



Distributionally robust optimization with polynomial densities: theory, models and algorithms

Etienne de Klerk¹ · Daniel Kuhn² · Krzysztof Postek³

Received: 9 May 2018 / Accepted: 30 August 2019 / Published online: 6 September 2019

© Springer-Verlag GmbH Germany, part of Springer Nature and Mathematical Optimization Society 2019

Abstract

In distributionally robust optimization the probability distribution of the uncertain problem parameters is itself uncertain, and a fictitious adversary, e.g., nature, chooses the worst distribution from within a known *ambiguity set*. A common shortcoming of most existing distributionally robust optimization models is that their ambiguity sets contain pathological discrete distributions that give nature too much freedom to inflict damage. We thus introduce a new class of ambiguity sets that contain only distributions with sum-of-squares (SOS) polynomial density functions of known degrees. We show that these ambiguity sets are highly expressive as they conveniently accommodate distributional information about higher-order moments, conditional probabilities, conditional moments or marginal distributions. Exploiting the theoretical properties of a measure-based hierarchy for polynomial optimization due to Lasserre (SIAM J Optim 21(3):864–885, 2011), we prove that certain worst-case expectation constraints are polynomial-time solvable under these new ambiguity sets. We also show how SOS densities can be used to approximately solve the *general problem of moments*. We showcase the applicability of the proposed approach in the context of a stylized portfolio optimization problem and a risk aggregation problem of an insurance company.

Keywords Distributionally robust optimization · Semidefinite programming · Sum-of-squares polynomials · Generalized eigenvalue problem

✉ Daniel Kuhn
daniel.kuhn@epfl.ch

Etienne de Klerk
E.deKlerk@uvt.nl

Krzysztof Postek
postek@ese.eur.nl

¹ Tilburg University, Tilburg, The Netherlands

² EPFL, Lausanne, Switzerland

³ Erasmus University Rotterdam, Rotterdam, The Netherlands

1 Introduction

Since George Dantzig's 1955 paper on linear programming under uncertainty [11], the field of stochastic programming has developed numerous methods for solving optimization problems that depend on uncertain parameters governed by a *known* probability distribution, see, e.g., [5, 41, 47]. Stochastic programming usually aims to minimize a probability functional such as the expected value, a percentile or the conditional value-at-risk of a given cost function. In practice, however, the distribution needed to evaluate this probability functional is at best indirectly observable through independent training samples. Thus, the stochastic programming approach is primarily useful when there is abundant training data. If data is scarce or absent, on the other hand, it may be more adequate to use a robust optimization approach, which models the uncertainty through the set of all possible (or sufficiently likely) uncertainty realizations and minimizes the worst-case costs. Robust optimization is the appropriate modeling paradigm for safety-critical applications with little tolerance for failure and has been popularized in the late 1990's, when it was discovered that robust optimization models often display better tractability properties than stochastic programming models [1]. Distributionally robust optimization is a hybrid approach that attempts to salvage the tractability of robust optimization while maintaining the benefits of (limited) distributional information. In this context, uncertainty is modeled through an *ambiguity set*, that is, a family of typically infinitely many different distributions that are consistent with the available training data or any prior distributional information, and the objective is to minimize the worst-case expected costs across all distributions in the ambiguity set. A distributionally robust newsvendor model that admits an analytical solution has been investigated as early as in 1958 [44], and the theoretical properties of distributionally robust linear programs were first studied in 1966 [52]. Interest in distributionally robust optimization has also been fuelled by important applications in finance [38, 39]. However, only recently it was recognized that many distributionally robust optimization problems of practical relevance can actually be solved in polynomial time. Tractability results are available both for *moment ambiguity sets*, which contain all distributions that satisfy a finite number of moment conditions [12, 17, 51], as well as for *metric-based ambiguity sets*, which contain all distributions within a prescribed distance from a nominal distribution with respect to some probability metric [7, 36]. In all these cases, the extremal distributions that determine the worst-case expectation are discrete, and the number of their discretization points is often surprisingly small, e.g., proportional to the number of moment constraints. As these unnatural discrete distributions are almost always inconsistent with the available training samples, distributionally robust optimization models with moment and metric-based ambiguity sets are often perceived as overly pessimistic.

In an attempt to mitigate the over-conservatism of traditional distributionally robust optimization, several authors have studied moment ambiguity sets that require their member distributions to satisfy additional structural properties such as symmetry, unimodality, monotonicity or smoothness etc. By leveraging ideas from Choquet theory

and polynomial optimization, it has been shown that the resulting distributionally robust optimization problems admit hierarchies of increasingly accurate semidefinite programming (SDP) bounds [40]. An exact SDP reformulation for the worst-case probability of a polytope with respect to all unimodal distributions with known first and second moments is derived in [50], while second-order conic reformulations of distributionally robust individual chance constraints with moment and unimodality information are reported in [31]. For a survey of recent results on distributionally robust uncertainty quantification and chance constrained programming problems with moment and structural information we refer to [22]. Even though unimodality or monotonicity conditions eliminate all discrete distributions from a moment ambiguity set, the extremal distributions that critically determine all worst-case expectations remain pathological. For example, all extremal *unimodal* distributions are supported on line segments emanating from a single point in space (the mode) and thus fail to be absolutely continuous with respect to the Lebesgue measure. Thus, the existing distributionally robust optimization models with structural information remain overly conservative. This observation motivates us to investigate a new class of ambiguity sets that contain only distributions with non-degenerate polynomial density functions.

This paper aims to study *worst-case expectation constraints* of the form

$$\inf_{\mathbb{P} \in \mathcal{P}} \mathbb{E}_{\mathbb{P}} f(x, z) \geq 0, \quad (1)$$

where $x \in \mathbb{R}^n$ is a decision vector, $z \in \mathbb{R}^m$ is an uncertain parameter governed by an ambiguous probability distribution $\mathbb{P} \in \mathcal{P}$, and $f(x, z)$ is an uncertainty-affected constraint function that can be interpreted as a cost. In words, the constraint (1) requires that the expected value of the function $f(x, z)$ given the decision x be nonnegative for every distribution in the ambiguity set \mathcal{P} . Throughout the paper we will assume that $f(x, z)$ depends polynomially on z and that each distribution $\mathbb{P} \in \mathcal{P}$ admits a sum-of-squares (hence nonnegative) polynomial density function $h(z)$ with respect to some prescribed reference measure μ on \mathbb{R}^m (e.g., the Lebesgue measure). Imposing an upper bound on the polynomial degree of $h(z)$ thus yields a finite-dimensional parameterization of the ambiguity set \mathcal{P} . Moreover, many popular distributional properties can be expressed through linear constraints on the coefficients of $h(z)$ and are thus conveniently accounted for in the definition of \mathcal{P} . Examples include moment bounds, probability bounds for certain subsets of \mathbb{R}^m , bounds on conditional tail probabilities and marginal distribution conditions. Note that by fixing the marginal distributions of all components of z , the worst-case expectation problem on the left-hand side of (1) reduces to a Fréchet problem that seeks the worst-case copula of the uncertain parameters.

By leveraging a measure-based hierarchy for polynomial optimization due to Lasserre [28], we will demonstrate that the subordinate worst-case probability problem in (1) admits an exact SDP reformulation. Under mild additional conditions on $f(x, z)$, we will further prove that the feasible set of the constraint (1) admits a polynomial-time separation oracle. Moreover, we will analyze the convergence of the worst-case expectation in (1) as the polynomial degree of $h(z)$ tends to infinity, and we will illustrate the practical use of the proposed approach through numerical examples.

More succinctly, the main contributions of this paper can be summarized as follows:

- (i) *Modeling power*: We introduce a new class of ambiguity sets containing distributions that admit sum-of-squares polynomial density functions of degree at most $2r$, $r \in \mathbb{N}$, with respect to a given reference measure. Ambiguity sets of this type are highly expressive as they conveniently accommodate distributional information about higher-order moments, conditional probabilities or conditional moments. They also allow the modeler to prescribe (not necessarily discrete) marginal distributions that must be matched exactly by all distributions in the ambiguity set.
- (ii) *Computational tractability*: A key advantage of working with sum-of-squares polynomial density functions is computational tractability in the sense that one may often formulate the emerging problems as SDPs. Specifically, we identify general conditions under which the worst-case expectations over the new ambiguity sets can be reformulated *exactly* as tractable SDPs with $\mathcal{O}(\binom{n+r}{r})$ variables. Finally, we delineate conditions under which the feasible sets of the worst-case expectation constraints admit a polynomial-time separation oracle and thus lend themselves to efficient optimization via the ellipsoid method.
- (iii) *Convergence analysis*: We demonstrate that, as r tends to infinity, the worst-case expectations over the new ambiguity sets converge monotonically to classical worst-case expectations over larger ambiguity sets that relax the polynomial density requirement. At the same time, we show that the extremal density functions converge to pathological discrete worst-case distributions characteristic for classical moment ambiguity sets without restrictions on the density functions. This convergence analysis showcases that our approach naturally embeds classical stochastic programming (for $r = 0$) and the traditional approach to (distributionally) robust optimization (for $r \rightarrow \infty$) into a unifying framework.
- (iv) *Numerical results*: We showcase the practical applicability of the proposed approach in the context of a stylized portfolio optimization problem and a simple Fréchet problem inspired by [49] that models the risk aggregation problem of an insurance company.

The intimate relation between polynomial optimization and the problem of moments has already been exploited in several papers on distributionally robust optimization. For example, ideas from polynomial optimization give rise to SDP bounds on the probability of a semi-algebraic set [3] or the expected value of a piecewise polynomial [53] across all probability distributions satisfying a given set of moment constraints. These SDP bounds are tight in the univariate case or if only marginal moments are specified. Otherwise, one may obtain hierarchies of asymptotically tight SDP bounds. As an application, these techniques can be used to derive bounds on the prices of options with piecewise polynomial payoff functions, based solely on the knowledge of a few moments of the underlying asset prices [2]. Moreover, asymptotically tight SDP bounds that account for both moment and structural information are proposed in [40]. All these approaches differ from our work in that the ambiguity sets have discrete or otherwise degenerate extremal distributions.

Distributionally robust polynomial optimization problems over non-degenerate polynomial density functions that are close to a nominal density estimate (obtained,

e.g., via a Legendre series density estimator) in terms of the L_2 -distance are considered in [35]. In this work the non-negativity of the candidate density functions is not enforced explicitly, which considerably simplifies the problem and may be justified if the distance to the nominal density is sufficiently small. It is shown that the emerging distributionally robust optimization problems are equivalent to deterministic polynomial optimization problems that are not significantly harder than the underlying nominal problem and can be addressed by solving a sequence of tractable SDP relaxations.

Distributionally robust chance constraints with ambiguity sets containing all possible mixtures of a given parametric distribution family are studied in [30]. The mixtures are encoded through a probability density function on a compact parameter space. The authors propose an asymptotically tight SDP hierarchy of inner approximations for the feasible set of the distributionally robust chance constraint. In contrast, we explicitly represent all probability distributions in the ambiguity set through polynomial density functions that can capture a wide range of distributional features.

The paper is structured as follows. Section 2 reviews the measure-based approach to polynomial optimization due to Lasserre [28], which is central to this paper. Section 3 develops SDP hierarchies for worst-case expectations over ambiguity sets that contain probability measures with polynomial densities and investigates the convergence as the polynomial degree tends to infinity. Section 4 extends this analysis to ambiguity sets imposing moment conditions. Section 5 highlights the modeling power of the proposed approach, while Sect. 6 reports on numerical results for a portfolio design and a risk aggregation problem of an insurance company. Conclusions are drawn in Sect. 7.

2 Lasserre's measure-based hierarchy for polynomial optimization

In what follows, we denote by $x^\alpha := \prod_{i=1}^n x_i^{\alpha_i}$ the monomial of the variables $x = (x_1, \dots, x_n)$ with respective exponents $\alpha = (\alpha_1, \dots, \alpha_n) \in \mathbb{N}_0^n$, and we define $N(n, r) := \{\alpha \in \mathbb{N}_0^n : \sum_{i=1}^n \alpha_i \leq r\}$ as the set of all exponents that give rise to monomials with degrees of at most r . We let $\Sigma[x]$ denote the set of all sum-of-squares (SOS) polynomials in the variables x , and we define $\Sigma[x]_r$ as the subset of all SOS polynomials with degrees of at most $2r$.

