# Optimization of Polynomials with Sparsity Encoded in a Few Linear Forms

Jean-Bernard Lasserre

# Optimization of Polynomials with Sparsity Encoded in a Few Linear Forms ⋆

**Jean B. Lasserre** *

* *LAAS-CNRS & Institute of Mathematics, University of Toulouse, France (e-mail: lasserrer@ laas.fr).*

**Abstract:** We consider polynomials of a few linear forms and show how exploit this type of sparsity for optimization on some particular domains like the Euclidean sphere or a polytope. Moreover, a simple procedure allows to detect this form of sparsity and also allows to provide an approximation of any polynomial by such sparse polynomials.

*Keywords:* Optimization – Sparsity in Optimization

## 1. INTRODUCTION

In this paper we discuss the optimization problems

$$\mathbf{P}: \quad \min\{h(\mathbf{x}) : \mathbf{x} \in \mathbf{\Omega}\},$$

where the polynomial [1] $h \in \mathbb{R}[x_1, \ldots, x_n]$ is a defined in terms of a few linear forms, that is,

$$\mathbf{x} \mapsto h(\mathbf{x}) = f(\boldsymbol{\ell}^T\mathbf{x}), \quad \mathbf{x} \in \mathbb{R}^n, \tag{1}$$

for some polynomial $f \in \mathbb{R}[X_1, \ldots, X_m]$ and some real matrix $\boldsymbol{\ell} \in \mathbb{R}^{n\times m}$. We are also interested in approximating an arbitrary polynomial by polynomials in the form (1).

*Motivation.* When $m \ll n$, formulation (1) exhibits some sort of *sparsity* as only a few linear forms are involved in $h$. Indeed such a sparsity has been explored in several contexts like e.g. statistical learning in Roweis and Saul (2000), Camastra (2003), to learn a low-dimensional manifold (where $h$ is called a low-rank function), in Constantine et al. (2014) for contracting response surfaces on a low-dimensional subspace, in Baldoni et al. (2011) for multivariate integration on the simplex, and in Barvinok (2007) for integration with respect to the Gaussian measure. Therefore one also expects that it can be exploited for an efficient computation of the (local or global) minimum on $\mathbf{\Omega}$. But notice that in general $\mathbf{\Omega}$ is *not* expressed in terms of the $\boldsymbol{\ell}_j$'s so that exploiting this sparsity to optimize $h$ on $\mathbf{\Omega}$ may not be easy. For the sphere $\mathbb{S}^{n-1}$ Barvinok (2007) has shown that the maximum (but not the minimum) of certain sparse homogeneous polynomials can be approximated well by a properly scaled maximum on the unit sphere of a random low-dimensional subspace. This class of homogeneous polynomials contains some polynomials of the form (1). Notice also that the sparsity (1) (when $m \ll n$) is different from the various sparsity patterns exploited for polynomial optimization in Ahmadi and Majumdar (2019), Lasserre (2006), Wang et al. (2021a), and Wang et al. (2021b).

If $h$ is not directly available in sparse form (1), its detection is quite important in view of the potential resulting benefits for optimization. It turns out that the detection issue has been already addressed in engineering and data science, in the more general context of approximating an arbitrary continuous differentiable function $h(\boldsymbol{\ell}\mathbf{y} + \mathbf{s}\,\mathbf{v})$ where the columns of $\boldsymbol{\ell} \in \mathbb{R}^{n\times m}$, (resp. $\mathbf{s} \in \mathbb{R}^{n\times(n-m)}$) are eigenvectors of $\mathrm{E}_\mu[\nabla h \nabla h^T]$ associated with the $m$ largest (resp. the remaining $n-m$) eigenvalues, and $\mu$ is an appropriate probability measure. In Constantine et al. (2014) the authors discuss methods to obtain and evaluate an approximation based on the function $G(\mathbf{y}) := \mathrm{E}_\mu[\,h|\mathbf{y}\,]$; see below. (In Constantine et al. (2014) $h$ in (1) is called a **z**-invariant function.) Notice that if even if $h$ is a polynomial, the resulting approximation $G$ is *not*.

*Contribution.* Our contribution is threefold:

(i) We show that the sparsity in (1) can be exploited in optimization on the Euclidean sphere $\mathbb{S}^{n-1}$ and arbitrary polytopes. Solving the original problem reduces to solving an explicit optimization problem in $\mathbb{R}^m$, simply related and similar to $\mathbf{P}$, but with a drastic reduction in difficulty. We thus extend Lasserre (2021) who considered the case $\mathbf{\Omega} = \mathbb{S}^{n-1}$ and showed that solving $\mathbf{P}$ is equivalent to minimizing the $m$-variables polynomial $X \mapsto f(\mathcal{L}_1 \cdot X_1, \ldots, \mathcal{L}_m \cdot X_m)$ on the Euclidean ball $\mathcal{E}_m$, (where $\mathcal{L}_i$ is the $i$-th column of $\boldsymbol{\ell}^T\boldsymbol{\ell}$).

(ii) A second contribution is with respect to *detection* of a sparsity (1). When $h$ is a polynomial we provide two procedures. We first choose $\mu$ to be the uniform distribution on $\mathcal{E}_n$. Then we build a matrix $\mathbf{H}_k\mathbf{H}_k^T$ where the columns of $\mathbf{H}_k$ are just the gradient of $h$ evaluated at points $(\mathbf{x}(1), \ldots, \mathbf{x}(k)) \subset \mathcal{E}_n$ (randomly generated according to $\mu$), until the condition $\mathrm{rank}(\mathbf{H}_k\mathbf{H}_k^T) = \mathrm{rank}(\mathbf{H}_{k-1}\mathbf{H}_{k-1}^T)$ is satisfied, say for $k = m+1$. Then a sparsity as in (1) for some explicit $\boldsymbol{\ell} \in \mathbb{R}^{n\times m}$, is detected with probability 1. A second possibility that gets rid of "*with prob.* 1" is to follow Constantine et al. (2014) and perform the SVD decomposition of $\mathcal{R} := \mathrm{E}_\mu[\nabla h\, \nabla h^T]$. But in pour context, as $h$ is a polynomial and integration of polynomials on $\mathcal{E}_n$ is easy, $\mathcal{R}$ can be computed *exactly*. Then $\boldsymbol{\ell}$ in (1) is obtained from eigenvectors associated with the $m$ non-zero eigenvalues of $\mathcal{R}$.

