \beta) := \min \left\{ N \in \mathbb{N} : \sum_{i=0}^{n_w - 1} \binom{N}{i} \epsilon^i (1-\epsilon)^{N-i} \leq \beta \right\}.$$

If problem $\mathcal{SB}_N$ *is solvable and* $N \geq N(\epsilon, \beta)$*, then we have*

$$\mathbb{P}^N \left(V\left(\boldsymbol{w}_N^*\right) > \epsilon \right) \leq \beta,$$

where $\boldsymbol{w}_N^*$ *denotes any optimal solution of* $\mathcal{SB}_N$*, and* $\mathbb{P}^N := \mathbb{P} \times \mathbb{P} \times \cdots \times \mathbb{P}$ *(*N *times) is the probability distribution of the sample* $(\boldsymbol{\xi}^{(1)}, \boldsymbol{\xi}^{(2)}, \ldots, \boldsymbol{\xi}^{(N)})$.

This result guarantees that any random solution of $\mathcal{SB}_N$ satisfies most of the original constraints with high confidence provided that the sample size N is chosen large enough.

The following corollary is an immediate consequence of Theorem 1 and generalizes a result of Bertsimas and Caramanis [2007] for polynomial decision rules.

Corollary 1 (Complexity of problem $\mathcal{SB}_N$) *For any fixed probability level* $\epsilon \in (0,1)$ *and confidence level* $\beta \in (0,1)$*, the number of samples* N *needed such that the optimal solution* $\boldsymbol{w}_N^*$ *of* $\mathcal{SB}_N$ *satisfies* $\mathbb{P}^N(V(\boldsymbol{w}_N^*) > \epsilon) \leq \beta$ *remains polynomially bounded in* n *and* k.

Remark 2 (Computational tractability) *Corollary 1 implies that problem* $\mathcal{SB}_N$ *can be solved in polynomial time with respect to the size of the input parameters, provided that for any fixed* $\boldsymbol{\xi} \in \mathbb{R}^k$*, the set* $\{\boldsymbol{w} \in \mathbb{R}^n : f(\boldsymbol{w}, \boldsymbol{\xi}) \leq 0\}$ *admits an efficient separation oracle, see Grötschel et al. [1981].*

Proof of Corollary 1 In Calafiore [2009] it was shown that

$$N(\epsilon, \beta) \leq \frac{2}{\epsilon} \left(\ln \frac{1}{\beta} + n_w \right).$$

By construction of problem $\mathcal{RB}$ (see Section 3.1), we have that

$$n_w = 1 + \sum_{t \in \mathbb{T}} n_t s_d \left(k^t\right) \leq 1 + \sum_{t \in \mathbb{T}} n_t s_d (k) = 1 + n s_d (k),$$

where the inequality holds since s_d is increasing and $k \geq k_t$ for each $t \in \mathbb{T}$. As s_d is polynomially bounded by condition (C2), the number of decision variables n_w, and hence the number of samples $N(\epsilon, \beta)$ required for the prescribed violation probability ϵ and confidence level β, remain polynomially bounded in n and k. This concludes the proof. $\square$

Remark 3 (Scalability) *Typically, n and k are both linear in T, in which case the number of samples N needed to ensure $\mathbb{P}^N(V(\boldsymbol{w}_N^*) > \epsilon) \leq \beta$ is also polynomially bounded with T.*

Remark 4 *The basis size $s_d(k^t)$ can be exponential in the complexity parameter d for fixed values of k^t. In such cases, the number of samples required to sustain a maximum violation probability of ϵ at confidence $1-\beta$ is exponential in d. If $\mathbb{P}$ is unknown and only $\tilde{N}$ samples from $\mathbb{P}$ are available, one may opt for "simpler" decision rules, i.e., low values of d, in order to guarantee the required level of feasibility. Indeed, the maximum admissible value of d for given values of ϵ, β and $\tilde{N}$ amounts to*

$$\overline{d} := \max\left\{ d \in \mathbb{N} : \min\left\{ N \in \mathbb{N} : \sum_{i=0}^{\sum_{t\in\mathbb{T}} n_t s_d(k^t)} \binom{N}{i} \epsilon^i (1-\epsilon)^{N-i} \leq \beta \right\} \leq \tilde{N} \right\}.$$

In particular, any d with $s_d(k) \leq n^{-1}(\epsilon\tilde{N}/2 - \ln(1/\beta) - 1)$ satisfies $d \leq \overline{d}$ and ensures that no more than $\tilde{N}$ samples are needed to sustain a maximum violation of ϵ with confidence $1-\beta$.

5 Convergence analysis

The first part of this section analyzes the convergence of the optimal value of $\mathcal{RB}$ to the optimal value of $\mathcal{R}$ as the complexity parameter d of the basis vector tends to infinity. The second part establishes the almost sure convergence of the optimal value of the scenario problem $\mathcal{SB}_N$ to the optimal value of $\mathcal{RB}$ as the number of samples N increases. Before introducing a few technical assumptions, we define the concept of strict feasibility.

Definition 2 (Strict feasibility) *A decision rule $\boldsymbol{x} : \mathbb{R}^k \to \mathbb{R}^n$ is said to be strictly feasible for the robust optimization problem $\mathcal{R}$ if $\boldsymbol{x} \in \mathcal{N}$ and there exists $\delta > 0$ with*

$$f_i\left(\boldsymbol{x}\left(\boldsymbol{\xi}\right), \boldsymbol{\xi}\right) \leq -\delta \quad \forall \boldsymbol{\xi} \in \Xi,\ i = 1, \ldots, I.$$

If there exists a strictly feasible decision rule for $\mathcal{R}$, then the problem is said to be strictly feasible.

The subsequent convergence results are based on the following mild assumptions.

(A1) The robust problem $\mathcal{R}$ is strictly feasible.

(A2) The functions $f_i : \mathbb{R}^n \times \mathbb{R}^k \to \mathbb{R}$ are convex in $\boldsymbol{x}$ for each $\boldsymbol{\xi} \in \mathbb{R}^k$ and continuous in $(\boldsymbol{x}, \boldsymbol{\xi})$, $i = 0, \ldots, I$.

(A3) There exists $R \in \mathbb{R}$ such that $|\boldsymbol{x}(\boldsymbol{\xi})| \leq R$ for each $\boldsymbol{\xi} \in \Xi$ and for each $\boldsymbol{x}$ feasible in $\mathcal{R}$.

(A4) An arbitrary number of independent samples from $\mathbb{P}$ can be obtained.

(A5) The uncertainty set Ξ is convex, compact, fully dimensional and rectangular in the sense that $\Xi = \bigtimes_{t\in\mathbb{T}} \Xi_t$.

Several comments are in order. Firstly, assumption (A1) is satisfied for all problems of practical interest if equality constraints are systematically eliminated by using them to reduce the number of decision variables. Moreover, assumption (A3) is not restrictive since we are ultimately interested in numerical solutions for problem $\mathcal{R}$ which are necessarily bounded. The condition that Ξ be fully dimensional (meaning that there exist $\boldsymbol{\xi}_0 \in \Xi$ and $\epsilon > 0$ such that $B_\epsilon(\boldsymbol{\xi}_0) \subseteq \Xi$) is also non-restrictive. It can always be enforced at the cost of reducing the dimension of $\boldsymbol{\xi}$ if necessary. Finally, we note that assumption (A5) could be relaxed to require the existence of a non-anticipative homeomorphism $\boldsymbol{g} : \mathbb{R}^k \to \mathbb{R}^k$, such that $\Xi = \boldsymbol{g}(\Xi')$ for some set $\Xi' \subseteq \mathbb{R}^k$ that satisfies (A5). This generalization would allow us to consider even many non-convex uncertainty sets. For notational simplicity, however, we will use the simpler assumption (A5) in the sequel. The assumptions (A1)-(A5) are always assumed to hold in the remainder of the paper.

Throughout what follows, we denote the infimum of a problem $\mathcal{P}$ by $\inf \mathcal{P}$, the closed n-ball $B_R(\mathbf{0})$ by X and the Lebesgue measure on $\mathbb{R}^k$ by μ.

5.1 Convergence of the decision rule approximation

The main result of this section is provided in Theorem 2. Before embarking on its proof, we provide definitions and technical background results which will facilitate our exposition.

Definition 3 (Mollifier) *A continuous function $\phi : \mathbb{R}^s \to \mathbb{R}$ is called an s-dimensional mollifier if*

(a) $\phi(\boldsymbol{z}) \geq 0$,

(b) $\phi(\boldsymbol{z}) = 0$ *if* $|\boldsymbol{z}| \geq 1$ *and*

(c) $\int_{\mathbb{R}^s} \phi(\boldsymbol{z})\, \mathrm{d}\boldsymbol{z} = 1$.

Let $\phi : \mathbb{R} \to \mathbb{R}$ be a unimodal one-dimensional mollifier for which there exist $\zeta \in (0,1)$ and $\kappa > 0$ such that $\phi(z) \geq \kappa$ for all z with $|z| \leq \zeta$. Let $C := \max_{z\in\mathbb{R}} \phi(z)$ and define the k-dimensional mollifiers $\phi_m : \mathbb{R}^k \to \mathbb{R}$ corresponding to ϕ through

$$\phi_m(\boldsymbol{z}) := m^k \prod_{i=1}^{k} \phi(m z_i), \quad m \in \mathbb{N}. \tag{5.1}$$

Lemma 1 *The mollifier ϕ_m satisfies*

$$\int_\Xi \phi_m(\boldsymbol{\xi} - \boldsymbol{z})\, \mathrm{d}\boldsymbol{z} > 0 \quad \forall \boldsymbol{\xi} \in \Xi.$$

Proof Select any $\boldsymbol{\xi} \in \Xi$. By our assumption on ϕ, we have $\phi_m(\boldsymbol{z}) \geq (m\kappa)^k$ for all $\boldsymbol{z}$ with $|\boldsymbol{z}| \leq \zeta/m$, which implies that

$$\int_\Xi \phi_m(\boldsymbol{\xi} - \boldsymbol{z})\, \mathrm{d}\boldsymbol{z} \ \geq (m\kappa)^k \mu\left(\Xi \cap B_{\zeta/m}(\boldsymbol{\xi})\right). \tag{5.2}$$

