A "JOINT+MARGINAL" APPROACH TO PARAMETRIC POLYNOMIAL OPTIMIZATION

JEAN B. LASSERRE*

Abstract. Given a compact parameter set $\mathbf{Y} \subset \mathbb{R}^p$, we consider polynomial optimization problems $(\mathbf{P_y})$ on $\mathbb{R}^n$ whose description depends on the parameter $\mathbf{y} \in \mathbf{Y}$. We assume that one can compute all moments of some probability measure φ on $\mathbf{Y}$, absolutely continuous with respect to the Lebesgue measure (e.g. $\mathbf{Y}$ is a box or a simplex and φ is uniformly distributed). We then provide a hierarchy of semidefinite relaxations whose associated sequence of optimal solutions converges to the moment vector of a probability measure that encodes all information about all global optimal solutions $\mathbf{x}^*(\mathbf{y})$ of $\mathbf{P_y}$, as $\mathbf{y} \in \mathbf{Y}$. In particular, one may approximate as closely as desired any polynomial functional of the optimal solutions, like e.g. their φ-mean. In addition, using this knowledge on moments, the measurable function $\mathbf{y} \mapsto x_k^*(\mathbf{y})$ of the k-th coordinate of optimal solutions, can be estimated, e.g. by maximum entropy methods. Also, for a boolean variable x_k, one may approximate as closely as desired its persistency $\varphi(\{\mathbf{y} : x_k^*(\mathbf{y}) = 1\}$, i.e. the probability that in an optimal solution $\mathbf{x}^*(\mathbf{y})$, the coordinate $x_k^*(\mathbf{y})$ takes the value 1. At last but not least, from an optimal solution of the dual semidefinite relaxations, one provides a sequence of polynomial (resp. piecewise polynomial) lower approximations with $L_1(\varphi)$ (resp. φ-almost uniform) convergence to the optimal value function.

Key words. Parametric and polynomial optimization; semidefinite relaxations

AMS subject classifications. 65 D15, 65 K05, 46 N10, 90 C22

1. Introduction.

Roughly speaking, given a set of parameters $\mathbf{Y}$ and an optimization problem whose description depends on $\mathbf{y} \in \mathbf{Y}$ (call it $\mathbf{P_y}$), *parametric optimization* is concerned with the behavior and properties of the optimal value as well as primal (and possibly dual) optimal solutions of $\mathbf{P_y}$, when $\mathbf{y}$ varies in $\mathbf{Y}$. This is quite a challenging problem and in general one may only obtain information locally around some nominal value $\mathbf{y}_0$ of the parameter. There is a vast and rich literature on the topic and for a detailed treatment, the interested reader is referred to e.g. Bonnans and Shapiro [5] and the many references therein.

Sometimes, in the context of optimization with data uncertainty, some probability distribution φ on the parameter set $\mathbf{Y}$ is available and in this context one is also interested in e.g. the distribution of the optimal value, optimal solutions, all viewed as random variables. In fact, very often only some moments of φ are known (typically first- and second-order moments) and the goal of this more realistic moment-based approach is to obtain optimal bounds over all distributions that share this moment information. In particular, for discrete optimization problems where coefficients of the cost vector (of the objective function to optimize) are random variables with joint distribution φ, some bounds on the expected optimal value have been obtained. More recently Natarajan et al. [20] extended the earlier work in [4] to even provide a convex optimization problem for computing the so-called *persistency* values[1] of (discrete) variables, for a particular distribution φ^* in a certain set Φ of distributions that share some common moment information. However, the resulting second-order cone program requires knowledge of the convex hull of some discrete set of points, which

* LAAS-CNRS and Institute of Mathematics, University of Toulouse, 7 Avenue du Colonel Roche, 31077 Toulouse Cédex 4, France. Tel: 33 +5 61336415; Fax: 33 +5 61336936; email: lasserre@laas.fr

[1] Given a $0-1$ optimization problem $\max\{\mathbf{c}'\mathbf{x} : \mathbf{x} \in \mathcal{X} \cap \{0,1\}^n\}$ and a distribution φ on $\mathbf{c}$, the persistency value of the variable x_i is $\text{Prob}_\varphi(x_i^* = 1)$ at an optimal solution $\mathbf{x}^*(\mathbf{c}) = (x_i^*)$.

is possible when the number of vertices is small. The approach is nicely illustrated on a discrete choice problem and a stochastic knapsack problem. For more details on persistency in discrete optimization, the interested reader is referred to [20] and the references therein. Recently Natarajan et al. [21] have considered mixed zero-one linear programs with uncertainty in the objective function and first- and second-order moment information. They show that computing the supremum of the expected optimal value (where the supremum is over all distributions sharing this moment information) reduces to solving a completely positive program.

Moment-based approaches also appear in robust optimization and stochastic programming under data uncertainty, where the goal is different since in this context, decisions of interest must be taken *before* the uncertain data is known. For these *min-max* type problems, a popular approach initiated in Arrow et al. [2] is to optimize decisions against the worst possible distribution θ (on uncertain data) taken in some set Φ of candidate distributions that share some common moment information (in general first- and second-order moments). Recently Delage and Ye [9]) have even considered the case where only a confidence interval is known for first- and second-order moments. For a nice discussion the interested reader is referred to [9] and the many references therein.

In the context of solving systems of polynomial equations whose coefficients are themselves polynomials of some parameter $\mathbf{y} \in \mathbf{Y}$, specific "parametric" methods exist. For instance, one may compute symbolically once and for all, what is called a *comprehensive* Gröbner basis, i.e., a fixed basis that, for all instances of $\mathbf{y} \in \mathbf{Y}$, is a Gröbner basis of the ideal associated with the polynomials in the system of equations; see Weispfenning [29] and more recently Rostalski [23] for more details. Then when needed, and for any specific value of the parameter $\mathbf{y}$, one may compute all complex solutions of the system of equations, e.g. by the eigenvalue method of Möller and Stetter [19]. However, one still needs to apply the latter method for *each* value of the prameter $\mathbf{y}$. A similar two-step approach is also proposed for homotopy (instead of Gröbner bases) methods in [23].

The purpose of this paper devoted to parametric optimization is to show that in the case of *polynomial* optimization, all information about the optimal value and optimal solutions can be obtained, or at least, approximated as closely as desired.

Contribution. We here restrict our attention to parametric polynomial optimization, that is, when $\mathbf{P_y}$ is described by *polynomial* equality and inequality constraints on both the parameter vector $\mathbf{y}$ and the optimization variables $\mathbf{x}$. Moreover, the set $\mathbf{Y}$ is restricted to be a compact basic semi-algebraic set of $\mathbb{R}^p$, and preferably a set sufficiently simple so that one may obtain the moments of some probability measure on $\mathbf{Y}$, absolutely continuous with respect to the Lebesgue measure (or with respect to the counting measure if $\mathbf{Y}$ is discrete). For instance if $\mathbf{Y}$ is a simple set (like a simplex, a box) one may choose φ to be the probability measure uniformly distributed on $\mathbf{Y}$; typical $\mathbf{Y}$ candidates are polyhedra. Or sometimes, in the context of optimization with data uncertainty, φ is already specified. In this specific context we are going to show that one may get insightful information on the set of all global minimizers of $\mathbf{P_y}$ and on the global optimum, via what we call a *"joint+marginal"* approach. Our contribution is as follows:

(a) Call $J(\mathbf{y})$ (resp. $\mathbf{X}^*_\mathbf{y} \in \mathbb{R}^n$) the optimal value (resp. the set of optimal solutions) of $\mathbf{P_y}$ for the value $\mathbf{y} \in \mathbf{Y}$ of the parameter. We first define an infinite-

dimensional optimization problem $\mathbf{P}$ whose optimal value is exactly $\rho = \int_{\mathbf{Y}} J(\mathbf{y})d\varphi(\mathbf{y})$, i.e. the φ-mean of the global optimum. Any optimal solution of $\mathbf{P}$ is a probability measure μ^* on $\mathbb{R}^n \times \mathbb{R}^p$ with marginal φ on $\mathbb{R}^p$. It turns out that μ^* encodes all information on the set of global minimizers $\mathbf{X}^*_{\mathbf{y}}$, $\mathbf{y} \in \mathbf{Y}$. Whence the name *"joint+marginal"* as μ^* is a *joint* distribution of $\mathbf{x}$ and $\mathbf{y}$, and φ is the *marginal* of μ^* on $\mathbf{Y} \subset \mathbb{R}^p$.

(b) Next, we provide a hierarchy of semidefinite relaxations of $\mathbf{P}$ with associated sequence of optimal values $(\rho_i)_i$, in the spirit of the hierarchy defined in [15]. An optimal solution of the i-th semidefinite relaxation is a sequence $\mathbf{z}^i = (z^i_{\alpha\beta})$ indexed in the monomial basis $(\mathbf{x}^\alpha \mathbf{y}^\beta)$ of the subspace $\mathbb{R}[\mathbf{x}, \mathbf{y}]_{2i}$ of polynomials of degree at most $2i$. If for φ-almost all $\mathbf{y} \in \mathbf{Y}$, $\mathbf{P}_\mathbf{y}$ has a unique *global* minimizer $\mathbf{x}^*(\mathbf{y}) \in \mathbb{R}^n$, then as $i \to \infty$, $\mathbf{z}^i$ converges pointwise to the sequence of moments of μ^* defined in (a). In particular, one obtains the distribution of the optimal solution $\mathbf{x}^*(\mathbf{y})$, and therefore, one may approximate as closely as desired any polynomial functional of the solution $\mathbf{x}^*(\mathbf{y})$, like e.g. the φ-mean or variance of $\mathbf{x}^*(\mathbf{y})$.

In addition, if the optimization variable x_k is boolean then one may approximate as closely as desired its *persistency* $\varphi(\{\mathbf{y} : x^*_k(\mathbf{y}) = 1\}$ (i.e., the probability that $x^*_k(\mathbf{y}) = 1$ in an optimal solution $\mathbf{x}^*(\mathbf{y})$), as well as a a necessary and sufficient condition for this persistency to be 1.

(c) Let $e(k) \in \mathbb{N}^n$ be the vector with $e_j(k) = \delta_{j=k}$, $j = 1, \ldots, n$. Then as $i \to \infty$, and for every $\beta \in \mathbb{N}^p$, the sequence $(z^i_{e(k)\beta})$ converges to $z^*_{k\beta} := \int_{\mathbf{Y}} \mathbf{y}^\beta g_k(\mathbf{y})d\varphi(\mathbf{y})$ for the measurable function $\mathbf{y} \mapsto g_k(\mathbf{y}) := x^*_k(\mathbf{y})$. In other words, the sequence $(z^*_{k\beta})_{\beta \in \mathbb{N}^p}$ is the moment sequence of the measure $d\psi(y) := x^*_k(\mathbf{y})d\varphi(\mathbf{y})$ on $\mathbf{Y}$. And so, the k-th coordinate function $\mathbf{y} \mapsto x^*_k(\mathbf{y})$ of the global minimizer of $\mathbf{P}_\mathbf{y}$, $\mathbf{y} \in \mathbf{Y}$, can be estimated, e.g. by maximum entropy methods. Of course, the latter estimation is not pointwise but it still provides useful information on optimal solutions, e.g. the shape of the function $\mathbf{y} \mapsto x^*_k(\mathbf{y})$, especially if the function $x^*_k(\cdot)$ is continuous, as illustrated on some simple examples. For instance, for parametric polynomial equations, one may use this estimation of $\mathbf{x}^*(\mathbf{y})$ as an initial point for Newton's method for any given value of the parameter $\mathbf{y}$.

(d) At last but not least, from an optimal solution of the dual of the i-th semidefinite relaxation, one obtains a piecewise polynomial approximation of the optimal value function $\mathbf{y} \mapsto J(\mathbf{y})$, that converges φ-almost surely to J.

Finally, the computational complexity of the above methodology is roughly the same as the moment approach described in [15] for an optimization problem with $n + p$ variables since we consider the joint distribution of the n variables $\mathbf{x}$ and the p parameters $\mathbf{y}$. Hence, the approach is particularly interesting when the number of parameters is small, say 1 or 2. In addition, in the latter case the max-entropy estimation has been shown to be very efficient in several examples in the literature; see e.g. [6, 26, 27]. However, in view of the present status of SDP solvers, if no sparsity or symmetry is taken into account as proposed in e.g. [17], the approach is limited to small to medium size polynomial optimization problems. Alternatively one may also use LP-relaxations which can handle larger size problems but with probably less precise results because of their poor convergence properties in general.

But this computational price may not seem that high in view of the ambitious goal of the approach. After all, keep in mind that by applying the moment approach to a single $(n + p)$-variables problem, one obtains information on global optimal solutions of an n-variables problem that depends on p parameters, that is, one approximates n *functions* of p variables!

2. A related linear program. For a Borel space X let $\mathcal{B}(X)$ denote the Borel σ-field associated with X. Let $\mathbb{R}[\mathbf{x},\mathbf{y}]$ denote the ring of polynomials in the variables $\mathbf{x} = (x_1,\ldots,x_n)$, and the variables $\mathbf{y} = (y_1,\ldots,y_p)$, whereas $\mathbb{R}[\mathbf{x},\mathbf{y}]_k$ denotes its subspace of polynomials of degree at most k. Let $\Sigma[\mathbf{x},\mathbf{y}] \subset \mathbb{R}[\mathbf{x},\mathbf{y}]$ denote the subset of polynomials that are sums of squares (in short s.o.s.). For a real symmetric matrix $\mathbf{A}$ the notation $\mathbf{A} \succeq 0$ stands for $\mathbf{A}$ is positive semidefinite.

The parametric optimization problem. Let $\mathbf{Y} \subset \mathbb{R}^p$ be a compact set, called the *parameter* set, and let $f, h_j : \mathbb{R}^n \times \mathbb{R}^p \to \mathbb{R}$, $j = 1,\ldots,m$, be continuous. For each $\mathbf{y} \in \mathbf{Y}$, fixed, consider the following optimization problem:

$$J(\mathbf{y}) := \inf_{\mathbf{x}} \{\, f_{\mathbf{y}}(\mathbf{x}) \,:\, h_{j\mathbf{y}}(\mathbf{x}) \geq 0,\ j = 1,\ldots,m \,\} \tag{2.1}$$

where the functions $f_{\mathbf{y}}, h_{j\mathbf{y}} : \mathbb{R}^n \to \mathbb{R}$ are defined via:

$$\left.\begin{array}{rcl} \mathbf{x} \mapsto f_{\mathbf{y}}(\mathbf{x}) & := & f(\mathbf{x},\mathbf{y}) \\ \mathbf{x} \mapsto h_{j\mathbf{y}}(\mathbf{x}) & := & h_j(\mathbf{x},\mathbf{y}),\ j = 1,\ldots,m \end{array}\right\} \quad \forall\, \mathbf{x} \in \mathbb{R}^n,\ \forall\, \mathbf{y} \in \mathbb{R}^p.$$

Next, let $\mathbf{K} \subset \mathbb{R}^n \times \mathbb{R}^p$ be the set:

$$\mathbf{K} := \{\, (\mathbf{x},\mathbf{y}) \,:\, \mathbf{y} \in \mathbf{Y}\,; \quad h_j(\mathbf{x},\mathbf{y}) \geq 0, \quad j = 1,\ldots,m \,\}, \tag{2.2}$$

and for each $\mathbf{y} \in \mathbf{Y}$, let

$$\mathbf{K}_{\mathbf{y}} := \{\, \mathbf{x} \in \mathbb{R}^n \,:\, h_{j\mathbf{y}}(\mathbf{x}) \geq 0, \quad j = 1,\ldots,m \,\}. \tag{2.3}$$

The interpretation is as follows: $\mathbf{Y}$ is a set of parameters and for each instance $\mathbf{y} \in \mathbf{Y}$ of the parameter, one wishes to compute an optimal *decision* vector $\mathbf{x}^*(\mathbf{y})$ that solves problem (2.1). Let φ be a Borel probability measure on $\mathbf{Y}$, with a positive density with respect to the Lebesgue measure on the smallest affine variety that contains $\mathbf{Y}$. For instance choose for φ the probability measure

$$\varphi(B) := \left(\int_{\mathbf{Y}} d\mathbf{y}\right)^{-1} \int_B d\mathbf{y}, \qquad \forall B \in \mathcal{B}(\mathbf{Y}),$$

uniformly distributed on $\mathbf{Y}$ (assuming of course that $\mathbf{Y}$ has nonempty interior). Of course, one may also treat the case of a *discrete* set of parameters $\mathbf{Y}$ (finite or countable) by taking for φ a discrete probability measure on $\mathbf{Y}$ with strictly positive weight at each point of the support. Sometimes, e.g. in the context of optimization with data uncertainty, φ is already specified.