Now consider the polynomial global optimization problem

$$p_{\min, \mathbf{K}} := \min_{x \in \mathbf{K}} p(x) = \min_{x \in \mathbf{K}} \sum_{\alpha \in N(n, d)} p_\alpha x^\alpha, \quad (2)$$

where $p(x) = \sum_{\alpha \in N(n, d)} p_\alpha x^\alpha$ is an n -variate polynomial of degree d , and $\mathbf{K} \subset \mathbb{R}^n$ a closed set with nonempty interior. In the subsequent discussion we always assume that a global minimizer exists. We also assume that the moments of a finite Borel measure μ supported on \mathbf{K} are known in the sense that they are either available in closed form or efficiently computable. Recall that a finite Borel measure μ on \mathbb{R}^n as a nonnegative set function defined on the Borel σ -algebra of \mathbb{R}^n , that is, the σ -algebra generated by all open sets in \mathbb{R}^n . By definition, μ must satisfy $\mu(\emptyset) = 0$

and $\mu(\cup_{i=1}^{\infty} S_i) = \sum_{i=1}^{\infty} \mu(S_i)$ for any countable collection of disjoint, measurable sets $S_i \subset \mathbb{R}^n$, $i \in \mathbb{N}$, and $\mu(\mathbb{R}^n) < \infty$. The support of μ , denoted by $\text{supp}(\mu)$, is defined as the smallest closed set \mathbf{K} with $\mu(\mathbb{R}^n \setminus \mathbf{K}) = 0$.

In the following we denote the (known) moments of μ by

$$m_{\alpha}(\mathbf{K}) := \int_{\mathbf{K}} x^{\alpha} d\mu(x) \quad \text{for } \alpha \in \mathbb{N}_0^n. \quad (3)$$

Lasserre [28] introduced the following upper bound on $p_{\min, \mathbf{K}}$,

$$\begin{aligned} \underline{p}_{\mathbf{K}}^{(r)} &:= \min_{h \in \Sigma[x]_r} \left\{ \int_{\mathbf{K}} p(x) h(x) d\mu(x) : \int_{\mathbf{K}} h(x) d\mu(x) = 1 \right\} \\ &= \min_{h \in \Sigma[x]_r} \mathbb{E}_{x \sim (\mathbf{K}, h)}[p(x)], \end{aligned} \quad (4)$$

where r is a fixed integer, and $x \sim (\mathbf{K}, h)$ indicates that x is a random vector supported on \mathbf{K} that is governed by the probability measure $h \cdot d\mu$. It is known that $\underline{p}_{\mathbf{K}}^{(r)}$ is equal to the smallest generalized eigenvalue of the system

$$Av = \lambda Bv, \quad (5)$$

where $v \neq 0$ is the generalized eigenvector corresponding to λ , while the symmetric matrices A and B are of size $\binom{n+r}{r}$ with rows and columns indexed by $N(n, r)$, and

$$A_{\alpha, \beta} = \sum_{\delta \in N(n, d)} p_{\delta} m_{\alpha+\beta+\delta}(\mathbf{K}), \quad B_{\alpha, \beta} = m_{\alpha+\beta}(\mathbf{K}) \quad \text{for } \alpha, \beta \in N(n, r). \quad (6)$$

A review of solution techniques for the generalized eigenvalue problem may be found in [18, §8.7.2]. Lasserre [28] established conditions on μ and \mathbf{K} ensuring that $\lim_{r \rightarrow \infty} \underline{p}_{\mathbf{K}}^{(r)} = p_{\min, \mathbf{K}}$, and the rate of convergence was subsequently studied in [8–10] for special choices of μ and \mathbf{K} . The most general condition under which convergence is known to hold, as shown in [29, Theorem 2.2], is when μ is supported on a closed, basic semi-algebraic set \mathbf{K} with nonempty interior, all moments of μ are finite, and there is $M > 0$ such that

$$\int_{\mathbf{K}} x_i^{2k} d\mu(x) \leq (2k)! M \quad \forall i \in \{1, \dots, n\}, k \in \mathbb{N}. \quad (7)$$

For example, if one defines μ in terms of a finite Borel measure φ with $\text{supp}(\varphi) = \mathbf{K}$ via

$$d\mu(x) = \exp(-c|x_1| - \dots - c|x_n|) d\varphi(x) \quad (8)$$

for some fixed $c > 0$, then μ satisfies the conditions (7); see [28, §3.2].

We summarize the known convergence results in Table 1.

Table 1 Known rates of convergence for the Lasserre hierarchy

$\mathbf{K} \subset \mathbb{R}^n$	$p_{\mathbf{K}}^{(r)} - p_{\min, \mathbf{K}}$	Measure μ , $\text{supp}(\mu) = \mathbf{K}$	References
Closed, basic semi-algebraic, nonempty interior	$o(1)$	Satisfies (7) and all moments finite	[29]
Compact, nonempty interior	$o(1)$	Finite Borel measure	[28]
Compact, satisfies interior cone condition	$O\left(\frac{1}{\sqrt{r}}\right)$	Lebesgue measure	[10]
Convex body	$O\left(\frac{1}{r}\right)$	Lebesgue measure	[8]
$[-1, 1]^n$	$\Theta\left(\frac{1}{r^2}\right)$	$d\mu(x) = \prod_{i=1}^n (1-x_i^2)^{-1/2} dx_i$	[9]

2.1 Examples of known moments

The moments (3) are available in closed form, for example, if μ is the Lebesgue measure and \mathbf{K} is an ellipsoid or triangulated polytope; see, e.g., [10, 28]. For the canonical simplex, $\Delta_n = \{x \in \mathbb{R}_+^n : \sum_{i=1}^n x_i \leq 1\}$, we have

$$m_\alpha(\Delta_n) = \frac{\prod_{i=1}^n \alpha_i!}{(\sum_{i=1}^n \alpha_i + n)!}, \quad (9)$$

see, e.g., [26, Equation (2.4)] or [21, Equation (2.2)]. One may trivially verify that the moments for the hypercube $\mathbf{Q}_n = [0, 1]^n$ are given by

$$m_\alpha(\mathbf{Q}_n) = \int_{\mathbf{Q}_n} x^\alpha dx = \prod_{i=1}^n \int_0^1 x_i^{\alpha_i} dx_i = \prod_{i=1}^n \frac{1}{\alpha_i + 1}.$$

The moments for the unit Euclidean ball are given by

$$m_\alpha(B_1(0)) = \begin{cases} \frac{\pi^{(n-1)/2} 2^{(n+1)/2} \prod_{i=1}^n (\alpha_i - 1)!!}{(n + \sum_{i=1}^n \alpha_i)!!} & \text{if } \alpha_i \text{ is even for all } i, \\ 0 & \text{otherwise,} \end{cases} \quad (10)$$

where the double factorial of any integer k is defined through

$$k!! = \begin{cases} k \cdot (k-2) \cdots 3 \cdot 1 & \text{if } k > 0 \text{ is odd,} \\ k \cdot (k-2) \cdots 4 \cdot 2 & \text{if } k > 0 \text{ is even,} \\ 1 & \text{if } k = 0 \text{ or } k = -1. \end{cases}$$

When \mathbf{K} is an ellipsoid, one may obtain the moments from (10) via an affine variable transformation. Another tractable support set that will become relevant in Sect. 6.1 of this paper is the knapsack polytope, that is, the intersection of a hypercube and a half-space; the moments for this and other related polytopes are derived in [34]. Finally, in Sect. 6.3 we will work with the nonnegative orthant $\mathbf{K} = \mathbb{R}_+^n$. Since \mathbf{K} is

unbounded in this case, we need to introduce a measure of the form (8). A suitable choice that corresponds to $c = 1/2$ and $d\varphi(x) = \exp(-\frac{1}{2} \sum_{i=1}^n x_i) dx$ in (8) is

$$d\mu(x) = \exp\left(-\sum_{i=1}^n x_i\right) dx.$$

This is the exponential measure associated with the orthogonal Laguerre polynomials.

We will also use another choice of measure for $\mathbf{K} = \mathbb{R}_+^n$ in Sect. 6.3, namely the lognormal measure,

$$d\mu(x) = \prod_{i=1}^n \frac{1}{x_i v_i \sqrt{2\pi}} \exp\left(-\frac{(\ln(x_i) - \bar{z}_i)^2}{2v_i^2}\right) dx_i, \quad (11)$$

where \bar{z}_i and v_i represent prescribed location and scale parameters for all $i = 1, \dots, n$. In this case the moments of μ are given by

$$m_\alpha(\mathbf{K}) = \prod_{i=1}^n \exp(\alpha_i \bar{z}_i + (\alpha_i v_i)^2/2). \quad (12)$$

One may verify that these moments do *not* obey the bounds in (7). When using the lognormal measure we are therefore not guaranteed convergence of the Lasserre hierarchy.

We stress that, even though these examples of known moments are limited, they include typical sets that are routinely used in (distributionally) robust optimization to represent uncertainty sets or supports, most notably budget uncertainty sets and ellipsoids.

3 Distributionally robust constraints with ambiguous polynomial density functions

Consider now a worst-case feasibility expectation of the form (1), where $z \in \mathbb{R}^m$ represents a random vector with support $\mathbf{K} \subset \mathbb{R}^m$, assumed to be closed and with nonempty interior. Suppose that the constraint function $f(x, z)$ displays a polynomial dependence on z . In particular, assume that $f(x, z) = \sum_{\beta \in N(m, d)} f_\beta(x) z^\beta$ has degree d in z , where the $f_\beta : \mathbb{R}^n \rightarrow \mathbb{R}$ are functions of x only.

If the ambiguity set \mathcal{P} contains all distributions that have an SOS polynomial density of degree at most $2r$, $r > 1$, with respect to a fixed, finite Borel measure μ supported on \mathbf{K} , then the worst-case expectation constraint (1) reduces to

$$f_{\mathbf{K}}^{(r)}(x) := \inf_{h \in \Sigma[z]_r} \left\{ \int_{\mathbf{K}} f(x, z) h(z) d\mu(z) : \int_{\mathbf{K}} h(z) d\mu(z) = 1 \right\} \geq 0. \quad (13)$$

Formally speaking, we consider an ambiguity set of the form

$$\mathcal{P} = \left\{ h \cdot d\mu : h \in \Sigma[z]_r, \int_{\mathbf{K}} h(z) d\mu(z) = 1 \right\}. \quad (14)$$

We assume that the moments of the measure μ on \mathbf{K} are available, and we again use the notation

$$m_\alpha(\mathbf{K}) := \int_{\mathbf{K}} z^\alpha d\mu(z) \quad \text{for } \alpha \in \mathbb{N}_0^m.$$

Expressing $h \in \Sigma[z]_r$ as $h(z) = \sum_{\alpha \in N(m, 2r)} h_\alpha z^\alpha$, the left-hand-side of (13) may be re-written as

$$\begin{aligned} & \inf_{h_\alpha: \alpha \in N(m, 2r)} \sum_{\beta \in N(m, d)} f_\beta(x) \sum_{\alpha \in N(m, 2r)} h_\alpha m_{\alpha+\beta}(\mathbf{K}) \\ \text{s.t.} \quad & \sum_{\alpha \in N(m, 2r)} h_\alpha m_\alpha(\mathbf{K}) = 1, \\ & \sum_{\alpha \in N(m, 2r)} h_\alpha z^\alpha \in \Sigma[z]_r. \end{aligned} \quad (15)$$

Since the condition that h be a sum-of-squares polynomial is equivalent to a linear matrix inequality in the coefficients of h , problem (15) constitutes a tractable SDP in $h_\alpha, \alpha \in N(m, 2r)$, if x is fixed. The next theorem establishes that we can also efficiently optimize over the feasible set of the constraint (13) whenever the coefficient functions f_β are concave and $\mathbf{K} \subset \mathbb{R}_+^m$.

Theorem 1 *Assume that all f_β are concave functions of x whose supergradients are efficiently computable. Moreover, assume that $\mathbf{K} \subset \mathbb{R}_+^m$. Then, the set of $x \in \mathbb{R}^n$ that satisfy the worst-case expectation constraint (13) is convex and admits a polynomial-time separating hyperplane oracle.*

Proof We have to show that the function $f_{\mathbf{K}}^{(r)}(x)$ defined in (13) is concave in x . To this end, we may rewrite this function as

$$f_{\mathbf{K}}^{(r)}(x) = \inf_{h \in \Sigma[z]_r} \sum_{\beta \in N(m, d)} \mathbb{E}_{z \sim (\mathbf{K}, h)} [z^\beta] f_\beta(x).$$

For each $h \in \Sigma[z]_r$, the function $\mathbb{E}_{z \sim (\mathbf{K}, h)} [z^\beta] f_\beta(x)$ is concave in x because $\mathbf{K} \subset \mathbb{R}_+^m$ implies that $\mathbb{E}_{z \sim (\mathbf{K}, h)} [z^\beta] \geq 0$. Thus, $f_{\mathbf{K}}^{(r)}(x)$ is the point-wise infimum of an infinite collection of concave functions and is therefore itself concave (see, e.g., [42, Theorem 5.5]). This allows us to conclude that the set $\mathcal{C} := \{x \in \mathbb{R}^n \mid f_{\mathbf{K}}^{(r)}(x) \geq 0\}$ is convex.