Assumption (A5) guarantees that there exists $\boldsymbol{\xi}_0 \in \Xi$ and $\epsilon > 0$ such that $B_\epsilon(\boldsymbol{\xi}_0) \subseteq \Xi$. Next, we introduce the set

$$C(\boldsymbol{\xi}) \;:=\; \lambda \mathcal{B}_\epsilon(\boldsymbol{\xi}_0) + (1-\lambda)\boldsymbol{\xi} \;=\; \mathcal{B}_{\lambda\epsilon}(\lambda\boldsymbol{\xi}_0 + (1-\lambda)\boldsymbol{\xi}),$$

where

$$\lambda := \min\left\{ \frac{\zeta}{m(|\boldsymbol{\xi}-\boldsymbol{\xi}_0| + \epsilon)}, 1 \right\},$$

see Figure 2. In the following, we show that $C(\boldsymbol{\xi}) \subseteq \Xi \cap B_{\zeta/m}(\boldsymbol{\xi})$. Firstly, the set $C(\boldsymbol{\xi})$ was constructed as a convex combination of two sets contained in Ξ, and therefore $C(\boldsymbol{\xi}) \subseteq \Xi$. Secondly, note that

$$\begin{aligned} C(\boldsymbol{\xi}) \subseteq B_{\zeta/m}(\boldsymbol{\xi}) \quad &\Leftrightarrow \quad |\boldsymbol{\xi}' - \boldsymbol{\xi}| \le \frac{\zeta}{m} \quad \forall \boldsymbol{\xi}' \in C(\boldsymbol{\xi}) \\ &\Leftrightarrow \quad |\lambda(\boldsymbol{\xi}' - \boldsymbol{\xi})| \le \frac{\zeta}{m} \quad \forall \boldsymbol{\xi}' \in \mathcal{B}_\epsilon(\boldsymbol{\xi}_0) \\ &\Leftrightarrow \quad \begin{cases} |\boldsymbol{\xi}' - \boldsymbol{\xi}| \le \frac{\zeta}{m} \quad \forall \boldsymbol{\xi}' \in \mathcal{B}_\epsilon(\boldsymbol{\xi}_0) & \text{if } \lambda = 1, \\ |\boldsymbol{\xi}' - \boldsymbol{\xi}| \le |\boldsymbol{\xi} - \boldsymbol{\xi}_0| + \epsilon \quad \forall \boldsymbol{\xi}' \in \mathcal{B}_\epsilon(\boldsymbol{\xi}_0) & \text{else.} \end{cases} \end{aligned}$$

Both cases above are trivially satisfied. Therefore, $C(\boldsymbol{\xi}) \subseteq B_{\zeta/m}(\boldsymbol{\xi})$. In conclusion, we have $C(\boldsymbol{\xi}) \subseteq \Xi \cap B_{\zeta/m}(\boldsymbol{\xi})$ and consequently $\mu(\Xi \cap B_{\zeta/m}(\boldsymbol{\xi})) \ge \mu(C(\boldsymbol{\xi})) > 0$ since $C(\boldsymbol{\xi})$ is a k-ball of strictly positive radius. The claim now follows from (5.2). □

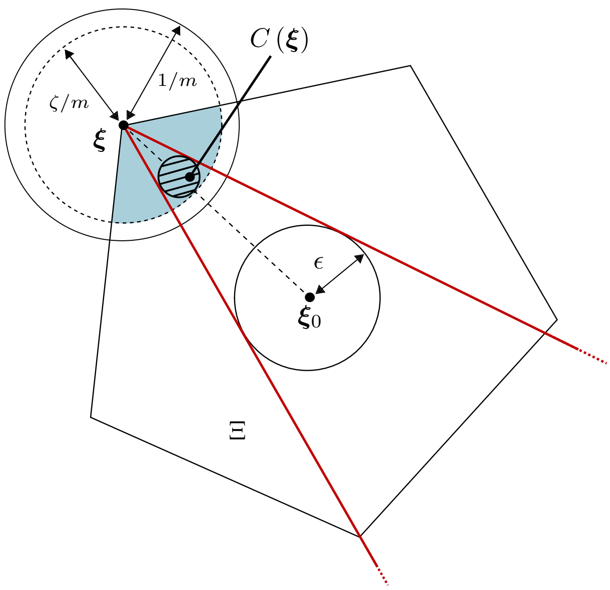


Figure 2: The measure of $C(\boldsymbol{\xi})$ (striped area) provides a lower bound for the measure of $\Xi \cap B_{\zeta/m}(\boldsymbol{\xi})$ (shaded area).

Lemma 2 *Let $\boldsymbol{x}$ be feasible for problem $\mathcal{R}$ and define the function $\boldsymbol{x}^m : \mathbb{R}^k \to \mathbb{R}^n$ through*

$$\boldsymbol{x}^m(\boldsymbol{\xi}) := \frac{\int_\Xi \boldsymbol{x}(\boldsymbol{z})\, \phi_m(\boldsymbol{\xi} - \boldsymbol{z})\, \mathrm{d}\boldsymbol{z}}{\int_\Xi \phi_m(\boldsymbol{\xi} - \boldsymbol{z})\, \mathrm{d}\boldsymbol{z}}. \tag{5.3}$$

Then,

(a) $\boldsymbol{x}^m$ *is continuous on* Ξ, *and*

(b) $\boldsymbol{x}^m \in \mathcal{N}$.

Proof (a) Select $\boldsymbol{\xi}' \in \Xi$. Then, we have

$$\lim_{\boldsymbol{\xi}\to\boldsymbol{\xi}'} \int_{\Xi} \boldsymbol{x}(\boldsymbol{z})\, \phi_m(\boldsymbol{\xi}-\boldsymbol{z})\, \mathrm{d}\boldsymbol{z} \;=\; \int_{\Xi} \lim_{\boldsymbol{\xi}\to\boldsymbol{\xi}'} \boldsymbol{x}(\boldsymbol{z})\, \phi_m(\boldsymbol{\xi}-\boldsymbol{z})\, \mathrm{d}\boldsymbol{z} \;=\; \int_{\Xi} \boldsymbol{x}(\boldsymbol{z})\, \phi_m(\boldsymbol{\xi}'-\boldsymbol{z})\, \mathrm{d}\boldsymbol{z}.$$

The interchange of the limit and the integral is justified by the dominated convergence theorem, which applies because $|\boldsymbol{x}(\boldsymbol{z})\, \phi_m(\boldsymbol{\xi}-\boldsymbol{z})| \le m^k CR$ uniformly for all $\boldsymbol{\xi}, \boldsymbol{z} \in \Xi$, see assumption (A3). Thus, the numerator of $\boldsymbol{x}^m$ in (5.3) is continuous on Ξ. Following similar arguments, it is possible to prove that the denominator of $\boldsymbol{x}^m$ is also continuous on Ξ. Thus, $\boldsymbol{x}^m$ is continuous on Ξ as it is a ratio of two continuous functions whose denominator is strictly positive, see Lemma 1.

(b) For each $t \in \mathbb{T}$, we introduce a new sequence of mollifiers $\phi_{m,t} : \mathbb{R}^{k_t} \to \mathbb{R}$, $m \in \mathbb{N}$, which are defined in terms of ϕ through

$$\phi_{m,t}(\boldsymbol{z}_t) := \prod_{j=1}^{k_t} m\phi(m z_{t,j}).$$

Using this definition, ϕ_m can be written as

$$\phi_m(\boldsymbol{z}) = \prod_{t=1}^{T} \phi_{m,t}(\boldsymbol{z}_t), \; m \in \mathbb{N}. \tag{5.4}$$

Select an arbitrary $t \in \mathbb{T}$. As $\boldsymbol{x}_t$ is $\mathcal{F}^t$-measurable, there exists a function $\boldsymbol{\chi}_t : \mathbb{R}^{k^t} \to \mathbb{R}$ such that $\boldsymbol{x}_t(\boldsymbol{\xi}) = \boldsymbol{\chi}_t(\boldsymbol{\xi}^t)$ for all $\boldsymbol{\xi} \in \mathbb{R}^k$. Thus, we have

$$\begin{aligned}\boldsymbol{x}_t^m(\boldsymbol{\xi}) &= \frac{\int_{\Xi} \boldsymbol{x}_t(\boldsymbol{z})\, \phi_m(\boldsymbol{\xi}-\boldsymbol{z})\, \mathrm{d}\boldsymbol{z}}{\int_{\Xi} \phi_m(\boldsymbol{\xi}-\boldsymbol{z})\, \mathrm{d}\boldsymbol{z}} \\ &= \frac{\int_{\Xi_1} \cdots \int_{\Xi_t} \boldsymbol{\chi}_t(\boldsymbol{z}^t) \prod_{\tau=1}^{t} \phi_{m,\tau}(\boldsymbol{\xi}_\tau - \boldsymbol{z}_\tau)\, \mathrm{d}\boldsymbol{z}_1 \cdots \mathrm{d}\boldsymbol{z}_t}{\int_{\Xi_1} \cdots \int_{\Xi_t} \prod_{\tau=1}^{t} \phi_{m,\tau}(\boldsymbol{\xi}_\tau - \boldsymbol{z}_\tau)\, \mathrm{d}\boldsymbol{z}_1 \cdots \mathrm{d}\boldsymbol{z}_t},\end{aligned}$$

where the second equality holds due to (5.4) and since $\Xi = \times_{\tau=1}^{T} \Xi_\tau$, see assumption (A5). The last expression for $\boldsymbol{x}_t^m$ is independent of $\boldsymbol{\xi}_{t+1}, \ldots, \boldsymbol{\xi}_T$. Therefore $\boldsymbol{x}_t^m$ is $\mathcal{F}^t$-measurable. As the choice of $t \in \mathbb{T}$ was arbitrary, the decision rule $\boldsymbol{x}^m$ is non-anticipative. Part (a) further implies that $\boldsymbol{x}^m$ is continuous and bounded on the compact set Ξ. Therefore $\boldsymbol{x}^m \in \mathcal{N}$ as postulated. □

Theorem 2 (Convergence of the decision rule approximation) *The infimum of the robust problem* $\mathcal{RB}$ *with basis functions converges to the infimum of the original problem* $\mathcal{R}$ *as the complexity parameter* d *tends to infinity.*

Proof The general strategy of our proof is to choose an arbitrary tolerance level $\epsilon > 0$ and to construct a sequence of four ϵ-optimal decision rules feasible in $\mathcal{R}$ with increasingly regular structure. The four decision rules will be (a) measurable, (b) measurable and strictly feasible, (c) continuous and strictly feasible and (d) representable as a finite linear combination of basis functions.