⋆ Research sponsored by the Artificial and Natural Intelligence Institute (ANITI) of Toulouse, and ANR-NuSCAP-20-CE48-0014

[1] Most of what follows also applies to continuously differentiable functions

(iii) A third contribution is with respect to detection of an *approximate sparsity* and is directly inspired by the active set method, as described in e.g. Constantine et al. (2014). Write $h$ in the form

$$h(\mathbf{x}) \,=\, h(\boldsymbol{\ell}\,\mathbf{y} + \mathbf{s}\,\mathbf{z})\,, \quad \boldsymbol{\ell} \in \mathbb{R}^{n\times m}\ \mathbf{s} \in \mathbb{R}^{n\times(n-m)}\,,$$

where the columns of $\boldsymbol{\ell}, \mathbf{s}$ are the eigenvectors of $\mathcal{R}$ (with norm 1), and where the $n-m$ eigenvalues associated with $\mathbf{s}$ are much smaller than the $m$ eigenvalues associated with $\boldsymbol{\ell}$. In Constantine et al. (2014) the authors propose to approximate $h$ with the function $G(\boldsymbol{\ell}^T\mathbf{x})$ defined by:

$$G(\mathbf{y}) \,:=\, \mathrm{E}_\mu[\,h\,|\,\mathbf{y}] \,=\, \int h(\boldsymbol{\ell}\mathbf{y} + \mathbf{s}\,\mathbf{z})\,\pi(d\mathbf{z}|\mathbf{y})\,, \qquad (2)$$

where $\pi(d\mathbf{z}|\mathbf{y})$ is the conditional probability on $\mathbf{z}$ given $\mathbf{y}$. They propose to evaluate the integral (2) by Monte-Carlo sampling on $\mathbf{z}$. But this sample *depends* on $\mathbf{y}$ and therefore a sample has to be generated for each $\mathbf{y}$.

Our third contribution and novelty is to exploit that if $h$ is a polynomial and $\mu$ is the uniform distribution on $\mathcal{E}_n$, then after the simple scaling $\mathbf{v} \to \mathbf{z}/\sqrt{1-\|\mathbf{y}\|^2}$, $\pi(d\mathbf{v}|\mathbf{y})$ in (2) is the uniform distribution on $\mathcal{E}_{n-m}$. Therefore as the integrand is a polynomial in $\mathbf{v}$ of fixed degree, $G(\mathbf{y})$ is a polynomial in the $m+1$ variables $\mathbf{y}$ and $\sqrt{1-\|\mathbf{y}\|^2}$. Its coefficients can be obtained exactly, e.g. by direct integration term by term after expansion of the integrand in the monomial basis. Alternatively, $G$ can be expressed directly in terms of $h$ via a cubature formula on $\mathcal{E}_{n-m}$. Importantly, the cubature *does not depend* on $\mathbf{y}$. Then for optimization on say $\mathbb{S}^{n-1}$ or $\mathcal{E}_n$, instead of minimizing $h$, one proposes to minimize $G(X) = G(\boldsymbol{\ell}^T\mathbf{x})$ on $\mathcal{E}_m$. This in turn is equivalent to minimizing a related function $\hat{f}(X,|Y|)$ on $(X,Y) \in \mathbb{S}^m$ for some polynomial $\hat{f} \in \mathbb{R}[X_1,\ldots X_m, Y]$.

## 2. EXPLOITING SPARSITY FOR OPTIMIZATION

### *2.1 Notation and definitions*

Let $C^1(\mathbb{R}^n)$ be the space of continuously differentiable functions on $\mathbb{R}^n$. For any two vector $\mathbf{x},\mathbf{y} \in \mathbb{R}^n$ denote by $\mathbf{x}\cdot\mathbf{y}$ their usual scalar product. Given a vector space $V \subset \mathbb{R}^n$ denote by $V^\perp$ its orthogonal complement, i.e., $V^\perp = \{\,\mathbf{y} \in \mathbb{R}^n : \mathbf{x}\cdot\mathbf{y} = 0\,,\ \forall \mathbf{x} \in V\,\}$.

The following result is relatively straightforward and its proof is omitted.

*Proposition 1.* Let $f : \mathbb{R}^n \times \mathbb{R}^m :\to \mathbb{R}$, $(\mathbf{x},\mathbf{y}) \mapsto f(\mathbf{x},\mathbf{y})$, be continuously differentiable, and assume that $\nabla_{\mathbf{y}} f(\mathbf{x},\mathbf{y}) = 0$ for all $(\mathbf{x},\mathbf{y})$. Then with $\mathbf{y}_0 \in \mathbb{R}^m$ fixed, arbitrary:

$$f(\mathbf{x},\mathbf{y}) \,=\, f(\mathbf{x},\mathbf{y}_0) \,=:\, g(\mathbf{x})\,, \quad \forall (\mathbf{x},\mathbf{y}) \in \mathbb{R}^n \times \mathbb{R}^m\,, \quad (3)$$

and $g : \mathbb{R}^n \to \mathbb{R}$ is continuously differentiable.

### *2.2 Exploiting sparsity*

Let $h \in \mathbb{R}[\mathbf{x}] = \mathbb{R}[x_1,\ldots,x_n]$ and let

$$\boldsymbol{\ell} \,:=\, [\boldsymbol{\ell}_1, \boldsymbol{\ell}_2, \ldots, \boldsymbol{\ell}_m] \,\in\, \mathbb{R}^{n\times m}\,,$$

for some $m$ linearly independent column vectors $\boldsymbol{\ell}_1,\ldots\boldsymbol{\ell}_m \in \mathbb{R}^n$. Let $\mathbf{x}\cdot\mathbf{y}$ denote the usual scalar product of $\mathbf{x},\mathbf{y} \in \mathbb{R}^n$.

*Theorem 2.* Let $h \in C^1(\mathbb{R}^n)$. Then the two statements below are equivalent:

(a) There exists $f \in C^1(\mathbb{R}^m)$ and $(\boldsymbol{\ell}_i)_{i=1,\ldots,m} \subset \mathbb{R}^n$ such that $h(\mathbf{x}) = f(\boldsymbol{\ell}^T\mathbf{x}) = f(\boldsymbol{\ell}_1\cdot\mathbf{x},\ldots,\boldsymbol{\ell}_m\cdot\mathbf{x})$ for all $\mathbf{x} \in \mathbb{R}^n$.