If $\bar{x} \notin \mathcal{C}$, i.e., $f_{\mathbf{K}}^{(r)}(\bar{x}) < 0$, then we may construct a hyperplane that separates \bar{x} from \mathcal{C} as follows. Let $\bar{h} \in \Sigma[z]_r$ be such that

$$f_{\bar{h}}(\bar{x}) := \sum_{\beta \in N(m, d)} \mathbb{E}_{z \sim (\mathbf{K}, \bar{h})} [z^\beta] f_\beta(\bar{x}) < 0.$$

In particular, we may choose \bar{h} to be the worst-case density obtained by solving the SDP (15) with fixed $x = \bar{x}$. Now let $\partial f_{\bar{h}}(\bar{x})$ denote a supergradient of $f_{\bar{h}}$ at \bar{x} . Note that, by assumption, such a supergradient is available in polynomial time. By the definition of a supergradient, we now have

$$\partial f_{\bar{h}}(\bar{x})^T(x - \bar{x}) \geq f_{\bar{h}}(x) - f_{\bar{h}}(\bar{x}) \geq -f_{\bar{h}}(\bar{x}) > 0 \quad \forall x \in \mathcal{C}.$$

Thus, the hyperplane $\{x \in \mathbb{R}^n \mid \partial f_{\bar{h}}(\bar{x})^T(x - \bar{x}) = 0\}$ separates \bar{x} from \mathcal{C} . \square

Theorem 1 implies that if all coefficient functions f_β are concave, one may minimize a convex function of x over a feasible set given by constraints of the type (13) in polynomial time, e.g., by using the ellipsoid method, provided that an initial ellipsoid is known that contains an optimal solution [20].

Finally, we point out that, due to the convergence properties of the Lasserre hierarchy, one recovers the usual robust counterpart (robust against the single worst-case realization of z as in [1]) in the limit as r tends to infinity.

Theorem 2 *Assume that $\mathbf{K} \subset \mathbb{R}^n$ is closed with nonempty interior. Then, in the limit as $r \rightarrow \infty$, the constraint (13) reduces to the usual robust counterpart constraint*

$$\min_{z \in \mathbf{K}} f(x, z) \geq 0.$$

More precisely, if $x \in \mathbf{K}$ is fixed, and (\mathbf{K}, μ) satisfies one of the assumptions in Table 1, one has

$$\lim_{r \rightarrow \infty} f_{\mathbf{K}}^{(r)}(x) = \min_{z \in \mathbf{K}} f(x, z).$$

Moreover, the rate of convergence is as given in Table 1, depending on the choice of (\mathbf{K}, μ) .

Proof For fixed x , the computation of $f_{\mathbf{K}}^{(r)}(x)$ is an SDP problem of the form (4), and the required convergence result therefore follows from the convergence of the Lasserre hierarchy (4), as summarized in Table 1. \square

4 Approximate solution of the general problem of moments

In applications it is often possible to reduce the ambiguity set \mathcal{P} by including information about the moments of the unknown probability distribution at play. For example, if there is prior information about the location or the dispersion of the random vector z , one can include constraints on its mean vector or its covariance matrix into the definition of the ambiguity set. Specifically, if it is known that $\mathbb{E}_{z \sim (\mathbf{K}, h)}[z^{\beta_i}] = \gamma_i$ for some $\beta_i \in \mathbb{N}_0^n$ and $\gamma_i \in \mathbb{R}$ for $i = 1, \dots, p$, one can restrict the ambiguity set (14) by including the moment constraints

$$\int_{\mathbf{K}} z^{\beta_i} h(z) d\mu(z) = \sum_{\alpha \in N(m, 2r)} h_\alpha m_{\alpha + \beta_i}(\mathbf{K}) = \gamma_i \quad \forall i = 1, \dots, p,$$

which reduce to simple linear equations for the coefficients h_α , $\alpha \in N(m, 2r)$, of the density function h . In this setup, the maximization over the ambiguity set corresponds to a general problem of moments (see, e.g., [46]), where one optimizes the expectation of a function of a random variable over a set of probability measures with known generalized moments. To formalize this problem and to showcase how our approach relates to it, we assume throughout this section that $\mathbf{K} \subset \mathbb{R}^n$ is a nonempty closed set, while f_0, f_1, \dots, f_p are real-valued Borel-measurable functions on \mathbf{K} . Moreover, we assume that μ is a finite Borel measure on \mathbf{K} such that f_0, \dots, f_p are μ -integrable.

Definition 1 The *general problem of moments* is defined as the optimization problem

$$b_0 := \inf_{\nu \in \mathcal{P}_0} \left\{ \int_{\mathbf{K}_0} f_0(z) d\nu(z) : \int_{\mathbf{K}_i} f_i(z) d\nu(z) = b_i \quad \forall i = 1, \dots, p \right\}, \quad (16)$$

where \mathcal{P}_0 is the set of all Borel probability measures supported on \mathbf{K} , and \mathbf{K}_i is a Borel-measurable subset of \mathbf{K} for each $i = 0, \dots, p$.

We now showcase that in the presence of moment information our polynomial-based approach may be used to approximately solve problem (16), and we will illustrate this through concrete examples in Sect. 6. We begin by recalling the following result due to Rogosinsky [43].

Theorem 3 Let \mathbf{K}_i , $i = 0, \dots, p$, be Borel-measurable subsets of \mathbf{K} . Then there exists an atomic Borel measure μ' on \mathbf{K} with a finite support of at most $p + 2$ points so that

$$\int_{\mathbf{K}_i} f_i(z) d\mu(z) = \int_{\mathbf{K}_i} f_i(z) d\mu'(z) \quad \forall i = 0, \dots, p.$$

Proof An elementary proof is given by Shapiro [46, Lemma 3.1]; see also Lasserre [27]. \square

The following corollary is an immediate consequence of Theorem 3.

Corollary 1 If the problem (16) has a solution, it has a solution that is an atomic measure supported on at most $p + 2$ points in \mathbf{K} , i.e., a convex combination of at most $p + 2$ Dirac delta measures supported on \mathbf{K} .

In what follows we show how the atomic measure solution, whose existence is guaranteed by Corollary 1, may be approximated arbitrarily well by SOS polynomial density functions.

Theorem 4 Assume that problem (16) has an optimal solution, $\mathbf{K} \subset \mathbb{R}^n$ has nonempty interior, the functions f_0, f_1, \dots, f_p are polynomials, and that the sets \mathbf{K}_i are closed ($i = 0, \dots, p$). Also assume that \mathbf{K} and μ satisfy one of the assumptions in Table 1. Then, as r tends to infinity,

$$\varepsilon(r) := \inf_{h \in \Sigma_r} \max_{i \in \{0, \dots, p\}} \left| \int_{\mathbf{K}_i} f_i(z) h(z) d\mu(z) - b_i \right|$$

converges to zero at the rate (see also Table 1):

1. $\varepsilon(r) = o(1)$ if \mathbf{K} is basic semi-algebraic, all moments of μ are finite, and (7) holds;
2. $\varepsilon(r) = o(1)$ if \mathbf{K} is compact and μ a finite Borel measure;
3. $\varepsilon(r) = O(r^{-1/4})$ if \mathbf{K} is compact and satisfies an interior cone condition, and μ is the Lebesgue measure;
4. $\varepsilon(r) = O(r^{-1/2})$ if \mathbf{K} is a convex body, and μ is the Lebesgue measure;
5. $\varepsilon(r) = O(r^{-1})$ if $\mathbf{K} = [-1, 1]^n$, and $d\mu(x) = \prod_{i=1}^n (1 - x_i^2)^{-1/2} dx_i$.

Proof We first outline the proof strategy. The high-level idea is to approximate the discrete solution of the moment problem (16) by a convex combination of probability distributions with SOS density functions. Each density is obtained from the Lasserre hierarchy (4), applied to a suitable polynomial (denoted by \hat{p} below).

Define the functions

$$\hat{f}_i(z) = \begin{cases} f_i(z) & \text{if } z \in \mathbf{K}_i, \\ 0 & \text{else,} \end{cases}$$

for $i = 0, \dots, p$. Thus, each \hat{f}_i is Borel-measurable. Moreover, fix $a \in \mathbf{K}$, and let \hat{p} be a polynomial with global minimizer a such that $\hat{p}(a) = 0$ and

$$\hat{p}(z) \geq [\hat{f}_i(z) - \hat{f}_i(a)]^2 \quad \forall z \in \mathbf{K}, \quad i \in \{0, \dots, p\}. \quad (17)$$

For example, \hat{p} may be defined through

$$\begin{aligned} \hat{p}(z) &= \sum_{i=0}^p [f_i(z) - f_i(a)]^2 + \sum_{i:a \notin \mathbf{K}_i} \frac{f_i(z)^2 \|z - a\|^2}{(\text{dist}(a, \mathbf{K}_i))^2} \\ &\quad + \sum_{i:a \in \mathbf{K}_i} \frac{f_i(a)^2 \|z - a\|^2}{(\min_{j:a \notin \mathbf{K}_j} \text{dist}(a, \mathbf{K}_j))^2}, \end{aligned}$$

where $\text{dist}(a, \mathbf{K}_i) = \inf_{z \in \mathbf{K}_i} \|z - a\|$ is the Euclidean distance of a to the closed set \mathbf{K}_i . Indeed, the polynomial \hat{p} is nonnegative and has the global minimizer $z = a$ with $\hat{p}(a) = 0$. Moreover, one may verify that (17) holds by distinguishing the four cases (for fixed $i \in \{0, 1, \dots, p\}$):

- (i) $a \in \mathbf{K}_i$ and $z \in \mathbf{K}_i$,
- (ii) $a \in \mathbf{K}_i$ and $z \notin \mathbf{K}_i$,
- (iii) $a \notin \mathbf{K}_i$ and $z \in \mathbf{K}_i$,
- (iv) $a \notin \mathbf{K}_i$ and $z \notin \mathbf{K}_i$.

It is immediate to verify that (17) holds for cases (i) and (iv). For case (ii), note that

$$\begin{aligned} \hat{p}(z) &\geq [f_i(z) - f_i(a)]^2 + \frac{f_i(a)^2 \|z - a\|^2}{(\min_{j:a \notin \mathbf{K}_j} \text{dist}(a, \mathbf{K}_j))^2} \\ &\geq f_i(a)^2 - 2f_i(z)f_i(a) + f_i(z)^2 + f_i(a)^2 \\ &= [\hat{f}_i(z) - \hat{f}_i(a)]^2 + (f_i(z) - f_i(a))^2 \\ &\geq [\hat{f}_i(z) - \hat{f}_i(a)]^2. \end{aligned}$$

Case (iii) may be verified similarly.

The idea is now to apply the Lasserre hierarchy to minimize \widehat{p} . For a given probability density $h \in \Sigma[z]_r$ with $\int_{\mathbf{K}} h(z)d\mu(z) = 1$, we write as before

$$\mathbb{E}_{z \sim (\mathbf{K}, h)}[\widehat{p}(z)] = \int_{\mathbf{K}} \widehat{p}(z)h(z)d\mu(z).$$

Recalling the notation of the Lasserre hierarchy from (4), we set $\underline{p}_{\mathbf{K}}^{(r)} = \min_{h \in \Sigma[z]_r} \mathbb{E}_{z \sim (\mathbf{K}, h)}[\widehat{p}(z)]$. If μ and \mathbf{K} satisfy any of the conditions in Table 1, we have $\lim_{r \rightarrow \infty} \underline{p}_{\mathbf{K}}^{(r)} = 0$, with the rate of convergence as indicated in the table. Thus, for any $\varepsilon > 0$ there is a sufficiently large $r_a \in \mathbb{N}$ and a density $h_a \in \Sigma[z]_{r_a}$ such that

$$\mathbb{E}_{z \sim (\mathbf{K}, h_a)}[(\widehat{f}_i(z) - \widehat{f}_i(a))^2] \leq \underline{p}_{\mathbf{K}}^{(d)} \leq \varepsilon^2 \quad \forall i \in \{0, \dots, p\}. \quad (18)$$

The asymptotics of the polynomial degree r_a needed to drive the error below ε follows from the convergence rates reported in Table 1. For example, if \mathbf{K} is a convex body and μ is the Lebesgue measure, we may assume that $r_a = O\left(\frac{1}{\varepsilon^2}\right)$, or equivalently that $\varepsilon = \varepsilon(r_a) = O(r_a^{-1/2})$.