Select $\epsilon > 0$. Since problem $\mathcal{R}$ is feasible, there exists a feasible decision rule $\boldsymbol{x}^{(1)} \in \mathcal{N}$ with

$$|\sup_{\boldsymbol{\xi}\in\Xi} f_0(\boldsymbol{x}^{(1)}(\boldsymbol{\xi}), \boldsymbol{\xi}) - \inf \mathcal{R}| \leq \frac{\epsilon}{4}. \tag{5.5}$$

Similarly, since the robust optimization problem $\mathcal{R}$ is strictly feasible, there exists a strictly feasible decision rule $\boldsymbol{x}^{(2)} \in \mathcal{N}$ and $\delta^{(2)} > 0$ such that

$$f_i(\boldsymbol{x}^{(2)}(\boldsymbol{\xi}), \boldsymbol{\xi}) \leq -\delta^{(2)} \quad \forall \boldsymbol{\xi} \in \Xi,\ i = 1, \ldots, I.$$

Set $\boldsymbol{x}^{\lambda} := (1-\lambda)\boldsymbol{x}^{(1)} + \lambda \boldsymbol{x}^{(2)}$ for $\lambda \in [0,1]$. Since the functions f_i are convex in $\boldsymbol{x}$, we have

$$f_i(\boldsymbol{x}^{\lambda}(\boldsymbol{\xi}), \boldsymbol{\xi}) \leq (1-\lambda)\, f_i(\boldsymbol{x}^{(1)}(\boldsymbol{\xi}), \boldsymbol{\xi}) + \lambda f_i(\boldsymbol{x}^{(2)}(\boldsymbol{\xi}), \boldsymbol{\xi}) \leq -\lambda\delta^{(2)}\ \forall \boldsymbol{\xi} \in \Xi,\ i = 1, \ldots, I.$$

Therefore, the causal decision rule $\boldsymbol{x}^{\lambda} \in \mathcal{N}$ is strictly feasible in $\mathcal{R}$ for all $\lambda > 0$. The function f_0 in the objective of $\mathcal{R}$ is uniformly continuous on the compact set $X \times \Xi$. Thus, there exists $\eta > 0$ such that

$$|f_0(\boldsymbol{x}, \boldsymbol{\xi}) - f_0(\boldsymbol{x}', \boldsymbol{\xi})| \leq \frac{\epsilon}{4} \quad \forall \boldsymbol{x},\ \boldsymbol{x}' \in X \text{ with } |\boldsymbol{x} - \boldsymbol{x}'| \leq \eta \text{ and } \boldsymbol{\xi} \in \Xi. \tag{5.6}$$

Since $\boldsymbol{x}^{(1)}$ and $\boldsymbol{x}^{(2)}$ are both feasible for $\mathcal{R}$, assumption (A3) implies

$$|\boldsymbol{x}^{\lambda}(\boldsymbol{\xi}) - \boldsymbol{x}^{(1)}(\boldsymbol{\xi})| = |\lambda \boldsymbol{x}^{(2)}(\boldsymbol{\xi}) - \lambda \boldsymbol{x}^{(1)}(\boldsymbol{\xi})| \leq 2\lambda R \quad \forall \boldsymbol{\xi} \in \Xi. \tag{5.7}$$

Setting $\lambda_0 := \frac{\eta}{2R}$ and $\boldsymbol{x}^{(3)} := \boldsymbol{x}^{\lambda_0} \in \mathcal{N}$, we conclude via (5.6) and (5.7) that

$$|f_0(\boldsymbol{x}^{(3)}(\boldsymbol{\xi}), \boldsymbol{\xi}) - f_0(\boldsymbol{x}^{(1)}(\boldsymbol{\xi}), \boldsymbol{\xi})| \leq \frac{\epsilon}{4} \quad \forall \boldsymbol{\xi} \in \Xi,$$

and therefore

$$|\sup_{\boldsymbol{\xi}\in\Xi} f_0(\boldsymbol{x}^{(3)}(\boldsymbol{\xi}), \boldsymbol{\xi}) - \sup_{\boldsymbol{\xi}\in\Xi} f_0(\boldsymbol{x}^{(1)}(\boldsymbol{\xi}), \boldsymbol{\xi})| \leq \frac{\epsilon}{4}. \tag{5.8}$$

We note that our choice of λ_0 implies

$$f_i(\boldsymbol{x}^{(3)}(\boldsymbol{\xi}), \boldsymbol{\xi}) \leq -\delta^{(3)} \quad \forall \boldsymbol{\xi} \in \Xi,\ i = 1, \ldots, I, \tag{5.9}$$

where $\delta^{(3)} := \eta\delta^{(2)}/2R$. Thus, $\boldsymbol{x}^{(3)}$ constitutes a strictly feasible, near optimal decision rule.

Now, let the mollifier sequence $\{\phi_m\}_{m\in\mathbb{N}}$ be as in (5.1) and define the mollified decision rules $\{\boldsymbol{x}^m\}_{m\in\mathbb{N}}$ through

$$\boldsymbol{x}^m(\boldsymbol{\xi}) = \frac{\int_{\Xi} \boldsymbol{x}^{(3)}(\boldsymbol{z})\phi_m(\boldsymbol{\xi} - \boldsymbol{z})\mathrm{d}\boldsymbol{z}}{\int_{\Xi} \phi_m(\boldsymbol{\xi} - \boldsymbol{z})\mathrm{d}\boldsymbol{z}}.$$

Lemma 2 implies that each $\boldsymbol{x}^m$ is non-anticipative and continuous on Ξ. Since the objective and constraint functions f_i, $i = 0, \ldots, I$ are convex in $\boldsymbol{x}$, we can apply Jensen's inequality to obtain

$$f_i(\boldsymbol{x}^m(\boldsymbol{\xi}), \boldsymbol{\xi}) \;\; \leq \;\; \int_\Xi f_i(\boldsymbol{x}^{(3)}(\boldsymbol{z}), \boldsymbol{\xi}) \frac{\phi_m(\boldsymbol{\xi} - \boldsymbol{z})}{\int_\Xi \phi_m(\boldsymbol{\xi} - \boldsymbol{y}) \mathrm{d}\boldsymbol{y}} \mathrm{d}\boldsymbol{z}. \tag{5.10}$$

Next, set $\delta^{(4)} := \delta^{(3)}/2$. By the uniform continuity of f_i on $X \times \Xi$, there exists $m_0 \in \mathbb{N}$ such that

$$|f_i(\boldsymbol{x}^{(3)}(\boldsymbol{z}), \boldsymbol{z}) - f_i(\boldsymbol{x}^{(3)}(\boldsymbol{z}), \boldsymbol{\xi})| \leq \min\left\{\delta^{(4)}, \frac{\epsilon}{4}\right\} \quad \forall \boldsymbol{\xi},\ \boldsymbol{z} \in \Xi \text{ with } |\boldsymbol{\xi} - \boldsymbol{z}| \leq \frac{1}{m_0},\ i = 0, \ldots, I. \tag{5.11}$$

For $i = 1, \ldots, I$ (constraint functions) and $m \geq m_0$, we thus have

$$f_i(\boldsymbol{x}^m(\boldsymbol{\xi}), \boldsymbol{\xi}) \;\leq\; \int_\Xi \left(f_i(\boldsymbol{x}^{(3)}(\boldsymbol{z}), \boldsymbol{z}) + \delta^{(4)}\right) \frac{\phi_m\left(\boldsymbol{\xi} - \boldsymbol{z}\right)}{\int_\Xi \phi_m\left(\boldsymbol{\xi} - \boldsymbol{y}\right) \mathrm{d}\boldsymbol{y}} \mathrm{d}\boldsymbol{z} \;\leq\; \delta^{(4)} - \delta^{(3)} \;=\; -\delta^{(4)} \quad \forall \boldsymbol{\xi} \in \Xi,$$

where the first inequality follows from (5.10) and (5.11), while the second inequality holds because of (5.9). For $i = 0$ (objective function) and $m \geq m_0$, we can make a similar argument to obtain

$$f_0(\boldsymbol{x}^m(\boldsymbol{\xi}), \boldsymbol{\xi}) \;\leq\; \int_\Xi f_0(\boldsymbol{x}^{(3)}(\boldsymbol{z}), \boldsymbol{z}) \frac{\phi_m(\boldsymbol{\xi} - \boldsymbol{z})}{\int_\Xi \phi_m(\boldsymbol{\xi} - \boldsymbol{y}) \mathrm{d}\boldsymbol{y}} \mathrm{d}\boldsymbol{z} + \frac{\epsilon}{4} \;\leq\; \sup_{\boldsymbol{z} \in \Xi} f_0(\boldsymbol{x}^{(3)}(\boldsymbol{z}), \boldsymbol{z}) + \frac{\epsilon}{4} \quad \forall \boldsymbol{\xi} \in \Xi. \tag{5.12}$$

Next, define $\boldsymbol{x}^{(4)} := \boldsymbol{x}^{m_0}$. By construction, $\boldsymbol{x}^{(4)}$ is non-anticipative and continuous on Ξ. Moreover, it is near optimal and strictly feasible in $\mathcal{R}$.

