A SHORT INTRODUCTION TO MOMENT-SOS HIERARCHIES

YOHANN DE CASTRO

Abstract. In this article, we present optimization techniques using "Moment-Sum-of-Squares hierarchies". These techniques have recently been deployed in a successful manner in several Learning problems. They are based on the decomposition into Sum-of-Squares of nonnegative polynomials and they encode the moments of nonnegative measure on compact basic semi-algebraic sets.

1. A First Approach on Moment-SoS Hierarchies

1.1. **Global Optimization.** One important task in Applied Mathematics is to assess procedures that can find a global minimizer $x^\star$ of a function f over a space $\mathcal{X}$. This can be simply written into the form of the optimization program

$$\min_{x\in\mathcal{X}} f(x)$$

A standard approach, dating back to Cauchy [11], is to use a local search of the minima thanks to a gradient descent. This approach converges to critical points of the objective function and one needs additional efforts to hopefully compute a global minimum.

An other approach, dating back to Hilbert [20], may rather focus on a suitable decomposition of the objective function as a "*sum-of-squares*" (SoS), namely

$$f = \lambda^\star + \underbrace{\sum_{k=1}^{K} g_k\, h_k^2}_{p^\star}\,, \tag{1}$$

where $\lambda^\star \in \mathbb{R}$ is a constant, g_k some non-negative functions and h_k some functions. Furthermore, if one can provide $x^\star \in \mathcal{X}$ such that

$$p^\star(x^\star) := \sum_{k=1}^{K} g_k(x^\star)\, h_k^2(x^\star) = 0\,, \tag{2}$$

then $x^\star$ is a global minimum of f and the minimal value of f over $\mathcal{X}$ is $\lambda^\star$. We understand that this discussion exhibits two important tasks: finding decompositions of the objective function as in (1) and finding roots $x^\star$ as in (2). These two tasks are closely related as we will see in the next subsection.

1.2. **A Problem of Moments.** Assume that $\mathcal{X}$ is a compact subset of $\mathbb{R}^d$ and assume that f is a multivariate polynomial, namely it holds that

$$f(x) = \sum_{\alpha\,:\,|\alpha|\leq r} f_a x^\alpha\,,$$

for $f_\alpha \in \mathbb{R}$ and denoting $x^\alpha := x_1^{\alpha_1}\ldots x_d^{\alpha_d}$ the monomials and

$$|\alpha| = \sum_{k=1}^{d} \alpha_k\,,$$

the degree of a monomial which allows to define the degree of a polynomial. Now, one can write

$$\min_{x\in\mathcal{X}} f(x) = \min_{\left\{\mu\in\mathcal{M}(\mathcal{X})\,:\,\mu(\mathcal{X})=1\right\}} \int_{\mathcal{X}} f\mathrm{d}\mu = \min_{\left\{(m_\alpha)\in\boldsymbol{M}_r(\mathcal{X})\,:\,m_0=1\right\}} \sum_{\alpha} f_\alpha m_\alpha\,, \tag{3}$$

Date: April 3, 2019.

where

- $\mathcal{M}(\mathcal{X})$ is the set of nonnegative measure over $\mathcal{X}$;
- $\boldsymbol{M}_r(\mathcal{X})$ is the set of $(m_\alpha)_{|\alpha|\leq r}$ of moments of nonnegative measure over $\mathcal{X}$, see Section 2 for further details;
- the condition $\mu(\mathcal{X}) = 1$ (namely $m_0 = 1$) ensures that the first moment is one, *i.e.*, μ is a probability measure over $\mathcal{X}$.

Now, let us have a look of the aforementioned equality. On the left hand side we have a "hard" objective function f that has to be optimized over a "simple" space $\mathcal{X}$, while on the right hand side we have a "simple" linear objective function, namely

$$(m_\alpha)_\alpha \mapsto \sum_\alpha f_\alpha m_\alpha \,,$$

that has to be optimized over a "hard" space $\boldsymbol{M}_r(\mathcal{X})$. This "trick" has been successfully used in Optimal Transport theory by Kantorovich [22]. In our case, it reduces the problem of global optimization of multivariate polynomials to finding the most correlated sequence of moments to the coefficients of the objective polynomial.

1.3. **Lagrangian Duality.** The global optimization problem (3) is in fact a linear program over the cone of truncated moments $\boldsymbol{M}_r(\mathcal{X})$. From this point, one can consider the following Lagrangian expression

$$\mathcal{L}((m_\alpha),(c_\alpha),\lambda) := \sum_\alpha f_\alpha m_\alpha - \sum_\alpha c_\alpha m_\alpha + \lambda(1-m_0) - \imath_{\boldsymbol{C}_r(\mathcal{X})}((c_\alpha))$$

where the dual variables are $(c_\alpha)_{|\alpha|\leq r}$ with $c_\alpha \in \mathbb{R}$ and $\lambda \in \mathbb{R}$. We denote by

- $\boldsymbol{C}_r(\mathcal{X})$ the set of coefficients $(c_\alpha)_{|\alpha|\leq r}$ of nonnegative polynomials of degree at most r over $\mathcal{X}$;
- $\imath_{\boldsymbol{C}(\mathcal{X})}((c_\alpha))$ the indicator function of this set, namely it is 0 if $(c_\alpha)_{|\alpha|\leq r}$ belongs to $\boldsymbol{C}_r(\mathcal{X})$ and ∞ otherwise.

Remark that

$$\forall (m_\alpha)_{|\alpha|\leq r} \in \boldsymbol{M}_r(\mathcal{X})\,,\ \forall (c_\alpha)_{|\alpha|\leq r} \in \boldsymbol{C}_r(\mathcal{X})\,,\quad \sum_\alpha c_\alpha m_\alpha = \int_{\mathcal{X}} \big(\sum_\alpha x^\alpha\big)\mathrm{d}\mu\,,$$

where $(m_\alpha)_{|\alpha|\leq r}$ is represented by μ. Then, it is standard to deduce that the "*dual cone*" (see [5, Page 96] for instance) of $\boldsymbol{C}_r(\mathcal{X})$ is exactly $\boldsymbol{M}_r(\mathcal{X})$ and vice versa, namely

- If $\sum_\alpha c_\alpha m_\alpha \geq 0$ for all $(m_\alpha)_{|\alpha|\leq r} \in \boldsymbol{M}_r(\mathcal{X})$ then $(c_\alpha)_{|\alpha|\leq r} \in \boldsymbol{C}_r(\mathcal{X})$;
- If $\sum_\alpha c_\alpha m_\alpha \geq 0$ for all $(c_\alpha)_{|\alpha|\leq r} \in \boldsymbol{C}_r(\mathcal{X})$ then $(m_\alpha)_{|\alpha|\leq r} \in \boldsymbol{M}_r(\mathcal{X})$.