We will use φ (or more precisely, its moments) to get information on the distribution of optimal solutions $\mathbf{x}^*(\mathbf{y})$ of $\mathbf{P}_{\mathbf{y}}$, viewed as random vectors.

In the rest of the paper we assume that for every $\mathbf{y} \in \mathbf{Y}$, the set $\mathbf{K}_{\mathbf{y}}$ in (2.3) is nonempty.

2.1. A related infinite-dimensional linear program. Let $\mathbf{M}(\mathbf{K})$ be the set of finite Borel measures on $\mathbf{K}$, and consider the following infinite-dimensional linear program $\mathbf{P}$:

$$\rho := \inf_{\mu \in \mathbf{M}(\mathbf{K})} \{\, \int_{\mathbf{K}} f\, d\mu \,:\, \pi\mu = \varphi \,\} \tag{2.4}$$

where $\pi\mu$ denotes the marginal of μ on $\mathbb{R}^p$, that is, $\pi\mu$ is a probability measure on $\mathbb{R}^p$ defined by

$$\pi\mu(B) := \mu(\mathbb{R}^n \times B), \qquad \forall\, B \in \mathcal{B}(\mathbb{R}^p).$$

Notice that $\mu(\mathbf{K}) = 1$ for any feasible solution μ of $\mathbf{P}$. Indeed, as φ is a probability measure and $\pi\mu = \varphi$ one has $1 = \varphi(\mathbf{Y}) = \mu(\mathbb{R}^n \times \mathbb{R}^p) = \mu(\mathbf{K})$.

Recall that for two Borel spaces X, Y, the graph $\mathrm{Gr}\,\psi \subset X \times Y$ of a set-valued mapping $\psi : X \to Y$ is the set

$$\mathrm{Gr}\,\psi := \{(\mathbf{x}, \mathbf{y}) \,:\, \mathbf{x} \in X\,;\, \mathbf{y} \in \psi(\mathbf{x})\,\}.$$

If ψ is measurable then any measurable function $h : X \to Y$ with $h(\mathbf{x}) \in \psi(\mathbf{x})$ for every $\mathbf{x} \in X$, is called a *measurable selector*. (See §4.1 for more details.)

LEMMA 2.1. *Let both $\mathbf{Y} \subset \mathbb{R}^n$ and $\mathbf{K}$ in (2.2) be compact. Then the set-valued mapping $\mathbf{y} \mapsto \mathbf{K}_\mathbf{y}$ is Borel-measurable. In addition:*

(a) *The mapping $\mathbf{y} \mapsto J(\mathbf{y})$ is measurable.*

(b) *There exists a measurable selector $g : \mathbf{Y} \to \mathbf{K}_y$ such that $J(\mathbf{y}) = f(g(\mathbf{y}), \mathbf{y})$ for every $\mathbf{y} \in \mathbf{Y}$.*

Proof. As $\mathbf{K}$ and $\mathbf{Y}$ are both compact, the set valued mapping $\mathbf{y} \mapsto \mathbf{K}_\mathbf{y} \subset \mathbb{R}^n$ is compact-valued. Moreover, the graph of $\mathbf{K}_\mathbf{y}$ is by definition the set $\mathbf{K}$, which is a Borel subset of $\mathbb{R}^n \times \mathbb{R}^p$. Next, since $\mathbf{x} \mapsto f_\mathbf{y}(\mathbf{x})$ is continuous for every $\mathbf{y} \in \mathbf{Y}$, (a) and (b) follow from Proposition 4.3 and 4.4. □

THEOREM 2.2. *Let both $\mathbf{Y} \subset \mathbb{R}^p$ and $\mathbf{K}$ in (2.2) be compact and assume that for every $\mathbf{y} \in \mathbf{Y}$, the set $\mathbf{K}_\mathbf{y} \subset \mathbb{R}^n$ in (2.3) is nonempty. Let $\mathbf{P}$ be the optimization problem (2.4) and let $\mathbf{X}^*_\mathbf{y} := \{\mathbf{x} \in \mathbb{R}^n \,:\, f(\mathbf{x}, \mathbf{y}) = J(\mathbf{y})\}$, $\mathbf{y} \in \mathbf{Y}$. Then:*

(a) $\rho = \displaystyle\int_\mathbf{Y} J(\mathbf{y})\, d\varphi(\mathbf{y})$ *and $\mathbf{P}$ has an optimal solution.*

(b) *For every optimal solution μ^* of $\mathbf{P}$, and for φ-almost all $\mathbf{y} \in \mathbf{Y}$, there is a probability measure $\psi^*(d\mathbf{x}\,|\,\mathbf{y})$ on $\mathbb{R}^n$, concentrated on $\mathbf{X}^*_\mathbf{y}$, such that:*

$$\mu^*(C \times B) = \int_B \psi^*(C\,|\,\mathbf{y})\, d\varphi(\mathbf{y}), \qquad \forall B \in \mathcal{B}(\mathbf{Y}),\, C \in \mathcal{B}(\mathbb{R}^n). \tag{2.5}$$

(c) *Assume that for φ-almost all $\mathbf{y} \in \mathbf{Y}$, the set of minimizers of $\mathbf{X}^*_\mathbf{y}$ is the singleton $\{\mathbf{x}^*(\mathbf{y})\}$ for some $\mathbf{x}^*(\mathbf{y}) \in \mathbf{K}_y$. Then there is a measurable mapping $g : \mathbf{Y} \to \mathbf{K}_\mathbf{y}$ such that*

$$g(\mathbf{y}) = \mathbf{x}^*(\mathbf{y}) \;\; \textit{for every} \;\; \mathbf{y} \in \mathbf{Y}\,; \quad \rho = \int_\mathbf{Y} f(g(\mathbf{y}), \mathbf{y})\, d\varphi(\mathbf{y}), \tag{2.6}$$

and for every $\alpha \in \mathbb{N}^n$, and $\beta \in \mathbb{N}^p$:

$$\int_\mathbf{K} \mathbf{x}^\alpha \mathbf{y}^\beta\, d\mu^*(\mathbf{x}, \mathbf{y}) = \int_\mathbf{Y} \mathbf{y}^\beta\, g(\mathbf{y})^\alpha\, d\varphi(\mathbf{y}). \tag{2.7}$$

Proof. (a) As $\mathbf{K}$ is compact then so is $\mathbf{K}_\mathbf{y}$ for every $\mathbf{y} \in \mathbf{Y}$. Next, as $\mathbf{K}_\mathbf{y} \neq \emptyset$ for every $\mathbf{y} \in \mathbf{Y}$ and f is continuous, the set $\mathbf{X}^*_\mathbf{y} := \{\mathbf{x} \in \mathbb{R}^n \,:\, f(\mathbf{x}, \mathbf{y}) = J(\mathbf{y})\}$ is nonempty for every $\mathbf{y} \in \mathbf{Y}$. Let μ be any feasible solution of $\mathbf{P}$ and so by definition, its marginal on $\mathbb{R}^p$ is just φ. Since $\mathbf{X}^*_\mathbf{y} \neq \emptyset$, $\forall \mathbf{y} \in \mathbf{Y}$, one has $f_\mathbf{y}(\mathbf{x}) \geq J(\mathbf{y})$ for all $\mathbf{x} \in \mathbf{K}_\mathbf{y}$ and all $\mathbf{y} \in \mathbf{Y}$. So, $f(\mathbf{x}, \mathbf{y}) \geq J(\mathbf{y})$ for all $(\mathbf{x}, \mathbf{y}) \in \mathbf{K}$ and therefore

$$\int_\mathbf{K} f\, d\mu \geq \int_\mathbf{K} J(\mathbf{y})\, d\mu = \int_\mathbf{Y} J(\mathbf{y})\, d\varphi,$$

which proves that $\rho \geq \int_{\mathbf{Y}} J(\mathbf{y})\, d\varphi$.

On the other hand, recall that $\mathbf{K}_\mathbf{y} \neq \emptyset, \forall \mathbf{y} \in \mathbf{Y}$. Consider the set-valued mapping $\mathbf{y} \mapsto \mathbf{X}^*_\mathbf{y} \subset \mathbf{K}_\mathbf{y}$. As f is continuous and $\mathbf{K}$ is compact, then $\mathbf{X}^*_\mathbf{y}$ is compact-valued. In addition, as $f_\mathbf{y}$ is continuous, by Proposition 4.4 there exists a measurable selector $g : \mathbf{Y} \to \mathbf{X}^*_y$ (and so $f(g(\mathbf{y}), \mathbf{y}) = J(\mathbf{y})$). Therefore, for every $\mathbf{y} \in \mathbf{Y}$, let $\psi^*_\mathbf{y} := \delta_{g(\mathbf{y})}$ be the Dirac probability measure with support on the singleton $g(\mathbf{y}) \in \mathbf{X}^*_\mathbf{y}$, and let μ be the probability measure on $\mathbf{K}$ defined by:

$$\mu(C \times B) \,:=\, \int_B 1_C(g(\mathbf{y}))\, \varphi(d\mathbf{y}), \qquad \forall B \in \mathcal{B}(\mathbb{R}^p),\, C \in \mathcal{B}(\mathbb{R}^n).$$

(The measure μ is well-defined because g is measurable.) Then μ is feasible for $\mathbf{P}$ and

$$\begin{aligned}\rho \leq \int_\mathbf{K} f\, d\mu &= \int_\mathbf{Y} \left[\int_{\mathbf{K}_\mathbf{y}} f(\mathbf{x}, \mathbf{y})\, d\delta_{g(\mathbf{y})} \right] d\varphi(\mathbf{y}) \\ &= \int_\mathbf{Y} f(g(\mathbf{y}), \mathbf{y})\, d\varphi(\mathbf{y}) = \int_\mathbf{Y} J(\mathbf{y})\, d\varphi(\mathbf{y}),\end{aligned}$$

which shows that μ is an optimal solution of $\mathbf{P}$ and $\rho = \int_\mathbf{Y} J(\mathbf{y}) d\varphi(\mathbf{y})$.

(b) Let μ^* be an arbitrary optimal solution of $\mathbf{P}$, hence on $\mathbf{Y} \times \mathbb{R}^n$ and concentrated on $\mathrm{Gr}\, \mathbf{K}_\mathbf{y} = \mathbf{Y} \times \mathbf{K}_\mathbf{y}$. Therefore by Proposition 4.6 the probability measure μ^* can be disintegrated as

$$\mu^*(C \times B) \,:=\, \int_B \psi^*(C \,|\, \mathbf{y})\, d\varphi(\mathbf{y}), \qquad \forall B \in \mathcal{B}(\mathbf{Y}),\, C \in \mathcal{B}(\mathbb{R}^n),$$

where for all $\mathbf{y} \in \mathbf{Y}$, $\psi^*(\cdot \,|\, \mathbf{y})$ is a probability measure on $\mathbf{K}_\mathbf{y}$. (The object $\psi^*(\cdot|\cdot)$ is called a stochastic kernel; see Proposition 4.6.) Hence from (a),

$$\begin{aligned}\rho = \int_\mathbf{Y} J(\mathbf{y})\, d\varphi(\mathbf{y}) &= \int_\mathbf{K} f(\mathbf{x}, \mathbf{y})\, d\mu^*(\mathbf{x}, \mathbf{y}) \\ &= \int_\mathbf{Y} \left(\int_{\mathbf{K}_\mathbf{y}} f(\mathbf{x}, \mathbf{y})\, \psi^*(d\mathbf{x} \,|\, \mathbf{y}) \right) d\varphi(\mathbf{y}).\end{aligned}$$

Therefore, using $f(\mathbf{x}, \mathbf{y}) \geq J(\mathbf{y})$ on $\mathbf{K}$,

$$0 = \int_\mathbf{Y} \left(\int_{\mathbf{K}_\mathbf{y}} \underbrace{J(\mathbf{y}) - f(\mathbf{x}, \mathbf{y})}_{\leq 0}\, \psi^*(d\mathbf{x} \,|\, \mathbf{y}) \right) d\varphi(\mathbf{y}),$$

which implies $\psi^*(\mathbf{X}^*(\mathbf{y}) \,|\, \mathbf{y}) = 1$ for φ-almost all $\mathbf{y} \in \mathbf{Y}$.

(c) Let $g : \mathbf{Y} \to \mathbf{K}_\mathbf{y}$ be the measurable mapping of Lemma 2.1(b). As $J(\mathbf{y}) = f(g(\mathbf{y}), \mathbf{y})$ and $(g(\mathbf{y}), \mathbf{y}) \in \mathbf{K}$ then necessarily $g(\mathbf{y}) \in \mathbf{X}^*_\mathbf{y}$ for every $\mathbf{y} \in \mathbf{Y}$. Next, let μ^* be an optimal solution of $\mathbf{P}$, and let $\alpha \in \mathbb{N}^n$, $\beta \in \mathbb{N}^p$. Then

$$\begin{aligned}\int_\mathbf{K} \mathbf{x}^\alpha \mathbf{y}^\beta\, d\mu^*(\mathbf{x}, \mathbf{y}) &= \int_\mathbf{Y} \mathbf{y}^\beta \left(\int_{\mathbf{X}^*_\mathbf{y}} \mathbf{x}^\alpha\, \psi^*(d\mathbf{x} | \, \mathbf{y}) \right) d\varphi(\mathbf{y}) \\ &= \int_\mathbf{Y} \mathbf{y}^\beta\, g(\mathbf{y})^\alpha\, d\varphi(\mathbf{y}),\end{aligned}$$

the desired result. □

An optimal solution μ^* of $\mathbf{P}$ encodes *all* information on the optimal solutions $\mathbf{x}^*(\mathbf{y})$ of $\mathbf{P}_\mathbf{y}$. For instance, let $\mathbf{B}$ be a given Borel set of $\mathbb{R}^n$. Then from Theorem 2.2,

$$\text{Prob}\,(\mathbf{x}^*(\mathbf{y}) \in \mathbf{B}) \;=\; \mu^*(\mathbf{B} \times \mathbb{R}^p) \;=\; \int_\mathbf{Y} \psi^*(\mathbf{B}\,|\,\mathbf{y})\, d\varphi(\mathbf{y}),$$

with ψ^* as in Theorem 2.2(b).