(b) There exists an $m$-dimensional vector space $V \subset \mathbb{R}^n$ such that $\nabla h(\mathbf{x}) \in V$ for all $\mathbf{x} \in \mathbb{R}^n$.

**Proof.** (a) $\Rightarrow$ (b) is straightforward as

$$\nabla h(\mathbf{x}) \,=\, \boldsymbol{\ell}\,\nabla f(\boldsymbol{\ell}^T\mathbf{x}) \,=\, \sum_{i=1}^m \boldsymbol{\ell}_i\,\frac{\partial f(X)}{\partial X_i}, \quad \forall \mathbf{x} \in \mathbb{R}^n\,.$$

where $X_i = \boldsymbol{\ell}_i\cdot\mathbf{x}$, $i = 1,\ldots,m$. Equivalently, $\nabla h(\mathbf{x}) \in V$ for all $\mathbf{x} \in \mathbb{R}^n$, where $V := \mathrm{Span}(\boldsymbol{\ell}_1,\ldots,\boldsymbol{\ell}_m) \subset \mathbb{R}^n$, which is clearly statement (b).

(b) $\Rightarrow$ (a). Let $V \subset \mathbb{R}^n$ have dimension $m < n$ and let $(\boldsymbol{\ell}_i)_{i=1,\ldots,m}$ be a basis of $V$. Similarly, let $(\mathbf{s}_j)_{j=1,\ldots n-m} \subset \mathbb{R}^n$ be a basis of $V^\perp$ and write $\mathbf{x} = \boldsymbol{\ell}\,\mathbf{u} + \mathbf{s}\,\mathbf{v}$, with matrices $\boldsymbol{\ell} = [\boldsymbol{\ell}_1,\ldots\,\boldsymbol{\ell}_m] \in \mathbb{R}^{n\times m}$ and $\mathbf{s} = [\mathbf{s}_1,\ldots\,\mathbf{s}_{n-m}] \in \mathbb{R}^{n\times(n-m)}$, and such that $\mathbf{s}_i^T\boldsymbol{\ell} = 0$ for all $i = 1,\ldots,n-m$. Notice that

$$\mathbf{u} \,=\, (\boldsymbol{\ell}^T\,\boldsymbol{\ell})^{-1}\boldsymbol{\ell}^T\,\mathbf{x}\,; \quad \mathbf{v} \,=\, (\mathbf{s}^T\,\mathbf{s})^{-1}\mathbf{s}^T\,\mathbf{x}\,. \qquad (4)$$

Hence write $h(\mathbf{x})$ as

$$h(\boldsymbol{\ell}\,\mathbf{u}+\mathbf{s}\,\mathbf{v}) \,=:\, \phi(\mathbf{u}\,,\mathbf{v}) \,=\, \phi((\boldsymbol{\ell}^T\,\boldsymbol{\ell})^{-1}\boldsymbol{\ell}^T\,\mathbf{x}\,,\,(\mathbf{s}^T\,\mathbf{s})^{-1}\mathbf{s}^T\,\mathbf{x}\,),$$

for some function $\phi : \mathbb{R}^n \to \mathbb{R}$. Then $\phi \in C^1(\mathbb{R}^n)$ follows from $h \in C^1(\mathbb{R}^n)$. Next, by the chain rule of differentiation:

$$\nabla h(\mathbf{x}) \,=\, \boldsymbol{\ell}\,(\boldsymbol{\ell}^T\,\boldsymbol{\ell})^{-1}\nabla_{\mathbf{u}}\phi(\mathbf{u},\mathbf{v}) + \mathbf{s}\,(\mathbf{s}^T\,\mathbf{s})^{-1}\nabla_{\mathbf{v}}\phi(\mathbf{u},\mathbf{v})\,.$$

Observe that $\mathbf{s}^T\cdot\nabla h(\mathbf{x}) = 0$ for all $\mathbf{x} \in \mathbb{R}^n$ and all $i = 1,\ldots,n-m$, because $\nabla h(\mathbf{x}) \in V$ for all $\mathbf{x} \in \mathbb{R}^n$. Hence

$$0 \,=\, \mathbf{s}^T\cdot\nabla h(\mathbf{x}) \,=\, \nabla_{\mathbf{v}}\phi(\mathbf{u},\mathbf{v})\,, \quad \forall \mathbf{x} \in \mathbb{R}^n\,,$$

and therefore $\nabla_{\mathbf{v}}\phi(\mathbf{u},\mathbf{v}) = 0$, for all $\mathbf{v} \in \mathbb{R}^n$. By Proposition 1 applied to $\phi$, $\phi(\mathbf{u},\mathbf{v}) = \phi(\mathbf{u},\mathbf{v}_0)$ for all $\mathbf{u},\mathbf{v}$, where where $\mathbf{v}_0$ is arbitrary. Letting $\mathbf{v}_0 := 0$ yields

$$\begin{aligned} h(\mathbf{x}) \,=\, \phi(\mathbf{u},\mathbf{v}) \,=\, \phi(\mathbf{u},0) &= \phi((\boldsymbol{\ell}^T\boldsymbol{\ell})^{-1}\boldsymbol{\ell}^T\mathbf{x},0) \\ &= f(\tilde{\boldsymbol{\ell}}_1\cdot\mathbf{x},\ldots,\tilde{\boldsymbol{\ell}}_m\cdot\mathbf{x})\,, \end{aligned}$$

where $\tilde{\boldsymbol{\ell}}_i \in \mathbb{R}^n$ is the $i$-th row of $(\boldsymbol{\ell}^T\boldsymbol{\ell})^{-1}\boldsymbol{\ell}^T$, $i = 1,\ldots,m$, and $f(X_1,\ldots,X_m) := \phi(X_1,\ldots,X_m,0)$ for all $X \in \mathbb{R}^m$.

### *2.3 Detection of the sparse form*

In this section we suppose that $h \in \mathbb{R}[\mathbf{x}]$ is a sparse polynomial but *not* given in its sparse form $\mathbf{x} \mapsto f(\boldsymbol{\ell}^T\mathbf{x})$ for some real matrix $\boldsymbol{\ell} \in \mathbb{R}^{n\times m}$. In view of Theorem 2, it suffices to determine a basis of the $m$-dimensional subspace $V \subset \mathbb{R}^n$ to which $\nabla h$ belongs.