Using Jensen's inequality, we have, for each $i \in \{0, \dots, p\}$,

$$\begin{aligned} \mathbb{E}_{z \sim (\mathbf{K}, h_a)}[(\widehat{f}_i(z) - \widehat{f}_i(a))^2] &\geq (\mathbb{E}_{z \sim (\mathbf{K}, h_a)}[(\widehat{f}_i(z) - \widehat{f}_i(a))])^2 \\ &= \begin{cases} \left[\int_{\mathbf{K}_i} f_i(z)h_a(z)d\mu(z) - f_i(a) \right]^2 & \text{if } a \in \mathbf{K}_i, \\ \left[\int_{\mathbf{K}_i} f_i(z)h_a(z)d\mu(z) \right]^2 & \text{if } a \notin \mathbf{K}_i. \end{cases} \end{aligned}$$

We conclude that, for $i \in \{0, \dots, p\}$,

$$\left| \int_{\mathbf{K}_i} f_i(z)h_a(z)d\mu(z) - f_i(a) \right| \leq \varepsilon \text{ if } a \in \mathbf{K}_i, \quad \left| \int_{\mathbf{K}_i} f_i(z)h_a(z)d\mu(z) \right| \leq \varepsilon \text{ if } a \notin \mathbf{K}_i. \quad (19)$$

In this sense, the measure $h_a(z)d\mu(z)$ can be viewed as an approximation of the Dirac point measure δ_a that concentrates unit mass at a .

In the following we denote by $\nu = \sum_{j=0}^{p+1} \lambda_j \delta_{a_j}$ an atomic measure that is optimal in (16) with atoms $a_j \in \mathbf{K}$ and corresponding probabilities $\lambda_j \geq 0$, $j = 0, \dots, p+1$, where $\sum_{j=0}^{p+1} \lambda_j = 1$. The atomic measure ν exists due to Corollary 1 and because the generalized moment problem (16) is solvable by assumption.

By the optimality and feasibility of ν in (16) we have

$$\int_{\mathbf{K}_i} f_i d\nu = \sum_{j: a_j \in \mathbf{K}_i} \lambda_j f_i(a_j) = b_i \quad \forall i \in \{0, \dots, p\}.$$

We now approximate the atomic measure ν by a measure of the form $\sum_j \lambda_j h_{a_j}(z)d\mu(z)$, where h_{a_j} is an SOS density of degree r_{a_j} , say, as in (18) with $a = a_j$. Setting $r = \max_j r_{a_j}$ and $h = \sum_j \lambda_j h_{a_j} \in \Sigma[z]_r$, we may use (19) to conclude that

$$\begin{aligned}
& \left| \int_{\mathbf{K}_i} f_i(z) h(z) d\mu(z) - b_i \right| \\
& \leq \sum_{j: a_j \in \mathbf{K}_i} \lambda_j \left| \int_{\mathbf{K}_i} f_i(z) h_{a_j}(z) d\mu(z) - f_i(a_j) \right| \\
& + \sum_{j: a_j \notin \mathbf{K}_i} \lambda_j \left| \int_{\mathbf{K}_i} f_i(z) h_{a_j}(z) d\mu(z) \right| \leq \sum_j \lambda_j \varepsilon = \varepsilon,
\end{aligned}$$

for each $i \in \{0, \dots, p\}$.

Note again that the relation between ε and r follows from Table 1. For example, if \mathbf{K} is a convex body and μ is the Lebesgue measure, we may assume that $\varepsilon = \varepsilon(r) = O(r^{-1/2})$. The other cases listed in the theorem statement can be proved analogously. \square

As a consequence of Theorem 4, we may obtain approximate solutions to the generalized problem of moments (16) by solving SDPs of the form:

$$\min_{h \in \Sigma[z]_r} \left\{ \int_{\mathbf{K}_0} f_0(z) h(z) d\mu(z) : \int_{\mathbf{K}_i} f_i(z) h(z) d\mu(z) \in [b_i - \varepsilon(r), b_i + \varepsilon(r)] \right. \\
\left. \forall i = 1, \dots, p \right\} \quad (20)$$

for $r \in \mathbb{N}$, and where we may assume $\varepsilon(r) > 0$ depends on r as indicated in the theorem statement.

In analogy to Theorem 2, Theorem 4 has interesting implications for distributionally robust optimization. Indeed, if we replace the set \mathcal{P}_0 of all probability measures used in (16) by the ambiguity set (14) of all probability measures with SOS densities of degree at most $2r$, then we essentially obtain the SDP problem (20) (modulo the tolerance $\varepsilon(r)$ that is needed to ensure feasibility). By Theorem 4, the solution of the SDP (20) converges to the solution of (16) as $r \rightarrow \infty$. Thus we may again view r as the degree of conservatism (or risk-aversion), to be chosen by a user for the problem at hand. We will see detailed examples of such modeling in Sect. 6.

Finally, we emphasize that the tolerance $\varepsilon(r)$ is unavoidable. For example, the constraint

$$\int_0^1 z^2 d\mu(z) = 0$$

for μ is solved by the Dirac delta centered at zero, but it admits no solution of the form $d\mu(z) = h(z)dz$ with $h \in \Sigma_r$ for any r .

We remark that the SDP hierarchy (20) is different from the one studied by Lasserre [27], where an *outer* approximation of the cone of finite Borel measures supported on \mathbf{K} is used, whereas we use an inner approximation. The advantage of the inner approximation is that it allows us to derive the convergence result of Theorem 4, which quantifies the error in terms of the SOS density degrees.

5 Modeling power

The ambiguity set \mathcal{P} defined in (14) contains all distributions supported on a convex body \mathbf{K} that have an SOS polynomial density $h \in \Sigma[z]_r$ with respect to a prescribed reference measure μ .

In stochastic programming, it is common to model μ as a probability measure that best reflects one's information about the stochastics of the uncertain problem parameter z . Thus, μ is typically constructed by fitting a parametric model to a given dataset by using some methods of statistics. Classical stochastic programming takes this estimated model at face value and ignores the possibility that the true data-generating distribution might deviate from μ . In this paper, we suggest to construct μ as one would do in stochastic programming, which implies that stochastic programming actually emerges as a special case of our distributionally robust optimization approach if we set $r = 0$. Setting $r > 0$, on the other hand, implies that we do not place full trust in μ but take into consideration other probability measures that have a polynomial density function $h \in \Sigma[z]_r$ with respect to μ . Intuitively, the less we trust in μ , the higher we should set r . Thus, r can be interpreted as a regularization parameter to be tuned via cross-validation, for instance.

Another possible motivation for the choice of r comes from approximation theory. If one wants to safeguard against a class of distributions, e.g., truncated multivariate normal distributions on some convex body \mathbf{K} with norm of the covariance matrix in some range and mean in some interval, then there are theorems from approximation theory that tell one what the degree of the polynomial SOS density should be to get an approximation of the truncated normal density within ϵ on K .

For example, the following result is implicit in [8].

Theorem 5 Consider a convex body \mathbf{K} and a truncated exponential-type distribution on \mathbf{K} with probability density given by

$$P_f(x) = \frac{\exp(-f(x))}{\int_{\mathbf{K}} \exp(-f(x)) dx},$$

where f is a polynomial of degree d that is nonnegative on \mathbf{K} . Then, there exists a probability density $h \in \Sigma_{rd}$ such that

$$P_f(x) \leq h(x) \leq P_f(x) + \frac{(f(x))^{2r+1}}{(2r+1)! \int_{\mathbf{K}} \exp(-f(x)) dx}.$$

One may therefore choose r such that h is a sufficiently good approximation of P_f , for some class of f where the maximal value of f on \mathbf{K} may be bounded.

We have seen that for any fixed r , the worst-case expectation $f_{\mathbf{K}}^{(r)}(x)$ on the left-hand-side of the worst-case feasibility constraint (13) can be computed efficiently for any fixed x by solving the SDP (15). As decision-makers may have access to prior information that restricts the set of admissible density functions $h \in \Sigma[z]_r$, we demonstrate now that the ambiguity set \mathcal{P} admits several generalizations that preserve the SDP-representability of the worst-case expectation.

Moment information As discussed at length in Sect. 4, conditions on (mixed) moment values of different random variables give rise to simple linear conditions on the polynomial coefficients of h .

Confidence information If the random vector z is known to materialize inside a given Borel set $\mathbf{C} \subset \mathbb{R}^m$ with probability $\gamma \in [0, 1]$, we can add the condition $\mathbb{P}_{z \sim (\mathbf{K}, h)}[z \in \mathbf{C}] = \gamma$ to the definition of the ambiguity set \mathcal{P} . Moreover, if the moments $m_\alpha(\mathbf{K} \cap \mathbf{C})$ of the reference measure μ over $\mathbf{K} \cap \mathbf{C}$ are either available analytically or efficiently computable for all $\alpha \in N(m, 2r)$, then this condition can be re-expressed as the following simple linear equation in the polynomial coefficients of h .

$$\int_{\mathbf{K}} \mathbf{1}_{\mathbf{C}} h(z) d\mu(z) = \sum_{\alpha \in N(m, 2r)} h_\alpha m_\alpha(\mathbf{K} \cap \mathbf{C}) = \gamma$$

Upper and lower bounds on $\mathbb{P}_{z \sim (\mathbf{K}, h)}[z \in \mathbf{C}]$ can be handled similarly in the obvious manner. In the context of purely moment-based ambiguity sets, such probability bounds have been studied in [51].

Conditional probabilities Given any two Borel sets $\mathbf{C}_1, \mathbf{C}_2 \subset \mathbb{R}^m$ and a probability $\gamma \in [0, 1]$, we can also enforce the condition $\mathbb{P}_{z \sim (\mathbf{K}, h)}[z \in \mathbf{C}_2 | z \in \mathbf{C}_1] = \gamma$ in the definition of \mathcal{P} . If the moments $m_\alpha(\mathbf{K} \cap \mathbf{C}_1)$ and $m_\alpha(\mathbf{K} \cap \mathbf{C}_1 \cap \mathbf{C}_2)$ of the reference measure μ are either available analytically or efficiently computable for all $\alpha \in N(m, 2r)$, then this condition can be re-expressed as

$$\begin{aligned} \int_{\mathbf{K}} \mathbf{1}_{\mathbf{C}_1 \cap \mathbf{C}_2} h(z) d\mu(z) &= \gamma \int_{\mathbf{K}} \mathbf{1}_{\mathbf{C}_1} h(z) d\mu(z) \\ \iff \sum_{\alpha \in N(m, 2r)} h_\alpha (m_\alpha(\mathbf{K} \cap \mathbf{C}_1 \cap \mathbf{C}_2) - \gamma m_\alpha(\mathbf{K} \cap \mathbf{C}_1)) &= 0, \end{aligned}$$

which is again linear in the coefficients of h . Upper and lower bounds on conditional probabilities can be handled similarly.