Consequently, if one knows an optimal solution μ^* of $\mathbf{P}$ then one may evaluate functionals on the solutions of $\mathbf{P}_\mathbf{y}$, $\mathbf{y} \in \mathbf{Y}$. That is, assuming that for φ-almost all $\mathbf{y} \in \mathbf{Y}$, problem $\mathbf{P}_\mathbf{y}$ has a unique optimal solution $\mathbf{x}^*(\mathbf{y})$, and given a measurable mapping $h : \mathbb{R}^n \to \mathbb{R}^q$, one may evaluate the functional

$$\int_\mathbf{Y} h(\mathbf{x}^*(\mathbf{y}))\, d\varphi(\mathbf{y}).$$

For instance, with $\mathbf{x} \mapsto h(\mathbf{x}) := \mathbf{x}$ one obtains the *mean* vector $\mathrm{E}_\varphi(\mathbf{x}^*(\mathbf{y})) := \int_\mathbf{Y} \mathbf{x}^*(\mathbf{y}) d\varphi(\mathbf{y})$ of optimal solutions $\mathbf{x}^*(\mathbf{y})$, $\mathbf{y} \in \mathbf{Y}$.

Corollary 2.3. *Let both $\mathbf{Y} \subset \mathbb{R}^p$ and $\mathbf{K}$ in (2.2) be compact. Assume that for every $\mathbf{y} \in \mathbf{Y}$, the set $\mathbf{K}_\mathbf{y} \subset \mathbb{R}^n$ in (2.3) is nonempty, and for φ-almost all $\mathbf{y} \in \mathbf{Y}$, the set $\mathbf{X}^*_y := \{\mathbf{x} \in \mathbf{K}_\mathbf{y} \;:\; J(\mathbf{y}) = f(\mathbf{x}, \mathbf{y})\}$ is the singleton $\{\mathbf{x}^*(\mathbf{y})\}$. Then for every measurable mapping $h : \mathbb{R}^n \to \mathbb{R}^q$,*

$$\int_\mathbf{Y} h(\mathbf{x}^*(\mathbf{y}))\, d\varphi(\mathbf{y}) \;=\; \int_\mathbf{K} h(\mathbf{x})\, d\mu^*(\mathbf{x}, \mathbf{y}). \tag{2.8}$$

where μ^* is an optimal solution of $\mathbf{P}$.

Proof. By Theorem 2.2(c)

$$\int_\mathbf{K} h(\mathbf{x})\, d\mu^*(\mathbf{x}, \mathbf{y}) \;=\; \int_\mathbf{Y} \left[\int_{\mathbf{X}^*_\mathbf{y}} h(\mathbf{x}) \psi^*(d\mathbf{x}\,|\,\mathbf{y}) \right] d\varphi(\mathbf{y}) \;=\; \int_\mathbf{Y} h(\mathbf{x}^*(\mathbf{y}))\, d\varphi(\mathbf{y}).$$

□

Remark 2.4. If the set $\mathbf{X}^*(\mathbf{y})$ is not a singleton on some set with positive φ-measure then Theorem 2.2(c) now becomes: For each $\alpha \in \mathbb{N}^n$, there exists a measurable mapping $g_\alpha : \mathbf{Y} \to \mathbb{R}$ such that:

$$\int_\mathbf{K} \mathbf{x}^\alpha \mathbf{y}^\beta d\mu^* = \int_\mathbf{Y} \mathbf{y}^\beta \left(\int_{\mathbf{X}^*_\mathbf{y}} \mathbf{x}^\alpha \psi^*(d\mathbf{x}|\mathbf{y}) \right) d\mathbf{y} = \int_\mathbf{Y} \mathbf{y}^\beta\, g_\alpha(\mathbf{y})\, d\mathbf{y},$$

where

$$\mathbf{y} \mapsto g_\alpha(\mathbf{y}) = \int_{\mathbf{X}^*_\mathbf{y}} \mathbf{x}^\alpha \psi^*(d\mathbf{x}|\mathbf{y}) = \mathrm{E}\,[\mathbf{x}^\alpha\,|\,\mathbf{y}], \quad \mathbf{y} \in \mathbf{Y},$$

and $\mathrm{E}\,[\cdot\,|\,\mathbf{y}]$ denotes the conditional expectation operator associated with μ^*.

2.2. Duality. Consider the following infinite-dimensional linear program $\mathbf{P}^*$:

$$\begin{aligned} \rho^* \;:=\; \sup_{p \in \mathbb{R}[\mathbf{y}]} \quad & \int_\mathbf{Y} p\, d\varphi \\ & f(\mathbf{x}, \mathbf{y}) - p(\mathbf{y}) \;\geq\; 0 \quad \forall (\mathbf{x}, \mathbf{y}) \in \mathbf{K}. \end{aligned} \tag{2.9}$$

Then $\mathbf{P}^*$ is a *dual* of $\mathbf{P}$.

LEMMA 2.5. *Let both $\mathbf{Y} \subset \mathbb{R}^p$ and $\mathbf{K}$ in (2.2) be compact and let $\mathbf{P}$ and $\mathbf{P}^*$ be as in (2.4) and (2.9) respectively. Then there is no duality gap, i.e., $\rho = \rho^*$.*

Proof. For a topological space $\mathcal{X}$ denote by $C(\mathcal{X})$ the space of bounded continuous functions on $\mathcal{X}$. Let $\mathcal{M}(\mathbf{K})$ be the vector space of finite signed Borel measures on $\mathbf{K}$ (and so $\mathbf{M}(\mathbf{K})$ is its positive cone). Let $\pi : \mathcal{M}(\mathbf{K}) \to \mathcal{M}(\mathbf{Y})$ be defined by $(\pi\mu)(B) = \mu((\mathbb{R}^n \times B) \cap \mathbf{K})$ for all $B \in \mathcal{B}(\mathbf{Y})$, with adjoint mapping $\pi^* : C(\mathbf{Y}) \to C(\mathbf{K})$ defined as

$$(\mathbf{x}, \mathbf{y}) \mapsto (\pi^* h)(\mathbf{x}, \mathbf{y}) := h(\mathbf{y}), \qquad \forall h \in C(\mathbf{Y}).$$

Put (2.4) in the framework of infinite-dimensional linear programs on vector spaces, as described in e.g. [1]. That is:

$$\rho = \inf_{\mu \in \mathcal{M}(\mathbf{K})} \{\langle f, \mu\rangle \ : \ \pi\mu = \varphi, \ \mu \geq 0\},$$

with dual:

$$\tilde{\rho} = \sup_{h \in C(\mathbf{Y})} \{\langle h, \varphi\rangle \ : \ f - \pi^* h \geq 0 \quad \text{on } \mathbf{K}\}.$$

Endow $\mathcal{M}(\mathbf{K})$ (respectively $\mathcal{M}(\mathbf{Y})$) with the weak $\star$ topology $\sigma(\mathcal{M}(\mathbf{K}), C(\mathbf{K}))$ (respectively $\sigma(\mathcal{M}(\mathbf{Y}), C(\mathbf{Y}))$. One first proves that $\rho = \tilde{\rho}$ and then $\tilde{\rho} = \rho^*$.

By [1, Theor. 3.10], to get $\rho = \tilde{\rho}$, it suffices to prove that the set $D := \{(\pi\mu, \langle f, \mu\rangle) : \mu \in \mathbf{M}(\mathbf{K})\}$ is closed for the respective weak $\star$ topologies $\sigma(\mathcal{M}(\mathbf{Y}) \times \mathbb{R}, C(\mathbf{Y}) \times \mathbb{R})$ of $\mathbf{M}(\mathbf{Y}) \times \mathbb{R}$ and $\sigma(\mathcal{M}(\mathbf{K}), C(\mathbf{K}))$ of $\mathbf{M}(\mathbf{K})$. Therefore consider a converging sequence $\pi\mu_n \to a$ with $\mu_n \in \mathbf{M}(\mathbf{K})$. The sequence (μ_n) is uniformly bounded because

$$\mu_n(\mathbf{K}) = (\pi\mu_n)(\mathbf{Y}) \ = \ \langle 1, \pi\mu_n\rangle \ \to \ \langle 1, a\rangle \ = \ a(\mathbf{Y}).$$

But by the Banach-Alaoglu Theorem (see e.g. [3, Theor. 3.5.16]), the bounded closed sets of $\mathbf{M}(\mathbf{K})$ are compact in the weak $\star$ topology. And so $\mu_{n_k} \to \mu$ for some $\mu \in \mathbf{M}(\mathbf{K})$ and some subsequence (n_k). Next, observe that for $h \in C(\mathbf{Y})$ arbitrary,

$$\langle h, \pi\mu_{n_k}\rangle \ = \ \langle \pi^* h, \mu_{n_k}\rangle \ \to \ \langle \pi^* h, \mu\rangle \ = \ \langle h, \pi\mu\rangle,$$

where we have used that $\pi^* h \in C(\mathbf{K})$. Hence combining the above with $\pi\mu_{n_k} \to a$, we obtain $\pi\mu = a$. Similarly, $\langle f, \mu_{n_k}\rangle \to \langle f, \mu\rangle$ because $f \in C(\mathbf{K})$. Hence D is closed and the desired result $\rho = \tilde{\rho}$ follows.

We next prove that $\tilde{\rho} = \rho^*$. Given $\epsilon > 0$ fixed arbitrary, there is a function $h_\epsilon \in C(\mathbf{Y})$ such that $f - h_\epsilon \geq 0$ on $\mathbf{K}$ and $\int_{\mathbf{Y}} h_\epsilon d\varphi \geq \tilde{\rho} - \epsilon$. By compactness of $\mathbf{Y}$ and the Stone-Weierstrass theorem, there is $p_\epsilon \in \mathbb{R}[\mathbf{y}]$ such that $\sup_{\mathbf{y} \in \mathbf{Y}} |h_\epsilon(\mathbf{y}) - p_\epsilon(\mathbf{y})| \leq \epsilon$. Hence the polynomial $\tilde{p}_\epsilon := p_\epsilon - \epsilon$ is feasible with value $\int_{\mathbf{Y}} \tilde{p}_\epsilon d\varphi \geq \tilde{\rho} - 3\epsilon$, and as ϵ was arbitrary, the result $\tilde{\rho} = \rho^*$ follows. □

As next shown, optimal or nearly optimal solutions of $\mathbf{P}^*$ provide us with polynomial lower approximations of the optimal value function $\mathbf{y} \mapsto J(\mathbf{y})$ that converges to $J(\cdot)$ in the $L_1(\varphi)$ norm. Moreover, one may also obtain a piecewise polynomial approximation that converges to $J(\cdot)$ φ-almost uniformly. (Recall that a sequence of measurable functions (g_n) on a measure space $(\mathbf{Y}, \mathcal{B}(\mathbf{Y}), \varphi)$ converges to g *φ-almost uniformly* if and only if for every $\epsilon > 0$, there is a set $A \in \mathcal{B}(\mathbf{Y})$ such that $\varphi(A) < \epsilon$ and $g_n \to g$ uniformly on $\mathbf{Y} \setminus A$.)

COROLLARY 2.6. *Let both* $\mathbf{Y} \subset \mathbb{R}^p$ *and* $\mathbf{K}$ *in (2.2) be compact and assume that for every* $\mathbf{y} \in \mathbf{Y}$*, the set* $\mathbf{K}_\mathbf{y}$ *is nonempty. Let* $\mathbf{P}^*$ *be as in (2.9). If* $(p_i)_{i\in\mathbb{N}} \subset \mathbb{R}[\mathbf{y}]$ *is a maximizing sequence of (2.9) then*

$$\int_\mathbf{Y} |\, J(\mathbf{y}) - p_i(\mathbf{y})\,|\, d\varphi \;\to\; 0 \quad \text{as } i \to \infty. \tag{2.10}$$

Moreover, define the functions $(\tilde{p}_i)$ *as follows:*

$$\tilde{p}_0 := p_0, \quad \mathbf{y} \mapsto \tilde{p}_i(\mathbf{y}) \;:=\; \max[\,\tilde{p}_{i-1}(\mathbf{y}), p_i(\mathbf{y})\,], \quad i = 1, 2, \ldots$$

Then $\tilde{p}_i \to J(\cdot)$ φ*-almost uniformly.*

Proof. By Lemma 2.5, we already know that $\rho^* = \rho$ and so

$$\int_\mathbf{Y} p_i(\mathbf{y})\, d\varphi(\mathbf{y}) \;\uparrow\; \rho^* \;=\; \rho \;=\; \int_\mathbf{Y} J(\mathbf{y})\, d\varphi.$$

Next by feasibility of p_i in (2.9)

$$f(\mathbf{x},\mathbf{y}) \geq p_i(\mathbf{y}) \quad \forall (\mathbf{x},\mathbf{y}) \in \mathbf{K} \;\Rightarrow\; \inf_{\mathbf{x}\in\mathbf{K}_\mathbf{y}} f(\mathbf{x},\mathbf{y}) = J(\mathbf{y}) \;\geq\; p_i(\mathbf{y}) \quad \forall\, \mathbf{y} \in \mathbf{Y}.$$

Hence (2.10) follows from $p_i(\mathbf{y}) \leq J(\mathbf{y})$ on $\mathbf{Y}$.

With $\mathbf{y} \in \mathbf{Y}$ fixed, the sequence $(\tilde{p}_i(\mathbf{y}))_i$ is obviously monotone nondecreasing and bounded above by $J(\mathbf{y})$, hence with a limit $p^*(\mathbf{y}) \leq J(\mathbf{y})$. Therefore $\tilde{p}_i$ has the pointwise limit $\mathbf{y} \mapsto p^*(\mathbf{y}) \leq J(\mathbf{y})$. Also, by the Montone convergence theorem, $\int_\mathbf{Y} \tilde{p}_i(\mathbf{y}) d\varphi(\mathbf{y}) \to \int_\mathbf{Y} p^*(\mathbf{y}) d\varphi(\mathbf{y})$. This latter fact combined with (2.10) and $p_i(\mathbf{y}) \leq \tilde{p}_i(\mathbf{y}) \leq J(\mathbf{y})$ yields

$$0 \;=\; \int_\mathbf{Y} (J(\mathbf{y}) - p^*(\mathbf{y}))\, d\varphi(\mathbf{y}),$$

which in turn implies that $p^*(\mathbf{y}) = J(\mathbf{y})$ for φ-almost all $\mathbf{y} \in \mathbf{Y}$. Therefore $\tilde{p}_i(\mathbf{y}) \to J(\mathbf{y})$ for φ-almost all $\mathbf{y} \in \mathbf{Y}$. And so by Egoroff's Theorem [3, Theor. 2.5.5], $\tilde{p}_i \to J(\cdot)$, φ-almost uniformly. □

3. A hierarchy of semidefinite relaxations. In general, solving the infinite-dimensional problem $\mathbf{P}$ and getting an optimal solution μ^* is impossible. One possibility is to use numerical discretization schemes on a box containing $\mathbf{K}$; see for instance [14]. But in the present context of parametric optimization, if one selects finitely many *grid* points $(\mathbf{x},\mathbf{y}) \in \mathbf{K}$, one is implicitly considering solving (or rather approximating) $\mathbf{P}_\mathbf{y}$ for finitely many points $\mathbf{y}$ in a grid of $\mathbf{Y}$, which we want to avoid. To avoid this numerical discretization scheme we will use specific features of $\mathbf{P}$ when its data f (resp. $\mathbf{K}$) is a polynomial (resp. a compact basic semi-algebraic set).

Therefore in this section we are now considering a *polynomial* parametric optimization problem, a special case of (2.1) as we assume the following:

- $f \in \mathbb{R}[\mathbf{x},\mathbf{y}]$ and $h_j \in \mathbb{R}[\mathbf{x},\mathbf{y}]$, for every $j = 1, \ldots, m$.
- $\mathbf{K}$ is compact and $\mathbf{Y} \subset \mathbb{R}^p$ is a compact basic semi-algebraic set.