By the uniform continuity of f_i on $X \times \Xi$, there exists $\eta > 0$ such that

$$|f_i(\boldsymbol{x}, \boldsymbol{\xi}) - f_i(\boldsymbol{x}', \boldsymbol{\xi})| \leq \min\left\{\frac{\epsilon}{4}, \delta^{(4)}\right\} \quad \forall \boldsymbol{x},\ \boldsymbol{x}' \in X \text{ with } |\boldsymbol{x} - \boldsymbol{x}'| \leq \eta,\ \boldsymbol{\xi} \in \Xi,\ i = 0, \ldots, I. \tag{5.13}$$

By construction of the basis vectors $\boldsymbol{b}_t^d$ and since $\boldsymbol{x}^{(4)}$ is continuous on Ξ, there exists a complexity parameter $d \in \mathbb{N}$ and matrices $\boldsymbol{X}_t \in \mathbb{R}^{n_t \times s_d(k^t)}$, $t \in \mathbb{T}$, such that

$$|\boldsymbol{x}^{(5)}(\boldsymbol{\xi}) - \boldsymbol{x}^{(4)}(\boldsymbol{\xi})| \leq \eta \quad \forall \boldsymbol{\xi} \in \Xi,$$

where $\boldsymbol{x}^{(5)} \in \mathcal{N}$ is defined through $\boldsymbol{x}_t^{(5)}(\boldsymbol{\xi}) := \boldsymbol{X}_t\, \boldsymbol{b}_t^d(\boldsymbol{\xi}^t)$, see condition (C1). The estimate (5.13) and the strict feasibility of $\boldsymbol{x}^{(4)}$ imply that

$$|\sup_{\boldsymbol{\xi} \in \Xi} f_0(\boldsymbol{x}^{(5)}(\boldsymbol{\xi}), \boldsymbol{\xi}) - \sup_{\boldsymbol{\xi} \in \Xi} f_0(\boldsymbol{x}^{(4)}(\boldsymbol{\xi}), \boldsymbol{\xi})| \leq \frac{\epsilon}{4} \tag{5.14}$$

and

$$f_i(\boldsymbol{x}^{(5)}(\boldsymbol{\xi}), \boldsymbol{\xi}) \leq 0 \quad \forall \boldsymbol{\xi} \in \Xi,\ i = 1, \ldots, I.$$

We have thus shown that $\boldsymbol{x}^{(5)}$ is feasible for $\mathcal{R}$, and we also obtain

$$\begin{aligned}
0 \;\leq\; \sup_{\boldsymbol{\xi}\in\Xi} f_0(\boldsymbol{x}^{(5)}(\boldsymbol{\xi}),\boldsymbol{\xi}) - \inf\mathcal{R} \;&=\; \sup_{\boldsymbol{\xi}\in\Xi}\; f_0(\boldsymbol{x}^{(1)}(\boldsymbol{\xi}),\boldsymbol{\xi}) - \inf\mathcal{R} \\
&+\; \sup_{\boldsymbol{\xi}\in\Xi}\; f_0(\boldsymbol{x}^{(3)}(\boldsymbol{\xi}),\boldsymbol{\xi}) - \sup_{\boldsymbol{\xi}\in\Xi}\; f_0(\boldsymbol{x}^{(1)}(\boldsymbol{\xi}),\boldsymbol{\xi}) \\
&+\; \sup_{\boldsymbol{\xi}\in\Xi}\; f_0(\boldsymbol{x}^{(4)}(\boldsymbol{\xi}),\boldsymbol{\xi}) - \sup_{\boldsymbol{\xi}\in\Xi}\; f_0(\boldsymbol{x}^{(3)}(\boldsymbol{\xi}),\boldsymbol{\xi}) \\
&+\; \sup_{\boldsymbol{\xi}\in\Xi} f_0(\boldsymbol{x}^{(5)}(\boldsymbol{\xi}),\boldsymbol{\xi}) - \sup_{\boldsymbol{\xi}\in\Xi} f_0(\boldsymbol{x}^{(4)}(\boldsymbol{\xi}),\boldsymbol{\xi}) \\
&\leq\; \epsilon,
\end{aligned}$$

where the first inequality follows from the feasibility of $\boldsymbol{x}^{(5)}$ in $\mathcal{R}$, and the last inequality follows from (5.5), (5.8), (5.12) and (5.14). We thus conclude that the optimal value of problem $\mathcal{RB}$ differs at most by ϵ from the optimal value of $\mathcal{R}$. As the choice of ϵ was arbitrary, the claim follows. □

5.2 Convergence of the sampling approximation

In this section, we demonstrate that the optimal value of $\mathcal{SB}_N$ converges with probability 1 (w.p.1) to the optimal value of $\mathcal{RB}$ as the number of samples tends to infinity whenever problem $\mathcal{RB}$ is feasible. Thus, we will henceforth assume that $\mathcal{RB}$ is indeed feasible. Note that this can always be enforced by choosing the complexity parameter d large enough, see Theorem 2. The general strategy of our proof is as follows. We will first consider a fixed realization of the stochastic process $\{\boldsymbol{\xi}^{(l)}\}_{l\in\mathbb{N}}$ that is dense in Ξ. For this particular sequence of samples, we will prove that $\inf\mathcal{SB}_N$ converges to $\inf\mathcal{RB}$ as N tends to infinity. To prove the main result, we will then argue that $\{\boldsymbol{\xi}^{(l)}\}_{l\in\mathbb{N}}$ is dense in Ξ w.p.1.

For the further argumentation, we introduce the function $\psi:\mathbb{R}^{n_w}\to\mathbb{R}$, $\psi(\boldsymbol{w}):=\max_{\boldsymbol{\xi}\in\Xi} f(\boldsymbol{w},\boldsymbol{\xi})$ and the corresponding random function $\psi_N:\mathbb{R}^{n_w}\to\mathbb{R}$, $\psi_N(\boldsymbol{w}):=\max_{\boldsymbol{\xi}\in\Xi_N} f(\boldsymbol{w},\boldsymbol{\xi})$, where the random set $\Xi_N\subseteq\mathbb{R}^k$ is defined through $\Xi_N:=\{\boldsymbol{\xi}^{(1)},\ldots,\boldsymbol{\xi}^{(N)}\}$, $N\in\mathbb{N}$.

Remark 5 *By construction, the feasible set of problem $\mathcal{RB}$ is given by $\{\boldsymbol{w}\in\mathbb{R}^{n_w}:\psi(\boldsymbol{w})\leq 0\}$, while the feasible set of $\mathcal{SB}_N$ is representable as $\{\boldsymbol{w}\in\mathbb{R}^{n_w}:\psi_N(\boldsymbol{w})\leq 0\}$.*

Lemma 3 *If $\{\boldsymbol{\xi}^{(l)}\}_{l\in\mathbb{N}}$ is dense in Ξ, then ψ_N converges continuously to ψ on $\mathbb{R}^{n_w}$.*

Proof We first demonstrate that ψ_N converges pointwise to ψ on $\mathbb{R}^{n_w}$. By the continuity of f and the compactness of Ξ, $\psi(\boldsymbol{w})$ exists and is finite for each $\boldsymbol{w}\in\mathbb{R}^{n_w}$. Also, $\{\psi_N(\boldsymbol{w})\}_{N\in\mathbb{N}}$ is a non-decreasing sequence bounded above by $\psi(\boldsymbol{w})$. We thus conclude that

$$\lim_{N\to\infty}\psi_N(\boldsymbol{w}) \;=\; \sup_{N\in\mathbb{N}}\max_{\boldsymbol{\xi}\in\Xi_N} f(\boldsymbol{w},\boldsymbol{\xi}) \;=\; \max_{\boldsymbol{\xi}\in\Xi} f(\boldsymbol{w},\boldsymbol{\xi}) \;=\; \psi(\boldsymbol{w}),$$

where the second equality holds since f is continuous and $\{\boldsymbol{\xi}^{(l)}\}_{l\in\mathbb{N}}$ is dense in Ξ. Thus, ψ_N converges pointwise to ψ on $\mathbb{R}^{n_w}$.

We now prove the main result. Choose $\boldsymbol{w} \in \mathbb{R}^{n_w}$ and $\epsilon > 0$. Since f is continuous and Ξ is compact, there exists $\delta > 0$ such that

$$|f(\boldsymbol{w}, \boldsymbol{\xi}) - f(\boldsymbol{w}', \boldsymbol{\xi})| \leq \epsilon \ \ \forall \boldsymbol{w}' \in B_\delta(\boldsymbol{w}), \ \boldsymbol{\xi} \in \Xi.$$

Thus, we find

$$|\psi_N(\boldsymbol{w}) - \psi_N(\boldsymbol{w}')| = |\max_{\boldsymbol{\xi} \in \Xi_N} f(\boldsymbol{w}, \boldsymbol{\xi}) - \max_{\boldsymbol{\xi} \in \Xi_N} f(\boldsymbol{w}', \boldsymbol{\xi})| \leq \epsilon \ \ \forall N \in \mathbb{N}, \ \forall \boldsymbol{w}' \in B_\delta(\boldsymbol{w}). \tag{5.15}$$

The sequence $\{\psi_N\}_{N \in \mathbb{N}}$ is locally bounded at $\boldsymbol{w}$. As the choice of ϵ was arbitrary, we thus conclude from (5.15) that the sequence is equicontinuous at $\boldsymbol{w}$, see e.g. pp.248–249 in Rockafellar and Wets [1997]. The above argument holds for all $\boldsymbol{w} \in \mathbb{R}^{n_w}$, and thus $\{\psi_N\}_{N \in \mathbb{N}}$ is equicontinuous on $\mathbb{R}^{n_w}$. The pointwise convergence of ψ_N to ψ on $\mathbb{R}^{n_w}$ and Theorem 7.10 in Rockafellar and Wets [1997] ensure that ψ_N epiconverges to ψ, which implies via Theorem 7.11 in Rockafellar and Wets [1997] that ψ_N converges continuously to ψ. □

Lemma 4 *If $\{\boldsymbol{\xi}^{(l)}\}_{l \in \mathbb{N}}$ is dense in Ξ, then there exist $N_0 \in \mathbb{N}$ and a non-empty compact set $W^* \subseteq \mathbb{R}^{n_w}$ such that W^* contains at least one optimal solution of problem $\mathcal{SB}_N$ for each $N \geq N_0$.*

Remark 6 *We emphasize that W^* and N_0 may depend on the particular realization of $\{\boldsymbol{\xi}^{(l)}\}_{l \in \mathbb{N}}$. We further note that Lemma 4 implicitly guarantees that $\mathcal{SB}_N$ is solvable (i.e., the minimum is attained) and has a finite optimal value for all $N \geq N_0$.*

Proof of Lemma 4 See Appendix A. □

Lemma 5 *If $\{\boldsymbol{\xi}^{(l)}\}_{l \in \mathbb{N}}$ is dense in Ξ, then the infimum of the scenario problem $\mathcal{SB}_N$ converges to the infimum of $\mathcal{RB}$ as the number of samples N tends to infinity.*

The following proof is inspired by the proof of Proposition 2.2 in Pagnoncelli et al. [2009], where almost sure convergence of sample average approximations to chance constrained programs is demonstrated.