Conditional moment information If it is known that $\mathbb{E}_{z \sim (\mathbf{K}, h)}[z^\beta | \mathbf{C}] = \gamma$ for some $\beta \in \mathbb{N}_0^n$, Borel set $\mathbf{C} \subset \mathbb{R}^m$ and $\gamma \in \mathbb{R}$, while the moments $m_{\alpha+\beta}(\mathbf{K} \cap \mathbf{C})$ of the reference measure μ over set $\mathbf{K} \cap \mathbf{C}$ are either available analytically or efficiently computable for all $\alpha \in N(m, 2r)$, then one can add the following condition to the ambiguity set \mathcal{P} , which is linear in the coefficients of h .

$$\begin{aligned} \int_{\mathbf{K}} z^\beta \mathbf{1}_{\mathbf{C}} h(z) d\mu(z) &= \gamma \int_{\mathbf{K}} \mathbf{1}_{\mathbf{C}} h(z) d\mu(z) \\ \iff \sum_{\alpha \in N(m, 2r)} h_\alpha (m_{\alpha+\beta}(\mathbf{K} \cap \mathbf{C}) - \gamma m_\alpha(\mathbf{K} \cap \mathbf{C})) &= 0, \end{aligned}$$

Multiple reference measures The distributions in the ambiguity set \mathcal{P} defined in (14) depend both on the reference measure μ as well as the density function h . A richer ambiguity set can be constructed by specifying multiple reference measures μ_i with

corresponding density functions $h^i \in \Sigma[z]_r$, $i = 1, \dots, p$. The distributions in the resulting ambiguity set are of the form $\sum_{i=1}^p h^i \cdot d\mu_i$. If the moments $m_\alpha^i(\mathbf{K})$ of the reference measure μ_i over \mathbf{K} are either available analytically or efficiently computable for all $\alpha \in N(m, 2r)$ and $i = 1, \dots, p$, then the normalization constraint can be recast as

$$\sum_{\alpha \in N(m, 2r)} h_\alpha^i m_\alpha^i(\mathbf{K}) = \gamma_i \quad \forall i = 1, \dots, p \quad \text{and} \quad \sum_{i=1}^p \gamma_i = 1,$$

where $\gamma = (\gamma_1, \dots, \gamma_p) \geq 0$ constitutes an auxiliary decision vector. The resulting ambiguity set can be interpreted as a convex combination of p ambiguity sets of the form (14) and thus lends itself for modeling multimodality information; see, e.g., [24]. In this case, γ_i captures the probability of the i -th mode, which may itself be uncertain. Thus, γ should range over a subset of the probability simplex, e.g., a ϕ -divergence uncertainty set of the type studied in [7].

Marginal distributions It is often easier to estimate the marginal distributions of all m components of a random vector z instead of the full joint distribution. Marginal distribution information can also be conveniently encoded in ambiguity sets of the type (14). To see this, assume that the marginal distribution of z_i is given by μ_i and is supported on a compact interval $\mathbf{K}_i \subset \mathbb{R}$, $i = 1, \dots, m$. In this case it makes sense to set $\mathbf{K} = \times_{i=1}^m \mathbf{K}_i$ and to define the reference measure μ through $d\mu = \prod_{i=1}^m d\mu_i$. Thus, μ coincides with the product of the known marginals. The requirement

$$\int_{\times_{j \neq i} \mathbf{K}_j} h(z) \prod_{j \neq i} d\mu_j(z_j) = 1 \quad \forall z_i \in \mathbf{K}_i, \quad \forall i = 1, \dots, m$$

then ensures that the marginal distribution of z_i under $h \cdot d\mu$ exactly matches μ_i . If the moments $m_{\alpha_i}(\mathbf{K}_i)$ of the marginal distribution μ_i over \mathbf{K}_i are either available analytically or efficiently computable for all $\alpha_i = 1, \dots, 2r$, then the above condition simplifies to the linear equations

$$\begin{aligned} \sum_{\substack{\alpha \in N(m, 2r) \\ \alpha_i=0}} h_\alpha \prod_{j \neq i} m_{\alpha_j}(\mathbf{K}_j) &= 1 \quad \text{and} \\ \sum_{\substack{\alpha \in N(m, 2r) \\ \alpha_i=\ell}} h_\alpha \prod_{j \neq i} m_{\alpha_j}(\mathbf{K}_j) &= 0 \quad \forall \ell = 1, \dots, 2r, \quad \forall i = 1, \dots, m. \end{aligned} \quad (21)$$

Situations where the marginals of groups of random variables are known can be handled analogously. Note that when all marginals are known, there is only ambiguity about the dependence structure or *copula* of the components of z [45]. Quantifying the worst-case copula amounts to solving a so-called Fréchet problem. In distributionally robust optimization, Fréchet problems with discrete marginals or approximate marginal matching conditions have been studied in [13, 14, 49].

We emphasize that, in contrast to the existing approaches in the literature, our method allows us to match any (continuous or discrete) marginal distributions *exactly* instead of matching only some of their moments.

We illustrate the proposed construction in the following example, where, for ease of exposition, μ_1 and μ_2 are chosen to be uniform measures.

Example 1 Consider a bivariate random vector $z = (z_1, z_2)$ supported on $\mathbf{K} = [0, 1]^2$, and assume that the marginal distributions of z_1 and z_2 are known to be governed by the uniform probability measures μ_1 and μ_2 supported on $[0, 1]$, respectively. In this case it makes sense to define the reference measure μ through $d\mu = d\mu_1 \cdot d\mu_2$. For ease of exposition, we model the probability density function of z with respect to μ as a polynomial h of degree 2, that is, we set

$$h(z) = h_{00} + h_{10}z_1 + h_{01}z_2 + h_{11}z_1z_2 + h_{20}z_1^2 + h_{02}z_2^2.$$

Integrating h with respect to μ_2 and μ_1 along z_2 and z_1 yields the marginal densities

$$\begin{aligned} h_{z_1}(z_1) &= h_{00} + h_{10}z_1 + \frac{1}{2}h_{01} + \frac{1}{2}h_{11}z_1 + h_{20}z_1^2 + \frac{1}{3}h_{02}, \\ h_{z_2}(z_2) &= h_{00} + \frac{1}{2}h_{10} + h_{01}z_2 + \frac{1}{2}h_{11}z_2 + \frac{1}{3}h_{20} + h_{02}z_2^2, \end{aligned}$$

respectively. Here we used the assumption that μ_1 and μ_2 are uniform measures on $[0, 1]$, which implies that their (non-central) first and second moments are given by $\frac{1}{2}$ and $\frac{1}{3}$, respectively. For z_1 to be distributed according to μ_1 , its marginal density must satisfy $h_{z_1}(z_1) = 1$ for all $z_1 \in [0, 1]$, which is equivalent to the conditions $h_{20} = 0$, $h_{10} + h_{11}/2 = 0$ and $h_{00} + h_{01}/2 + h_{02}/3 = 1$. Similarly, z_2 is distributed according to μ_2 if and only if $h_{02} = 0$, $h_{01} + h_{11}/2 = 0$ and $h_{00} + h_{10}/2 + h_{20}/3 = 1$. If we set $h_{11} = \epsilon$, these conditions imply that $h_{20} = h_{02} = 0$, $h_{10} = h_{01} = -\epsilon/2$ and $h_{00} = 1 + \epsilon/4$. In summary, any admissible quadratic density function of z can be represented as

$$h(z) = 1 + \frac{\epsilon}{4} - \frac{\epsilon}{2}(z_1 + z_2) + \epsilon z_1 z_2$$

and is thus determined by the single parameter ϵ . Note that h is nonnegative on \mathbf{K} if $\epsilon \in [-4, 4]$.

Besides the ambiguity set \mathcal{P} , the constraint function f also admits some generalizations that preserve the SDP-representability of the worst-case expectation in (13).

Uncertainty quantification problems If the constraint function f in (13) is given by $f(x, z) = \mathbf{1}_{\mathbf{C}}$ for some Borel set $\mathbf{C} \subset \mathbb{R}^m$, then the worst-case expectation reduces to the worst-case probability of the set \mathbf{C} . Moreover, if the moments $m_\alpha(\mathbf{K} \cap \mathbf{C})$ of the reference measure μ over $\mathbf{K} \cap \mathbf{C}$ are either available analytically or efficiently computable for all $\alpha \in N(m, 2r)$, then the worst-case probability can be computed by solving a variant of the SDP (15) with the alternative objective function

$$\sum_{\alpha \in N(m, 2r)} h_\alpha m_\alpha(\mathbf{K} \cap \mathbf{C}).$$

6 Numerical experiments

In the following we will exemplify the proposed approach to distributionally robust optimization in the context of financial portfolio analysis (Sect. 6.1), portfolio selection (Sect. 6.2) and risk aggregation (Sect. 6.3).

6.1 Portfolio analysis

Consider a portfolio optimization problem, where the decision vector $x \in \mathbb{R}^n$ captures the percentage weights of the initial capital allocated to n different assets. By definition, one thus has $x_i \in [0, 1]$ for all $i = 1, \dots, n$ and $\sum_i x_i = 1$. We assume that the asset returns $r_i = (u_i + l_i)/2 + z_i(u_i - l_i)/2$ depend linearly on some uncertain risk factors $z_i \in [-1, 1]$ for all $i = 1, \dots, n$, where u_i and l_i represent known upper and lower bounds on the i -th return, respectively. In this framework, we denote by $z \in \mathbb{R}^n$ the vector of all risk factors and by $\mathbf{K} = [-1, 1]^n$ its support. Moreover, the portfolio return can be expressed as

$$f(x, z) = \sum_{i=1}^n x_i \cdot ((u_i + l_i)/2 + z_i(u_i - l_i)/2).$$

Unless otherwise stated, we set μ to the Lebesgue measure on \mathbb{R}^n . Modeling the probability density functions as SOS polynomials allows us to account for various statistical properties and stylized facts of real asset returns as described in [6]. For example, the proposed approach can conveniently capture *gain loss asymmetry*, *i.e.*, the observation that large drawdowns in stock prices and stock index values are more common than equally large upward movements. This feature can be modeled by assigning a higher probability to an individual asset's large upward returns than to its low downward returns. Specifically, the ambiguity set may include the conditions $\mathbb{P}_{z \sim (\mathbf{K}, h)}(z_i \leq a_i) = \gamma_1$ and $\mathbb{P}_{z \sim (\mathbf{K}, h)}(z_i \geq b_i) = \gamma_2$ for some thresholds $a_i < b_i$ and confidence levels $\gamma_1 > \gamma_2$.

Similarly, our approach can handle *correlations of extreme returns*. As pointed out in [6], in spite of the widespread use of the covariance matrix, ‘in circumstances when stock prices undergo large fluctuations [...], a more relevant quantity is the conditional probability of a large (negative) return in one stock given a large negative movement in another stock.’ An example constraint on the conditional probability of one asset's low performance given another assets' lower performance is $\mathbb{P}_{z \sim (\mathbf{K}, h)}(z_i \leq r_i | z_j \leq r_j) \leq \gamma$, where r_i and r_j are given thresholds, while γ is a confidence level.

In this numerical experiment we evaluate the probability that the return of a fixed portfolio x materializes below a prescribed threshold r , that is, we evaluate the worst case of the probability

$$\mathbb{P}_{z \sim (\mathbf{K}, h)} (r(x, z) \leq \underline{r})$$

over an ambiguity set \mathcal{P} of the form (14) with the additional moment constraints $\mathbb{E}_{z \sim (\mathbf{K}, h)}[z^{\beta_i}] = \gamma_i$ for some given exponents $\beta_i \in \mathbb{N}_0^n$ and targets $\gamma_i \in \mathbb{R}$ for $i = 1, \dots, p$. This corresponds to computing the integral of the density function over the knapsack polytope $\mathbf{K} \cap \mathbf{A}(x, u, l, \underline{r})$, where

$$\mathbf{A}(x, u, l, \underline{r}) = \left\{ z \in \mathbb{R}^n : \sum_{i=1}^n x_i(u_i - l_i)z_i/2 \leq \underline{r} - \sum_{i=1}^n x_i(u_i + l_i)/2 \right\}$$

represents a halfspace in \mathbb{R}^n that depends on the fixed portfolio x , the return bounds $l = (l_1, \dots, l_n)$ and $u = (u_1, \dots, u_n)$, and the threshold \underline{r} . To formulate this problem as an SDP, we first need to compute the moments of the monomials with respect to the Lebesgue measure over the given knapsack polytope by using the results of [34]. The worst-case probability problem can then be reformulated as the SDP

$$\begin{aligned} & \sup_{h(z)} \sum_{\alpha \in N(n, 2r)} h_\alpha m_\alpha(\mathbf{K} \cap \mathbf{A}(x, u, l, \underline{r})) \\ & \text{s.t. } \sum_{\alpha \in N(n, 2r)} h_\alpha m_\alpha(\mathbf{K}) = 1, \\ & \quad \sum_{\alpha \in N(n, 2r)} h_\alpha m_{\alpha + \beta_i}(\mathbf{K}) = \gamma_i \quad \forall i = 1, \dots, p, \\ & \quad \sum_{\alpha \in N(n, 2r)} h_\alpha z^\alpha \in \Sigma[z]_r. \end{aligned} \tag{22}$$

In the numerical experiment we assume that there are $n = 2$ assets with lower and upper return bounds $l = (0.8, 0.7)^\top$ and $u = (1.2, 1.3)^\top$, respectively. We evaluate the probability that the return of the fixed portfolio $x = (0.75, 0.25)^\top$ falls below the threshold $\underline{r} = 0.9$ (the minimum possible return of the portfolio is 0.775). We assume that the only known moment information about the asset returns is that their means both vanish, that is, we set $p = 2$, $\beta_1 = (1, 0)$, $\beta_2 = (0, 1)$ and $\gamma_1 = \gamma_2 = 0$. Table 2 reports the *exact* optimal values of the SDP (22) for $r = 1, \dots, 12$. The last value in the table (labeled $r = \infty$) provides the worst-case probability across *all* distributions satisfying the prescribed moment conditions (not only those with a polynomial density) and was computed using the methods described in [23]. In this case, one can also show that there exists a worst-case distribution with only two atoms. It assigns probability 0.31 to the scenario $z = (1, 1)^\top$ and probability 0.69 to the scenario $z = (-0.44, -0.44)^\top$. All SDPs are solved using SeDuMi [48] via the CVX interface [19].