We deduce that the primal expression is given by

$$\inf_{(m_\alpha)} \sup_{(c_\alpha),\lambda} \mathcal{L}((m_\alpha),(c_\alpha),\lambda) = \inf_{(m_\alpha)} \Big\{ \sum_\alpha f_\alpha m_\alpha \,:\, m_0 = 1 \text{ and } (m_\alpha)_{|\alpha|\leq r} \in \boldsymbol{M}_r(\mathcal{X}) \Big\}\,; \tag{4}$$

and the dual expression is

$$\begin{aligned} \sup_{(c_\alpha),\lambda} \inf_{(m_\alpha)} \mathcal{L}((m_\alpha),(c_\alpha),\lambda) &= \sup_{\lambda\in\mathbb{R}} \big\{ \lambda \,:\, (f_\alpha - \lambda \mathbb{1}_{\alpha=0})_{|\alpha|\leq r} \in \boldsymbol{C}_r(\mathcal{X}) \big\}\,, \\ &= \max_{\{\lambda\,:\, f(x)\geq\lambda\,,\ \forall x\in\mathcal{X}\}} \lambda\,. \end{aligned} \tag{5}$$

The optimal value $\lambda^\star$ of λ is the minimum of f over $\mathcal{X}$ and there is no duality gap in view of (3). The complementary conditions between the primal optimal solutions $(m_\alpha^\star)$ and the dual optimal solutions $(c_\alpha^\star), \lambda^\star$ give that

$$f = \lambda^\star + p^\star\,, \tag{6}$$

$$\sum_\alpha (f_\alpha - \lambda^\star \mathbb{1}_{\{\alpha=0\}}) m_\alpha^\star = \int_{\mathcal{X}} (f-\lambda^\star)\mathrm{d}\mu^\star = \int_{\mathcal{X}} p^\star \mathrm{d}\mu^\star = 0\,, \tag{7}$$

where $\mu^\star$ reprensents the moments $(m^\star_\alpha)_{|\alpha|\leq r}$ and $p^\star(x)=\sum_{|\alpha|\leq r} c^\star_\alpha x^\alpha$ is a nonnegative polynomial over $\mathcal{X}$.

Now, one can look at a discrete probability measure $\mu^\star$ representing $(m^\star_\alpha)_{|\alpha|\leq r}$. Indeed, the set of moments $(m_\alpha)_{|\alpha|\leq r}$ of probability measures over $\mathcal{X}$, namely

$$\boldsymbol{M}^0_r(\mathcal{X}) := \boldsymbol{M}_r(\mathcal{X}) \cap \big\{(m_\alpha)_{|\alpha|\leq r}\,:\, m_0=1\big\} \subset \mathbb{R}^{r(d)}\,,$$

is a compact convex set, where $r(d)=\binom{r+d}{r}$. By Caratheodory's theorem, $(m^\star_\alpha)_{|\alpha|\leq r}$ is a convex combination of $r(d)$ extreme points of $\boldsymbol{M}^0_r(\mathcal{X})$. One can remark that $\boldsymbol{M}^0_r(\mathcal{X})$ is the convex hull of the moment curve $\big\{(x^\alpha)_{|\alpha|\leq r}\,:\, x\in\mathcal{X}\big\}$, namely

$$\boldsymbol{M}^0_r(\mathcal{X}) = \mathrm{conv\,hull}\big(\big\{(x^\alpha)_{|\alpha|\leq r}\,:\, x\in\mathcal{X}\big\}\big)\,,$$

and hence the extreme points of $\boldsymbol{M}^0_r(\mathcal{X})$ are included in the moment curve, which is exactly the moments of the Dirac masse δ_x at point $x\in\mathcal{X}$. We understand that there exists a discrete measure $\mu^\star$ representing the moments $(m^\star_\alpha)_{|\alpha|\leq r}$, so that

$$\mu^\star := \sum_{k=1}^{K} a_k \delta_{x^\star_k}\,, \tag{8}$$

where $1\leq k\leq r(d)$, $x^\star_k\in\mathcal{X}$ distincts and $a_k\geq 0$ so that $\sum_{k=1}^K a_k=1$. Note that (7) shows that the support $\{x^\star_k\,:\,k\in[K]\}$ of $\mu^\star$ is included in the set of roots $x^\star$ of $p^\star$.

Remark 1. *As a conclusion, one can remark that solving the Lagrangian formulation of the problem of global optimization leads to optimal solutions giving a decomposition* (6) *of the same flavor of Hilbert decomposition* (1) *and solution points $x^\star$ (as introduced in* (2)*) given by* (7) *and a well chosen solution $\mu^\star$ as in* (8)*.*

1.4. **Outline: Moment-SoS Hierarchies and their applications.** The issue in practice is to represent the cone $\boldsymbol{M}_r(\mathcal{X})$. In Section 2, we will show that $\boldsymbol{M}_r(\mathcal{X})$ is characterized by countably many inequalities that can be represented using countably many (indexed by $\delta\geq 0$) semidefinite matrices of growing sizes $r+\delta$. These constraints can be written as Semi-Definite Programs (SDP) which can be solved approximately using an interior point method, see for instance [1] on this latter aspect.

The strategy is to certify a finite number $\delta=0,\ldots,\Delta-1$ of these constraints and to forget the other constraints for $\delta\geq\Delta$. This leads to an outer approximation of $\boldsymbol{M}_r(\mathcal{X})$ which is SDP representable, see for instance Section 2.3 et Section 2.5. Given an arbitrarily small positive real ε, we call "ε-approximate solution" any point m^Δ of $\mathbb{R}^r$ that is admissible for these Δ constraints and such that the objective function at this point, namely $\sum_\alpha f_\alpha m^\Delta_\alpha$, is ε close to the objective function at the true solution point, namely $\sum_\alpha f_\alpha m^\star_\alpha$.

Remark 2. *We call a polynomial time algorithm for this kind of problems an algorithm such that, given the bit length description of the input, say L,*

- *the number a certified constraints Δ satisfies $\Delta=\mathcal{O}(\log L)$;*
- *the SDP matrices involved in describing the constraints at point m^Δ have Frobenius norm bounded by $\mathcal{O}(\exp(\mathrm{poly}(L))$;*
- *an ε-approximate solution up to $\varepsilon=\Omega(\exp(-\mathrm{poly}(L))$ is sufficient for the purpose of interest;*

as presented for instance in [1, Section 3.3].

As we will detail in Section 2, on a "*compact basic semi-algebraic set*" (see (9) for a definition) the Putinar's Positivstellensatz [29] shows that positive polynomials are Sum-of-Squares (SoS) polynomials, see Section 2.2. Then we will see that SoS polynomials can be parametrized by semidefinite matrices, see Section 2.4. Using this characterization, one can show that sequences of moments can be equivalently described using "*Hierarchies*" of semidefinite matrices referred to as "*Moment-SoS Hierarchies*". These hierarchies corresponds to SDP programs leading to a construction of an approximate solution point m^Δ. Polynomial time solvability (in the sense defined above) of various combinatorial problems using Moment-SoS Hierarchies has been studied for at least 3 decades, see for instance [4] and references therein.