Proof of Lemma 5 Since $\mathcal{SB}_N$ is a relaxation of $\mathcal{RB}$, we have

$$\limsup_{N \to \infty} \ \inf \mathcal{SB}_N \ \leq \ \inf \mathcal{RB}. \tag{5.16}$$

For all $N \geq N_0$, let $\boldsymbol{w}_N^* \in W^*$ denote an optimal solution to $\mathcal{SB}_N$, whose existence is guaranteed by Lemma 4. Consider now a subsequence $\{\boldsymbol{w}_{N_j}^*\}_{j \in \mathbb{N}}$ with the property that $\liminf_{N \to \infty} \boldsymbol{e}_1^\top \boldsymbol{w}_N^* = \lim_{j \to \infty} \boldsymbol{e}_1^\top \boldsymbol{w}_{N_j}^*$. Since any $\boldsymbol{w}_N^*$ is an element of the compact set W^*, we may assume without loss of generality that $\lim_{j \to \infty} \boldsymbol{w}_{N_j}^* = \boldsymbol{w}^*$ for some $\boldsymbol{w}^* \in W^*$. The continuous convergence of ψ_N to ψ, which is ensured by Lemma 3, then implies

$$\lim_{j \to \infty} \psi_N(\tilde{\boldsymbol{w}}_{N_j}^*) \ = \ \psi(\boldsymbol{w}^*).$$

As $\psi_{N_j}(\boldsymbol{w}^*_{N_j}) \le 0$ for all $j \in \mathbb{N}$, we conclude that $\psi(\boldsymbol{w}^*) \le 0$, that is, $\boldsymbol{w}^*$ is feasible in $\mathcal{RB}$, and thus $\boldsymbol{e}_1^\top \boldsymbol{w}^* \ge \inf \mathcal{RB}$. Hence, we find

$$\liminf_{N\to\infty} \inf \mathcal{SB}_N \;=\; \liminf_{N\to\infty} \boldsymbol{e}_1^\top \boldsymbol{w}^*_N \;=\; \lim_{j\to\infty} \boldsymbol{e}_1^\top \boldsymbol{w}^*_{N_j} \;=\; \boldsymbol{e}_1^\top \boldsymbol{w}^* \;\ge\; \inf \mathcal{RB}. \tag{5.17}$$

It follows from (5.16) and (5.17) that $\inf \mathcal{SB}_N$ converges to $\inf \mathcal{RB}$. □

Theorem 3 (Convergence of the sampling approximation) *The infimum of the scenario problem $\mathcal{SB}_N$ converges w.p.1 to the infimum of $\mathcal{RB}$ as the number of samples N tends to infinity.*

Proof It can be shown that $\{\boldsymbol{\xi}^{(l)}\}_{l\in\mathbb{N}}$ is dense in Ξ w.p.1, see Appendix B. Thus, Theorem 3 is a simple corollary of Lemma 5. □

6 Examples from inventory management

We assess the convergence and scalability properties of our approach in the context of two single-echelon, multi-period supply chain models from the literature.

6.1 Inventory management with cumulative order constraints

The first problem considered here was originally proposed by Ben-Tal et al. [2005]. We discuss a simplified version due to Bertsimas et al. [2010], which we denote by $\mathcal{R}^{\mathrm{COC}}$.

At the beginning of each period $t \in \mathbb{T}$, a retailer receives orders for $\xi_{\mathrm{d},t}$ units of a product. This demand needs to be satisfied from the on-hand inventory, whose filling level is denoted by $s_{\mathrm{inv},t}$. In order to replenish the inventory, the retailer may place orders $x_{\mathrm{o},t}$ with a supplier, thereby incurring shipping costs d_t per unit of the product. Unsatisfied demand may be backlogged at cost p_t and inventory may be held on the premises at cost h_t per unit of the product. Furthermore, there are prescribed limits on the orders placed at each period as well as on the cumulative orders $s_{\mathrm{co},t}$ placed up to period t.

The dynamics of the inventory level and the cumulative orders are governed by

$$\left.\begin{aligned} s_{\mathrm{inv},t+1}(\boldsymbol{\xi}^{t+1}) &= s_{\mathrm{inv},t}(\boldsymbol{\xi}^{t}) + x_{\mathrm{o},t}(\boldsymbol{\xi}^{t}) - \xi_{\mathrm{d},t+1} \\ s_{\mathrm{co},t+1}(\boldsymbol{\xi}^{t+1}) &= s_{\mathrm{co},t}(\boldsymbol{\xi}^{t}) + x_{\mathrm{o},t}(\boldsymbol{\xi}^{t}) \end{aligned}\right\} \quad t = 1, \ldots, T-1,$$

where $s_{\mathrm{inv},1}$ denotes the initial inventory and $s_{\mathrm{co},1}$ the initial cumulative order level. We impose the box constraints

$$\begin{aligned} \underline{x}_{\mathrm{o},t} \;\le\; x_{\mathrm{o},t}(\boldsymbol{\xi}^t) \;\le\; \overline{x}_{\mathrm{o},t},\; t = 1, \ldots, T-1 \\ \underline{s}_{\mathrm{co},t} \;\le\; s_{\mathrm{co},t}(\boldsymbol{\xi}^t) \;\le\; \overline{s}_{\mathrm{co},t},\; t = 1, \ldots, T, \end{aligned}$$

where $\underline{x}_{\mathrm{o},t}$, $\overline{x}_{\mathrm{o},t}$ and $\underline{s}_{\mathrm{co},t}$, $\overline{s}_{\mathrm{co},t}$ denote the lower and upper bounds on the instantaneous and cumulative orders, respectively.

For simplicity, we assume that there is no demand at time $t = 1$. The future demands are independent and uniformly distributed as $\xi_{\mathrm{d},t} \sim \mathcal{U}\left(\overline{\xi}_{\mathrm{d},t}\left(1-\rho_{\mathrm{d}}\right), \overline{\xi}_{\mathrm{d},t}\left(1+\rho_{\mathrm{d}}\right)\right)$, where $\overline{\xi}_{\mathrm{d},t}$ denotes the nominal demand in period t, and $\rho_{\mathrm{d}} \in [0,1]$ quantifies the degree of uncertainty. The support of $\boldsymbol{\xi}$ thus corresponds to a box uncertainty set of the form $\Xi = \times_{t\in\mathbb{T}}\{\xi_{\mathrm{d},t} \in \mathbb{R} : |\xi_{\mathrm{d},t} - \overline{\xi}_{\mathrm{d},t}| \leq \rho_{\mathrm{d}}\overline{\xi}_{\mathrm{d},t}\}$.

The retailer's objective is to minimize the worst-case cumulative cost $\max_{\boldsymbol{\xi}\in\Xi}\sum_{t\in\mathbb{T}} x_{\mathrm{c},t}(\boldsymbol{\xi}^t)$, where the stage-wise costs $x_{\mathrm{c},t}$ satisfy

$$\begin{aligned} x_{\mathrm{c},t}(\boldsymbol{\xi}^t) &\geq d_t x_{\mathrm{o},t}(\boldsymbol{\xi}^t) + \max\left\{h_t s_{\mathrm{inv},t}(\boldsymbol{\xi}^t), -p_t s_{\mathrm{inv},t}(\boldsymbol{\xi}^t)\right\}, \ t = 1,\dots,T-1 \\ x_{\mathrm{c},T}(\boldsymbol{\xi}^T) &\geq \max\left\{h_T s_{\mathrm{inv},T}(\boldsymbol{\xi}^T), -p_T s_{\mathrm{inv},T}(\boldsymbol{\xi}^T)\right\}. \end{aligned} \tag{6.1}$$

6.2 Inventory management with random yield

As an extension to the basic model from Section 6.1, we assume now that the quality of the shipments received from the supplier is uncertain in the sense that only a fraction $\xi_{\mathrm{y},t}$ of the total number of products ordered in period t is of satisfactory quality. We denote this modified problem by $\mathcal{R}^{\mathrm{RY}}$. We assume that the yields $\xi_{\mathrm{y},t}$ are mutually independent and uniformly distributed as $\xi_{\mathrm{y},t} \sim \mathcal{U}(\overline{\xi}_{\mathrm{y},t}(1-\rho_{\mathrm{y}}), \overline{\xi}_{\mathrm{y},t}(1+\rho_{\mathrm{y}}))$, with $[\overline{\xi}_{\mathrm{y},t}(1-\rho_{\mathrm{y}}), \overline{\xi}_{\mathrm{y},t}(1+\rho_{\mathrm{y}})] \subseteq [0,1]$. We further assume that the retailer only pays for the shipments of sufficient quality and that the payment for orders placed at time t is made at time $t+1$, *after* the shipment quality has been observed. These amendments to the basic model result in the following modified system dynamics

$$\left.\begin{aligned} s_{\mathrm{inv},t+1}(\boldsymbol{\xi}^{t+1}) &= s_{\mathrm{inv},t}(\boldsymbol{\xi}^t) + \xi_{\mathrm{y},t+1} x_{\mathrm{o},t}(\boldsymbol{\xi}^t) - \xi_{\mathrm{d},t+1} \\ s_{\mathrm{co},t+1}(\boldsymbol{\xi}^{t+1}) &= s_{\mathrm{co},t}(\boldsymbol{\xi}^t) + x_{\mathrm{o},t}(\boldsymbol{\xi}^t) \end{aligned}\right\} \quad t = 1,\dots,T-1,$$

and modified cost constraints

$$\begin{aligned} x_{\mathrm{c},1} &\geq \max\{h_1 s_{\mathrm{inv},1}, -p_1 s_{\mathrm{inv},1}\} \\ x_{\mathrm{c},t}(\boldsymbol{\xi}^t) &\geq d_{t-1}\xi_{\mathrm{y},t} x_{\mathrm{o},t-1}(\boldsymbol{\xi}^{t-1}) + \max\left\{h_t s_{\mathrm{inv},t}(\boldsymbol{\xi}^t), -p_t s_{\mathrm{inv},t}(\boldsymbol{\xi}^t)\right\}, \ t = 2,\dots,T. \end{aligned}$$

An illustration of the problem flow is provided in Figure 3.