Table 2 Worst-case probability of the portfolio return falling below \underline{r} computed by solving the SDP (22)

r	0	1	2	3	4	5	6	7	8	9	10	11	12	∞
0	0.17	0.39	0.48	0.50	0.53	0.55	0.56	0.58	0.59	0.59	0.60	0.61	0.61	0.69

6.2 Mean-variance portfolio selection

We now study a portfolio optimization problem that minimizes the worst-case variance subject to a lower bound on the worst-case mean of the portfolio return. This amounts to solving a distributionally robust optimization problem where an investor selects the portfolio allocation vector x , while a fictitious adversary chooses the density function $h(z)$ of the asset return distribution with the goal to inflict maximum damage to the investor. Specifically, we solve the optimization problem

$$\begin{aligned} \min_x \quad & \sup_{\mathbb{P} \in \mathcal{P}} \mathbb{E}_{\mathbb{P}} \left[\left(\sum_{i=1}^n x_i \cdot ((u_i + l_i)/2 + z_i(u_i - l_i)/2) \right. \right. \\ & \quad \left. \left. - \mathbb{E}_{\mathbb{P}} \left[\sum_{i=1}^n x_i \cdot ((u_i + l_i)/2 + z_i(u_i - l_i)/2) \right] \right)^2 \right] \\ \text{s.t.} \quad & \inf_{\mathbb{P} \in \mathcal{P}} \mathbb{E}_{\mathbb{P}} \left[\sum_{i=1}^n x_i \cdot ((u_i + l_i)/2 + z_i(u_i - l_i)/2) \right] \geq \underline{r} \\ & 1^\top x = 1 \\ & x \geq 0, \end{aligned} \tag{23}$$

where the asset returns are explained by a vector of risk factors $z \in \mathbf{K} = [-1, 1]^n$ as in Sect. 6.1, while the ambiguity set containing all possible distributions of z is defined as

$$\mathcal{P} = \left\{ h \cdot dz : h \in \Sigma[z]_r, \int_{\mathbf{K}} h(z) dz = 1, \int_{\mathbf{K}} z h(z) dz = 0 \right\}.$$

In the numerical experiments we will assume that there are $n = 3$ assets whose returns have known upper and lower bounds $u = (1, 1.2, 1.3)^\top$ and $l = (1, 0.85, 0.8)^\top$, respectively, and known mean values $(u + l)/2$. This choice implies that asset 1 is risk-free. For ease of exposition, we henceforth set $v_i^\pm = (u_i \pm l_i)/2$ for $i = 1, \dots, n$. As z has zero mean under every $\mathbb{P} \in \mathcal{P}$, the return target constraint reduces to the linear inequality $\sum_{i=1}^n x_i v_i^+ \geq \underline{r}$, while the objective function simplifies to

$$\sup_{h \in \Sigma[z]_r} \left\{ \int_{\mathbf{K}} \left(\sum_{i=1}^n x_i z_i v_i^- \right)^2 h(z) dz : \int_{\mathbf{K}} h(z) dz = 1, \int_{\mathbf{K}} z h(z) dz = 0 \right\}. \tag{24}$$

Next, we define $m_r(z)$ as the vector of all monomials of degree at most r in the variables z . Then, any SOS density function $h \in \Sigma[z]_r$ can be represented as $h(z) = m_r(z)^\top H m_r(z)$ for some positive semidefinite coefficient matrix H . Moreover, it is useful to introduce the matrices

$$M^0 = \int_{\mathbf{K}} m_r(z) m_r(z)^\top dz, \quad M^i = \int_{\mathbf{K}} z_i m_r(z) m_r(z)^\top dz \quad \forall i = 1, \dots, n,$$

and

$$A^{ij} = \int_{\mathbf{K}} z_i v_i^- z_j v_j^- m_r(z) m_r(z)^\top dz \quad \forall i, j = 1, \dots, n.$$

Using this notation, the worst-case variance problem (24) can be reformulated as the SDP

$$\begin{aligned} & \sup_{H \succeq 0} \sum_{i,j=1}^n x_i x_j \langle H, A^{ij} \rangle \\ & \text{s.t. } \langle H, M^0 \rangle = 1, \quad \langle H, M^i \rangle = 0 \quad \forall i = 1, \dots, n, \end{aligned}$$

where $\langle \cdot, \cdot \rangle$ stands for the trace inner product. The SDP dual to the above problem is given by

$$\begin{aligned} & \inf_{\gamma} \gamma_0 \\ & \text{s.t. } \sum_{i=0}^n \gamma_i M^i \succeq \sum_{i,j=1}^n x_i x_j A^{ij}. \end{aligned} \tag{25}$$

Strong duality holds because $M^0 \succ 0$ as an outer product of the vector $m_r(z)$ of all monomials of degree at most r in the variables z , integrated over a convex body with non-empty interior. Therefore, the dual SDP is strictly feasible. Substituting (25) into (23) converts the distributionally robust portfolio selection problem to the nonlinear SDP

$$\begin{aligned} & \inf_{x, \gamma} \gamma_0 \\ & \text{s.t. } \sum_{i=0}^n \gamma_i M^i \succeq \sum_{i,j=1}^n x_i x_j A^{ij} \\ & \quad \sum_{i=1}^n x_i v_i \geq r_- \\ & \quad 1^\top x = 1 \\ & \quad x \geq 0. \end{aligned} \tag{26}$$

In the following we will demonstrate that (26) is a convex optimization problem. Indeed, by recalling the definition of the matrices A^{ij} , introducing a Borel-measurable matrix-valued function $Y(z)$ and using Schur complements, the nonlinear matrix inequality in (26) can be reformulated as

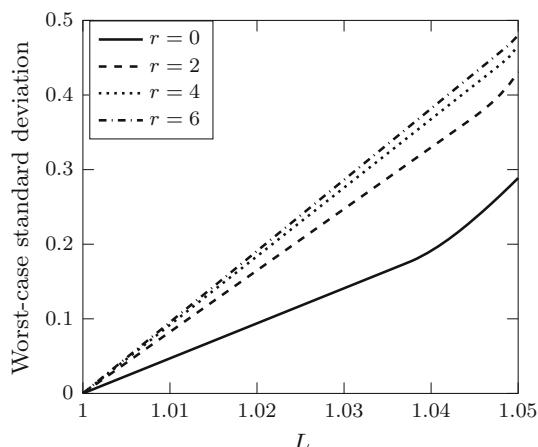
$$\begin{aligned} & \sum_{i=0}^n \gamma_i H^i \succeq \int_{\mathbf{K}} \left(\sum_{i=1}^n x_i z_i v_i \right)^2 m_r(z) m_r(z)^\top dz \\ & \iff \begin{cases} \exists Y(\cdot) \text{ Borel-measurable : } \sum_{i=0}^n \gamma_i H^i \succeq \int_{\mathbf{K}} Y(z) dz, \\ Y(z) \succeq \left(\sum_{i=1}^n x_i z_i v_i \right)^2 m_r(z) m_r(z)^\top \quad \forall z \in \mathbf{K} \end{cases} \end{aligned}$$

$$\iff \begin{cases} \exists Y(\cdot) \text{ Borel-measurable : } \sum_{i=0}^n \gamma_i H^i \succeq \int_{\mathbf{K}} Y(z) dz, \\ \left(\begin{array}{cc} 1 & \left(\sum_{i=1}^n x_i z_i v_i \right) m_r(z)^\top \\ \left(\sum_{i=1}^n x_i z_i v_i \right) m_r(z) & Y(z) \end{array} \right) \succeq 0 \quad \forall z \in \mathbf{K}. \end{cases}$$

This reasoning shows that the feasible set of the first constraint in (26) can be expressed through uncountably many linear matrix inequalities. Thus, the feasible set of (26) can be viewed as a finite-dimensional projection of an infinite-dimensional convex set. As convexity is preserved under projections, we conclude that (26) is indeed a convex optimization problem. We can thus solve (26) to global optimality using the PENLAB solver [15] via the YALMIP modeling language [32].

Figures 1 and 2 visualize the worst-case standard deviation of the portfolio return and the corresponding optimal portfolio weights as the return threshold r is swept. The following observations are in order. First, if r is smaller or equal to the risk-free return, it is optimal to allocate all the capital to the risk-free asset, in which case the standard deviation of the portfolio return vanishes. As r increases, the investor is forced to shift capital to the risky assets, which leads to an increase of the worst-case standard deviation. Moreover, increasing the degree $2r$ of the polynomial density functions amounts to giving the adversary more flexibility to inflict damage. Thus, the worst-case standard deviation increases with r for any fixed return target r . We further observe that the portfolio weights as well as the worst-case standard deviation display a distinct kink when the structure of the optimal portfolio changes, that is, when the weight of the risk-free asset drops to 0, in which case the corresponding no-short-sales constraint becomes active.

Fig. 1 Mean-standard deviation efficient frontiers for $r \in \{0, 2, 4, 6\}$



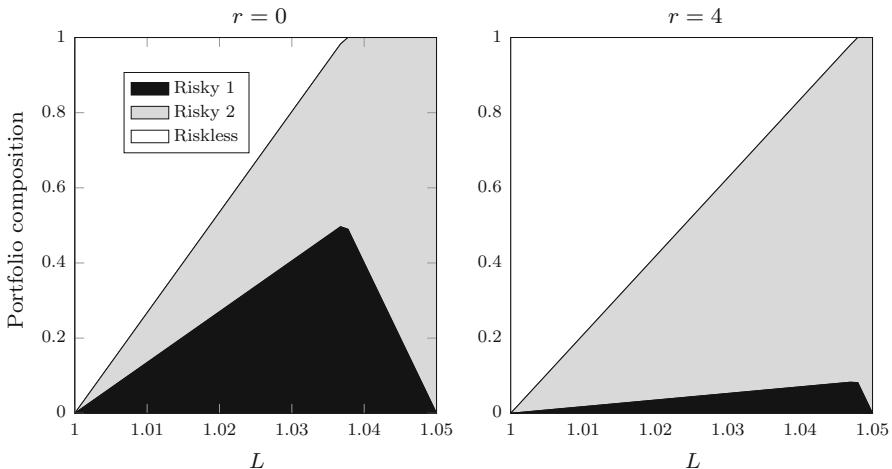


Fig. 2 Optimal portfolio compositions for $r \in \{0, 4\}$

6.3 Risk aggregation

In the third experiment we study the risk aggregation problem of an insurer holding a portfolio of different random losses z_i , $i = 1, \dots, n$, corresponding to different types of insurance claims, e.g., life, vehicle, health or home insurance policies, etc. Inspired by [49, § 6], we aim to estimate the worst-case probability that the sum of the n losses exceeds a critical threshold $b = 10$ beyond which the insurance company would be driven into illiquidity. Formally, we aim to maximize

$$\mathbb{P}_{z \sim (\mathbf{K}, h)} (z_1 + \dots + z_n \geq b) \quad (27)$$

across all distributions in an ambiguity set \mathcal{P} , which reflects the prior distributional information available to the insurer. We will consider different models for the domain \mathbf{K} of $z = (z_1, \dots, z_n)$, the reference measure μ on \mathbf{K} and the ambiguity set \mathcal{P} . Throughout the experiments we will always assume that the reference measure is separable with respect to the losses, that is, we assume that

$$d\mu(z) = \varrho_1(z_1) \cdots \varrho_n(z_n) dz,$$

where ϱ_i denotes a given density function (with respect to the Lebesgue measure) of the random variables z_i for each $i = 1, \dots, n$. We will consider the following complementary settings:

1. *Lognormal densities:* We set $\mathbf{K} = \mathbb{R}_+^n$ and let ϱ_i be a lognormal density function defined earlier in (11), but repeated here for convenience:

$$\varrho_i(z_i) = \frac{1}{z_i v_i \sqrt{2\pi}} \exp\left(-\frac{(\log(z_i) - \bar{z}_i)^2}{2v_i^2}\right), \quad (28)$$

- where \bar{z}_i and v_i represent prescribed location and scale parameters, $i = 1, \dots, n$.
2. *Exponential densities:* We set $\mathbf{K} = \mathbb{R}_+^n$ and let ϱ_i be the exponential density function with unit rate parameter defined through $\varrho_i(z_i) = \exp(-z_i)$, $i = 1, \dots, n$. The resulting reference measure is intimately related to the orthogonal Laguerre polynomials.
 3. *Uniform densities:* We set $\mathbf{K} = [0, M]^n$ for some constant $M > 0$ and let ϱ_i be the uniform density function defined through $\varrho_i(z_i) = 1/M$, $i = 1, \dots, n$. Note that under this choice the reference measure is proportional to the Lebesgue measure.