Besides, these Moment-SoS hierarchies have been used in Statistical Learning recently and we will present these developments in Section 3.

1.5. **Related works.** Over the last 20 years, Moment-SoS hierarchies have been popularized since they also give a measure of the complexity of some optimization programs, if not the limit of polynomial time algorithms. In particular, a famous conjecture (the "Unique Game Conjecture" [4]) implies that no polynomial algorithm can improve the bound given by the SoS hierarchies.

We would like to point out that the developments in Moment-SoS hierarchies as been presented in [25] with the first proof of convergence of the hierarchiy and it has been used in optimal control in [27]. Problems in optimal control for linear systems, formulated as a primal LP (as in (3)) on measures and with dual LP on continuous functions (as in (5)), can be solved numerically with primal-dual moment-SoS SDP hierarchies as shown in [12, 13]. Formulating optimal control problems as moment problems was a classical research topic in the 1960s, where optimal control laws were sought in measures spaces (completions of Lebesgue spaces) to allow for oscillations and concentrations, see e.g. [24] or the overview in [19, Section III]. In the case of linear optimal control of an ordinary differential equation of order n, it was proved in [28] that there is always an n-atomic optimal measure solving problem primal LP (as in (3)). The idea of using SoS relaxation in optimization can be traced back to the 80's (see [25] and [9] for historical account). Furthermore, the use of Linear Matrix Inequalities in control has been studied in [8] for instance.

Recently, these hierarchies have been deployed in problems such as tensor decomposition [3], dictionary learning [3], robustness of matrix completion [14], tensor PCA [6], among others. In addition, the SoS hierarchies have been introduced by [17] to solve the problem of "*Super-Resolution*" [10] in all dimensions, which can be lead by minimizing the L^1-norm (total variation norm of a signed measure) on the space of signed measures. The Super-Resolution framework can be understood as recovering a spiked signal from few linear measurements and the SoS hierarchies allow to compute an estimate of these spikes in a grid-less manner (referred to as "Off-The-Grid"). Companion problems are "Line Spectral Estimation" [30] or "Off-The-Grid Compressed Sensing" [31]. Independently, SoS hierarchies have been used to solve the problem of optimal design of experiments in Statistics by [16].

2. Representing Multivariate Moments

2.1. **The "*full*" moment cone, its dual and its projections.** Given a positive measure $\mu \in \mathcal{M}(\mathcal{X})$ and $\alpha \in \mathbb{N}^n$, we may call the sequence $\mathbf{m} = (m_\alpha)_{\alpha\in\mathbb{N}^d}$ the "*full*" moment sequence of μ. Conversely, we say that $\mathbf{m} = (m_\alpha)_{\alpha\in\mathbb{N}^d}$ has a *representing measure*, if there exists a measure $\mu \in \mathcal{M}(\mathcal{X})$ such that $\mathbf{m}$ is its moment sequence. The "*full*" moment cone is given by

$$\mathbf{M}(\mathcal{X}) := \Big\{\mathbf{m} = (m_\alpha)_{\alpha\in\mathbb{N}^d} \quad \text{s.t.} \quad \forall \alpha \in \mathbb{N}^d,\ m_\alpha = \int_{\mathcal{X}} x^\alpha \,\mathrm{d}\mu,\ \mu \geq 0\Big\}.$$

One can characterize full sequences of algebraic moments thanks to the Riesz-Haviland extension theorem, see [26, Theorem 3.1] for example. This representation theorem shows that the dual cone of $\mathbf{M}(\mathcal{X})$ is exactly the cone $\mathbf{C}(\mathcal{X})$ of (sequence of coefficients of) nonnegative polynomials over $\mathcal{X}$.

In the sequel we will not distinguish between a polynomial and its sequence of coefficients. Note that we assimilate polynomials p of degree at most r with a vector of dimension $r(d) = \binom{r+d}{r}$, which contains the coefficients of p in the chosen basis.

2.2. **Putinar's Positivstellensatz.** Recall that $r(d) = \binom{r+d}{r}$. The *moment matrix* of a truncated sequence $\mathbf{m}_{2r} = (m_\alpha)_{|\alpha|\leq 2r}$ is the $r(d) \times r(d)$-matrix $M_r(\mathbf{m})$ with rows and columns respectively indexed by integer d-tuples $\alpha, \beta \in \mathbb{N}^d, |\alpha|, |\beta| \leq r$ and whose entries are given by

$$M_r(\mathbf{m})(\alpha, \beta) = m_{\alpha+\beta}.$$

It is symmetric and linear in $\mathbf{m}$. Further, if $\mathbf{m}$ has a representing measure, then $M_r(\mathbf{m})$ is *positive semidefinite* (written $M_r(\mathbf{y}) \succcurlyeq 0$). Similarly, we define the *localizing matrix* of a polynomial

$$g = \sum_{|\alpha|\leq n} g_\alpha x^\alpha \in \mathbb{R}[x]_n\,,$$

of degree n and a sequence $\mathbf{m}_{2r+n} = (m_\alpha)_{|\alpha|\leq 2r+n}$ as the $r(d)\times r(d)$ matrix $M_r(g\,\mathbf{m})$ with rows and columns respectively indexed by $\alpha,\beta\in\mathbb{N}^d$, $|\alpha|,|\beta|\leq r$ and whose entries are given by

$$M_r(g\,\mathbf{m})(\alpha,\beta) = \sum_{\gamma\in\mathbb{N}^d} g_\gamma m_{\gamma+\alpha+\beta}.$$

We deduce the following proposition.

Proposition 1. *If $\mathbf{m}$ has a representing measure whose support is contained in $\{x\in\mathbb{R}^n : g(x)\geqslant 0\}$, then $M_r(\mathbf{m})\succcurlyeq 0$ and $M_r(g\mathbf{m})\succcurlyeq 0$ for $g\in\mathbb{R}[x]_n$.*

We would like to obtain the converse of Proposition 1, this would lead to a characterization of the set of sequences of moments of nonnegative measures. Consider m polynomials $g_1,\ldots,g_m$ and assume that

$$\mathcal{X} := \Big\{x\in\mathbb{R}^d \quad \text{s.t.} \quad \forall i\in[m],\ g_i(x)\geq 0\Big\} \tag{9}$$

has an "*algebraic certificate of compactness*" meaning that there exists finite sets $J_0,J_1,\ldots,J_m$ and polynomials $(h_j^{(0)})_{j\in J_0},(h_j^{(1)})_{j\in J_1},\ldots,(h_j^{(m)})_{j\in J_m}$ so that

$$\Big\{x\in\mathbb{R}^d \quad \text{s.t.} \sum_{j\in J_0}(h_j^{(0)})^2(x) + \sum_{i=1}^m\big[\sum_{j\in J_i}(h_j^{(i)})^2(x)\big]g_i(x)\geq 0\Big\} \text{ is compact.}$$

When $\mathcal{X}$ is compact, such algebraic certificate of compactness can be enforced adding the polynomial $g_{m+1}(x) = R - \|x\|_2^2$ to the g_i's, with $R>0$ sufficiently large. The set defined by (9) is referred to as a compact basic semi-algebraic set. Using Putinar's theorem [29], one may prove the following important result, see for instance the book [26, Theorem 3.8].