We consider a methodology inspired from Constantine et al. Constantine et al. (2014) but we here exploit that $h$ is a polynomial. Introduce a probability measure $\mu$ on a certain domain, e.g. the uniform distribution on $\mathcal{E}_n$:

(i) A first possibility consists in computing the $n\times n$ real symmetric matrix

$$\mathbf{M}_\mu \,:=\, \mathrm{E}_\mu[\,\nabla h(\mathbf{x})\,\nabla h(\mathbf{x})^T\,]\,,$$

and compute its SVD decomposition. In Constantine et al. (2014) the function $h$ is *not* a polynomial and therefore $\mathrm{E}_\mu$ must be approximated. Moreover $h$ is not necessarily sparse and the authors are interested in approximating $h$ in the subspace $V$ generated by the eigenvectors associated

with the largest eigenvalues of $\mathbf{M}_\mu$. In our setting, $\mathbf{M}_\mu$ can be computed *exactly* as one knows how to integrate *exactly* a polynomial on $\mathcal{E}_n$, and $V$ is spanned by the eigenvectors associated with the zero-eigenvalues of $\mathbf{M}_\mu$.

(ii) Another possibility is to consider a sample of $(m+1)$ i.i.d. vectors $(\nabla h(\mathbf{x}(i))_{i\le m+1} \subset \mathcal{E}_n$ randomly generated according to $\mu$, and construct the empirical matrix

$$\mathbf{H}_{m+1} := [\nabla h(\mathbf{x}(1)), \cdots, \nabla h(\mathbf{x}(m+1))] \in \mathbb{R}^{n\times(m+1)} \quad (5)$$

until one observes that $\text{rank}(\mathbf{H}_\ell^T\,\mathbf{H}_\ell) = m$, $\ell = m, m+1$.

*Theorem 3.* Let $\mathbf{H}_k := [\nabla h(\mathbf{x}(1)), \cdots, \nabla h(\mathbf{x}(k))] \in \mathbb{R}^{n\times k}$ be as in (5), and let $V := \text{span}\{\nabla h(\mathbf{x}) : \mathbf{x} \in \mathcal{E}_n\}$. Then $\dim(V) = m$ if and only, with probability 1:

$$\text{rank}(\mathbf{H}_\ell^T\,\mathbf{H}_\ell) \;=\; m\,, \quad \ell = m, m+1\,. \quad (6)$$

**Proof.** The *Only if part* is straightforward. Indeed in view of the definition of $V$, suppose that $\dim(V) = m$, and let $V^\perp$ denote its direct complement (hence of dimension $n-m$). Then $\mathbf{u}^T\nabla h(\mathbf{x}(i)) = 0$ for all $\mathbf{u} \in V^\perp$ and all $i = 1,\ldots,m$, which implies $\text{rank}(\mathbf{H}_\ell^T\,\mathbf{H}_\ell) \le m$, $\ell = m, m+1$.

Next, observe that

$$(\mathbf{H}_m^T\,\mathbf{H}_m)_{i,j} \;=\; \nabla h(\mathbf{x}(i))^T\nabla h(\mathbf{x}(j))\,, \quad i,j \le m\,,$$

and therefore

$$\det(\mathbf{H}_m^T\,\mathbf{H}_m) \;=\; p_m(\mathbf{x}(1),\mathbf{x}(2),\ldots,\mathbf{x}(m))\,, \quad (7)$$

for some polynomial $p_m \in \mathbb{R}[\mathbf{u}_1,\ldots,\mathbf{u}_m]$. As $\dim(V) = m$ then necessarily $p_m \neq 0$. Next, let $\mu^{\otimes m}$ be the product measure $\underbrace{\mu\otimes\mu\cdots\otimes\mu}_{m\ times}$ on $(\mathcal{E}_n)^m$. As $p_m \neq 0$ then $\mu^{\otimes m}(\{\mathbf{u} : p_m(\mathbf{u}_1,\ldots,\mathbf{u}_m) = 0\}) = 0$, or equivalently, with probability 1, $p_m(\mathbf{u}_1,\ldots,\mathbf{u}_m) \neq 0$, i.e., $\det(\mathbf{H}_m^T\,\mathbf{H}_m) \neq 0$, and so $\text{rank}(\mathbf{H}_m^T\mathbf{H}_m) = m$. Next consider the case $\ell = m+1$. As $\dim(V) = m$ then necessarily the family $(\nabla h(\mathbf{x}(i)))_{i\le m+1}$ is *not* linearly independent and therefore $\text{rank}(\mathbf{H}_{m+1}^T\mathbf{H}_{m+1}) < m+1$, which from what precedes, yields $\text{rank}(\mathbf{H}_{m+1}^T\mathbf{H}_{m+1}) = m$ with probability 1.

*If part.* As above, let

$$\det(\mathbf{H}_k^T\,\mathbf{H}_k) \;=:\; p_k(\mathbf{x}(1),\mathbf{x}(2),\ldots,\mathbf{x}(k))\,, \quad k\in\mathbb{N}\,, \quad (8)$$

for some polynomial $p_k \in \mathbb{R}[\mathbf{u}_1,\ldots,\mathbf{u}_k]$. The condition

"with probability 1, $\text{rank}(\mathbf{H}_\ell^T\,\mathbf{H}_\ell) = m\,, \quad \ell = m, m+1$",

is equivalent to

$$\text{"with probability 1: } \begin{cases} \det(\mathbf{H}_m^T\,\mathbf{H}_m) > 0\,, \text{ and } \text{„} \\ \det(\mathbf{H}_{m+1}^T\,\mathbf{H}_{m+1}) = 0\,, \end{cases}$$

which in turn is equivalent to

$$p_m \neq 0 \text{ and } p_{m+1} = 0\,, \quad (9)$$

with $p_m$ as in (8). The condition $p_{m+1} = 0$, i.e.,

$$\det(\mathbf{H}_{m+1}(\mathbf{u}_1,\ldots,\mathbf{u}_{m+1})^T\mathbf{H}_{m+1}(\mathbf{u}_1,\ldots,\mathbf{u}_{m+1})) = 0\,,$$

for all $\mathbf{u} := (\mathbf{u}_1,\ldots\mathbf{u}_{m+1}) \in (\mathcal{E}_n)^{m+1}$, implies that there exists a vector $0 \neq q^{\mathbf{u}} \in \mathbb{R}^{m+1}$ such that

$$\mathbf{H}_{m+1}(\mathbf{u}_1,\ldots,\mathbf{u}_{m+1})\,q^{\mathbf{u}} \;=\; 0\,, \quad \forall \mathbf{u} \in (\mathcal{E}_n)^{m+1}\,.$$