In order to reformulate the risk aggregation problem as a tractable SDP, we need the moments of the reference measure μ over the hypercube \mathbf{K} and over the knapsack polytope $\mathbf{K} \cap \mathbf{C}$, where

$$\mathbf{C} = \{z \in \mathbb{R}^n : z_1 + \dots + z_n \geq b\}.$$

For all classes of density functions described above, the moments of μ are indeed accessible. Specifically, under the lognormal densities, the moments of μ over \mathbf{K} are given by (12) and are repeated here for convenience:

$$\begin{aligned} m_\alpha(\mathbf{K}) &= \int_{\mathbf{K}} \left(\prod_{i=1}^n z_i^{\alpha_i} \right) \prod_{i=1}^n \frac{1}{z_i v_i \sqrt{2\pi}} \exp \left(-\frac{(\log(z_i) - \bar{z}_i)^2}{2v_i^2} \right) dz \\ &= \prod_{i=1}^n \exp(\alpha_i \bar{z}_i + (\alpha_i v_i)^2 / 2). \end{aligned}$$

Moreover, the moments of μ over $\mathbf{K} \cap \mathbf{C}$ can be expressed as

$$\begin{aligned} m_\alpha(\mathbf{K} \cap \mathbf{C}) &= m_\alpha(\mathbf{K}) - m_\alpha(\mathbf{K} \setminus \mathbf{C}) \\ &= m_\alpha(\mathbf{K}) - \int_{\mathbf{K} \setminus \mathbf{C}} \left(\prod_{i=1}^n z_i^{\alpha_i} \right) \prod_{i=1}^n \frac{1}{z_i v_i \sqrt{2\pi}} \exp \left(-\frac{(\log(z_i) - \bar{z}_i)^2}{2v_i^2} \right) dz. \end{aligned}$$

To evaluate the integral in the last line, we use the cubature routine `adsimp()` in MATLAB, which greedily refines a simplicial partition of the integration domain by iteratively subdividing the simplices with the largest estimated integration errors [16]. For the polynomial degrees considered in our experiments, the moments computed with this routine are sufficiently precise to result in numerically stable optimization problems. For higher degrees, however, computing sufficiently precise moments takes too much time or is not achievable. Note that for the exponential and the uniform densities, the moments of the reference measure μ over \mathbf{K} and $\mathbf{K} \cap \mathbf{C}$ are available in closed form.

We assume that the insurance company is able to estimate the marginal distributions of the individual losses either exactly or approximately by using a combination of statistical analysis and probabilistic modeling. However, the insurer has no information about the underlying copula. This type of distributional information is often

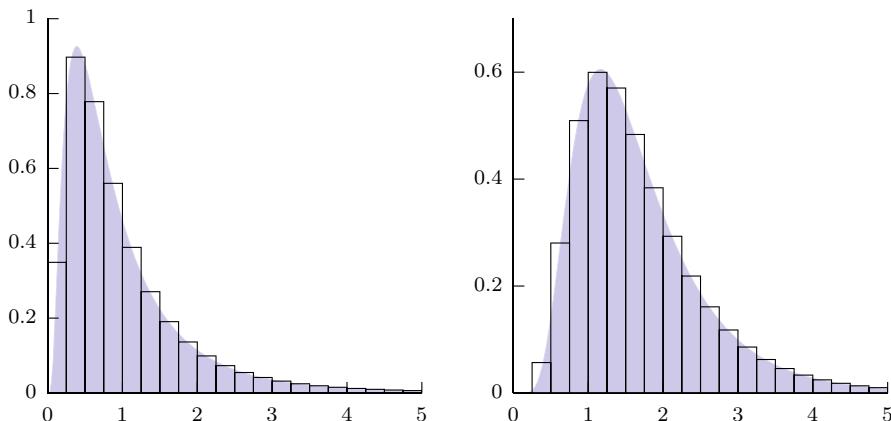


Fig. 3 Histograms of the lognormal marginal distributions of z_1 (left) and z_2 (right)

justified in practice because obtaining reliable marginal information requires significantly less data than obtaining exact dependence structures; see, e.g., [33]. Throughout the experiment we assume that there are $n = 2$ random losses governed by lognormal probability density functions of the form (28) with parameters $\bar{z}_1 = -0.3$, $\bar{z}_2 = 0.4$, $v_1 = 0.8$ and $v_2 = 0.5$. The ambiguity set \mathcal{P} then contains all distributions of the form $h \cdot d\mu$, $h \in \Sigma[z]_r$, under which the marginals of the losses follow the prescribed lognormal distributions either exactly or approximately. More precisely, we model the marginal distributional information as follows:

1. *Marginal distribution matching*: The lognormal distributions of the individual losses are matched *exactly* by any distribution $h \cdot d\mu$ in the ambiguity set. This can be achieved by defining the reference measure μ as the product of the marginal lognormal distributions and by requiring that h satisfies (21). Note that under the alternative reference measures corresponding to the uniform or exponential density functions, lognormal marginals cannot be matched exactly with polynomial densities of any degrees. This is because the lognormal density function can neither be expressed as a polynomial nor as the product of a polynomial with the exponential density. Note also that an exact matching of (non-discrete) marginal distributions cannot be enforced with the existing numerical techniques for solving Fréchet problems proposed in [13, 14, 49].
2. *Marginal moment matching*: The marginals of the individual losses have the same moments of order 0, 1 or 2 as the prescribed lognormal distributions. Note that this kind of moment matching can be enforced under any of the reference measures corresponding to lognormal, exponential or uniform density functions. Moreover, moment matching is also catered for in [49] bar the extra requirement that the joint distribution of the losses must have an SOS polynomial density.
3. *Marginal histogram matching*: We may associate a histogram with each marginal lognormal distribution as illustrated in Fig. 3 and require that the marginals of the losses under the joint distribution $h \cdot d\mu$ have the same histograms. This condition can be enforced under any of the reference measures corresponding to lognormal,

exponential or uniform density functions. In the numerical experiments, we use histograms with 20 bins of width 0.25 starting at the origin. Histogram matching is also envisaged in [49]. Formally, we thus define the ambiguity set as

$$\mathcal{P} = \left\{ h \cdot d\mu : h \in \Sigma[z]_r, \int_{\mathbf{K}} h(z) d\mu(z) = 1, \int_{\mathbf{K}} \mathbf{1}_{B_l}(z) h(z) d\mu(z) = p_l \forall l = 1, \dots, L \right\},$$

where $B_l \subset \mathbf{K}$ is the l -th bin, $p_l = \int_{\mathbf{K}} \mathbf{1}_{B_l}(z) d\mu(z)$ is its probability under the reference measure and $L \in \mathbb{N}$ is the total number of bins. As indicated above, in the numerical experiments we set $L = 40$. Moreover, we define the bins as $B_l = \{z \in \mathbf{K} : (l-1)/4 \leq z_1 \leq l/4\}$ for $l = 1, \dots, 20$ and $B_l = \{z \in \mathbf{K} : (l-1)/4 \leq z_2 \leq l/4\}$ for $l = 21, \dots, 40$. Note that since lognormal distributions are supported on \mathbb{R}_+ , we have $\sum_{l=1}^{20} p_l = \sum_{l=21}^{40} p_l < 1$, and thus \mathcal{P} is empty unless the square $[0, 5]^2$ is a *strict* subset of \mathbf{K} . In the experiments this condition is enforced both for the lognormal and the exponential reference measures, in which case we set $\mathbf{K} = \mathbb{R}_+^2$, as well as for the uniform reference measure, in which case we set $\mathbf{K} = [0, 10]^2$.

For $\mathbf{K} = \mathbb{R}_+^2$ and the reference measure corresponding to the lognormal density functions, the worst-case values of the probability (27) are reported in Table 3. Results are shown for $r \leq 5$, which corresponds to polynomial densities of degrees at most 10. The last row of the table ($r = \infty$) provides the worst-case probabilities across *all* distributions satisfying the prescribed moment or histogram conditions (not only those with a polynomial density) and was computed using the methods described in [49]. Note that under moment matching up to order 2, the worst-case probability for $r = 5$ amounts to 0.0021, as opposed to the much higher probability of 0.0615 obtained with the approach from [49]. A similar observation holds for histogram matching. The requirement that the distributions in the ambiguity set be sufficiently regular in the sense that they admit a polynomial density function with respect to the reference measure is therefore restrictive and effectively rules out pathological discrete worst-case

Table 3 Worst-case probabilities for the lognormal reference measure

r	Moment matching up to order			Histogram matching	Distribution matching
	0	1	2		
0	0.0017	0.0017	0.0017	0.0017	0.0017
1	0.1432	0.0042	0.0017	0.0017	0.0017
2	0.8255	0.0106	0.0020	0.0019	0.0018
3	0.9982	0.0114	0.0021	0.0022	0.0019
4	1.0000	0.0117	0.0021	0.0026	0.0023
5	1.0000	0.0118	0.0021	0.0026	0.0023
∞	1.0000	1.0000	0.0615	0.0198	n/a

Table 4 Worst-case probabilities for the exponential reference measure

r	Moment matching up to order			Histogram matching		
	0	1	2	ℓ_1 -dist. ≤ 0.1	ℓ_1 -dist. ≤ 0.05	ℓ_1 -dist. ≤ 0.02
0	0.0005	—	—	—	—	—
1	0.0214	0.0147	—	—	—	—
2	0.2058	0.0823	—	—	—	—
3	0.6481	0.1484	—	—	—	—
4	0.9393	0.1497	0.0086	—	—	—
5	0.9953	0.1699	0.0104	—	—	—
6	0.9998	0.1709	0.0139	—	—	—
7	1.0000	0.1800	0.0158	—	—	—
8	1.0000	0.1860	0.0182	0.0802	—	—
9	1.0000	0.1862	0.0207	0.1076	—	—
10	1.0000	0.1928	0.0224	0.1144	0.0515	—
11	1.0000	0.1968	0.0244	0.1156	0.0633	0.0204
12	1.0000	0.1971	0.0262	0.1160	0.0652	0.0320
∞	1.0000	1.0000	0.0615	n/a	n/a	n/a

distributions. Moreover, the worst-case probabilities under exact distribution matching and under histogram matching are of the same order of magnitude for all $r \leq 5$ but significantly smaller than the worst-case probability under histogram matching for $r = \infty$. A key question to be asked in practice is thus whether one deems the class of distributions $h \cdot d\mu$ with $h \in \Sigma[z]_r$ to be rich enough to contain all ‘reasonable’ distributions.

Table 4 reports the worst-case probabilities corresponding to the reference measure on $\mathbf{K} = \mathbb{R}_+^2$ induced by the exponential density functions. For low values of r , the polynomial densities lack the necessary degrees of freedom to match all imposed moment constraints. In these situations, the worst-case probability problem becomes infeasible (indicated by ‘—’). When feasible, however, we managed to solve the problem for r up to 12. The density functions corresponding to large values of r are highly flexible and thus result in worst-case probabilities that are closer to those obtained by the benchmark method from [49], which relaxes the restriction to a subspace of polynomial densities. Similar phenomena are also observed in the context of histogram matching. It was impossible to match the prescribed histogram probabilities exactly for all $r \leq 12$. We thus relaxed the histogram matching conditions in the definition of the ambiguity set to allow for densities whose implied marginal histograms are within a prescribed ℓ_1 -distance from the target histograms.