◦ **First Important Representation:** The sequence $\mathbf{m} = (m_\alpha)_{\alpha\in\mathbb{N}^d}$ has a representing measure $\mu\in\mathcal{M}(\mathcal{X})$ *if and only if* for all $r\in\mathbb{N}$ the matrices $M_r(\mathbf{m})$ and $M_r(g_j\,\mathbf{m})$ for $j=1,\ldots,m$, are positive semidefinite.

2.3. **SDP Approximations of the Moment Cone.** Letting $v_j := \lceil d_j/2\rceil$, for $j=1,\ldots,m$, denote half the degree of the g_j, by Putinar's theorem [29], we can approximate the moment cone $\mathbf{M}_{2r}(\mathcal{X})$ by the following semidefinite representable cones for $\delta\in\mathbb{N}$:

$$\begin{aligned}\mathbf{M}^{\mathsf{SDP}}_{2(r+\delta)}(\mathcal{X}) := \Big\{\mathbf{m}_{r,\delta}\in\mathbb{R}^{\binom{d+2r}{d}} \ :\ &\exists\mathbf{m}_\delta\in\mathbb{R}^{\binom{d+2(r+\delta)}{d}} \text{ such that}\\ &\mathbf{m}_{r,\delta} = (m_{\delta,\alpha})_{|\alpha|\leq 2r} \text{ and}\\ &M_{r+\delta}(\mathbf{m}_\delta)\succcurlyeq 0,\ M_{r+\delta-v_j}(g_j\mathbf{m}_\delta)\succcurlyeq 0,\ j=1,\ldots,m\Big\}.\end{aligned}$$

By semidefinite representable we mean that the cones are projections of linear sections of semidefinite cones. Since $\mathbf{M}_{2d}(\mathcal{X})$ is contained in every $(\mathbf{M}^{\mathsf{SDP}}_{2(d+\delta)}(\mathcal{X}))_{\delta\in\mathbb{N}}$, they are outer approximations of the moment cone. Moreover, they form a nested sequence, so we can build the hierarchy

$$\mathbf{M}_{2r}(\mathcal{X})\subseteq\cdots\subseteq\mathbf{M}^{\mathsf{SDP}}_{2(r+2)}(\mathcal{X})\subseteq\mathbf{M}^{\mathsf{SDP}}_{2(r+1)}(\mathcal{X})\subseteq\mathbf{M}^{\mathsf{SDP}}_{2r}(\mathcal{X}). \tag{10}$$

This hierarchy actually converges, meaning

$$\mathbf{M}_{2r}(\mathcal{X}) = \bigcap_{\delta=0}^{\infty}\mathbf{M}^{\mathsf{SDP}}_{2(r+\delta)}(\mathcal{X})\,,$$

see for instance the book [26, Theorem 3.8].

2.4. **SOS Approximations of Nonnegative Polynomials.** Further, let $\Sigma[x]_r\subseteq\mathbb{R}[x]_{2r}$ be the set of all polynomials that are sums of squares of polynomials (SOS) of degree at most $2r$, *i.e.*,

$$\Sigma[x]_r = \Big\{\sigma\in\mathbb{R}[x]_{2r} : \sigma(x) = \sum_{i=1}^k h_i(x)^2 \text{ for some } h_i\in\mathbb{R}[x]_r \text{ and some } k\geq 1\Big\}\,.$$

The topological dual of $\mathbf{M}^{\mathsf{SDP}}_{2(d+\delta)}(\mathcal{X})$ is the cone of coefficients of the polynomials of a quadratic module, which we denote by $\mathbf{C}^{\mathsf{SOS}}_{2(d+\delta)}(\mathcal{X})$. It is given by

$$\mathbf{C}^{\mathsf{SOS}}_{2(r+\delta)}(\mathcal{X}) := \Big\{ h = \sigma_0 + \sum_{j=1}^{m} g_j \sigma_j : \sigma_0 \in \Sigma[x]_{r+\delta},\ \sigma_j \in \Sigma[x]_{r+\delta-\nu_j},\ j = 1, \ldots, m\,, \\ \text{and } \deg(h) \leq 2r \Big\}.$$

It is the set of polynomials of degree at most $2r$ which are Sum-of-Squares. Write

$$\mathbf{v}_n(x) := (\underbrace{1}_{\text{degree 0}}, \underbrace{x_1, \ldots, x_d}_{\text{degree 1}}, \underbrace{x_1^2, x_1x_2, \ldots, x_1x_d, x_2^2, \ldots, x_d^2}_{\text{degree 2}}, \ldots, \underbrace{\ldots, x_1^d, \ldots, x_d^n}_{\text{degree n}})^\top$$

for the column vector of the monomials ordered according to their degree, and where monomials of the same degree are ordered with respect to the lexicographic ordering. It holds that, see for instance [26, Proposition 2.1],

◦ **Second Important Representation:** One has $h \in \mathbf{C}^{\mathsf{SOS}}_{2(d+\delta)}(\mathcal{X})$ if and only if h has degree less than $2d$ and there exist real symmetric and positive semidefinite matrices Q_0 and $Q_j,\ j = 1, \ldots, m$ of size $\binom{d+r+\delta}{d} \times \binom{d+r+\delta}{d}$ and $\binom{d+r+\delta-\nu_j}{d} \times \binom{d+r+\delta-\nu_j}{d}$ respectively, such that for any $x \in \mathbb{R}^d$

$$\begin{aligned} h(x) &= \sigma_0(x) + \sum_{j=1}^{m} g_j(x)\sigma_j(x) \\ &= \mathbf{v}_{r+\delta}(x)^\top Q_0 \mathbf{v}_{r+\delta}(x) + \sum_{j=1}^{m} g_j(x)\, \mathbf{v}_{r+\delta-\nu_j}(x)^\top Q_j \mathbf{v}_{r+\delta-\nu_j}(x)\,. \end{aligned}$$

The elements of $\mathbf{C}^{\mathsf{SOS}}_{2(r+\delta)}(\mathcal{X})$ are polynomials of degree at most $2d$ which are non-negative on $\mathcal{X}$. Hence, it is a subset of $\mathbf{C}_{2r}(\mathcal{X})$, the set of nonnegative polynomials of degree at most $2r$, and it holds that