The extended model described here constitutes a multi-stage robust optimization problem with random recourse. Such problems are generically NP-hard even if they are solved in linear decision rules. However, in the case of linear decision rules, a tight tractable approximation can be obtained in the form of a conic-quadratic or semi-definite program (see [Ben-Tal et al., 2004, §4]). In what follows, we illustrate how our approach can be used to approximately solve this problem in polynomial decision rules. We remark that the problems described in Sections 6.1 and 6.2 can be brought to the form $\mathcal{R}$, without equality constraints, by eliminating the state variables $s_{\mathrm{inv},t}$ and $s_{\mathrm{co},t}$.

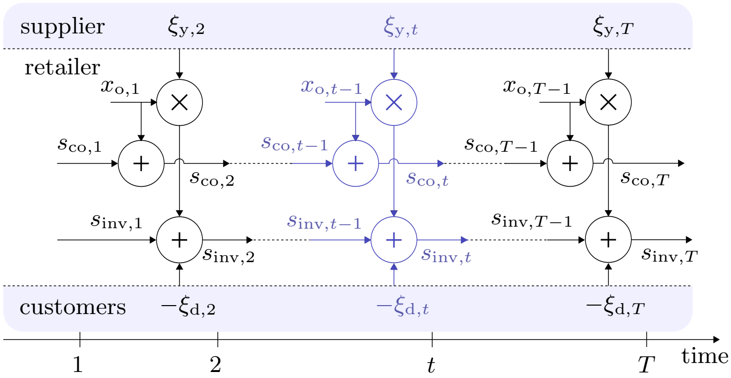


Figure 3: A single-echelon supply chain with T stages, uncertain customer demand and yield.

7 Numerical experiments

In this section, we consider specific instances of the problems $\mathcal{R}^{\text{COC}}$ and $\mathcal{R}^{\text{RY}}$. The input data for the two test problems is summarized in Table 1. We report on optimality gaps, empirical violation probabilities and solver times for the corresponding approximate problems $\mathcal{SB}_N^{\text{COC}}$ and $\mathcal{SB}_N^{\text{RY}}$. We define the optimality gaps of the problems $\mathcal{SB}_N$ and $\mathcal{RB}$ as $(\inf \mathcal{SB}_N - \inf \mathcal{R}) / \inf \mathcal{R}$ and $(\inf \mathcal{RB} - \inf \mathcal{R}) / \inf \mathcal{R}$, respectively. We remark that problem $\mathcal{R}^{\text{COC}}$ can be solved exactly by using a scenario tree that consists of the vertices of Ξ (see [Bertsimas et al., 2010, §3.3] for details). Therefore, the optimality gap for $\mathcal{SB}_N^{\text{COC}}$ can be computed. To the best of our knowledge, there exists no algorithm to obtain the exact solution of problem $\mathcal{R}^{\text{RY}}$ and thus the optimality gap for $\mathcal{SB}_N^{\text{RY}}$ is not accessible. The empirical violation probability of a given $\boldsymbol{w} \in \mathbb{R}^{n_w}$ is defined as

$$\hat{V}(\boldsymbol{w}) := \frac{1}{10,000} \sum_{s=1}^{10,000} \mathbb{1}\left(f(\boldsymbol{w}, \hat{\boldsymbol{\xi}}^{(s)}) > 10^{-4} \right),$$

where $\{\hat{\boldsymbol{\xi}}^{(s)}\}_{s\in\mathbb{N}}$ is a sequence of independent samples distributed according to $\mathbb{P}$. These samples are independent of the $\{\boldsymbol{\xi}^{(s)}\}_{s\in\mathbb{N}}$ which are used in problem $\mathcal{SB}_N$.

In our computational experiments we focus on polynomial decision rules, see Example 1, and use a fixed confidence level of $\beta = 0.1\%$ for the constraint sampling. The following procedure underlies all experiments:

- Select the degree d of the polynomial decision rules as well as the target violation probability ϵ and compute the corresponding sample size $N = N(\epsilon, \beta)$. Then, solve 100 instances of problem $\mathcal{SB}_N$, each based on a different set of N independent samples from $\mathbb{P}$. For each instance, compute the optimality gap and the empirical violation probability of the optimal solution. Moreover, record the solver run time[1].

- For each of the parameters recorded, compute statistics over the 100 problem instances.

[1] All computational experiments were run on a 2.66GHz Intel Core i7-920 processor machine with 12GB RAM and all optimization problems were solved with CPLEX 12.0.

Table 1: Input data for the two test problems

Parameter	$\mathcal{R}^{\text{COC}}$	$\mathcal{R}^{\text{RY}}$
T	5	7
(p_t, d_t, h_t)	$(11, 1, 10)$	$(11, 1, 10)$
$(s_{\text{inv},1}, s_{\text{co},1})$	$(0, 0)$	$(0, 0)$
$(\underline{x}_{\text{o},t}, \overline{x}_{\text{o},t})$	$(0, \infty)$	$(0, \infty)$
$(\underline{s}_{\text{co},1}, \ldots, \underline{s}_{\text{co},T})$	$(0, 47, 134, 188, 429)$	$(0, 32, 53, 223, 437, 437, 547)$
$(\overline{s}_{\text{co},1}, \ldots, \overline{s}_{\text{co},T})$	$(\infty, 94, 248, 370, 586)$	$(\infty, 98, 242, 461, 582, 838, 1068)$
$\overline{\xi}_{\text{d},t}$	$100\left(1 + \frac{1}{2}\sin\left(\pi\frac{t}{6}\right)\right)$	$100\left(1 + \frac{1}{2}\sin\left(\pi\frac{t}{6}\right)\right)$
ρ_{d}	30%	30%
$\overline{\xi}_{\text{y},t}$	n/a	0.90
ρ_{y}	n/a	10%

Figure 4 visualizes the empirical distribution of the optimality gap of problem $\mathcal{SB}_N^{\text{COC}}$ for different values of ϵ and d. The solid horizontal lines represent the optimality gap of problem $\mathcal{RB}^{\text{COC}}$. Recall that the structure and the relatively small size of problem $\mathcal{R}^{\text{COC}}$ enable us to compute $\inf \mathcal{R}^{\text{COC}}$ exactly. We can also compute $\inf \mathcal{RB}^{\text{COC}}$ exactly by solving a variant of problem $\mathcal{SB}_N^{\text{COC}}$ in which the sample set coincides with the vertices of the hypercube Ξ. Thus all optimality gaps are numerically accessible. By Theorem 3, the optimality gap of problem $\mathcal{SB}_N^{\text{COC}}$ converges to the optimality gap of problem $\mathcal{RB}^{\text{COC}}$ as the number of samples is driven to infinity (or equivalently, as ϵ goes to 0). Our numerical results are in agreement with this convergence result, see Figure 4. We also observe that the variance of the optimality gap for problem $\mathcal{SB}_N^{\text{COC}}$ decreases substantially as ϵ decreases and that the optimality gap of $\mathcal{RB}^{\text{COC}}$ decreases rapidly as d increases. We gain about 10% in optimality when passing from linear to quadratic and about 2% when passing from quadratic to cubic policies. Note that cubic policies are indeed optimal for $\mathcal{R}^{\text{COC}}$. This is not surprising since at each decision stage $t = 1, \ldots, 5$, the number of degrees of freedom offered by cubic policies exceeds the number of vertices of the uncertainty set Ξ^t.

Figure 5, shows the empirical distribution of the empirical violation probability in dependence of ϵ and d. As expected, it converges to 0 as ϵ decreases. We also note that the empirical violation probability is always substantially smaller than ϵ (as the confidence level β is sufficiently small, the retailer's target violation probability ϵ is never exceeded).

Our computational experiments illustrate the trade-off between optimality, feasibility and computational complexity. For less flexible policies corresponding to small values of d, ϵ can be made very small, thereby meeting the requirements of a risk averse retailer. For a retailer tolerating a higher violation probability, the costs can be significantly reduced by increasing d. For example, a retailer tolerating a violation probability of $\epsilon = 0.1\%$ can reduce the 99.9% worst-case costs from 2225.5 to 1941.4 (-12.7%) by passing from linear to cubic policies.

Figure 6 displays the average solver time for problem $\mathcal{SB}_N^{\text{COC}}$ as a function of d and ϵ. As expected, the solver times for any fixed d are polynomial in the sample size $N(\epsilon, \beta)$ (indeed, Figure 6 reveals that $N(\epsilon, \beta)$ is linear in $1/\epsilon$, while the solver times are polynomial in $1/\epsilon$).

Next, we discuss problem $\mathcal{R}^{\text{RY}}$. Since this problem has seven stages and random recourse, the optimal

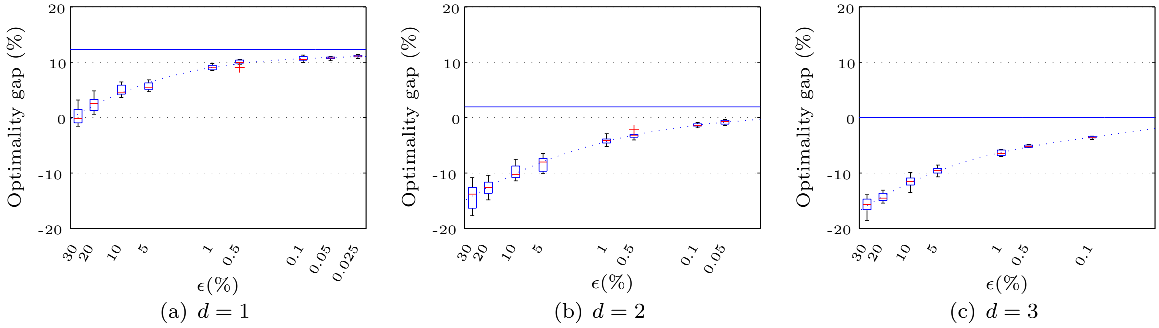


Figure 4: Optimality gaps for $\mathcal{RB}^{\text{COC}}$ (solid lines) and $\mathcal{SB}_N^{\text{COC}}$ (boxes and whiskers) for policies of degree 1, 2 and 3 in dependence of ϵ. The dotted curves represent a cubic fit of the average gaps as a function of $\log \epsilon$.