That is,

$$\sum_{i=1}^{m+1} q_i^{\mathbf{u}}\,\nabla h(\mathbf{u}_i) \;=\; 0\,, \quad \forall \mathbf{u} \in (\mathcal{E}_n)^{m+1}\,.$$

Next, let $S := \{\mathbf{u}_{m+1} \in \mathcal{E}_n : q_{m+1}^{\mathbf{u}} = 0\}$ and $\Theta = (\mathcal{E}_n)^m \times S$, so that $\mu^{\otimes(m+1)}(\Theta) = \mu(S)$. Hence

$$\sum_{i=1}^{m} q_i^{\mathbf{u}}\,\nabla h(\mathbf{u}_i) \;=\; 0\,, \quad \text{for all } \mathbf{u} \in \Theta\,.$$

Next, from $\det(\mathbf{H}_m(\mathbf{u}_1,\ldots,\mathbf{u}_m)^T\,\mathbf{H}_m(\mathbf{u}_1,\ldots,\mathbf{u}_m)) > 0$, we also deduce that

$$\sum_{i=1}^{m} q_i^{\mathbf{u}}\,\nabla h(\mathbf{u}_i) \;\neq\; 0\,, \quad \text{for a.a. } \mathbf{u} \in (\mathcal{E}_n)^{m+1}\,. \quad (10)$$

This yields $0 = \mu^{\otimes(m+1)}(\Theta) = \mu(S)$. Therefore, letting $\mathbf{u}(\mathbf{x}) := (\mathbf{u}_1,\ldots,\mathbf{u}_m,\mathbf{x}) \in (\mathcal{E}_n)^m \times (\mathcal{E}_n \setminus S)$,

$$\nabla h(\mathbf{x}) \;=\; \frac{1}{q_1^{\mathbf{u}(\mathbf{x})}} \sum_{i=1}^{m} q_i^{\mathbf{u}(\mathbf{x})}\,\nabla h(\mathbf{u}_i)\,,$$

for all $\mathbf{x} \in \mathcal{E}_n \setminus S$, and all $(\mathbf{u}_1,\ldots,\mathbf{u}_m) \in (\mathcal{E}_n)^m$. Hence with $(\mathbf{u}_2,\ldots,\mathbf{u}_{m+1}) \in (\mathcal{E}_n)^m$, fixed, arbitrary:

$$\nabla h(\mathbf{x}) \in \text{span}(\nabla h(\mathbf{u}_1),\ldots,\nabla h(\mathbf{u}_m)) =: V\,, \quad (11)$$

for all $\mathbf{x} \in \mathcal{E}_n \setminus S$, and $V$ is an $m$-dimensional vector space. To show that (11) holds for all $\mathbf{x} \in \mathbb{R}^n$, observe that

$$\mathbf{v}^T\nabla h(\mathbf{x}) \;=\; 0\,, \quad \forall \mathbf{x} \in \mathcal{E}_n \setminus S\,,\, \forall \mathbf{v} \in V^\perp. \quad (12)$$

Hence for fixed $\mathbf{v} \in V^\perp$, the polynomial $\mathbf{x} \mapsto \mathbf{v}^T\nabla h(\mathbf{x})$ vanishes on $\mathcal{E}_n \setminus S$ with $\mu(S) = 0$, which implies that $\mathbf{v}^T\nabla h(\mathbf{x})$ vanishes on the whole $\mathcal{E}_n$ and hence on the whole $\mathbb{R}^n$. As this is true for an arbitrary $\mathbf{v} \in V^\perp$, we obtain that $\nabla h(\mathbf{x}) \in V$ for all $\mathbf{x} \in \mathbb{R}^n$.

In practice, Theorem 3 is used as follows:

- Samples $k$ points $(\mathbf{x}(i))_{i\le k}$ according to $\lambda$ on $[0,1]^n$.
- Do the SVD decomposition of the real symmetric matrices $\mathbf{H}_{k-1}^T\mathbf{H}_{k-1}$. and $\mathbf{H}_k^T\mathbf{H}_k$.
- If $\text{rank}(\mathbf{H}_k^T\mathbf{H}_k) \neq \text{rank}(\mathbf{H}_{k-1}^T\mathbf{H}_{k-1})$ then set $k := k+1$ and repeat.
- If $\text{rank}(\mathbf{H}_k^T\mathbf{H}_k) = \text{rank}(\mathbf{H}_{k-1}^T\mathbf{H}_{k-1})$ then stop. Set $V := \text{span}\{\nabla h(\mathbf{x}(1),\ldots,\nabla h(\mathbf{x}(k-1))\}$.

### *2.4 Some applications in Optimization*

*Optimization on the Euclidean unit sphere* A first application was developed in Lasserre (2021) for optimization on the Euclidean unit sphere $\mathbb{S}^{n-1}$. Namely, let $h, f$ and $\boldsymbol{\ell}$ be as in Theorem 2(a). Then it was shown in that

$$\rho = \min_{\mathbf{x}}\{\, h(\mathbf{x}) \,:\, \mathbf{x} \in \mathbb{S}^{n-1}\,\} \quad (13)$$

$$= \min_{\mathbf{y}}\{\, f(\mathbf{L}_1\cdot\mathbf{y}_1,\ldots,\mathbf{L}_m\cdot\mathbf{y}_m) \,:\, \mathbf{y} \in \mathcal{E}_m\,\}\,, \quad (14)$$

with $\mathbf{L}_i$ is the $ith$-row of the matrix $(\boldsymbol{\ell}^T\boldsymbol{\ell})^{1/2}$, $i = 1,\ldots,m$. In fact all points $\mathbf{x}^* \in \mathbb{S}^{n-1}$ that satisfy the standard second-order necessary conditions of optimality for problem (13) are in one-to-one correspondence with the points $\mathbf{y}^* \in \mathcal{E}_m$ that satisfy the standard second-order necessary conditions of optimality for problem (14).