(11) $$\mathbf{C}_{2r}(\mathcal{X}) \supseteq \cdots \supseteq \mathbf{C}^{\mathsf{SOS}}_{2(r+2)}(\mathcal{X}) \supseteq \mathbf{C}^{\mathsf{SOS}}_{2(r+1)}(\mathcal{X}) \supseteq \mathbf{C}^{\mathsf{SOS}}_{2r}(\mathcal{X}).$$

Remark 3. *Note that* (10) *represents the Lasserre's hierarchy which a nested sequence of outer SDP approximations of the moment cone while its dual, namely* (11)*, represents the SoS hierarchy which a nested sequence of inner SOS representations of the nonnegative polynomials.*

2.5. **SDP Relaxations.** Using (10), one can substitute the cone of truncated moment $\mathbf{M}_{2r}(\mathcal{X})$ by an outer SDP approximation $\mathbf{M}^{\mathsf{SDP}}_{2(r+\delta)}(\mathcal{X})$. For instance, the primal program (4) can be approximate by the SDP program

$$\inf_{(m_\alpha)} \Big\{ \sum_\alpha f_\alpha m_\alpha :\ m_0 = 1 \text{ and } (m_\alpha)_{|\alpha|\leq 2r} \in \mathbf{M}^{\mathsf{SDP}}_{2(r+\delta)}(\mathcal{X}) \Big\}\,;$$

its dual is also a SDP program given by

$$\lambda^\star_\delta := \sup_{\lambda\in\mathbb{R}} \Big\{ \lambda\ :\ (\sum_\alpha f_\alpha x^\alpha - \lambda) \in \mathbf{C}^{\mathsf{SOS}}_{2(r+\delta)}(\mathcal{X}) \Big\}\,;$$

which is the SDP relaxation of (5). This latter can be understood as follows. The optimal value $\lambda^\star_\delta$, given by the aforementioned dual, is the larget value λ such that $(\sum_\alpha f_\alpha x^\alpha - \lambda) \in \mathbf{C}^{\mathsf{SOS}}_{2(r+\delta)}(\mathcal{X})$. Actually, it is straightforward that

$$\big\{ \lambda \in \mathbb{R}\ :\ (\sum_\alpha f_\alpha x^\alpha - \lambda) \in \mathbf{C}^{\mathsf{SOS}}_{2(r+\delta)}(\mathcal{X}) \big\} = (-\infty, \lambda^\star_\delta]$$

and that $\lambda^\star_0 \leq \lambda^\star_1 \leq \lambda^\star_2 \leq \ldots \leq \lim_\delta \lambda^\star_\delta = \lambda^\star$, see for instance [26, Theorem 5.6(a)].

Remark 4. *It can be important to notice that if*

(12) $$(\sum_\alpha f_\alpha x^\alpha - \lambda^\star) \in \mathbf{C}^{\mathsf{SOS}}_{2(r+\delta^\star)}(\mathcal{X}) \textit{ for some } \delta^\star\,,$$

then the relaxation of order $\delta^\star$ is exact, it holds that $\lambda^\star_{\delta^\star} = \lambda^\star$ and there exists a measure $\mu^\star$ representing the primal solution $(m^\star_\alpha)_{|\alpha|\leq 2r} \in \mathbf{M}^{\mathsf{SDP}}_{2(r+\delta^\star)}(\mathcal{X})$. In particular, one has $(m^\star_\alpha)_{|\alpha|\leq 2r} \in \mathbf{M}_{2r}(\mathcal{X})$.

If one can show by an ad hoc argument that (12) *holds true then*

$$\min_{\left\{(m_\alpha)\in \boldsymbol{M}_r(\mathcal{X})\,:\, m_0=1\right\}} \sum_\alpha f_\alpha m_\alpha = \min_{\left\{(m_\alpha)\in \mathbf{M}^{\mathsf{SDP}}_{2(r+\delta)}(\mathcal{X})\,:\, m_0=1\right\}} \sum_\alpha f_\alpha m_\alpha$$

and the SDP relaxation (*r.h.s.*)*returns the exact solution $(m^\star_\alpha)_{|\alpha|\leq r}$ of the moment problem* (*l.h.s.*). *The objective value at the solution point is $\lambda^\star$.*

3. Some examples from Statistical Learning

Recently, *Moment-SoS Hierarchies* have been deployed in Statistical Learning and we will briefly present some examples here.

3.1. **Spiked Tensor PCA.**

A central question in Statistical Learning seeks to identify the statistical limit of a problem, that is to describe the limit of the signal-to-noise ratio for which the maximum likelihood estimator converges to the parameter to be estimated. Once this limit is established, a fundamental question is whether a polynomial time algorithm reaches this bound, namely if there exists a polynomial time algorithm finding the maximum of the likelihood (or equivalently the minimum $x^\star$ of f, the opposite of the likelihood).

We will present an example here in which the *Moment-SoS Hierarchies* gives the computational limit. In a series of papers including [6], the authors are interested in detecting a rank one tensor $u^{\otimes d}$ from a noisy observation $Y \in \mathbb{R}^{n^d}$

$$Y = \lambda\, u^{\otimes d} + Z$$

where $\lambda > 0$ is a "signal-to-noise" ratio, $u \in \mathbb{R}^n$ such that $||u||_2 = 1$ and $Z_{i_1,\ldots,i_d} \sim_{iid} \mathcal{N}(0, 1/n)$ for $i_1 \leq \ldots \leq i_d$ and Z then complemented by symmetry. The maximum likelihood consists of calculating the rank one tensor $x^{\otimes d}$ most correlated with Y, which is the maximum on $x \in \mathbb{R}^n$ such that $||x||_2 = 1$ of likelihood (objective function, polynomial in x)

$$\langle Y, x^{\otimes d}\rangle = \lambda\langle u, x\rangle^d + \sum_{i_1,\ldots,i_d} Z_{i_1,\ldots,i_d} x_{i_1}\ldots x_{i_d}\,.$$

The d-homogeneous polynomial (indexed by the sphere) $\sum_{i_1,\ldots,i_d} Z_{i_1,\ldots,i_d} x_{i_1}\ldots x_{i_d}$ is centered (zero mean) and its maximum is $\mathcal{O}(1)$ with high probability. We deduce the statistical bound $\lambda = \mathcal{O}(1)$ for which we can estimate u consistently.

To calculate the maximum of the likelihood $x \mapsto \langle Y, x^{\otimes d}\rangle$ one has to calculate the maximum of a polynomial on the sphere. Lasserre hierarchies are designed to solve this problem. One can show [21] that the 4^{th} order relaxation of these hierarchies allows to compute a consistent estimator of u when $\lambda = \mathcal{O}(n^{\frac{d-2}{4}})$, it is the computational limit (we do not know if it is optimal). This upper bound can also be reached using a spectral estimator. We see that in order to find the computational limit we use the Lasserre hierarchies.