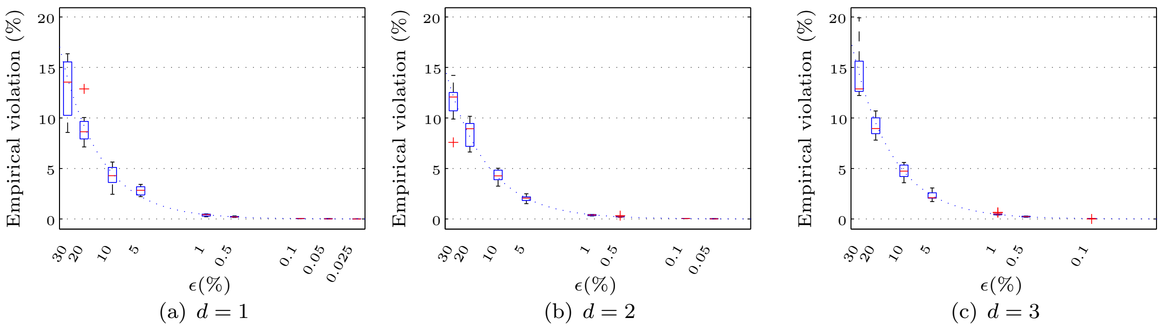


Figure 5: Empirical violation probabilities for problem $\mathcal{SB}_N^{\text{COC}}$ (boxes and whiskers) for policies of degree 1, 2 and 3 in dependence of ϵ. The dotted curves represent a linear fit of the logarithm of the average empirical violation probability as a function of $\log \epsilon$.

value of $\mathcal{SB}_N^{\text{RY}}$ does not saturate for manageable sample sizes ($N < 180,000$). However, for any fixed violation probability ϵ, there is a substantial gain in optimality when passing from linear to quadratic decision rules. For example, the 99% worst-case cost faced by a retailer tolerating a violation probability of $\epsilon = 1\%$ amounts to 3988.8 on average when using linear decision rules and to 3173.7 when using quadratic decision rules (i.e., a reduction of about 20.4%). Figure 7 illustrates the convergence of the empirical violation probability as ϵ tends to 0. As before, the empirical violation probability is always substantially smaller than the target violation ϵ.

The observations above testify to the attractiveness of our approximation scheme. Firstly, at fixed target violation probability, the gain from increasing the complexity of the decision rules can be significant. Secondly, the approximation can be tailored to the risk preferences of the retailer to trade off feasibility against optimality. Finally, our framework mitigates the over-conservatism of mainstream robust optimization models, which often result in very high costs that cater for scenarios that are unlikely to materialize.

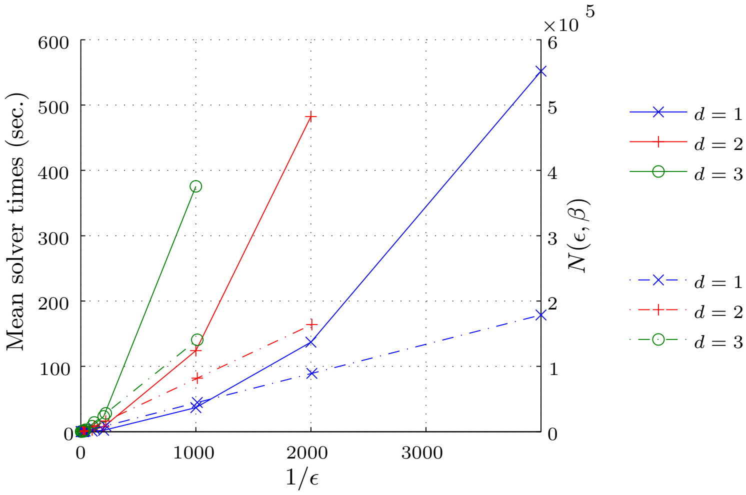
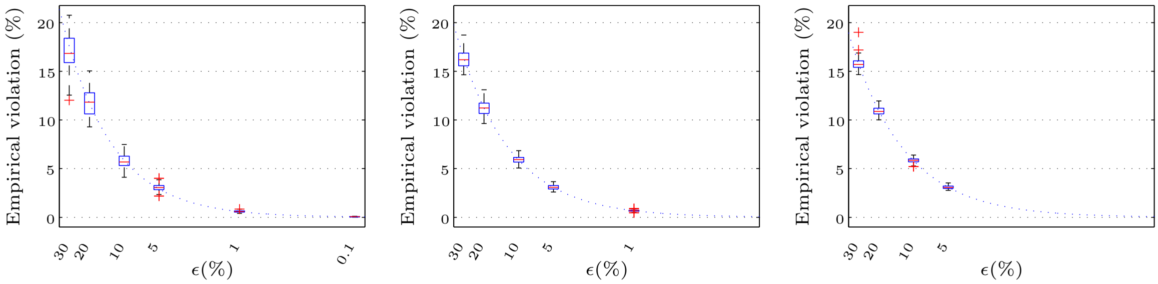


Figure 6: Mean solver times (solid lines) and numbers of samples (dashed lines) as a function of $1/\epsilon$ for problem $\mathcal{SB}_N^{\mathrm{COC}}$.



(a) $d = 1$ (b) $d = 2$ (c) $d = 3$

Figure 7: Empirical violation probabilities for problem $\mathcal{SB}_N^{\mathrm{RY}}$ for policies of degree 1, 2 and 3 in dependence of ϵ. The dotted curves represent a linear fit of the logarithm of the average empirical violation probability as a function of $\log \epsilon$.

Appendix

A Proof of Lemma 4

The proof of Lemma 4 involves the following definition.

Definition 4 (Linear dependence/independence on a set) *Let $S \subseteq \mathbb{R}^k$. If $\mathbf{c}^\top \boldsymbol{b}_t^d(\boldsymbol{\xi}^t) = 0$ for all $\boldsymbol{\xi} \in S$ implies $\mathbf{c} = \mathbf{0}$, we say that the component functions of the basis vector $\boldsymbol{b}_t^d$ are linearly independent on S. Conversely, if there exists $\mathbf{c} \neq \mathbf{0}$ such that $\mathbf{c}^\top \boldsymbol{b}_t^d(\boldsymbol{\xi}^t) = 0$ for all $\boldsymbol{\xi} \in S$, the basis functions are said to be linearly dependent on S.*

Proof of Lemma 4 We can assume without loss of generality that for each $t \in \mathbb{T}$ the components of $\boldsymbol{b}_t^d$ are linearly independent on Ξ. This assumption is non-restrictive since linearly dependent basis functions may always be removed without affecting the optimal value of problem $\mathcal{RB}$. Since $\{\boldsymbol{\xi}^{(l)}\}_{l\in\mathbb{N}}$ is dense in

Ξ, there exists $N_0 \in \mathbb{N}$ such that the components of $\boldsymbol{b}_t^d$ are linearly independent on Ξ_N for all $t \in \mathbb{T}$ and $N \geq N_0$, see Lemma 6 below. Choose $N \geq N_0$. $\mathcal{SB}_N$ is feasible since it is a relaxation of $\mathcal{RB}$, which is feasible by our choice of d. Furthermore, $\mathcal{SB}_N$ has a closed feasible set and a continuous objective function, see assumption (A2). We next show that $\mathcal{SB}_N$ is bounded.

Suppose that $\mathcal{SB}_{N_0}$ is unbounded. Then, there is a sequence $\{\boldsymbol{w}^{(i)}\}_{i\in\mathbb{N}}$ of points feasible in $\mathcal{SB}_{N_0}$ with $\lim_{i\to\infty} \theta^{(i)} = -\infty$, where $\boldsymbol{w}^{(i)} = (\theta^{(i)}, \mathrm{vec}(\boldsymbol{X}_1^{(i)}), \ldots, \mathrm{vec}(\boldsymbol{X}_T^{(i)}))$. This is only possible if $\lim_{i\to\infty} \|\boldsymbol{X}_t^{(i)}\|_2 = \infty$ for some $t \in \mathbb{T}$, where $\|\cdot\|_2$ denotes the Frobenius norm, see Section 3.1. However, by assumption (A3), we also have $|\boldsymbol{X}_t^{(i)}\, \boldsymbol{b}_t^d(\boldsymbol{\xi})| \leq R$ for all $\boldsymbol{\xi} \in \Xi_{N_0}$, $i \in \mathbb{N}$. The bounded sequence $\{\boldsymbol{X}_t^{(i)}/\|\boldsymbol{X}_t^{(i)}\|_2\}_{i\in\mathbb{N}}$ is confined to the compact unit sphere and therefore has an accumulation point $\boldsymbol{X}_t^*$ with $\|\boldsymbol{X}_t^*\|_2 = 1$. Let $\{\boldsymbol{X}_t^{(i_j)}/\|\boldsymbol{X}_t^{(i_j)}\|_2\}_{j\in\mathbb{N}}$ be a subsequence converging to $\boldsymbol{X}_t^*$. Then, we have

$$
\begin{aligned}
\lim_{j\to\infty} |\boldsymbol{X}_t^{(i_j)}\, \boldsymbol{b}_t^d(\boldsymbol{\xi})| \leq R\; \forall \boldsymbol{\xi} \in \Xi_{N_0} \quad &\Rightarrow \quad \lim_{j\to\infty} \|\boldsymbol{X}_t^{(i_j)}\|_2 \left| \frac{\boldsymbol{X}_t^{(i_j)}}{\|\boldsymbol{X}_t^{(i_j)}\|_2}\, \boldsymbol{b}_t^d(\boldsymbol{\xi}) \right| \leq R\; \forall \boldsymbol{\xi} \in \Xi_{N_0} \\
&\Rightarrow \quad \left| \boldsymbol{X}_t^*\, \boldsymbol{b}_t^d(\boldsymbol{\xi}) \right| = 0\; \forall \boldsymbol{\xi} \in \Xi_{N_0} \\
&\Rightarrow \quad \boldsymbol{X}_t^* = \boldsymbol{0},
\end{aligned}
$$

where the last implication follows from the linear independence of the basis functions on Ξ_{N_0}. This contradicts our earlier result that $\|\boldsymbol{X}_t^*\|_2 = 1$. We thus conclude that $\mathcal{SB}_{N_0}$ is bounded. Since $\mathcal{SB}_N$ is a restriction of $\mathcal{SB}_{N_0}$, it is also bounded. The above reasoning implies that $\mathcal{SB}_N$ is solvable and has a finite optimal value.