Hence in this case one has replaced optimization of the $n$-variate polynomial $h$ on the non convex set $\mathbb{S}^{n-1}$ by optimization of the $m$-variate polynomial $f$ of same degree on the (convex) unit Euclidean ball. If $m \ll n$ then it yields drastic computational savings.

*Optimization on a polytope* Next, let $\boldsymbol{\Omega} = \{\,\mathbf{x} \in \mathbb{R}^n_+ : \mathbf{A}\mathbf{x} = \mathbf{b}\,\}$ for some real matrix $\mathbf{A} \in \mathbb{R}^{s\times n}$, and consider the optimization problem:

$$\rho \;=\; \min_{\mathbf{x}}\{\, h(\mathbf{x}) \,:\, \mathbf{x} \in \boldsymbol{\Omega}\,\}\,. \quad (15)$$

*Theorem 4.* Let $h$ and $f$ be as in Theorem 2(a) with $\boldsymbol{\ell} \in \mathbb{R}^{n\times m}$, and let $(\boldsymbol{\lambda}_i, \mathbf{u}_i)_{i\in I}$ be a set of generators of the polyhedral convex cone

$$C := \{(\boldsymbol{\lambda}, \mathbf{u}) \in \mathbb{R}^s \times \mathbb{R}^m : \mathbf{A}^T\boldsymbol{\lambda} \geq \boldsymbol{\ell}\,\mathbf{u}\}. \tag{16}$$

Then with $\rho$ as in (15)

$$\rho = \min_{X\in\mathbb{R}^m} \{ f(X) : \mathbf{u}_i \cdot X \leq \boldsymbol{\lambda}_i \cdot \mathbf{b}, \quad \forall\, i \in I \} \tag{17}$$

**Proof.** Let $X$ be fixed. By Farkas Lemma,

$$\emptyset \neq \{\mathbf{x} : \boldsymbol{\ell}^T\mathbf{x} = X\,;\, \mathbf{x} \in \boldsymbol{\Omega}\} \Leftrightarrow \mathbf{u}\cdot X \leq \boldsymbol{\lambda}\cdot\mathbf{b},$$

for all $(\boldsymbol{\lambda}, \mathbf{u}) \in C$, which in turn is equivalent to $\mathbf{u}_i \cdot X \leq \boldsymbol{\lambda}_i \cdot \mathbf{b}$ for all $(\boldsymbol{\lambda}_i, \mathbf{u}_i)_{i\in I}$. Then observe that

$$\begin{aligned}
\mathbf{P}: \quad \rho &= \min_{\mathbf{x}} \{ h(\mathbf{x}) : \mathbf{x} \in \boldsymbol{\Omega} \} \\
&= \min_{\mathbf{x},X} \{ f(X) : X = \boldsymbol{\ell}^T\mathbf{x};\, \mathbf{x} \in \boldsymbol{\Omega} \} \\
&= \min_{X} \{ f(X) : \mathbf{u}_i \cdot X \leq \boldsymbol{\lambda}_i \cdot \mathbf{b}, \quad \forall\, i \in I \}.
\end{aligned} \tag{18}$$

Notice that one has replaced an $n$-dimensional optimization problem on the polyhedron $\boldsymbol{\Omega} \subset \mathbb{R}^n$ by an $m$-dimensional optimization problem on the polyhedron $\boldsymbol{\Omega}_m := \{X \in \mathbb{R}^m : \mathbf{u}_i \cdot X \leq \boldsymbol{\lambda}_i \cdot \mathbf{b},\, i \in I\} \subset \mathbb{R}^m$. Of course this transformation requires to compute as a prerequisite step, all generators of the convex cone $C$ in (16). If one wants to avoid this, one possibility is to proceed as follows:

• Start with a set $I_0 := \{(\boldsymbol{\lambda}_0, \mathbf{u}_0)\}$ for some $(\boldsymbol{\lambda}_0, \mathbf{u}_0) \in C$, and set $k = 0$.

• Step $k$. Solve

$$\mathbf{P}_k: \quad \tau_k = \min_X \{f(X) : \mathbf{u}_i \cdot X \leq \boldsymbol{\lambda}_i \cdot \mathbf{b}, \quad i \in I_k\},$$

to obtain $X_k^* \in \mathbb{R}^m$. Next, solve the linear program

$$\begin{aligned}
\tau = \min_{\boldsymbol{\lambda}^+,\boldsymbol{\lambda}^-,\mathbf{u}^+,\mathbf{u}^-} \; & \{(\boldsymbol{\lambda}^+ - \boldsymbol{\lambda}^-)\cdot\mathbf{b} - (\mathbf{u}^+ - \mathbf{u}^-)\cdot X_k^* : \\
& ((\boldsymbol{\lambda}^+ - \boldsymbol{\lambda}^-), (\mathbf{u}^+ - \mathbf{u}^-)) \in C\,; \\
& \sum_t \lambda_t^+ + \lambda_t^- + \sum_j (u_j^+ + u_j^-) = 1\}.
\end{aligned}$$

If $\tau = 0$ then stop. Otherwise set $I_{k+1} := I_k \cup \{(\boldsymbol{\lambda}_*, \mathbf{u}_*)\}$ for an optimal solution $(\boldsymbol{\lambda}_*^+ - \boldsymbol{\lambda}_*^-, \mathbf{u}_*^+ - \mathbf{u}_*^-)$, set $k := k+1$ and go to step $k$.

With this strategy one has to solve a sequence of optimization problems $(\mathbf{P}_k)_{k\in\mathbb{N}}$ with same criterion $f$, but on tighter and tighter outer approximations of the convex polyhedron $\{X \in \mathbb{R}^m : \mathbf{u}_i \cdot X \leq \boldsymbol{\lambda}_i \cdot \mathbf{b},\, ,\, i \in I\}$. So the overall complexity of this algorithm is governed by the computational complexity of problem $\mathbf{P}_k$.

For simple sets $\boldsymbol{\Omega}$ like the canonical simplex or the unit box, the cone $C$ has a simple expression.

*On the canonical simplex* $\boldsymbol{\Omega} = \{\mathbf{x} \in \mathbb{R}^n_+ : \mathbf{e}\cdot\mathbf{x} = 1\}$. $C = \{(\lambda, \mathbf{u}) : \lambda\,\mathbf{e} \geq \boldsymbol{\ell}\,\mathbf{u}\}$.