An interesting question then is whether this gap is optimal. In particular, it is relevant to find a heuristic that would tend to prove why a polynomial time algorithm can not exist below the limit $\lambda = \mathcal{O}(n^{\frac{d-2}{4}})$. The idea is then to look at the critical points for which the value of the objective function is close to its maximum. This analysis is the theory of "Landscapes" studied in particular by Gerard Ben Arous and Andrea Montanari [6]. For the problem of spiked tensor PCA it is possible to show that below the computational limit the likelihood has an exponentially large number of critical points close to the maximum and that these critical points have a Hessian of which almost all the eigenvalues are negative.

3.2. **Sparse Deconvolution.**

Another interesting avenue for Lasserre hierarchies is their recent developments in ℓ_1 minimization on the space of measures, referred to as "Off-The-Grid" methods [31, 2, 15]. Let us introduce some notation to present the framework. Consider the Banach space $E := (\mathcal{C}(\mathcal{X}, \mathbb{R}), \|\cdot\|_\infty)$ of real-valued continuous functions over $\mathcal{X}$ endowed with the supremum norm. Recall that its topological dual space $E^\star := (\mathcal{M}(\mathcal{X}), \|\cdot\|_1)$ is the Banach space of real Borel measures endowed with the total variation norm $\|\cdot\|_1$ that can be defined as

$$\forall \mu \in E^\star\,, \quad \|\mu\|_1 := \sup_{\|f\|_\infty \leq 1} \int_{\mathcal{X}} f \,\mathrm{d}\mu\,.$$

Consider $\Phi(x) = (\varphi_1(x), \ldots, \varphi_n(x)) \in \mathbb{R}[x]^n$ a multivariate polynomial function and $b \in \mathbb{R}^n$. Assume that

$$\int_{\mathcal{X}} \Phi \,\mathrm{d}\mu^0 = b^0\,,$$

and one would like to infer μ^0 from the observation of b. Note that there exists a matrix A with n rows and an integer r such that $\Phi(x) = A(x^\alpha)_{|\alpha|\leq r}$. Obviously, there exists an infinite number of solutions to the aforementioned equation but one would like to recover one solution that is "sparse" meaning that it is atomic with few atoms here. Indeed, assume that the target measure satisfies

$$\mu^0 = \sum_{k=1}^{K} a_k^0 \delta_{x_k^0}$$

then a powerful strategy is to consider the following primal program

$$(\mathcal{P}_{\mathsf{BME}}) \qquad \widehat{\mu} \in \arg \min_{\{\mu \in E^\star \,:\, \int_{\mathcal{X}} \Phi \,\mathrm{d}\mu = b^0\}} ||\mu||_1\,.$$

In this setting, one would like to recover a signed measure μ from a finite number of linear measurements. This extremal moment problem has been intensively studied in various fields of Mathematics at the beginning of the 20th century: Arne Beurling [7] initiated the theory of extension functions in Harmonic Analysis when studying the minimal total variation norm function among all bounded variation functions with prescribed Fourier transform on a given domain. The aforementioned estimator $\widehat{\mu}$ solution to ($\mathcal{P}_{\mathsf{BME}}$) was recently studied in [15] (referred to as "*Beurling Minimal Extrapolation*", BME for short) and [10] for instance.

Our ability to solve ($\mathcal{P}_{\mathsf{BME}}$) is intimately related to our capability to represent nonnegative measures. Indeed, observe that

$$\mathcal{P}_{\mathsf{BME}} = \arg \min_{\{\mu^+,\mu^- \geq 0 \,:\, \int_{\mathcal{X}} \Phi \,\mathrm{d}(\mu^+ - \mu^-) = b^0\}} \int_{\mathcal{X}} \mathrm{d}\mu^+ + \int_{\mathcal{X}} \mathrm{d}\mu^-\,,$$

using Borel-Jordan decomposition into positive and negative parts $\mu = \mu^+ - \mu^-$. Denoting $(m_\alpha^\pm)_{|\alpha|\leq r}$ the first moments of $\mu^\pm \geq 0$, it holds that

$$\mathcal{P}_{\mathsf{BME}} = \arg \min_{\{(m_\alpha^\pm)_{|\alpha|\leq r} \,:\, A((m_\alpha^+)_{|\alpha|\leq r} - (m_\alpha^+)_{|\alpha|\leq r}) = b^0\}} m_0^+ + m_0^-\,,$$

for which Moment-SoS hierarchies can be deployed. For further details, please refer to the paper [17]. In particular, we understand that $\mathcal{P}_{\mathsf{BME}}$ can be SDP relaxed using a Lasserre's hierarchy.

3.3. **The Optimal Design Problem.** The optimum experimental designs are computational and theoretical objects that minimize the variance of the best linear unbiased estimators in regression problems. In this frame, the experimenter models response y_i of a random experiment whose input parameters are represented by a vector $x_i \in \mathbb{R}^d$ with respect to known regression functions $\Phi := (\varphi_1, \ldots, \varphi_p)$, namely for all $i \in [N]$, one has $y_i = \sum_{j=1}^p \theta_j \varphi_j(x_i) + \varepsilon_i$ where $\theta \in \mathbb{R}^p$ are unknown parameters that the experimenter wants to estimate, ε_i is some noise and x_i is chosen by the experimenter in a *design space* $\mathcal{X} \subset \mathbb{R}^d$. Assume that the distinct points among $x_1, \ldots, x_N$ are the points $x_1, \ldots, x_s$, for some $s \in [N]$, and let N_i denote the number of times the particular point x_i occurs among $x_1, \ldots, x_N$, for all $i \in [s]$. This would be summarized by

$$\zeta := \begin{pmatrix} x_1 & \cdots & x_s \\ \frac{N_1}{N} & \cdots & \frac{N_s}{N} \end{pmatrix}, \tag{13}$$

whose first row gives the points in the *design space* $\mathcal{X}$ where the inputs parameters have to be taken and the second row tells the experimenter which proportion of experiments ("*frequencies*") have to be done at these points. The goal of the design of experiment theory is then to assess which inputs parameters x_i and frequencies $w_i := N_i/N$ the experimenter has to consider. For a given ζ, the standard analysis of the Gaussian linear model shows that the minimal covariance matrix (with respect to Loewner ordering) of unbiased estimators can be expressed in terms of the Moore-Penrose pseudoinverse of the *information matrix* which is defined by

$$\boldsymbol{I}(\zeta) := \sum_{i=1}^{s} w_i \Phi(x_i)\Phi^\top(x_i)\,. \tag{14}$$