Hence the set $\mathbf{K} \subset \mathbb{R}^n \times \mathbb{R}^p$ in (2.2) is a compact basic semi-algebraic set. We also assume that there is a probability measure φ on $\mathbf{Y}$, absolutely continuous with respect to the Lebesgue measure, whose moments $\gamma = (\gamma_\beta)$, $\beta \in \mathbb{N}^p$, are available. As already mentioned, if $\mathbf{Y}$ is a simple set (like e.g. a simplex or a box) then one may choose φ to be the probability measure uniformly distributed on $\mathbf{Y}$, for which all moments can be computed easily. Sometimes, in the context of optimization with data uncertainty, the probability measure φ is already specified and in this case we assume that its moments $\gamma = (\gamma_\beta)$, $\beta \in \mathbb{N}^p$, are available.

3.1. Notation and preliminaries. Let $\mathbb{N}^n_i := \{\alpha \in \mathbb{N}^n : |\alpha| \leq i\}$ with $|\alpha| = \sum_i \alpha_i$. With a sequence $\mathbf{z} = (z_{\alpha\beta})$, $\alpha \in \mathbb{N}^n, \beta \in \mathbb{N}^p$, indexed in the canonical basis $(\mathbf{x}^\alpha\,\mathbf{y}^\beta)$ of $\mathbb{R}[\mathbf{x},\mathbf{y}]$, let $L_\mathbf{z} : \mathbb{R}[\mathbf{x},\mathbf{y}] \to \mathbb{R}$ be the linear mapping:

$$f\,(=\sum_{\alpha\beta} f_{\alpha\beta}(\mathbf{x},\mathbf{y})) \;\mapsto\; L_\mathbf{z}(f) \;:=\; \sum_{\alpha\beta} f_{\alpha\beta}\,z_{\alpha\beta}, \qquad f \in \mathbb{R}[\mathbf{x},\mathbf{y}].$$

Moment matrix. The moment matrix $\mathbf{M}_i(\mathbf{z})$ associated with a sequence $\mathbf{z} = (z_{\alpha\beta})$, has its rows and columns indexed in the canonical basis $(\mathbf{x}^\alpha\,\mathbf{y}^\beta)$, and with entries.

$$\mathbf{M}_i(\mathbf{z})(\alpha,\beta),(\delta,\gamma)) \;=\; L_\mathbf{z}(\mathbf{x}^\alpha\mathbf{y}^\beta\,\mathbf{x}^\delta\mathbf{y}^\gamma) \;=\; z_{(\alpha+\delta)(\beta+\gamma)},$$

for every $\alpha,\delta \in \mathbb{N}^n_i$ and every $\beta,\gamma \in \mathbb{N}^p_i$.

Localizing matrix. Let q be the polynomial $(\mathbf{x},\mathbf{y}) \mapsto q(\mathbf{x},\mathbf{y}) := \sum_{u,v} q_{uv}\mathbf{x}^u\mathbf{y}^v$. The localizing matrix $\mathbf{M}_i(q\,\mathbf{z})$ associated with $q \in \mathbb{R}[\mathbf{x},\mathbf{y}]$ and a sequence $\mathbf{z} = (z_{\alpha\beta})$, has its rows and columns indexed in the canonical basis $(\mathbf{x}^\alpha\,\mathbf{y}^\beta)$, and with entries.

$$\begin{aligned}\mathbf{M}_i(q\,\mathbf{z})(\alpha,\beta),(\delta,\gamma)) &= L_\mathbf{z}(q(\mathbf{x},\mathbf{y})\mathbf{x}^\alpha\mathbf{y}^\beta\,x^\delta\mathbf{y}^\gamma)\\ &= \sum_{u\in\mathbb{N}^n, v\in\mathbb{N}^p} q_{uv}\,z_{(\alpha+\delta+u)(\beta+\gamma+v)},\end{aligned}$$

for every $\alpha,\delta \in \mathbb{N}^n_i$ and every $\beta,\gamma \in \mathbb{N}^p_i$.

A sequence $\mathbf{z} = (z_{\alpha\beta}) \subset \mathbb{R}$ has a *representing* finite Borel measure supported on $\mathbf{K}$ if there exists a finite Borel measure μ such that

$$z_{\alpha\beta} \;=\; \int_\mathbf{K} \mathbf{x}^\alpha\,\mathbf{y}^\beta\,d\mu, \qquad \forall\,\alpha \in \mathbb{N}^n,\,\beta \in \mathbb{N}^p.$$

The next important result states a necssary and sufficient condition when $\mathbf{K}$ is compact and its defining polynomials $(h_k) \subset \mathbb{R}[\mathbf{x},\mathbf{y}]$ satisfy some condition.

ASSUMPTION 3.1. *Let $(h_j)_{j=1}^t \subset \mathbb{R}[\mathbf{x},\mathbf{y}]$ be a given family of polynomials. There is some N such that the quadratic polynomial $(\mathbf{x},\mathbf{y}) \mapsto N - \|(\mathbf{x},\mathbf{y})\|^2$ can be written*

$$N - \|(\mathbf{x},\mathbf{y})\|^2 \;=\; \sigma_0 + \sum_{j=1}^t \sigma_j\,h_j,$$

for some s.o.s. polynomials $(\sigma_j)_{j=0}^t \subset \Sigma[\mathbf{x},\mathbf{y}]$.

THEOREM 3.2. *Let $\mathbf{K} := \{(\mathbf{x},\mathbf{y}) \,:\, h_k(\mathbf{x},\mathbf{y}) \geq 0,\; k = 1,\ldots t\}$ and let $(h_k)_{k=1}^t$ satisfy Assumption 3.1. A sequence $\mathbf{z} = (z_{\alpha\beta})$ has a representing measure on $\mathbf{K}$ if and only if, for all $i = 0,1,\ldots$*

$$\mathbf{M}_i(\mathbf{z}) \succeq 0\,; \quad \mathbf{M}_i(h_k\,\mathbf{z}) \succeq 0, \quad k = 1,\ldots,t.$$

Theorem 3.2 is a direct consequence of Putinar's Positivstellensatz [22] and [25]. Of course, when Assumption 3.1 holds then $\mathbf{K}$ is compact. On the other hand, if $\mathbf{K}$ is compact and one knows a bound N for $\|(\mathbf{x},\mathbf{y})\|$ on $\mathbf{K}$ then its suffices to add the redundant quadratic constraint $h_{t+1}(\mathbf{x},\mathbf{y})(:= N^2 - \|(\mathbf{x},\mathbf{y})\|^2) \geq 0$ to the definition of $\mathbf{K}$, and Assumption 3.1 holds.

3.2. Semidefinite relaxations. To compute (or at least, approximate) the optimal value ρ of problem **P** in (2.4), we now provide a hierarchy of semidefinite relaxations in the spirit of those defined in [15].

Let $\mathbf{K} \subset \mathbb{R}^n \times \mathbb{R}^p$ be as in (2.2), and let $\mathbf{Y} \subset \mathbb{R}^p$ be the compact semi-algebraic set defined by:

$$\mathbf{Y} := \{\, \mathbf{y} \in \mathbb{R}^p \,:\, h_k(\mathbf{y}) \geq 0, \quad k = m+1, \ldots, t\} \tag{3.1}$$

for some polynomials $(h_k)_{k=m+1}^t \in \mathbb{R}[\mathbf{y}]$; let $v_k := \lceil (\deg h_k)/2 \rceil]$ for every $k = 1, \ldots, t$. Next, let $\gamma = (\gamma_\beta)$ with

$$\gamma_\beta \;=\; \int_{\mathbf{Y}} \mathbf{y}^\beta \, d\varphi(\mathbf{y}), \qquad \forall\, \beta \in \mathbb{N}^p,$$

be the moments of a probability measure φ on $\mathbf{Y}$, absolutely continuous with respect to the Lebesgue measure, and let $i_0 := \max[\lceil (\deg f)/2 \rceil, \max_k v_k]$. For $i \geq i_0$, consider the following semidefinite relaxations:

$$\begin{array}{rll} \rho_i = & \inf\limits_{\mathbf{z}} & L_{\mathbf{z}}(f) \\ & \text{s.t.} & \mathbf{M}_i(\mathbf{z}) \succeq 0 \\ & & \mathbf{M}_{i-v_j}(h_j\, \mathbf{z}) \succeq 0, \quad j = 1, \ldots, t \\ & & L_{\mathbf{z}}(\mathbf{y}^\beta) = \gamma_\beta, \quad \forall\, \beta \in \mathbb{N}^p_i. \end{array} \tag{3.2}$$

THEOREM 3.3. *Let $\mathbf{K}, \mathbf{Y}$ be as (2.2) and (3.1) respectively, and let $(h_k)_{k=1}^t$ satisfy Assumption 3.1. Assume that for every $\mathbf{y} \in \mathbf{Y}$ the set $\mathbf{K}_\mathbf{y}$ is nonempty and consider the semidefinite relaxations (3.2). Then:*

(a) *$\rho_i \uparrow \rho$ as $i \to \infty$.*

(b) *Let $\mathbf{z}^i$ be a nearly optimal solution of (3.2), e.g. such that $L_{\mathbf{z}^i}(f) \leq \rho_i + 1/i$, and let $g : \mathbf{Y} \to \mathbf{K}_\mathbf{y}$ be the measurable mapping in Theorem 2.2(c). If for φ-almost all $\mathbf{y} \in \mathbf{Y}$, $J(\mathbf{y})$ is attained at a unique optimal solution $\mathbf{x}^*(\mathbf{y})$, then:*

$$\lim_{i\to\infty} z^i_{\alpha\beta} \;=\; \int_{\mathbf{Y}} \mathbf{y}^\beta\, g(\mathbf{y})^\alpha \, d\varphi(\mathbf{y}), \qquad \forall\, \alpha \in \mathbb{N}^n,\, \beta \in \mathbb{N}^p. \tag{3.3}$$

In particular, for every $k = 1, \ldots, n$,

$$\lim_{i\to\infty} z^i_{e(k)\beta} \;=\; \int_{\mathbf{Y}} \mathbf{y}^\beta\, g_k(\mathbf{y})\, d\varphi(\mathbf{y}), \qquad \forall\, \beta \in \mathbb{N}^p, \tag{3.4}$$

where $e(k)_j = \delta_{j=k}$, $j = 1, \ldots, n$ (with $\delta_\cdot$ being the Kronecker symbol).

The proof is postponed to §4.2.

REMARK 3.4. Observe that if $\rho_i = +\infty$ for some index i in the hierarchy (and hence for all $i' \geq i$), then the set $\mathbf{K}_\mathbf{y}$ is empty for all $\mathbf{y}$ in some Borel set B of $\mathbf{Y}$ with $\varphi(B) > 0$. Conversely, one may prove that if $\mathbf{K}_\mathbf{y}$ is empty for all $\mathbf{y} \in B$ with $\varphi(B) > 0$, then necessarily $\rho_i = +\infty$ for all i sufficiently large. In other words, the hierarchy of semidefinite relaxations (3.2) may also provide a certificate of emptyness of $\mathbf{K}_y$ for some Borel set of $\mathbf{Y}$ with positive φ-measure.

3.3. The dual semidefinite relaxations. The dual of the semidefinite relaxation (3.2) reads:

$$\begin{array}{rl} \rho_i^* = & \sup_{p,(\sigma_i)} \int_{\mathbf{Y}} p\, d\varphi \quad (= \sum_{\beta\in\mathbb{N}^p_i} p_\beta\, \gamma_\beta) \\ \text{s.t.} & f - p = \sigma_0 + \sum_{j=1}^t \sigma_j\, h_j \\ & p \in \mathbb{R}[\mathbf{y}];\ \sigma_j \subset \Sigma[\mathbf{x},\mathbf{y}], \quad j = 1,\ldots,t \\ & \deg p \le 2i,\ \deg \sigma_j h_j \le 2i, \quad j = 1,\ldots,t \end{array} \tag{3.5}$$

Observe that (3.5) is a strenghtening of (2.9) as one restricts to polynomials $p \in \mathbb{R}[\mathbf{y}]$ of degree at most $2i$ and the nonnegativity of $f - p$ in (2.9) is replaced with the stronger weighted s.o.s. representation in (3.5). Therefore $\rho_i^* \le \rho^*$ for every i.

THEOREM 3.5. *Let $\mathbf{K}, \mathbf{Y}$ be as (2.2) and (3.1) respectively, and let $(h_k)_{k=1}^t$ satisfy Assumption 3.1. Assume that for every $\mathbf{y} \in \mathbf{Y}$ the set $\mathbf{K}_\mathbf{y}$ is nonempty, and consider the semidefinite relaxations (3.5). Then:*

(a) *$\rho_i^* \uparrow \rho$ as $i \to \infty$.*

(b) *Let $(p_i, (\sigma_j^i))$ be a nearly optimal solution of (3.5), e.g. such that $\int_{\mathbf{Y}} p_i d\varphi \ge \rho_i^* - 1/i$. Then $p_i \le J(\cdot)$ and*

$$\lim_{i\to\infty} \int_{\mathbf{Y}} |J(\mathbf{y}) - p_i(\mathbf{y})|\, d\varphi(\mathbf{y}) = 0 \tag{3.6}$$

Moreover if one defines

$$\tilde{p}_0 := p_0, \quad \mathbf{y} \mapsto \tilde{p}_i(\mathbf{y}) := \max[\,\tilde{p}_{i-1}(\mathbf{y}), p_i(\mathbf{y})\,], \quad i = 1,2,\ldots,$$

then $\tilde{p}_i \to J(\cdot)$ φ-almost uniformly on $\mathbf{Y}$.

Proof. Recall that by Lemma 2.5, $\rho = \rho^*$. Moreover let $(p_k) \subset \mathbb{R}[\mathbf{y}]$ be a maximizing sequence of (2.9) as in Corollary 2.6 with value $s_k := \int_{\mathbf{Y}} p_k d\varphi$, and let $p'_k := p_k - 1/k$ for every k so that $f - p'_k > 1/k$ on $\mathbf{K}$. By Theorem 3.2, there exist s.o.s. polynomials $(\sigma_j^k) \subset \Sigma[\mathbf{x},\mathbf{y}]$ such that $f - p'_k = \sigma_0^k + \sum_j \sigma_j^k h_j$. Letting d_k be the maximum degree of σ_0 and $\sigma_j h_j$, $j = 1,\ldots,t$, it follows that $(s_k - 1/k, (\sigma_j^k))$ is a feasible solution of (3.5) with $i := d_k$. Hence $\rho^* \ge \rho_{d_k}^* \ge s_k - 1/k$ and the result (a) follows because $s_k \to \rho^*$, and the sequence ρ_i^* is monotone. Then (b) follows from Corollary 2.6. □

Hence in Theorem 3.5, $p_i \in \mathbb{R}[\mathbf{y}]$ provides a polynomial lower approximation of the optimal value function $J(\cdot)$, with degree at most $2i$ (the order of the moments (γ_β) of φ taken into account in the semidefinite relaxation (3.5)). Moreover one may even define a piecewise polynomial lower approximation $\tilde{p}_i$ that converges φ-almost uniformly to $J(\cdot)$ on $\mathbf{Y}$.