This approximate histogram matching condition is easily captured in our framework and gives rise to a few extra linear constraints on the coefficients of the polynomial density function. Table 4 reports the worst-case probabilities for three different tolerances on the histogram mismatch in terms of the ℓ_1 -distance. We observe that the resulting worst-case probabilities are significantly larger than those obtained under the lognormal reference measure and increase with the ℓ_1 -tolerance.

Table 5 Worst-case probabilities for the uniform reference measure

r	Moment matching up to order			Histogram matching		
	0	1	2	ℓ_1 -dist. ≤ 0.1	ℓ_1 -dist. ≤ 0.05	ℓ_1 -dist. ≤ 0.02
0	0.5000	–	–	–	–	–
1	0.9082	–	–	–	–	–
2	0.9933	–	–	–	–	–
3	0.9997	0.0304	–	–	–	–
4	1.0000	0.1035	–	–	–	–
5	1.0000	0.1340	–	–	–	–
6	1.0000	0.1612	0.0089	–	–	–
7	1.0000	0.1783	0.0166	–	–	–
8	1.0000	0.1935	0.0192	–	–	–
9	1.0000	0.2042	0.0216	0.0738	–	–
10	1.0000	0.2133	0.0274	0.1066	0.0407	–
11	1.0000	0.2202	0.0292	0.1142	0.0609	–
12	1.0000	0.2274	0.0311	0.1163	0.0653	0.0178
∞	1.0000	1.0000	0.0615	n/a	n/a	n/a

Finally, Table 5 reports the worst-case probabilities corresponding to the uniform reference measure on $\mathbf{K} = [0, 10]^2$. The results are qualitatively similar to those of Table 4, but they also show that the choice of the reference measure plays an important role when r is small.

7 Conclusions

In this paper, we present first steps towards using SOS polynomial densities in distributionally robust optimization for problems that display a polynomial dependence on the uncertain parameters. The main advantages of this approach may be summarized as follows:

1. The proposed framework is tractable (in the sense of polynomial-time solvability) for SOS density functions of any fixed degree.
2. The approach offers considerable modeling flexibility. Specifically, one may conveniently encode various salient features of the unknown distribution of the uncertain parameters through linear constraints and/or linear matrix inequalities.
3. In the limit as the degree of the SOS density functions tends to infinity, one recovers the usual robust counterpart or generalized moment problem. One may therefore view the degree of the density as a tuning parameter that captures the model's 'level of conservativeness.'

The approach also suffers from shortcomings that necessitate further work and insights:

1. The approach is not applicable to objective or constraint functions that display a general (decision-dependent) piecewise polynomial dependence on the uncertain parameters as is the case for the recourse functions of linear two-stage stochastic programs.
2. The proposed distributionally robust optimization problems can be reduced to generalized eigenvalue problems or even semidefinite programs of large sizes that are often poorly conditioned.

Acknowledgements Etienne de Klerk would like to thank Dorota Kurowicka and Jean Bernard Lasserre for valuable discussions and references. Daniel Kuhn gratefully acknowledges financial support from the Swiss National Science Foundation under grant BSCGI0_157733.

References

1. Ben-Tal, A., El Ghaoui, L., Nemirovski, A.: Robust Optimization. Princeton University Press, Princeton (2009)
2. Bertsimas, D., Popescu, I.: On the relation between option and stock prices: a convex optimization approach. *Oper. Res.* **50**(2), 358–374 (2002)
3. Bertsimas, D., Popescu, I.: Optimal inequalities in probability theory: a convex optimization approach. *SIAM J. Optim.* **15**(3), 780–804 (2005)
4. Bertsekas, D.P.: Convex Optimization Theory. Athena Scientific, Belmont (2009)
5. Birge, J.R., Louveaux, F.: Introduction to Stochastic Programming. Springer, New York (1997)
6. Cont, R.: Empirical properties of asset returns: stylized facts and statistical issues. *Quant. Finance* **1**, 223–236 (2001)
7. den Ben-Tal, A., den Hertog, D., de Waegenaere, A., Melenberg, B., Rennen, G.: Robust solutions of optimization problems affected by uncertain probabilities. *Manag. Sci.* **59**(2), 341–357 (2013)
8. De Klerk, E., Laurent, M.: Comparison of Lasserre’s measure-based bounds for polynomial optimization to bounds obtained by simulated annealing. *Math. Oper. Res.* **43**(4), 1317–1325 (2018)
9. De Klerk, E., Laurent, M.: Worst-case examples for Lasserre’s measure-based hierarchy for polynomial optimization on the hypercube (2018). Preprint available at [arXiv:1804.05524](https://arxiv.org/abs/1804.05524)
10. De Klerk, E., Laurent, M., Sun, Z.: Convergence analysis for Lasserre’s measure-based hierarchy of upper bounds for polynomial optimization. *Math. Program. Ser. A* **162**(1), 363–392 (2017)
11. Dantzig, G.B.: Linear programming under uncertainty. *Manag. Sci.* **1**(3–4), 197–206 (1955)
12. Delage, E., Ye, Y.: Distributionally robust optimization under moment uncertainty with application to data-driven problems. *Oper. Res.* **58**(3), 595–612 (2010)
13. Doan, X.V., Li, X., Natarajan, K.: Robustness to dependency in portfolio optimization using overlapping marginals. *Oper. Res.* **63**(6), 1468–1488 (2015)
14. Doan, X.V., Natarajan, K.: On the complexity of nonoverlapping multivariate marginal bounds for probabilistic combinatorial optimization problems. *Oper. Res.* **60**(1), 138–49 (2012)
15. Fiala, J., Kočvara, M., Michael Stingl, M.: PENLAB: a MATLAB solver for nonlinear semidefinite optimization. <http://arxiv.org/pdf/1311.5240v1.pdf> (2013)
16. Genz, A., Cools, R.: An adaptive numerical cubature algorithm for simplices. *ACM Trans. Math. Softw.* **29**(3), 297–308 (2003)
17. Goh, J., Sim, M.: Distributionally robust optimization and its tractable approximations. *Oper. Res.* **58**(4), 902–917 (2010)
18. Golub, G.H., Van Loan, C.F.: Matrix Computations, 3rd edn. The John Hopkins University Press, Baltimore (1996)
19. Grant, M., Boyd, S.: CVX: Matlab Software for Disciplined Convex Programming. <http://cvxr.com/cvx> (2014)
20. Grötschel, M., Lovász, L., Schrijver, A.: Geometric Algorithms and Combinatorial Optimization. Springer, New York (1988)
21. Grundmann, A., Moeller, H.M.: Invariant integration formulas for the n -simplex by combinatorial methods. *SIAM J. Numer. Anal.* **15**, 282–290 (1978)

22. Hanasusanto, G.A., Roitch, V., Kuhn, D., Wiesemann, W.: A distributionally robust perspective on uncertainty quantification and chance constrained programming. *Math. Program. Ser. B* **151**(1), 35–62 (2015)
23. Hanasusanto, G.A., Roitch, V., Kuhn, D., Wiesemann, W.: Ambiguous joint chance constraints under mean and dispersion information. *Oper. Res.* **65**(3), 751–767 (2017)
24. Hanasusanto, G.A., Kuhn, D., Wallace, S.W., Zymler, S.: Distributionally robust multi-item news-vendor problems with multimodal demand distributions. *Math. Program. Ser. A* **152**(1), 1–32 (2015)
25. Kroo, A., Szilárd, R.: On Bernstein and Markov-type inequalities for multivariate polynomials on convex bodies. *J. Approx. Theory* **99**(1), 134–152 (1999)
26. Lasserre, J.B., Zeron, E.S.: Solving a class of multivariate integration problems via Laplace techniques. *Appl. Math.* **28**(4), 391–405 (2001)
27. Lasserre, J.B.: A semidefinite programming approach to the generalized problem of moments. *Math. Program. Ser. B* **112**, 65–92 (2008)
28. Lasserre, J.B.: A new look at nonnegativity on closed sets and polynomial optimization. *SIAM J. Optim.* **21**(3), 864–885 (2011)
29. Lasserre, J.B.: The K -moment problem for continuous linear functionals. *Trans. Am. Math. Soc.* **365**(5), 2489–2504 (2012)
30. Lasserre, J.B., Weisser, T.: Representation of distributionally robust chance-constraints (2018). Preprint available at [arXiv:1803.11500](https://arxiv.org/abs/1803.11500)
31. Li, B., Jiang, R., Mathieu, J.L.: Ambiguous risk constraints with moment and unimodality information. *Mathematical Programming Series A* (to appear) (2017). Preprint available at http://www.optimization-online.org/DB_FILE/2016/09/5635.pdf
32. Löfberg, J.: YALMIP: a toolbox for modeling and optimization in MATLAB. In: Proceedings of the CACSD Conference (2004)
33. McNeil, A., Frey, R., Embrechts, P.: Quantitative Risk Management: Concepts, Techniques and Tools. Princeton University Press, Princeton (2015)
34. Marichal, J.-L., Mossinghof, M.J.: Slices, slabs, and sections of the unit hypercube. *Online J. Anal. Comb.* **3**, 1–11 (2008)
35. Mevissen, M., Ragnoli, E., Yu, J.Y.: Data-driven distributionally robust polynomial optimization. In: Burges, C.J.C., Bottou, L., Welling, M., Ghahramani, Z., Weinberger, K.Q. (eds.) *Advances in Neural Information Processing Systems 26*, pp. 37–45. Curran Associates, Inc. (2013)
36. Mohajerin Esfahani, P., Kuhn, D.: Data-driven distributionally robust optimization using the Wasserstein metric: performance guarantees and tractable reformulations. *Math. Program. Ser. A* **171**(1–2), 115–166 (2018)
37. Natarajan, K., Pachamanova, D., Sim, M.: Constructing risk measures from uncertainty sets. *Oper. Res.* **57**(5), 1129–1141 (2009)
38. Pflug, G.C., Pichler, A., Wozabal, D.: The $1/N$ investment strategy is optimal under high model ambiguity. *J. Bank. Finance* **36**(2), 410–417 (2012)
39. Pflug, G.C., Wozabal, D.: Ambiguity in portfolio selection. *Quant. Finance* **7**, 435–442 (2007)
40. Popescu, I.: A semidefinite programming approach to optimal-moment bounds for convex classes of distributions. *Math. Oper. Res.* **30**(3), 632–657 (2005)
41. Prékopa, A.: Stochastic Programming. Kluwer Academic Publishers, Berlin (1995)
42. Rockafellar, R.T.: Convex Analysis. Princeton University Press, Princeton (1970)
43. Rogosinski, W.W.: Moments of non-negative mass. *Proc. R. Soc. A* **245**, 1–27 (1958)
44. Scarf, H.: A min–max solution of an inventory problem. In: Scarf, H., Arrow, K., Karlin, S. (eds.) *Studies in the Mathematical Theory of Inventory and Production*, vol. 10, pp. 201–209. Stanford University Press, Redwood City (1958)
45. Sklar, A.: Fonctions de répartition à n dimensions et leurs marges. *Publications de l'Institut de Statistique de L'Université de Paris* **8**, 229–231 (1959)
46. Shapiro, A.: On duality theory of conic linear problems. In: Goberna, M.Á., López, M.A. (eds.) *Semi-Infinite Programming: Recent Advances*, pp. 135–165. Springer, New York (2001)
47. Shapiro, A., Dentcheva, D., Ruszczyński, A.: Lectures on Stochastic Programming: Modeling and Theory. SIAM, Philadelphia (2009)
48. Sturm, J.F.: Using SeDuMi 1.02, a MATLAB toolbox for optimization over symmetric cones. In: Optimization Methods and Software, pp. 11–12, 625–653 (1999)
49. Van Parys, B.P.G., Goulart, P.J., Embrechts, P.: Fréchet inequalities via convex optimization (2016). Preprint available at http://www.optimization-online.org/DB_FILE/2016/07/5536.pdf

50. Van Parys, B.P.G., Goulart, P.J., Kuhn, D.: Generalized Gauss inequalities via semidefinite programming. *Math. Program. Ser. A* **156**(1–2), 271–302 (2016)
51. Wiesemann, W., Kuhn, D., Sim, M.: Distributionally robust convex optimization. *Oper. Res.* **62**(6), 1358–1376 (2014)
52. Žáčková, J.: On minimax solutions of stochastic linear programming problems. *Časopis pro pěstování matematiky* **91**, 423–430 (1966)
53. Zuluaga, L., Peña, J.F.: A conic programming approach to generalized Tchebycheff inequalities. *Math. Oper. Res.* **30**(2), 369–388 (2005)

Publisher's Note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.