As a matter of fact, one major aspect of design of experiment theory seeks to maximize the information matrix over the set of all possible ζ. Notice the Loewner ordering $\succcurlyeq$ is partially ordered and, in general, there is no greatest element among all possible matrices $\boldsymbol{I}(\zeta)$. The standard approach is to consider some statistical criteria, namely the *Kiefer's ϕ_p-criteria* [23], in order to describe and construct the "*optimum designs*" with respect to those criteria. Observe that the *information matrix* $\boldsymbol{I}(\zeta)$ belongs to $\mathbb{S}_p^+$, the space of symmetric nonnegative definite matrices of size p, and define, for all $q \in [-\infty, 1]$, a criterion ϕ_q where for positive definite matrices M it holds

$$\phi_q(M) := \begin{cases} (\frac{1}{p}\mathrm{trace}(M^q))^{1/q} & \text{if } q \neq -\infty, 0 \\ \det(M)^{1/p} & \text{if } q = 0 \\ \lambda_{\min}(M) & \text{if } q = -\infty \end{cases}$$

and for nonnegative definite matrices M it reads $\phi_q(M) := (\frac{1}{p}\mathrm{trace}(M^q))^{1/q}$ if $q \in (0, 1]$, and zero otherwise. Those criteria are meant to be real valued, positively homogeneous, non constant, upper semi-continuous, isotonic (with respect to the Loewner ordering $\succcurlyeq$) and concave functions. In particular, we search for solutions to the following optimization problems

$$\zeta^\star \in \arg \max_{\zeta \text{ as in (13)}} \phi_q(\boldsymbol{I}(\zeta))\,, \tag{15}$$

where the maximum is taken over all design matrices ζ of the form (13) and $q \in [-\infty, 1]$.

Observe that the set of admissible designs described by (13) is any combination of s pairwise distinct support points x_i in the *design space* $\mathcal{X}$ and number of replications N_i at x_i such that $\sum_i N_i = N$. It appears that the set of admissible frequencies $w_i = N_i/N$ is discrete and contained in the set of rational numbers of the form a/N where a is an integer. Hence, notice that (15) is a discrete optimization problem with respect to frequencies w_i. To the best our of knowledge, this combinatorial problem is extremely difficult both analytically and computationally. A popular solution is then to consider "*approximate*" designs defined by

$$\zeta := \begin{pmatrix} x_1 & \cdots & x_s \\ w_1 & \cdots & w_s \end{pmatrix}, \tag{16}$$

where w_i are varying continuously from 0 to 1 and $\sum_{i=1}^s w_i = 1$. Accordingly, any solution to (15) where the maximum is taken over all matrices of type (16) is called "*approximate optimal design*".

Moreover, we assume that $\Phi \subset \mathbb{R}_n[x]^p$ where $\varphi_\ell(t) := \sum_{k\in\{0,\dots,n\}^d} \mathbf{a}_{\ell,k} t^k$. Notice that these assumptions cover a large class of problems in optimal design theory, see for instance [18, Chapter 5]. Define, for all $\mu \geq 0$, the information matrix (with an abuse of notation)

$$\boldsymbol{I}(\mu) = \Big(\int_{\mathcal{X}} \varphi_i\varphi_j \mathrm{d}\mu\Big)_{1\leq i,j\leq p} = \Big(\sum_{k,t\in\{0,\dots,d\}^n} \mathbf{a}_{i,k}\mathbf{a}_{j,t} m_{k+t}(\mu)\Big)_{1\leq i,j\leq p}.$$

Note that $\boldsymbol{I}(\mu) = \sum_{|\alpha|\leq 2d} m_\alpha(\mu)\,\mathbf{A}_\alpha$ where for all $\alpha \in \{0,\dots,2d\}^d$,

$$\mathbf{A}_\alpha := \Big(\sum_{k+\ell=\alpha} \mathbf{a}_{i,k}\mathbf{a}_{j,\ell}\Big)_{i,j}.$$

Further, set $\mu = \sum_{i=1}^\ell w_i\delta_{x_i}$ and observe that $\boldsymbol{I}(\mu) = \sum_{i=1}^\ell w_i\Phi(x_i)\Phi^\top(x_i)$ as in (14). Recall that the ϕ_q-criteria for $q \in [-\infty, 1]$ are isotonic with respect to the Loewner ordering $\succcurlyeq$ and then, for all $X \in \mathbb{S}_p^+$ and for all $\mu \in E^\star$,

$$\Big\{\sum_{\alpha\in\{0,\dots,2d\}^d} m_\alpha(\mu)\,\mathbf{A}_\alpha - X \succcurlyeq 0\Big\} \Rightarrow \Big\{\phi_q(\boldsymbol{I}(\mu)) \geq \phi_q(X)\Big\} \tag{17}$$

We deduce the following Linear Matrix Inequality (LMI) equivalent formulation of our problem

(18) $$\zeta^\star \in \arg \max_{X \in \mathcal{D}_0(\mathcal{X},\Phi)} \phi_q(X)\,,$$

where the feasible set $\mathcal{D}_0(\mathcal{X},\Phi)$ is given by

$$\mathcal{D}_0(\mathcal{X},\Phi) := \Big\{ X \in \mathbb{S}_p^+ \ :\ \sum_{|\alpha|\leq 2d} m_\alpha(\mu)\,\mathbf{A}_\alpha - X \succcurlyeq 0,\ \mu = \sum_{i=1}^{s} w_i \delta_{x_i} \geq 0,\ \sum_{i=1}^{s} w_i = 1 \Big\},$$

and designs ζ can be identified with atomic probabilities μ. In particular, note that $\zeta^\star$ is identified to $\mu^\star$ such that $X^\star = \sum_{\alpha\in\{0,\ldots,2d\}^d} m_\alpha(\mu^\star)\,\mathbf{A}_\alpha$, since that, by isotonicity, the constraint (17) is active at the solution point $X^\star$ of (18) .

Interestingly, one can show that (18) can be efficiently solved using moment-SoS hierarchies, see [16]. Indeed, one essentially need to represent $(m_\alpha)_{\alpha\in\{0,\ldots,2d\}^d}$ the first moments of measures $\mu = \sum_{i=1}^s w_i \delta_{x_i}$ that appears in the constraint of (18). In practice, we may witness finite convergence of the hierarchies so that the solution of the SDP relaxation is exactly the solution to (18). The optimal design points $x^\star$ are the roots of some SoS polynomial as in (2). This polynomial can be computed using Moment-SoS hierarchies in practice, see Figure 1.