Functionals of the optimal solutions. Theorem 3.3 provides a mean of approximating any polynomial functional on the global minimizers of $\mathbf{P}_\mathbf{y}$, $\mathbf{y} \in \mathbf{Y}$. Indeed,

COROLLARY 3.6. *Let $\mathbf{K}, \mathbf{Y}$ be as (2.2) and (3.1) respectively, and let $(h_k)_{k=1}^t$ satisfy Assumption 3.1. Assume that for every $\mathbf{y} \in \mathbf{Y}$ the set $\mathbf{K}_\mathbf{y}$ is nonempty, and for φ-almost all $\mathbf{y} \in \mathbf{Y}$, $J(\mathbf{y})$ is attained at a unique optimal solution $\mathbf{x}^*(\mathbf{y}) \in \mathbf{X}_\mathbf{y}^*$. Let*

$$\mathbf{x} \mapsto h(\mathbf{x}) := \sum_{\alpha\in\mathbb{N}^n} h_\alpha\, \mathbf{x}^\alpha,$$

and let $\mathbf{z}^i$ *be a nearly optimal solution of the semidefinite relaxations (3.2).*

Then, for i *sufficiently large,*

$$\int_{\mathbf{Y}} h(\mathbf{x}^*(\mathbf{y}))\, d\varphi(\mathbf{y}) \;\approx\; \sum_{\alpha\in\mathbb{N}^n} h_\alpha\, z^i_{\alpha 0}.$$

Proof. The proof is an immediate consequence of Theorem 3.3 and Corollary 2.3. □

3.4. Persistence for Boolean variables. One interesting and potentially useful application is in Boolean optimization. Indeed suppose that for some subset $I \subseteq \{1,\ldots,n\}$, the variables (x_i), $i \in I$, are boolean, that is, the definition of $\mathbf{K}$ in (2.2) includes the quadratic constraints $x_i^2 - x_i = 0$, for every $i \in I$.

Then for instance, one might be interested to determine whether in any optimal solution $\mathbf{x}^*(\mathbf{y})$ of $\mathbf{P}_\mathbf{y}$, and for some index $i \in I$, one has $x_i^*(\mathbf{y}) = 1$ (or $x_i^*(\mathbf{y}) = 0$) for φ-almost all values of the parameter $\mathbf{y} \in \mathbf{Y}$. In [4, 20] the probability that $x_k^*(\mathbf{y})$ is 1 is called the *persistency* of the boolean variable $x_k^*(\mathbf{y})$

COROLLARY 3.7. *Let* $\mathbf{K}, \mathbf{Y}$ *be as in (2.2) and (3.1) respectively. Let* $(h_k)_{k=1}^t$ *satisfy (3.1). Assume that for every* $\mathbf{y} \in \mathbf{Y}$ *the set* $\mathbf{K}_\mathbf{y}$ *is nonempty. Let* $\mathbf{z}^i$ *be a nearly optimal solution of the semidefinite relaxations (3.2). Then for* $k \in I$ *fixed:*

(a) $x_k^*(\mathbf{y}) = 1$ *in any optimal solution and for* φ*-almost all* $\mathbf{y} \in \mathbf{Y}$*, only if* $\lim_{i\to\infty} z^i_{e(k)0} = 1$.

(b) $x_k^*(\mathbf{y}) = 0$ *in any optimal solution and for* φ*-almost all* $\mathbf{y} \in \mathbf{Y}$*, only if* $\lim_{i\to\infty} z^i_{e(k)0} = 0$.

Assume that for φ*-almost all* $\mathbf{y} \in \mathbf{Y}$*,* $J(\mathbf{y})$ *is attained at a unique optimal solution* $\mathbf{x}^*(\mathbf{y}) \in \mathbf{X}^*_\mathbf{y}$*. Then* $\mathrm{Prob}\,(x_k^*(\mathbf{y}) = 1) = \lim_{i\to\infty} z^i_{e(k)0}$*, and so:*

(c) $x_k^*(\mathbf{y}) = 1$ *for* φ*-almost all* $\mathbf{y} \in \mathbf{Y}$*, if and only if* $\lim_{i\to\infty} z^i_{e(k)0} = 1$.

(d) $x_k^*(\mathbf{y}) = 0$ *for* φ*-almost all* $\mathbf{y} \in \mathbf{Y}$*, if and only if* $\lim_{i\to\infty} z^i_{e(k)0} = 0$.

Proof. (a) The *only if part.* Let $\alpha := e(k) \in \mathbb{N}^n$. From the proof of Theorem 3.3, for an arbitrary converging subsequence $(i_\ell)_\ell \subset (i)_i$ as in (4.5),

$$\lim_{l\to\infty} z^{i_\ell}_{e(k)0} \;=\; \int_{\mathbf{K}} x_k\, d\mu^*,$$

where μ^* is some optimal solution of $\mathbf{P}$. Hence, by Theorem 2.2(b), μ^* can be disintegrated into $\psi^*(d\mathbf{x}|\mathbf{y})d\varphi(\mathbf{y})$ where $\psi^*(d\mathbf{x}|\mathbf{y})$ is a probability measure on $\mathbf{X}^*_\mathbf{y}$ for every $\mathbf{y} \in \mathbf{Y}$. Therefore,

$$\begin{aligned}
\lim_{\ell\to\infty} z^{i_\ell}_{e(k)0} &= \int_{\mathbf{Y}} \left(\int_{\mathbf{X}^*_\mathbf{y}} x_k \psi^*(d\mathbf{x}\,|\,\mathbf{y}) \right) d\varphi(\mathbf{y}), \\
&= \int_{\mathbf{Y}} \psi^*(\mathbf{X}^*_\mathbf{y}\,|\,\mathbf{y})\, d\varphi(\mathbf{y}) \quad [\text{because } x_k^* = 1 \text{ in } \mathbf{X}^*_\mathbf{y}] \\
&= \int_{\mathbf{Y}} d\varphi(\mathbf{y}) \;=\; 1,
\end{aligned}$$

and as the converging subsequence $(i_\ell)_\ell$ in (4.5) was arbitrary, the whole sequence $(z^i_{e(k)0})$ converges to 1, the desired result. The proof of (b) being exactly the same is

omitted. Next, if for every $\mathbf{y} \in \mathbf{Y}$, $J(\mathbf{y})$ is attained at a singleton, then by Theorem 3.3(b),

$$\begin{aligned}\lim_{i\to\infty} z^i_{e(k)0} &= \int_{\mathbf{Y}} x^*_k(\mathbf{y})\, d\varphi(\mathbf{y}) \;=\; \varphi(\{\mathbf{y} \,:\, x^*_k(\mathbf{y}) = 1\}) \\ &= \mathrm{Prob}\,(x^*_k(\mathbf{y}) = 1),\end{aligned}$$

from which (c) and (d) follow. □

3.5. Estimating the density $g(\mathbf{y})$. By Corollary 3.6, one may approximate any polynomial functional of the optimal solutions, like for instance the mean, variance, etc. (with respect to the probability measure φ). However, one may also wish to approximate (in some sense) the function $\mathbf{y} \mapsto g_k(\mathbf{y})$, that is, the "curve" described by the k-th coordinate $\mathbf{x}^*_k(\mathbf{y})$ of the optimal solution $\mathbf{x}^*(\mathbf{y})$ when $\mathbf{y}$ varies in $\mathbf{Y}$.

So let $g : \mathbf{Y} \to \mathbb{R}^n$ be the measurable mapping in Theorem 3.3 and suppose that one knows some lower bound vector $\mathbf{a} = (a_k) \in \mathbb{R}^n$, where:

$$a_k \;\le\; \inf\{\, x_k \,:\, (\mathbf{x},\mathbf{y}) \,\in\, \mathbf{K}\,\}, \qquad k = 1,\ldots,n.$$

Then for every $k = 1,\ldots,n$, the measurable function $\hat{g}_k : \mathbf{Y} \to \mathbb{R}^n$ defined by

$$\mathbf{y} \;\mapsto\quad \hat{g}_k(\mathbf{y}) \;:=\; g_k(\mathbf{y}) - a_k, \qquad \mathbf{y} \in \mathbf{Y}, \tag{3.7}$$

is nonnegative and φ-integrable.

Hence for every $k = 1,\ldots,n$, one may consider $d\lambda := \hat{g}_k d\varphi$ as a Borel measure on $\mathbf{Y}$ with unknown density $\hat{g}_k$ with respect to φ, but with known moments $\mathbf{u} = (u_\beta)$. Indeed, using (3.4),

$$\begin{aligned}u_\beta \;:=\; \int_{\mathbf{Y}} \mathbf{y}^\beta\, d\lambda(\mathbf{y}) &= -a_k \int_{\mathbf{Y}} \mathbf{y}^\beta\, d\varphi(\mathbf{y}) + \int_{\mathbf{Y}} \mathbf{y}^\beta\, g_k(\mathbf{y})\, d\varphi(\mathbf{y}) \\ &= -a_k\gamma_\beta \;+\; z_{e(k)\beta}, \qquad \forall \beta \in \mathbb{N}^p,\end{aligned} \tag{3.8}$$

where for every $k = 1,\ldots,n$,

$$z_{e(k)\beta} \;=\; \lim_{i\to\infty} z^i_{e(k)\beta}, \qquad \forall \beta \in \mathbb{N}^n,$$

with $\mathbf{z}^i$ being an optimal (or nearly optimal) solution of the semidefinite relaxation (3.2).

Hence we are now faced with a density estimation problem, that is: Given the sequence of moments $u_\beta = \int_{\mathbf{Y}} \mathbf{y}^\beta g_k(\mathbf{y}) d\varphi$, $\beta \in \mathbb{N}^p$, of the unknown nonnegative measurable function $\hat{g}_k$ on $\mathbf{Y}$, "estimate" $\hat{g}_k$. One possibility is to use the so-called *maximum entropy* approach, briefly described in the next section.

Maximum-entropy estimation. We briefly describe the *maximum entropy* estimation technique in the univariate case. The multivariate case generalizes easily. Let $g \in L_1([0,1])$ be a nonnegative function[2] only known via the first $2d+1$ moments $\mathbf{u} = (u_j)_{j=0}^{2d}$ of its associated measure $d\varphi = g d\mathbf{x}$ on $[0,1]$. (In the context of previous section, the function g to estimate is $\mathbf{y} \mapsto \hat{g}_k(\mathbf{y})$ in (3.7) from the sequence $\mathbf{u}$ in (3.8) of its (multivariate) moments.)

[2] $L_1([0,1])$ denote the Banach space of integrable functions on the interval $[0,1]$ of the real line, equipped with the norm $\|g\|_1 = \int_0^1 |b(\mathbf{x})|\, d\mathbf{x}$.

From that partial knowledge one wishes (a) to provide an estimate h_d of g such that the first $2d+1$ moments of the measure $h_d d\mathbf{x}$ match those of $g d\mathbf{x}$, and (b) analyze the asymptotic behavior of h_d when $d \to \infty$. This problem has important applications in various areas of physics, engineering, and signal processing in particular. An elegant methodology is to search for h_d in a (finitely) parametrized family $\{h_d(\lambda, x)\}$ of functions, and optimize over the unknown parameters λ via a suitable criterion. For instance, one may wish to select an estimate h_d that maximizes some appropriate *entropy*. Several choices of entropy functional are possible as long as one obtains a convex optimization problem in the finitely many coefficients λ_i's. For more details the interested reader is referred to e.g. Borwein and Lewis [7, 8] and the many references therein.

We here choose the Boltzmann-Shannon entropy $\mathcal{H} : L_1([0,1]) \to \mathbb{R} \cup \{-\infty\}$:

$$h \mapsto \mathcal{H}[h] := -\int_0^1 h(x)\,\ln h(x)\,dx, \tag{3.9}$$

a strictly concave functional. Therefore, the problem reduces to:

$$\sup_h \left\{ \mathcal{H}[h] \;:\; \int_0^1 x^j\,h(x)\,dx \;=\; u_j, \quad j = 0,\ldots,2d \right\}. \tag{3.10}$$

The structure of this infinite-dimensional convex optimization problem permits to search for an optimal solution h_d^* of the form:

$$x \mapsto h_d^*(x) \;=\; \exp \sum_{j=0}^{2d} \lambda_j^*\,x^j, \tag{3.11}$$

and so λ^* is an optimal solution of the finite-dimensional unconstrained convex problem

$$\theta(\mathbf{u}) := \sup_\lambda \langle \mathbf{u}, \lambda \rangle - \int_0^1 \exp\left(\sum_{j=0}^{2d} \lambda_j x^j\right)\,dx.$$

Notice that the above function θ is just the Legendre-Fenchel transform of the convex function $\lambda \mapsto \int_0^1 \exp \sum_{j=0}^{2d} \lambda_j x^j\,dx$.

An optimal solution can be calculated by applying first-order methods, in which case the gradient ∇v_d of the function

$$\lambda \mapsto v_d(\lambda) := \langle \mathbf{u}, \lambda \rangle - \int_0^1 \exp\left(\sum_{j=0}^{2d} \lambda_j x^j\right)\,dx,$$

is provided by:

$$\frac{\partial v_d(\lambda)}{\partial \lambda_k} \;=\; u_k - \int_0^1 x^k\,\exp\left(\sum_{j=0}^{2d} \lambda_j x^j\right)\,d\mathbf{x}, \qquad k = 0,\ldots,2d.$$

If one applies second-order methods, e.g. Newton's method, then computing the Hessian $\nabla^2 v_d$ at current iterate λ, reduces to computing

$$\frac{\partial^2 v_d(\lambda)}{\partial \lambda_k \partial \lambda_j} \;=\; -\int_0^1 x^{k+j}\,\exp\left(\sum_{j=0}^{2d} \lambda_j x^j\right)\,dx, \qquad k,j = 0,\ldots,2d.$$

In such simple cases like a box $[a,b]$ (or $[a,b]^n$ in the multivariate case) such quantities can be approximated quite accurately via cubature formula as described in e.g. [11]. In particular, several cubature formula behave very well for exponentials of polynomials as shown in e.g. Bender et al. [6]. An alternative with no cubature formula is also proposed in [18]. One has the following convergence result which follows directly from [7, Theor. 1.7 and p. 259].

PROPOSITION 3.8. *Let $0 \le g \in L_1([0,1])$ and for every $d \in \mathbb{N}$, let h_d^* in (3.11) be an optimal solution of (3.10). Then, as $d \to \infty$,*

$$\int_0^1 \psi(x)\,(h_d^*(x) - g(x))\,dx \;\to\; 0,$$

for every bounded measurable function $\psi : [0,1] \to \mathbb{R}$ which is continuous φ-almost everywhere

Hence, the max-entropy estimate we obtain is not a pointwise estimate of g, and so, at some points of $[0,1]$ the max-entropy density h_d^* and the density g to estimate may differ significantly. However, for sufficiently large d, both curves of h_d^* and g are close to each other. For instance, in our context and with a one-dimensional parameter y in say $\mathbf{Y} = [0,1]$, recall that g is the mappping $y \mapsto x_k^*(y)$, and so in general, for fixed y, $h_d^*(y)$ is close to $x_k^*(y)$ and might be chosen for the k-coordinate of an initial point $\mathbf{x}$, input of a local minimization algorithm to find a local minimizer $\mathbf{x}^*(y)$ (with reasonable hope that it is a global minimizer).

3.6. LP-relaxations. Of course, in view of the present status of semidefinite programming solvers, the proposed approach is limited to small to medium size problems because the size of the semidefinite relaxations (3.2) grows like $O(n+p)^i$. On the other hand, if there is some structured sparsity pattern in the original problem, then one may use alternative sparse semidefinite relaxations as defined in e.g. [28], which increases significantly the size of problems that can be addressed by this technique. Alternatively, instead of the hierarchy of semidefinite relaxations (3.2), one may also use an analogous hierarchy of *LP-relaxations* as defined in [16], whose convergence is also guaranteed. In view of the status of LP solvers, and although the former relaxations have much better convergence properties, sometimes it may be better to use LP-relaxations, e.g., if one wishes to obtain only crude approximations but for larger size problems.