Moreover, the optimal solution of $\mathcal{SB}_N$ is contained in the set

$$
W^* := \left\{ \boldsymbol{w} = (\theta, \mathrm{vec}(\boldsymbol{X}_1), \ldots, \mathrm{vec}(\boldsymbol{X}_T)) \in \mathbb{R}^{n_w} : \inf \mathcal{SB}_{N_0} \leq \theta \leq \inf \mathcal{RB},\; |\boldsymbol{X}_t\, \boldsymbol{b}_t^d(\boldsymbol{\xi})| \leq R\; \forall \boldsymbol{\xi} \in \Xi_{N_0},\; t \in \mathbb{T} \right\},
$$

which is non-empty and compact. As the choice of $N \geq N_0$ was arbitrary and the definition of W^* is independent of N, the claim follows. □

Lemma 6 *If the components of* $\boldsymbol{b}_t^d$ *are linearly independent on* Ξ *and* $\{\boldsymbol{\xi}^{(l)}\}_{l\in\mathbb{N}}$ *is dense in* Ξ*, then there exists* $N_t \in \mathbb{N}$ *such that the components of* $\boldsymbol{b}_t^d$ *are linearly independent on* Ξ_N *for all* $N \geq N_t$.

Proof Suppose that the components of $\boldsymbol{b}_t^d$ are linearly dependent on Ξ_N for all $N \in \mathbb{N}$. Thus, there exist $\boldsymbol{c}_N \in \mathbb{R}^{s_d(k^t)}$ such that $|\boldsymbol{c}_N| = 1$ and $\boldsymbol{c}_N^\top\, \boldsymbol{b}_t^d(\boldsymbol{\xi}^t) = 0$ for all $\boldsymbol{\xi} \in \{\boldsymbol{\xi}^{(l)}\}_{l\in\mathbb{N}}$. The sequence $\{\boldsymbol{c}_N\}_{N\in\mathbb{N}}$ is confined to the compact unit sphere and therefore has an accumulation point $\boldsymbol{c}^*$ with $|\boldsymbol{c}^*| = 1$. Let $\{\boldsymbol{c}_{N_j}\}_{j\in\mathbb{N}}$ be a subsequence converging to $\boldsymbol{c}^*$. Then,

$$
\begin{aligned}
\lim_{j\to\infty} \boldsymbol{c}_{N_j}^\top\, \boldsymbol{b}_t^d(\boldsymbol{\xi}^t) = 0 \;\; \forall \boldsymbol{\xi} \in \{\boldsymbol{\xi}^{(l)}\}_{l\in\mathbb{N}} \quad &\Rightarrow \quad {\boldsymbol{c}^*}^\top\, \boldsymbol{b}_t^d(\boldsymbol{\xi}^t) = 0 \;\; \forall \boldsymbol{\xi} \in \{\boldsymbol{\xi}^{(l)}\}_{l\in\mathbb{N}} \\
&\Rightarrow \quad {\boldsymbol{c}^*}^\top\, \boldsymbol{b}_t^d(\boldsymbol{\xi}^t) = 0 \;\; \forall \boldsymbol{\xi} \in \Xi \\
&\Rightarrow \quad \text{the components of } \boldsymbol{b}_t^d \text{ are linearly dependent on } \Xi.
\end{aligned}
$$

The second implication follows from the continuity of the basis functions and since $\{\boldsymbol{\xi}^{(l)}\}_{l\in\mathbb{N}}$ is dense in Ξ, while the third implication holds since $\boldsymbol{c}^* \neq \mathbf{0}$. The last implication contradicts our assumption, and thus the claim follows. □

B Almost sure density of sample sequences

Lemma 7 *The sequence* $\{\boldsymbol{\xi}^{(l)}\}_{l\in\mathbb{N}}$ *is dense in* Ξ *w.p.1.*

Proof Let $\mathbb{P}^\infty := \bigtimes_{l=1}^{\infty} \mathbb{P}$ denote the probability distribution of the stochastic process $\{\boldsymbol{\xi}^{(l)}\}_{l\in\mathbb{N}}$. Then, for $\boldsymbol{z} \in \Xi$ and $\epsilon > 0$, we have

$$\mathbb{P}^\infty(B_\epsilon(\boldsymbol{z}) \cap \{\boldsymbol{\xi}^{(l)}\}_{l\in\mathbb{N}} = \emptyset) \;=\; \prod_{l=1}^{\infty} \mathbb{P}(\boldsymbol{\xi}^{(l)} \notin B_\epsilon(\boldsymbol{z})) \;=\; 0. \tag{B.1}$$

The last equality holds since $\boldsymbol{z}$ is an element of the support of $\mathbb{P}$ and thus $\mathbb{P}(\boldsymbol{\xi}^{(l)} \in B_\epsilon(\boldsymbol{z})) > 0$ independently of $l \in \mathbb{N}$. Since Ξ is convex and fully dimensional by assumption (A5), the sequence $\{\boldsymbol{\xi}^{(l)}\}_{l\in\mathbb{N}}$ is dense in Ξ if and only if for every $\boldsymbol{z} \in \Xi \cap \mathbb{Q}^k$ and $\epsilon \in \mathbb{Q}_+$, the set $B_\epsilon(\boldsymbol{z})$ contains at least one sample. Therefore,

$$\begin{aligned}
\mathbb{P}^\infty(\{\boldsymbol{\xi}^{(l)}\}_{l\in\mathbb{N}} \text{ is dense in } \Xi) &= 1 - \mathbb{P}^\infty(\exists \boldsymbol{z} \in \Xi \cap \mathbb{Q}^k,\ \epsilon \in \mathbb{Q}_+ \text{ with } B_\epsilon(\boldsymbol{z}) \cap \{\boldsymbol{\xi}^{(l)}\}_{l\in\mathbb{N}} = \emptyset) \\
&\geq 1 - \sum_{\boldsymbol{z}\in\Xi\cap\mathbb{Q}^k} \sum_{\epsilon\in\mathbb{Q}_+} \mathbb{P}^\infty(B_\epsilon(\boldsymbol{z}) \cap \{\boldsymbol{\xi}^{(l)}\}_{l\in\mathbb{N}} = \emptyset) \\
&= 1,
\end{aligned}$$

where the second line follows from the Bonferroni inequality and the last line follows from (B.1). □

Acknowledgments

The authors would like to thank the EPSRC for financial support under grant EP/H0204554/1.

References

A. Ben-Tal, A. Goryashko, E. Guslitzer, and A. Nemirovski. Adjustable robust solutions of uncertain linear programs. *Mathematical Programming*, 99(2):351–376, 2004.

A. Ben-Tal, B. Golany, A. Nemirovski, and J.-P. Vial. Retailer-supplier flexible commitments contracts: A robust optimization approach. *Manufacturing & Service Operations Management*, 7(3):248–271, 2005.

A. Ben-Tal, L. El Ghaoui, and A. Nemirovski. *Robust Optimization.* Princeton Series in Applied Mathematics. Princeton University Press, 2009.

D.P. Bertsekas. *Dynamic Programming and Optimal Control.* Athena Scientific, Belmont, MA, USA, 1995.

D. Bertsimas and C. Caramanis. Adaptability via sampling. In *Proceedings of the 46th IEEE Conference on Decision and Control*, 2007.

D. Bertsimas, D. Iancu, and P. Parrilo. A hierarchy of near-optimal policies for multi-stage adaptive optimization. Conditionally accepted in *IEEE Transactions on Automatic Control*, 2010.

J.R. Birge and F. Louveaux. *Introduction to Stochastic Programming*. Springer, 2000.

V.D. Blondel and J.N. Tsitsiklis. A survey of computational complexity results in systems and control. *Automatica*, 36:1249–1274, 2000.

G. Calafiore and M.C. Campi. Uncertain convex programs: randomized solutions and confidence levels. *Mathematical Programming*, 102(1):25–46, 2005.

G. Calafiore and A. Nilim. Multi-stage resource allocation under uncertainty. In *Proceedings of the 43rd IEEE Conference on Decision and Control*, volume 1, pages 417–423, 2004.

G.C. Calafiore. On the expected probability of constraint violation in sampled convex programs. *Journal of Optimization Theory and Applications*, 143(2):405–412, 2009.

G.C. Calafiore and M.C. Campi. The scenario approach to robust control design. *IEEE Transactions on Automatic Control*, 51(5):742–753, 2006.

M.C. Campi and S. Garatti. The exact feasibility of randomized solutions of robust convex programs. *SIAM Journal on Optimization*, 19(3):1211–1230, 2008.

G. Cybenko. Approximation by superpositions of a sigmoidal function. *Mathematics of Control, Signals, and Systems (MCSS)*, 2(4):303–314, 1989.

G.E. Dullerud and F. Paganini. *A Course in Robust Control Theory*. Springer, 2005.

A. Georghiou, W. Wiesemann, and D. Kuhn. Generalized decision rule approximations for stochastic programming via liftings. Available on Optimization Online, submitted for publication, 2010.

F. Girosi and T. Poggio. Networks and the best approximation property. *Biological Cybernetics*, 63(3): 169–176, 1990.

J. Goh and M. Sim. Distributionally robust optimization and its tractable approximations. *Operations Research*, 58:902–917, 2010.

M. Grötschel, L. Lovász, and A. Schrijver. The ellipsoid method and its consequences in combinatorial optimization. *Combinatorica*, 1(2):169–197, 1981.

R. Lobel and G. Perakis. Dynamic pricing through sampling based optimization. Technical report, Massachusetts Institute of Technology, 2010. Submitted for publication.

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, August 2009.

A. Prékopa. *Stochastic Programming*. Kluwer Academic Publishers, 1995.

R.T. Rockafellar and R.J.-B. Wets. *Variational Analysis*. Springer-Verlag, 1997.

A. Shapiro and A. Nemirovski. *Continuous Optimization: Current Trends and Applications*, chapter On complexity of stochastic programming problems, pages 111–114. Springer, 2005.

J. Skaf and S. Boyd. Nonlinear Q-design for convex stochastic control. *IEEE Transactions on Automatic Control*, 54(10):2426–2430, 2009.

S. Yu-Ting. A “natural” enumeration of non-negative rational numbers – an informal discussion. *The American Mathematical Monthly*, 87:25–29, 1980.