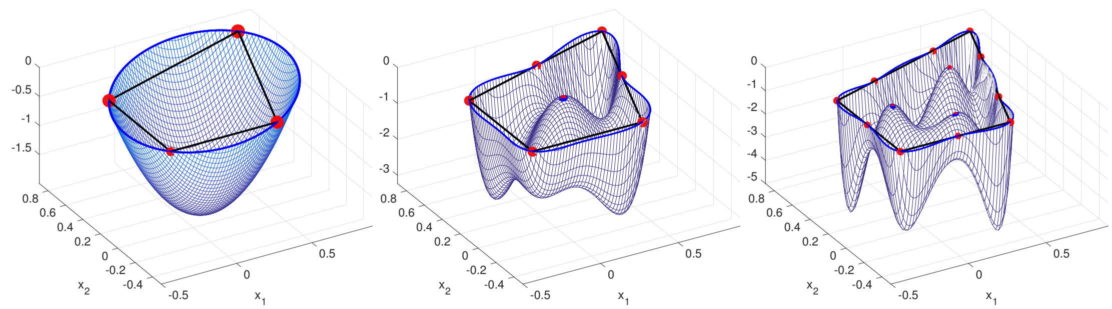

Figure 1. First SoS polynomials $p^\star$ (as in (2)) constructed on the Wynn polytope, see [16] for further details.

References

[1] F. Alizadeh. Interior point methods in semidefinite programming with applications to combinatorial optimization. *SIAM journal on Optimization*, 5(1):13–51, 1995.

[2] J.-M. Azaïs, Y. De Castro, and F. Gamboa. Spike detection from inaccurate samplings. *Applied and Computational Harmonic Analysis*, 38(2):177–195, 2015.

[3] B. Barak, J. A. Kelner, and D. Steurer. Dictionary learning and tensor decomposition via the sum-of-squares method. In *Proceedings of the forty-seventh annual ACM symposium on Theory of computing*, pages 143–151. ACM, 2015.

[4] B. Barak and D. Steurer. Sum-of-squares proofs and the quest toward optimal algorithms. *arXiv preprint arXiv:1404.5236*, 2014.

[5] H. H. Bauschke and P. L. Combettes. *Convex analysis and monotone operator theory in Hilbert spaces*. Springer Science & Business Media, 2011.

[6] G. Ben Arous, S. Mei, A. Montanari, and M. Nica. The landscape of the spiked tensor model. *arXiv preprint arXiv:1711.05424*, 2017.

[7] A. Beurling. Sur les intégrales de fourier absolument convergentes et leur application à une transformation fonctionnelle. In *Ninth Scandinavian Mathematical Congress*, pages 345–366, 1938.

[8] S. Boyd, L. El Ghaoui, E. Feron, and V. Balakrishnan. *Linear matrix inequalities in system and control theory*, volume 15. Siam, 1994.

[9] S. Boyd and L. Vandenberghe. Semidefinite programming relaxations of non-convex problems in control and combinatorial optimization. In *Communications, Computation, Control, and Signal Processing*, pages 279–287. Springer, 1997.

[10] E. J. Candès and C. Fernandez-Granda. Towards a mathematical theory of super-resolution. *Communications on Pure and Applied Mathematics*, 67(6):906–956, 2014.

[11] A. Cauchy. Méthode générale pour la résolution des systemes d'équations simultanées. *Comp. Rend. Sci. Paris*, 25(1847):536–538, 1847.

[12] M. Claeys, D. Arzelier, D. Henrion, and J. B. Lasserre. Moment lmi approach to ltv impulsive control. In *Decision and Control (CDC), 2013 IEEE 52nd Annual Conference on*, pages 5810–5815. IEEE, 2013.

[13] M. Claeys, D. Arzelier, D. Henrion, and J. B. Lasserre. Measures and lmis for impulsive nonlinear optimal control. *IEEE Transactions on Automatic Control*, 59(5):1374–1379, 2014.

[14] A. Cosse and L. Demanet. Stable rank one matrix completion is solved by two rounds of semidefinite programming relaxation. *arXiv preprint arXiv:1801.00368*, 2017.

[15] Y. De Castro and F. Gamboa. Exact reconstruction using beurling minimal extrapolation. *Journal of Mathematical Analysis and applications*, 395(1):336–354, 2012.

[16] Y. De Castro, F. Gamboa, D. Henrion, R. Hess, J.-B. Lasserre, et al. Approximate optimal designs for multivariate polynomial regression. *The Annals of Statistics*, 47(1):127–155, 2019.

[17] Y. De Castro, F. Gamboa, D. Henrion, and J.-B. Lasserre. Exact solutions to super resolution on semi-algebraic domains in higher dimensions. *IEEE Transactions on Information Theory*, 63(1):621–630, 2017.

[18] H. Dette and W. J. Studden. *The theory of canonical moments with applications in statistics, probability, and analysis*, volume 338. John Wiley & Sons, 1997.

[19] H. O. Fattorini. *Infinite dimensional optimization and control theory*, volume 54. Cambridge University Press, 1999.

[20] D. Hilbert. Über die darstellung definiter formen als summe von formenquadraten. *Mathematische Annalen*, 32(3):342–350, 1888.

[21] S. B. Hopkins, J. Shi, and D. Steurer. Tensor principal component analysis via sum-of-square proofs. In *Conference on Learning Theory*, pages 956–1006, 2015.

[22] L. V. Kantorovich. On the translocation of masses. In *Dokl. Akad. Nauk. USSR (NS)*, volume 37, pages 199–201, 1942.

[23] J. Kiefer. General equivalence theory for optimum designs (approximate theory). *The annals of Statistics*, pages 849–879, 1974.

[24] N. N. Krasovskii. Theory of motion control, 1968.

[25] J. B. Lasserre. Global optimization with polynomials and the problem of moments. *SIAM Journal on optimization*, 11(3):796–817, 2001.

[26] J.-B. Lasserre. *Moments, positive polynomials and their applications*, volume 1. World Scientific, 2009.

[27] J. B. Lasserre, D. Henrion, C. Prieur, and E. Trélat. Nonlinear optimal control via occupation measures and lmi-relaxations. *SIAM journal on control and optimization*, 47(4):1643–1666, 2008.

[28] L. W. Neustadt. Optimization, a moment problem, and nonlinear programming. *Journal of the Society for Industrial & Applied Mathematics, Series A: Control*, 2(1):33–53, 1964.

[29] M. Putinar. Positive polynomials on compact semi-algebraic sets. *Indiana University Mathematics Journal*, 42(3):969–984, 1993.

[30] G. Tang, B. N. Bhaskar, and B. Recht. Near minimax line spectral estimation. *Information Theory, IEEE Transactions on*, 61(1):499–512, 2015.

[31] G. Tang, B. N. Bhaskar, P. Shah, and B. Recht. Compressed sensing off the grid. *Information Theory, IEEE Transactions on*, 59(11):7465–7490, 2013.

École des Ponts ParisTech – CERMICS, 6 et 8 avenue Blaise Pascal, Cité Descartes – Champs sur Marne, FR-77455 Marne la Vallée Cedex 2

Email address: `yohann.de-castro at enpc.fr`