3.7. Illustrative examples. In this section we provide some simple illustrative examples. To show the potential of the approach we have voluntarily chosen very simple examples for which one knows the solutions exactly so as to compare the results we obtain with the exact optimal value and optimal solutions. The semidefinite relaxations (3.2) were implemented by using the software package Gloptipoly [12]. The max-entropy estimate h_d^* of g_k was computed by using Newton's method, where at each iterate $(\lambda^{(k)}, h_d(\lambda^{(k)}))$:

$$\lambda^{(k+1)} \;=\; \lambda^{(k)} - (\nabla^2 v_d(\lambda^{(k)}))^{-1} \nabla v_d(\lambda^{(k)}).$$

EXAMPLE 3.9. For illustration purpose, consider the toy example where $\mathbf{Y} := [0,1]$,

$$\mathbf{K} := \{(x,y) \,:\, 1 - x^2 - y^2 \ge 0;\, x,y \in \mathbf{Y}\} \subset \mathbb{R}^2, \quad (x,y) \mapsto f(x,y) := -x^2 y.$$

Hence for each value of the parameter $y \in \mathbf{Y}$, the unique optimal solution is $x^*(y) := \sqrt{1-y^2}$. And so in Theorem 3.3(b), $y \mapsto g(y) = \sqrt{1-y^2}$.

Let φ be the probability measure uniformly distributed on $[0,1]$. Therefore,

$$\rho = \int_0^1 J(y)\, d\varphi(y) \;=\; -\int_0^1 y(1-y^2)\, dy \;=\; -1/4.$$

Solving (3.2) with $i := 3$, that is, with moments up to order 6, one obtains the optimal value -0.250146. Solving (3.2) with $i := 4$, one obtains the optimal value -0.25001786 and the moment sequence

$$\mathbf{z} \;=\; (1, 0.7812, 0.5, 0.6604, 0.3334, 0.3333, 0.5813, 0.25, 0.1964, 0.25, 0.5244, 0.2, 0.1333,$$

$$0.1334, 0.2, 0.4810, 0.1667, 0.0981, 0.0833, 0.0983, 0.1667)$$

Observe that

$$z_{1k} - \int_0^1 y^k \sqrt{1-y^2}\, dy \approx O(10^{-6}), \qquad k = 0, \ldots 4,$$

$$z_{1k} - \int_0^1 y^k \sqrt{1-y^2}\, dy \approx O(10^{-5}), \qquad k = 5, 6, 7.$$

Using a max-entropy approach to approximate the density $y \mapsto g(y)$ on $[0,1]$, with the first 5 moments z_{1k}, $k = 0, \ldots, 4$, we find that the optimal function h_4^* in (3.11) is obtained with

$$\lambda^* = (-0.1564, 2.5316, -12.2194, 20.3835, -12.1867).$$

Both curves of g and h_4^* are displayed in Figure 3.1. Observe that with only 5 moments, the max-entropy solution h_4^* approximates g relatively well, even if it differs significantly at some points. Indeed, the shape of h_4^* resembles very much that of g.

Finally, from an optimal solution of (3.5) one obtains for $p \in \mathbb{R}[y]$, the degree-8 univariate polynomial

$$\begin{aligned} y \mapsto p(y) = &-0.0004 - 0.9909y - 0.0876y^2 + 1.4364y^3 - 1.2481y^4 \\ &+2.1261y^5 - 2.1309y^6 + 1.1593y^7 - 0.2641y^8 \end{aligned}$$

and Figure 3.2 displays the curve $y \mapsto J(y) - p(y)$ on $[0,1]$. One observes that $J \geq p$ and the maximum difference is about 3.10^{-4} close to 0 and much less for $y \geq 0.1$, a good precision with only 8 moments.

Example 3.10. Again with $\mathbf{Y} := [0,1]$, let

$$\mathbf{K} := \{(\mathbf{x}, y) \,:\, 1 - x_1^2 - x_2^2 \geq 0\} \subset \mathbb{R}^2, \quad (\mathbf{x}, y) \mapsto f(\mathbf{x}, y) := yx_1 + (1-y)x_2.$$

For each value of the parameter $y \in \mathbf{Y}$, the unique optimal solution $\mathbf{x}^* \in \mathbf{K}$ satisfies

$$(x_1^*(y))^2 + (x_2^*(y))^2 = 1; \quad (x_1^*(y))^2 = \frac{y^2}{y^2 + (1-y)^2}, \quad (x_2^*(y))^2 = \frac{(1-y)^2}{y^2 + (1-y)^2},$$

with optimal value

$$J(y) \;=\; -\frac{y^2}{\sqrt{y^2 + (1-y)^2}} - \frac{(1-y)^2}{\sqrt{y^2 + (1-y)^2}} \;=\; -\sqrt{y^2 + (1-y)^2}.$$

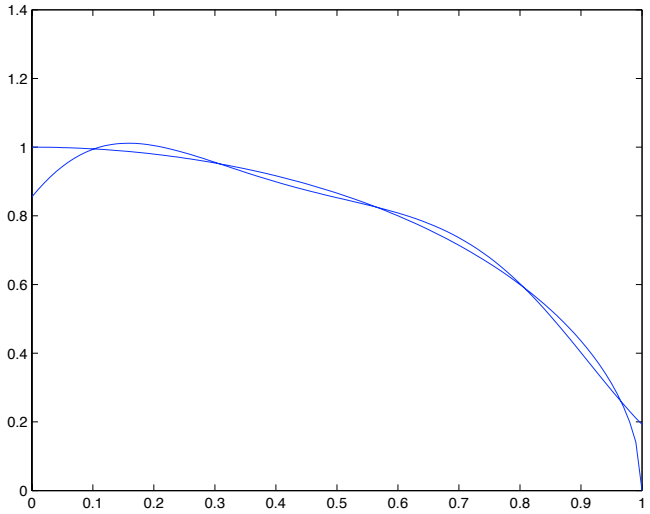

FIG. 3.1. *Example 3.9:* $g(y) = \sqrt{1-y^2}$ *versus* $h_4^*(y)$

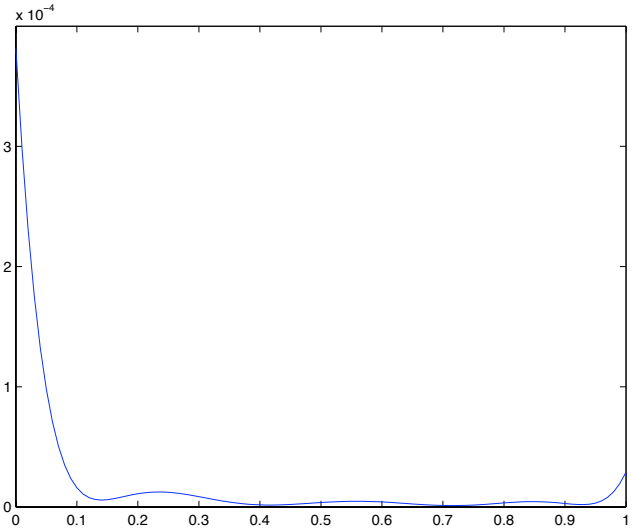

FIG. 3.2. *Example 3.9:* $J(y) - p(y)$ *on* $[0,1]$

So in Theorem 3.3(b),

$$y \mapsto g_1(y) = \frac{-y}{\sqrt{y^2+(1-y)^2}}, \quad y \mapsto g_2(y) = \frac{y-1}{\sqrt{y^2+(1-y)^2}},$$

and with φ being the probability measure uniformly distributed on $[0,1]$,

$$\rho = \int_0^1 J(y)\, d\varphi(y) \;=\; -\int_0^1 \sqrt{y^2+(1-y)^2}\, dy \;\approx\; -0.81162$$

Solving (3.2) with $i := 3$, that is, with moments up to order 6, one obtains $\rho_3 \approx -0.8117$ with $\rho_3 - \rho \approx O(10^{-5})$. Solving (3.2) with $i := 4$, one obtains $\rho_4 \approx -0.81162$ with $\rho_4 - \rho \approx O(10^{-6})$, and the moment sequence (z_{k10}), $k = 0,1,2,3,4$:

$$z_{k10} \;=\; (-0.6232,\, -0.4058,\, -0.2971,\, -0.2328,\, -0.1907),$$

and

$$z_{k10} - \int_0^1 y^k\, g_1(y)\, dy \;\approx\; O(10^{-5}), \quad k = 0,\ldots,4.$$

Using a max-entropy approach to approximate the density $y \mapsto -g_1(y)$ on $[0,1]$, with the first 5 moments z_{1k}, $k = 0,\ldots,4$, we find that the optimal function h_4^* in (3.11) is obtained with

$$\lambda^* = (-3.61284, 15.66153266 - 29.43090127.326347 - 9.9884452).$$

and we find that

$$z_{k10} \;+\; \int_0^1 y^k\, h_4^*(y)\, dy \;\approx\; O(10^{-11}), \quad k = 0,\ldots,4.$$

In Figure 3.3 are displayed the two functions $-g_1$ and h_4^*, and one observes a very good concordance.

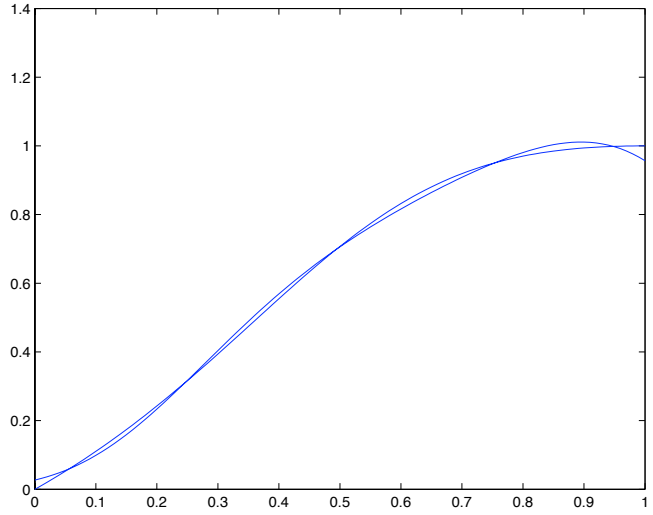

FIG. 3.3. *Example 3.10:* $h_4^*(y)$ *versus* $-g_1(y) = y/\sqrt{y^2+(1-y)^2}$

Finally, from an optimal solution of (3.5) one obtains for $p \in \mathbb{R}[y]$, the following degree-8 univariate polynomial

$$y \mapsto p(y) := -1.0000 + 0.9983y - 0.4537y^2 - 0.9941y^3 + 2.2488y^4 - 7.6739y^5 \\ +11.8448y^6 - 7.9606y^7 + 1.9903y^8$$

and Figure 3.4 displays the curve $y \mapsto J(y) - p(y)$ on $[0,1]$. One observes that $J \geq p$ and the maximum difference is about 10^{-4}, a good precision with only 8 moments.

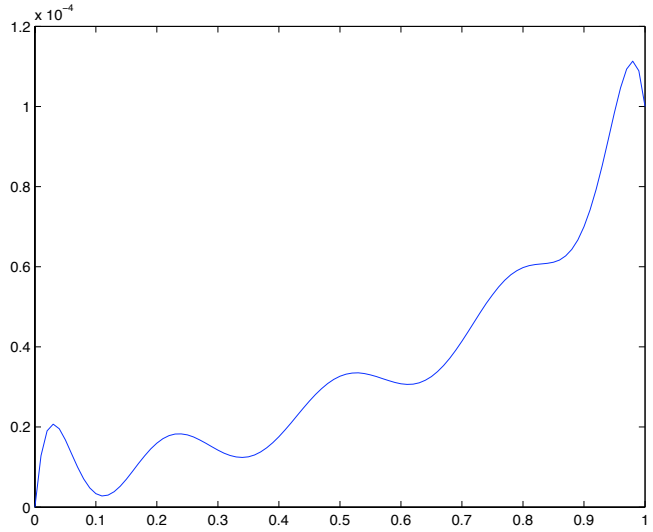

FIG. 3.4. *Example 3.10:* $J(y) - p(y)$ *on* $[0,1]$

EXAMPLE 3.11. In this example one has $\mathbf{Y} = [0,1]$, $(\mathbf{x}, y) \mapsto f(\mathbf{x}, y) := yx_1 + (1-y)x_2$, and

$$\mathbf{K} := \{(\mathbf{x}, y) \,:\, yx_1^2 + x_2^2 - y <= 0;\, x_1^2 + yx^2 - y <= 0\}.$$

That is, for each $y \in \mathbf{Y}$ the set $\mathbf{K}_y$ is the intersection of two ellipsoids. It is easy to chack that $1 + x_i^*(y) \geq 0$ for all $y \in \mathbf{Y}$, $i := 1, 2$. With $i = 4$, the max-entropy estimate $y \mapsto h_4^*(y)$ for $1 + x_1^*(y)$ is obtained with

$$\lambda^* = (-0.2894, 1.7192, -19.8381, 36.8285, -18.4828),$$

whereas the max-entropy estimate $y \mapsto h_4^*(y)$ for $1 + x_2^*(y)$ is obtained with

$$\lambda^* = (-0.1018, -3.0928, 4.4068, 1.7096, -7.5782).$$

Figure 3.5 displays the curves of $x_1^*(y)$ and $x_2^*(y)$, as well as the constraint $h_1(\mathbf{x}^*(y), y)$. Observe that $h_1(\mathbf{x}^*(y), y) \approx 0$ on $[0,1]$ which means that for φ-almost all $y \in [0,1]$, at an optimal solution $\mathbf{x}^*(y)$, the constraint $h_1 \leq 0$ is saturated. Figure 3.6 displays the curves of $h_1(\mathbf{x}^*(y), y)$ and $h_2(\mathbf{x}^*(y), y)$.

EXAMPLE 3.12. This time $\mathbf{Y} = [0,1]$, $(\mathbf{x}, y) \mapsto f(\mathbf{x}, y) := (1-2y)(x_1 + x_2)$, and

$$\mathbf{K} := \{(\mathbf{x}, y) \,:\, yx_1^2 + x_2^2 - y = 0;\, x_1^2 + yx^2 - y = 0\}.$$

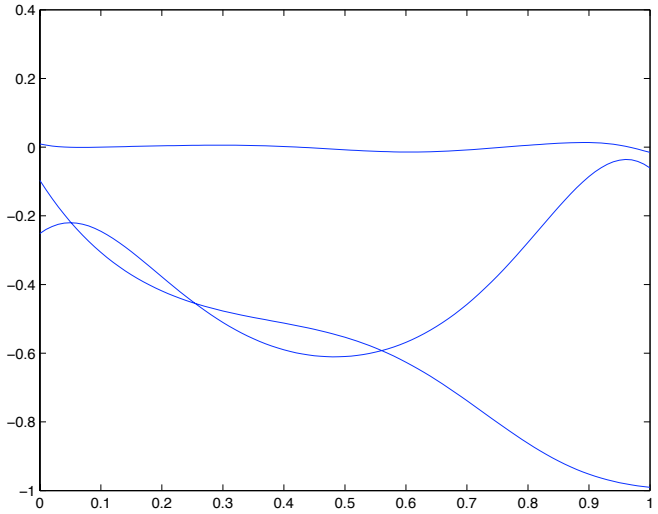

FIG. 3.5. *Example 3.11:* $x_1^*(y)$, $x_2^*(y)$ *and* $h_1(\mathbf{x}^*(y), y)$ *on* $[0,1]$

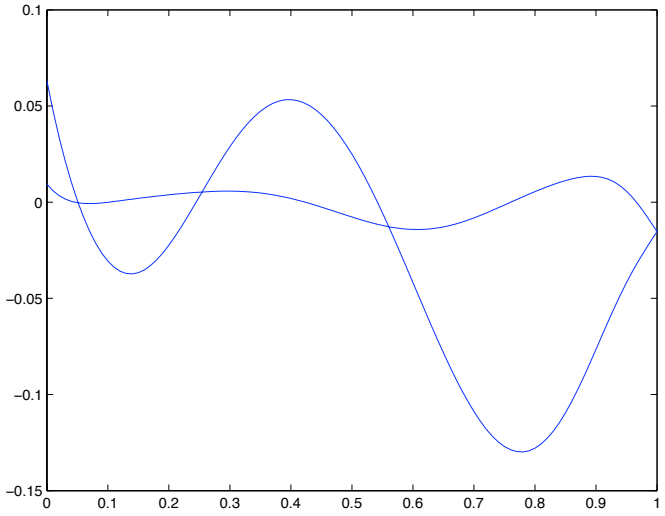

FIG. 3.6. *Example 3.11:* $h_1(\mathbf{x}^*(y), y)$ *and* $h_2(\mathbf{x}^*(y), y)$ *on* $[0,1]$

That is, for each $y \in \mathbf{Y}$ the set $\mathbf{K}_y$ is the intersection of two ellipses, and

$$\mathbf{x} = \left(\pm\sqrt{\frac{y}{1+y}}, \pm\sqrt{\frac{y}{1+y}}\right); \quad J(y) = -2|1-2y|\sqrt{\frac{y}{1+y}}.$$

With $i = 4$ the max-entropy estimate $y \mapsto h_4^*(y)$ for $1 + x_1^*(y)$ is obtained with

$$\lambda^* = (0.3071151, -12.51867, 43.215907, -46.985733, 16.395944).$$

In Figure 3.7 are displayed the curves $y \mapsto -p(y)$ and $y \mapsto -J(y)$, whereas in Figure 3.8 is displayed the curve $y \mapsto p(y) - J(y)$. One may see that p is a good lower approximation of J even with only 8 moments.