*On the Box* $\boldsymbol{\Omega} = [-1, 1]^n$. $C = \{(\boldsymbol{\lambda}^+, \boldsymbol{\lambda}^- \geq 0, \mathbf{u}) : \boldsymbol{\lambda}^+ - \boldsymbol{\lambda}^- = \boldsymbol{\ell}\,\mathbf{u}\}$ and so $(\boldsymbol{\lambda}^+ + \boldsymbol{\lambda}^-)\cdot\mathbf{e} = \|\boldsymbol{\ell}\mathbf{u}\|_1$.

## 3. APPROXIMATE SPARSITY

In this section $h \in \mathbb{R}[\mathbf{x}]$ and we now assume that $h$ is not exactly in the form $f(\boldsymbol{\ell}^T\mathbf{x})$ for some $\boldsymbol{\ell} \in \mathbb{R}^{n\times m}$. Let $\mu_n$ be the uniform distribution on $\mathcal{E}_n$ and let

$$\mathbf{M}(\mu) := \mathrm{E}_{\mu_n}[\,\nabla h(\mathbf{x})^T\,\nabla h(\mathbf{x})\,] = [\boldsymbol{\ell}, \mathbf{s}] \begin{bmatrix} \Lambda_1 & 0 \\ 0 & \Lambda_2 \end{bmatrix} [\boldsymbol{\ell}, \mathbf{s}]^T$$

where now $\boldsymbol{\ell} = [\boldsymbol{\ell}_1, \ldots, \boldsymbol{\ell}_m] \in \mathbb{R}^{n\times m}$ (resp. $\mathbf{s} = [\mathbf{s}_1, \ldots, \mathbf{s}_{n-m}] \in \mathbb{R}^{n\times(n-m)}$ is the matrix eigenvectors of $\mathbf{M}(\mu_n)$ associated with the first $m$ (nonnegative) eigenvalues $\lambda_1, \ldots, \lambda_m$ (resp. the remaining $n-m$ eigenvalues $\lambda_{m+1}, \ldots, \lambda_n$) arranged in decreasing order and which also form the diagonal elements of the diagonal matrices $\Lambda_1$ and $\Lambda_2$, respectively. The vectors $\boldsymbol{\ell}_j, \mathbf{s}_j$ form an orthonormal basis. Therefore if one writes $\mathbf{x} = \boldsymbol{\ell}\,\mathbf{y} + \mathbf{s}\,\mathbf{z}$ with $\mathbf{y} \in \mathbb{R}^m$ and $\mathbf{z} \in \mathbb{R}^{n-m}$, then

$$\|\mathbf{x}\|^2 = \|\mathbf{y}\|^2 + \|\mathbf{z}\|^2,$$

and so the support of the marginal $\pi_{\mathbf{y}}$ of $\mu_n$ is $\mathcal{E}_m$, with density (w.r.t. Lebesgue) the pushforward of $\mu_n$ by its projection on $\mathcal{E}_m$. The conditional $\pi(d\mathbf{z}|\mathbf{y})$ is the uniform probability distribution on the ball $\mathcal{E}_{n-m}(\mathbf{y}) := \{\mathbf{z} : \|\mathbf{z}\|^2 \leq 1 - \|\mathbf{y}\|^2\}$. Proceeding as in Constantine et al. (2014), introduce the function $f : \mathbb{R}^m \to \mathbb{R}$, defined by

$$\begin{aligned}
f(\mathbf{y}) &:= \mathrm{E}[h|\mathbf{y}] \\
&= \int_{\mathcal{E}_{n-m}(\mathbf{y})} h(\boldsymbol{\ell}\,\mathbf{y} + \mathbf{s}\,\mathbf{z})\,\pi(d\mathbf{z}|\mathbf{y}), \quad \forall \mathbf{y} \in \mathcal{E}_m.
\end{aligned} \tag{19}$$

Then the idea promoted in Constantine et al. (2014) for some applications, is to approximate $h$ on $\mathcal{E}_n$ with the function $\hat{h}(\mathbf{x}) := f(\boldsymbol{\ell}^T\mathbf{x})$. The rationale being:

*Theorem 5.* ((Constantine et al., 2014, Theorem 3.1)) With $\mu_n$ being the uniform probability distribution on $\mathcal{E}_n$, and $\hat{h}(\mathbf{x}) = f(\boldsymbol{\ell}^T\mathbf{x})$, with $f$ as in (19),

$$\mathrm{E}_{\mu_n}[\,(h - \hat{h})^2] \leq C\,(\lambda_{m+1} + \ldots, +\lambda_n). \tag{20}$$

where the constant $C$ does not depend on $h$.

So in view of (20), if the remaining eigenvalues $\lambda_{m+1}, \ldots, \lambda_n$ are small then $\hat{h}$ provides a good approximation of $h$ in $L^2(\mathcal{E}_n)$.

*Exact computation of the approximand f*

Observe that $\pi(d\mathbf{z}|\mathbf{y}) = d\mathbf{z}/C_{n-m}(1 - \|\mathbf{y}\|^2)^{(n-m)/2}$ on $\mathcal{E}_{n-m}(\mathbf{y})$, for a constant $C_{n-m}$. Therefore by doing the change of variable $\mathbf{v} := \mathbf{z}/\sqrt{1 - \|\mathbf{y}\|^2} \in \mathcal{E}_{n-m}$, and letting $\tau(\mathbf{y}) := 1 - \|\mathbf{y}\|^2$, (19) reads:

$$\begin{aligned}
\mathrm{E}[h|\mathbf{y}] &= \frac{1}{C_{n-m}} \int_{\mathcal{E}_{n-m}} h(\boldsymbol{\ell}\,\mathbf{y} + \tau(\mathbf{y})^{1/2}\,\mathbf{s}\,\mathbf{v})\,d\mathbf{v}, \\
&= \int_{\mathcal{E}_{n-m}} h(\boldsymbol{\ell}\,\mathbf{y} + \tau(\mathbf{y})^{1/2}\,\mathbf{s}\,\mathbf{v})\,d\mu_{n-m}(\mathbf{v}),
\end{aligned} \tag{21}$$