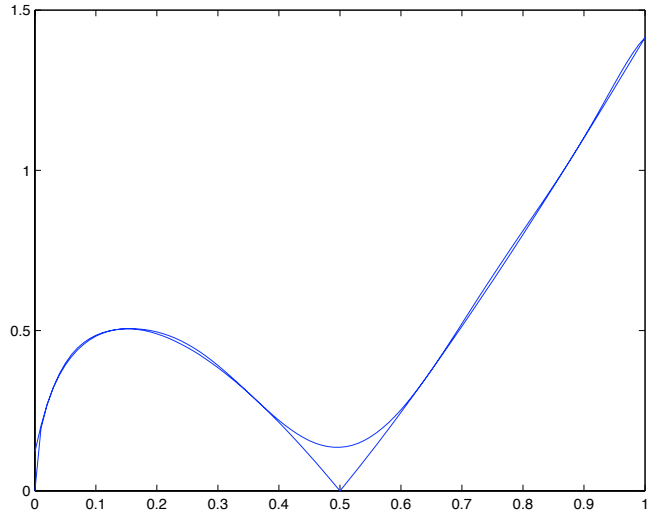

FIG. 3.7. *Example 3.12: $-p(y)$ and $-J(y)$ on $[0,1]$*

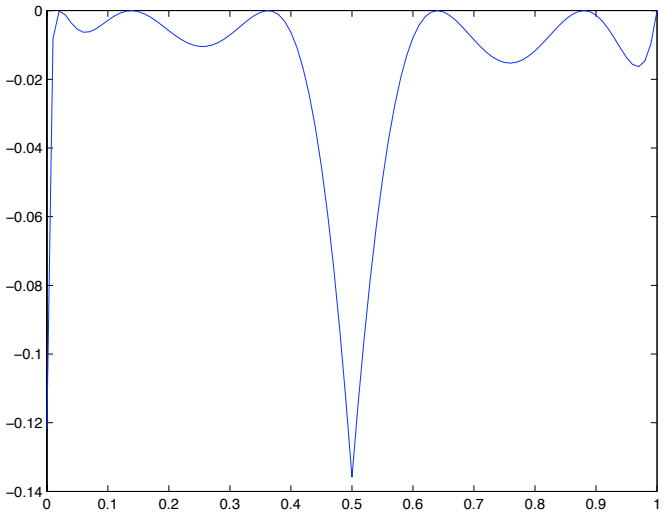

FIG. 3.8. *Example 3.12: the curve $p(y) - J(y)$ on $[0,1]$*

On the other hand, in Figure 3.9 is displayed $h_4^*(y)$ versus $x_1^*(y)$ where the latter is $-\sqrt{y/(1+y)}$ on $[0,1/2]$ and $\sqrt{y/(1+y)}$ on $[1/2,1]$. Here we see that the discontinuity

	y=0.1	y=0.5	y=1
$J(y)$	0.0400	1.0000	2.0000
$p_6(y)$	0.0384	0.9264	1.9887
$p_8(y)$	0.0390	0.9395	2.0000

TABLE 3.1
$J(y)$ versus $p_k(y)$ in Example 3.13

of $x_1^*(y)$ is difficult to approximate "pointwise" with few moments, and despite a very good precision on the five first moments. Indeed:

$$\int_0^1 y^k\,(h_4^*(y)-1)\,dx - \int_0^1 y^k\,x_1^*(y)\,dx \;=\; O(10^{-14}), \quad k=0,\ldots,4.$$

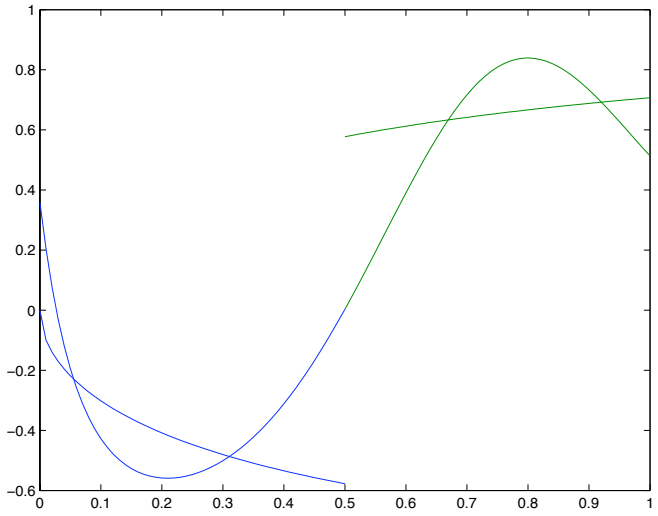

FIG. 3.9. *Example 3.12: $h_4^*(y)-1$ and $x_1^*(y)$ on $[0,1]$*

EXAMPLE 3.13. Consider the following system of 4 quadratic equations in 4 variables and one parameter $y \in \mathbf{Y} = [0,1]$:

$$\begin{aligned} x_1x_2 - x_1x_3 - x_4 = yx_2x_3 - x_2x_4 - x_1 = y \\ -x_1x_3 + x_3x_4 - x_2 = yx_1x_4 - x_2x_4 - x_3 = y, \end{aligned}$$

for which one wishes to compute the minimum norm $J(y)$ of real solutions as a function of $y \in \mathbf{Y}$. We have computed the polynomial $y \mapsto p_k(y)$ from (3.5) with degree $k=6$ and $k=8$, with φ being the probability with uniform distribution on $[0,1]$. The results displayed in Table 1 show that a pretty good approximation is obtained with few moments of φ.

We end up this section with the case where the density g_k to estimate is a step function which would be the case in an optimization problem $\mathbf{P}_y$ with boolean variables (e.g. the variable x_k takes values in $\{0,1\}$).

EXAMPLE 3.14. Assume that with a single parameter $y \in [0,1]$, the density g_k to estimate is the step function.

$$y \mapsto g_k(y) := \begin{cases} 1 & \text{if } y \in [0,1/3] \cup [2/3,1] \\ 0 & \text{otherwise.} \end{cases}$$

The max-entropy estimate h_4^* in (3.11) with 5 moments is obtained with

$$\lambda^* = [-0.6547367219.170724 - 115.39354192.4493171655 - 96.226948865],$$

and we have

$$\int_0^1 y^k h_4^*(y)\, dy - \int_0^1 y^k\, dg_k(y) \;\approx\; O(10^{-8}), \quad k = 0, \ldots, 4.$$

In particular, the persistency $\int_0^1 g_k(y)dy = 2/3$ of the variable $x_k^*(y)$, is very well approximated (up to 10^{-8} precision) by $\int h_4^*(y)dy$, with only 5 moments.

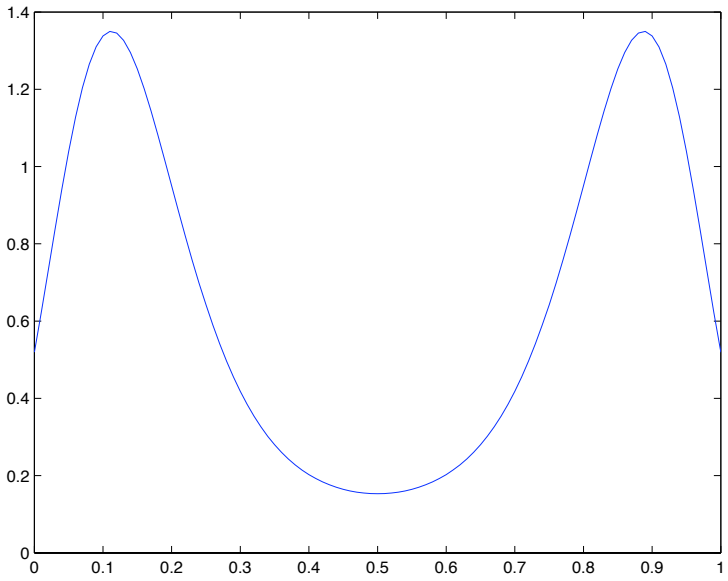

FIG. 3.10. *Example 3.14: Max-entropy estimate* $h_4^*(y)$ *of* $1_{[0,1/3]\cup[2/3,1]}$

Of course, in this case and with only 5 moments, the density h_4^* is not a good pointwise approximation of the step function $g_k(y) = 1_{[0,1/3]\cup[2/3,1]}$; however its "shape" in Figure 3.10 reveals the two steps of value 1 separated by a step of value 0. A better pointwise approximation would require more moments.

4. Appendix.

4.1. Set-valued functions and measurable selectors. The material of this subsection is taken from [13, D]. Let $X \subset \mathbb{R}^p$ and $Y \subset \mathbb{R}^n$ be Borel spaces. A set-valued mapping $\psi : X \to Y$ is a function such that $\psi(\mathbf{x})$ is a nonempty subset of Y for all $\mathbf{x} \in X$. The graph $\operatorname{Gr}\psi$ of the set-valued function ψ is the subset of $X \times Y$ defined by:

$$\operatorname{Gr}\psi \;:=\; \{\,(\mathbf{x},\mathbf{y}) \;:\; \mathbf{x} \in X,\; \mathbf{y} \in \psi(\mathbf{x})\,\}.$$

DEFINITION 4.1. *A set-valued function $\psi : X \to Y$ is said to be:*
(a) Borel-measurable if $\psi^{-1}(B)$ is a Borel subset of X for every open set $B \subset Y$.
(b) compact-valued if $\psi(\mathbf{x})$ is compact for every $\mathbf{x} \in X$.
(c) closed if its graph $\operatorname{Gr}\psi$ is closed.

DEFINITION 4.2. *Let $\psi : X \to Y$ be a Borel-measurable set-valued function, and let $\mathcal{F}$ be the set of all measurable functions $f : X \to Y$ with $f(\mathbf{x}) \in \psi(\mathbf{x})$ for all $\mathbf{x} \in X$. A function $f \in \mathcal{F}$ is called a measurable selector.* Let $v : \operatorname{Gr}\psi :\to \mathbb{R}$ be a measurable function and let $v^* : X \to \mathbb{R}$ be defined by:

$$v^*(\mathbf{x}) \;:=\; \inf_{\mathbf{y}}\{\, v(\mathbf{x},\mathbf{y}) \;:\; \mathbf{y} \in \psi(\mathbf{x})\,\}, \qquad \mathbf{x} \in X.$$

PROPOSITION 4.3. *([13, D.4]) Let $\psi : X \to Y$ be compact-valued. Then the following are equivalent:*
(a) *ψ is Borel-measurable.*
(b) *$\psi^{-1}(F)$ is a Borel subset of X for every closed set $F \subset Y$.*
(c) *$\operatorname{Gr}\psi$ is a Borel subset of $X \times Y$.*
(d) *ψ is a measurable function from X to the space of nonempty compact subsets of Y, topologized by the Hausdorff metric.*

PROPOSITION 4.4. *([13, D.5]) Suppose that $\psi : X \to Y$ is compact-valued and the function $\mathbf{y} \to v(\mathbf{x},\mathbf{y})$ is lower-semicontinuous on $\psi(\mathbf{x})$ for every $\mathbf{x} \in X$. Then there exists a measurable selector $f \in \mathcal{F}$ such that:*

$$v(\mathbf{x}, f(\mathbf{x})) \;=\; v^*(\mathbf{x}) \;=\; \min_{\mathbf{y}}\{v(\mathbf{x},\mathbf{y}) \;:\; \mathbf{y} \in \psi(\mathbf{x})\,\}, \qquad \forall \mathbf{x} \in X.$$

DEFINITION 4.5. *For two Borel spaces X, Y, a stochastic kernel on Y given X is a function $P(\cdot\,|\,\cdot)$ such that:*
(a) *$P(\cdot\,|\,\mathbf{x})$ is a probability measure on Y for each fixed $\mathbf{x} \in X$.*
(b) *$P(B\,|\,\cdot)$ is a measurable function on X for each fixed Borel subset B of Y.*

Let $\psi : X \to Y$ be a Borel-measurable set-valued function such that $\mathcal{F}$ is nonempty (equivalently, there is a measurable function $f : X \to Y$ whose graph is contained in $\operatorname{Gr}\psi$). Let Φ be the class of stochastic kernels φ on Y given Y such that $\varphi(\psi(\mathbf{x})\,|\,\mathbf{x}) = 1$ for every $\mathbf{x} \in X$. Finally, for a probability measure μ on $X \times Y$, denote by μ_1 the *marginal* (or projection) of μ on X, i.e.,

$$\mu_1(B) \;:=\; \mu(B \times Y), \qquad \forall B \in \mathcal{B}(X),$$

where $\mathcal{B}(X)$ is the Borel σ-field associated with X.

PROPOSITION 4.6. *([13, D.8]) If μ is a probability measure on $X \times Y$, concentrated on the graph $\operatorname{Gr}\psi$ of ψ, then there exists a stochastic kernel $\varphi \in \Phi$ such that*

$$\mu(B \times C) \;=\; \int_B \varphi(C\,|\,\mathbf{x})\,\mu_1(d\mathbf{x}), \qquad \forall B \in \mathcal{B}(X),\, C \in \mathcal{B}(Y).$$

See also [10, p. 88–89].

4.2. Proof of Theorem 3.3.. We already know that $\rho_i \le \rho$ for all $i \ge i_0$. We also need to prove that $\rho_i > -\infty$ for sufficiently large i. Let $Q \subset \mathbb{R}[\mathbf{x},\mathbf{y}]$ be the *quadratic module* generated by the polynomials $\{h_j\} \subset \mathbb{R}[\mathbf{x},\mathbf{y}]$ that define $\mathbf{K}$, i.e.,

$$Q := \{\, \sigma \in \mathbb{R}[\mathbf{x},\mathbf{y}] \,:\quad \sigma \,=\, \sigma_0 + \sum_{j=1}^{t} \sigma_j\, h_j \quad \text{with } \{\sigma_j\}_{j=0}^{t} \subset \Sigma[\mathbf{x},\mathbf{y}]\}.$$

In addition, let $Q(l) \subset Q$ be the set of elements $\sigma \in Q$ which have a representation $\sigma_0 + \sum_{j=0}^{t} \sigma_j\, h_j$ for some s.o.s. family $\{\sigma_j\} \subset \Sigma^2$ with $\deg \sigma_0 \le 2l$ and $\deg \sigma_j h_j \le 2l$ for all $j = 1,\ldots,t$.

Let $i \in \mathbb{N}$ be fixed. As $\mathbf{K}$ is compact, there exists N such that $N \pm \mathbf{x}^\alpha \mathbf{y}^\beta > 0$ on $\mathbf{K}$, for all $\alpha \in \mathbb{N}^n$ and $\beta \in \mathbb{N}^p$, with $|\alpha+\beta| \le 2i$. Therefore, under Assumption 3.1(ii), the polynomial $N \pm \mathbf{x}^\alpha \mathbf{y}^\beta$ belongs to Q; see Putinar [22]. But there is even some $l(i)$ such that $N \pm \mathbf{x}^\alpha \mathbf{y}^\beta \in Q(l(i))$ for every $|\alpha+\beta| \le 2i$. Of course we also have $N \pm \mathbf{x}^\alpha \mathbf{y}^\beta \in Q(l)$ for every $|\alpha+\beta| \le 2i$, whenever $l \ge l(i)$. Therefore, let us take $l(i) \ge i_0$. For every feasible solution $\mathbf{z}$ of $\mathbf{Q}_{l(i)}$ one has

$$|z_{\alpha\beta}| \,=\, |\, L_{\mathbf{z}}(\mathbf{x}^\alpha \mathbf{y}^\beta)\,| \le N, \qquad \forall\, |\alpha+\beta| \le 2i.$$

This follows from $z_0 = 1$, $\mathbf{M}_{l(i)}(\mathbf{z}) \succeq 0$ and $\mathbf{M}_{l(i)-v_j}(h_j\,\mathbf{z}) \succeq 0$, which implies

$$N z_0 \pm z_{\alpha\beta} \,=\, L_{\mathbf{z}}(N \pm \mathbf{x}^\alpha \mathbf{y}^\beta) \,=\, L_{\mathbf{z}}(\sigma_0) + \sum_{j=1}^{t} L_{\mathbf{z}}(\sigma_j\, h_j) \ge 0$$

for some $\{\sigma_j\} \subset \Sigma[\mathbf{x},\mathbf{y}]$ with $\deg \sigma_j\, h_j \le 2l(i)$. In particular, $L_{\mathbf{z}}(f) \,\ge\, -N \sum_{\alpha,\beta} |f_{\alpha\beta}|$, which proves that $\rho_{l(i)} > -\infty$, and so $\rho_i > -\infty$ for all sufficiently large i.

From what precedes, and with $k \in \mathbb{N}$ arbitrary, let $l(k) \ge k$ and N_k be such that

$$N_k \pm \mathbf{x}^\alpha \mathbf{y}^\beta \in Q(l(k)) \qquad \forall\, \alpha \,\in \mathbb{N}^n,\, \beta \in \mathbb{N}^p \;\text{ with } |\alpha+\beta| \le 2k. \tag{4.1}$$

Let $i \ge l(i_0)$, and let $\mathbf{z}^i$ be a nearly optimal solution of (3.2) with value

$$\rho_i \,\le\, L_{\mathbf{z}^i}(f) \,\le\, \rho_i + \frac{1}{i} \quad \left(\le \rho + \frac{1}{i}\right). \tag{4.2}$$

Fix $k \in \mathbb{N}$. Notice that from (4.1), for every $i \ge l(k)$, one has

$$|\, L_{\mathbf{z}^i}(\mathbf{x}^\alpha \mathbf{y}^\beta)\,| \,\le\, N_k z_0 = N_k, \quad \forall\, \alpha \in \mathbb{N}^n, \beta \in \mathbb{N}^p \;\text{ with } |\alpha+\beta| \,\le\, 2k.$$

Therefore, for all $i \ge l(i_0)$,

$$|z^i_{\alpha\beta}| \,=\, |\, L_{\mathbf{z}^i}(\mathbf{x}^\alpha \mathbf{y}^\beta)\,| \,\le\, N'_k, \quad \forall\, \alpha \in \mathbb{N}^n, \beta \in \mathbb{N}^p \;\text{ with } |\alpha+\beta| \,\le\, 2k, \tag{4.3}$$

where $N'_k = \max[N_k, V_k]$, with

$$V_k := \max_{\alpha,\beta,i} \{\, |z^i_{\alpha\beta}| \,:\quad |\alpha+\beta| \le 2k\,; \quad l(i_0) \le i \le l(k)\,\}.$$

Complete each vector $\mathbf{z}^i$ with zeros to make it an infinite bounded sequence in l_∞, indexed in the canonical basis $(\mathbf{x}^\alpha \mathbf{y}^\beta)$ of $\mathbb{R}[\mathbf{x},\mathbf{y}]$. In view of (4.3),

$$|z^i_{\alpha\beta}| \,\le N'_k \qquad \forall\, \alpha \in \mathbb{N}^n, \beta \in \mathbb{N}^p \;\text{ with } \quad 2k-1 \le |\alpha+\beta| \le 2k, \tag{4.4}$$

and for all $k = 1, 2, \ldots$.

Hence, let $\widehat{\mathbf{z}}^i \in l_\infty$ be the new sequence defined by

$$\widehat{z}^i_{\alpha\beta} := \frac{z^i_{\alpha\beta}}{N'_k}, \quad \forall\, \alpha \in \mathbb{N}^n, \beta \in \mathbb{N}^p \text{ with } \quad 2k-1 \le |\alpha+\beta| \le 2k, \quad \forall\, k = 1, 2, \ldots,$$

and in l_∞, consider the sequence $\{\widehat{\mathbf{z}}^i\}_i$, as $i \to \infty$.

Obviously, the sequence $\{\widehat{\mathbf{z}}^i\}_i$ is in the unit ball B_1 of ℓ_∞ which by the Banach-Alaoglu theorem (see e.g. Ash [3, Theor. 3.5.16]), is compact (and even sequentially compact) for the weak $\star$ topology $\sigma(\ell_\infty, \ell_1)$. Hence, there exists some $\widehat{\mathbf{z}} \in B_1$ and a subsequence i_ℓ, such that $\widehat{\mathbf{z}}^{i_\ell} \to \widehat{\mathbf{z}}$ as $l \to \infty$, for the weak $\star$ topology $\sigma(\ell_\infty, \ell_1)$ of ℓ_∞. So consider an arbitrary such converging subsequence. In particular, pointwise convergence holds, that is,

$$\lim_{\ell\to\infty} \widehat{z}^{i_\ell}_{\alpha\beta} \;\to\; \widehat{z}_{\alpha\beta} \qquad \forall\, \alpha \in \mathbb{N}^n, \beta \in \mathbb{N}^p.$$

Next, define

$$z_{\alpha\beta} := \widehat{z}_{\alpha\beta} \times N'_k \quad \forall\, \alpha \in \mathbb{N}^n, \beta \in \mathbb{N}^p \text{ with } \quad 2k-1 \le |\alpha+\beta| \le 2k, \quad \forall\, k = 1, 2, \ldots$$

The pointwise convergence $\widehat{\mathbf{z}}^{i_\ell} \to \widehat{\mathbf{z}}$ implies the pointwise convergence $\mathbf{z}^{i_\ell} \to \mathbf{z}$, i.e.,

$$\lim_{\ell\to\infty} z^{i_\ell}_{\alpha\beta} \;\to\; z_{\alpha\beta} \qquad \forall\, \alpha \in \mathbb{N}^n, \beta \in \mathbb{N}^p. \tag{4.5}$$

Next, let $s \in \mathbb{N}$ be fixed. From the pointwise convergence (4.5) we deduce that

$$\lim_{\ell\to\infty} \mathbf{M}_s(\mathbf{z}^{i_\ell}) \;=\; \mathbf{M}_s(\mathbf{z}) \;\succeq 0.$$

Similarly

$$\lim_{\ell\to\infty} \mathbf{M}_s(h_j\, \mathbf{z}^{i_\ell}) \;=\; \mathbf{M}_s(h_j\, \mathbf{z}) \;\succeq 0, \quad j = 1, \ldots, t.$$

As s was arbitrary, we obtain

$$\mathbf{M}_s(\mathbf{y}) \succeq 0; \quad \mathbf{M}_s(h_j\, \mathbf{z}) \succeq 0, \quad j = 1, \ldots, t; \quad s = 0, 1, 2, \ldots, \tag{4.6}$$

which by Theorem 3.2 implies that $\mathbf{z}$ is the sequence of moments of some finite measure μ^* with support contained in $\mathbf{K}$. Moreover, the pointwise convergence (4.5) also implies that

$$\int_{\mathbf{Y}} \mathbf{y}^\beta\, d\varphi(\mathbf{y}) \;=\; \gamma_\beta \;=\; \lim_{\ell\to\infty} z^{i_\ell}_{0\beta} \;=\; z_{0\beta} \;=\; \int_{\mathbf{K}} \mathbf{y}^\beta\, d\mu^*, \quad \forall \beta \in \mathbb{N}^p. \tag{4.7}$$

As measures on compacts sets are determinate, (4.7) implies that the marginal of μ^* on $\mathbb{R}^p$ is the probability measure φ, and so μ^* is feasible for $\mathbf{P}$. Finally, combining the pointwise convergence (4.5) with (4.2) yields

$$\rho \;\ge\; \lim_{l\to\infty} \rho_{i_\ell} \;=\; \lim_{i\to\infty} L_{\mathbf{z}^{i_\ell}}(f) \;=\; L_{\mathbf{z}}(f) \;=\; \int_{\mathbf{K}} f\, d\mu^*,$$

which in turn yields that μ^* is an optimal solution of $\mathbf{P}$. And so $\rho_{i_\ell} \to \rho$ as $l \to \infty$. As the sequence (ρ_i) is monotone this yields the desired result (a).

(b) Next, let $\alpha \in \mathbb{N}^n$ and $\beta \in \mathbb{N}^p$ be fixed, arbitrary. From (4.5), we have:

$$\lim_{\ell\to\infty} z^{i_\ell}_{\alpha\beta} \;=\; z_{\alpha\beta} \;=\; \int_{\mathbf{K}} \mathbf{x}^\alpha\,\mathbf{y}^\beta\,d\mu^*,$$

and by Theorem 2.2(c)

$$\lim_{\ell\to\infty} z^{i_\ell}_{\alpha\beta} \;=\; \int_{\mathbf{K}} \mathbf{x}^\alpha\,\mathbf{y}^\beta\,d\mu^* \;=\; \int_{\mathbf{Y}} \mathbf{y}^\beta\,g(\mathbf{y})^\alpha\,d\varphi(\mathbf{y}),$$

and as the converging subsequence was arbitrary, the above convergence holds for the whole sequence $(z^i_{\alpha\beta})$. □

REFERENCES

[1] E.J Anderson and P. Nash, *Linear Programming in Infinite-Dimensional Spaces*, John Wiley & Sons, Chichester, 1987.

[2] K. Arrow, S. Karlin, and H. Scarf, *Studies in the Mathematical Theory of Inventory and Production*, Stanford University Press, Stanford, CA, 1958.

[3] R. Ash, *Real Analysis and Probability*, Academic Press, San Diego, CA, 1972.

[4] D. Bertsimas, K. Natarajan and Chung-Piaw Teo, *Persistence in discrete optimization under data uncertainty*, Math. Prog. Ser. B, 108 (2005), pp. 251–274.

[5] J. F. Bonnans and A. Shapiro. *Perturbation Analysis of Optimization Problems*, Springer, New York, 2000.

[6] C.M. Bender, L. R. Mead and N. Papanicolaou, *Maximum entropy summation of divergent perturbation series*, J. Math. Phys. 28, (1987), pp. 1016-1018.

[7] J. Borwein and A.S. Lewis, *On the convergence of moment problems*, Trans. Am. Math. Soc. 325, (1991), pp. 249–271.

[8] J. Borwein and A.S. Lewis, *Convergence of best entropy estimates*, SIAM J. Optim. 1, (1991), pp. 191–205.

[9] E. Delage and Yinyu Ye, *Distributionally robust optimization under moment uncertainty with application to data-driven problems*, Oper. Res., to appear.

[10] E.G Dynkin and A.A. Yushkevich, *Controlled Markov Processes*, Springer-Verlag, New York, 1979.

[11] W. Gautschi, *Numerical Analysis: An Introduction*, Birkhäuser, Boston, 1997.

[12] D. Henrion, J. B. Lasserre and J. Löfberg, *GloptiPoly 3: moments, optimization and semidefinite programming*, Optim. Methods and Softw. 24, (2009), pp. 761–779.

[13] O. Hernández-Lerma and J. B. Lasserre, *Discrete-Time Markov Control Processes: Basic Optimality Criteria*, Springer-Verlag, New York, 1996.

[14] O. Hernández-Lerma and J.B. Lasserre, *Approximation schemes for infinite linear programs*, SIAM J. Optim. 8, (1998), pp. 973–988.

[15] J.B. Lasserre, *Global optimization with polynomials and the problem of moments*, SIAM J. Optim. 11, (2001), pp. 796–817.

[16] J.B. Lasserre, *Polynomial programming: LP-relaxations also converge*, SIAM J. Optim. 15, (2004), pp. 383–393.

[17] J.B. Lasserre, *Convergent SDP-relaxations in polynomial optimization with sparsity*, SIAM J. Optim. 17, (2006), pp. 822–843.

[18] J.B. Lasserre, *Semidefinite programming for gradient and Hessian computation in maximum entropy estimation*, Proc. 48th IEEE CDC Conference, New-Orleans (2007), pp. 3060–3064.

[19] H.M. Möller and H.J. Stetter, *Multivariate polynomial equations with multiple zeros solved by matrix eigenproblems*, Num. Math. 70, (1995), pp. 311–329.

[20] K. Natarajan, Miao Song and Chung-Piaw Teo, *Persistency and its applications in choice modelling*, Manag. Sci. 55, (2009), pp. 453–469.

[21] K. Natarajan, Chung-Piaw Teo and Z. Zheng, *Mixed zero-one linear programs under objective uncertainty: a completely positive representation*, Technical report, NUS Business School, National University of Singapore, August 2009.

[22] M. Putinar, *Positive polynomials on compact semi-algebraic sets*, Indiana Univ. Math. J. 42, (1993), pp. 969–984.

[23] P. Rostalski, *Algebraic Moments: Real root finding and related topics*, PhD Thesis, Automatic Control Laboratory, ETH Zurich, Switzerland, May 2009.

[24] M. Schäl, *Conditions for optimality and for the limit of n-stage optimal policies to be optimal*, Z. Wahrs. verw. Gerb. 32, (1975), pp. 179–196.
[25] M. Schweighofer, *Optimization of polynomials on compact semialgebraic sets*, SIAM J. Optim. 15, (2005), pp. 805–825.
[26] A. Tagliani, *Entropy estimate of probability densities having assigned moments: Hausdorff case*, Appl. Math. Lett. 15, (2002), pp. 309–314.
[27] A. Tagliani, *Entropy estimate of probability densities having assigned moments: Stieltjes case*, Appl. Math. Comput. 130, (2002), pp. 201–211.
[28] H. Waki, S. Kim, M. Kojima and M. Muramatsu, *Sums of squares and semidefinite program relaxations for polynomial optimization problems with structured sparsity*, SIAM J. Optim. 17, (2006), pp. 218–242.
[29] V. Weispfenning, *Comprehensive Gröbner bases*, J. Symb. Comp. 14, (1992), pp. 1–29.