for all $\mathbf{y} \in \mathcal{E}_m$. Observe that the integrand $\mathbf{v} \mapsto h(\boldsymbol{\ell}\,\mathbf{y} + \tau(\mathbf{y})^{1/2}\,\mathbf{s}\,\mathbf{v})$ is a polynomial of fixed degree, say $d$, in $\mathbf{v}$. Therefore it can be integrated exactly on $\mathcal{E}_{n-m}$. Equivalently one can also use a degree-$d$ cubature rule for Lebesgue measure on $\mathcal{E}_{n-m}$ to obtain:

$$f(\mathbf{y}) = \sum_{j=1}^r \theta_j\, h(\boldsymbol{\ell}\,\mathbf{y} + \tau(\mathbf{y})^{1/2}\mathbf{s}\,\mathbf{v}_j), \tag{22}$$

for some positive weights $(\theta_j)$ and cubature points $(\mathbf{v}_j) \subset \mathcal{E}_{n-m}$. Importantly, and in contrast to the function $G(\mathbf{y})$ in (Constantine et al., 2014, (3.10)), the cubature points $(\mathbf{v}_j)$ do *not* depend on $\mathbf{y}$ and so can be computed once and for

all [2]. Notice that $f$ is a polynomial in the $(m+1)$ variables $(y_1,\ldots,y_m,\sqrt{1-\|\mathbf{y}\|^2})$, i.e., $f\in\mathbb{R}[\mathbf{y},\sqrt{1-\|\mathbf{y}\|^2}]$. Next, again following Constantine et al. (2014) we approximate $h$ on $\mathcal{E}_n$ with $h(\mathbf{x})\approx\hat{h}(\mathbf{x}):=f(\boldsymbol{\ell}^T\mathbf{x})$, i.e.:

$$\hat{h}(\mathbf{x}) \;=\; \sum_{j=1}^{r}\theta_j\, h(\boldsymbol{\ell}\,\boldsymbol{\ell}^T\mathbf{x}+\tau(\boldsymbol{\ell}^T\mathbf{x})^{1/2}\mathbf{s}\,\mathbf{v}_j)\,. \qquad (23)$$

Hence letting $X:=\boldsymbol{\ell}^T\mathbf{x}$ and using the orthogonality of the $(\boldsymbol{\ell}_j)$, we obtain $X\in\mathcal{E}_m$, and

$$f(X) \;=\; \sum_{j=1}^{r}\theta_j\, h(\boldsymbol{\ell}\,X+(1-\|X\|^2)^{1/2}\,\mathbf{s}\,\mathbf{v}_j)\,. \qquad (24)$$

Next, introduce the polynomial $\hat{f}\in\mathbb{R}[X,Y]$ with

$$\hat{f}(X,Y) \;:=\; \sum_{j=1}^{r}\theta_j\, h(\boldsymbol{\ell}\,X+Y\,\mathbf{s}\,\mathbf{v}_j)\,, \qquad (25)$$

for all $(X,Y)\in\mathbb{R}^{m+1}$, and let $Y^2=1-\|X\|^2$ so that $(X,Y)\in\mathbb{S}^m$ whenever $X\in\mathcal{E}_m$. Hence whenever $\mathbf{x}\in\mathcal{E}_n$, then $(X,Y)\in\mathbb{S}^m$, and

$$\hat{h}(\mathbf{x}) \;=\; \hat{f}(\boldsymbol{\ell}^T\mathbf{x},\sqrt{1-\|\boldsymbol{\ell}^T\mathbf{x}\|^2}) \;=\; \hat{f}(X,|Y|)\ \text{on}\ \mathbb{S}^m\,. \qquad (26)$$

*Approximate sparse optimization on $\mathcal{E}_n$ or $\mathbb{S}^{n-1}$*

So when the $n-m$ remaining eigenvalues $(\lambda_{m+1},\ldots,\lambda_n)$ are small compared to the first $m$ ones, Theorem 5 suggests to consider replacing $h$ with $\hat{h}$ in the initial optimization problem $\mathbf{P}$. As we next show, when $\boldsymbol{\Omega}=\mathbb{S}^{n-1}$ or $\boldsymbol{\Omega}=\mathcal{E}_n$, the resulting problem is equivalent to solving:

$$\mathbf{Q}:\quad \rho=\min_{(X,Y)}\{\,\hat{f}(X,|Y|):\,(X,Y)\,\in\,\mathbb{S}^m\,\}\,, \qquad (27)$$

an $(m+1)$-variables optimization problem. Note that $\hat{f}(X,|Y|)$ is not a polynomial but $\rho=\min[\,\rho^+,\rho^-\,]$ with

$$\rho^+=\min_{X,Y}\{\,\hat{f}(X,Y):\,(X,Y)\,\in\,\mathbb{S}^m\,;\,Y\geq 0\}$$

$$\rho^-=\min_{X,Y}\{\,\hat{f}(X,-Y):\,(X,Y)\,\in\,\mathbb{S}^m\,;\,Y\leq 0\}\,.$$

So to solve $\mathbf{Q}$ and obtain $\rho$, one has to solve two *polynomial* optimization problems of same type as $\mathbf{P}$ but on $\mathbb{S}^m$, hence of much lower dimension when $m\ll n$.

*Lemma 6.* Let $\hat{h}$ be as in (23), $\hat{f}$ as in (25), and let $\rho=\min[\,\rho^+,\rho^-]$. Then

$$\min\{\,\hat{h}(\mathbf{x}):\mathbf{x}\in\mathbb{S}^{n-1}\}\;=\;\min\{\,\hat{h}(\mathbf{x}):\mathbf{x}\in\mathcal{E}_n\}\;=\;\rho \qquad (28)$$

**Proof.** Let $\tau:=\min\{\,\hat{h}(\mathbf{x}):\mathbf{x}\in\mathcal{E}_n\}$ and let $\mathbf{x}^*:=\arg\min\{\,\hat{h}(\mathbf{x}):\mathbf{x}\in\mathcal{E}_n\}$ so that $\hat{h}(\mathbf{x}^*)=\tau$. Write $\mathbf{x}^*=\boldsymbol{\ell}\,\mathbf{y}^*+\mathbf{s}\,\mathbf{z}^*$ so that $\|\mathbf{x}^*\|^2=\|\mathbf{y}^*\|^2+\|\mathbf{z}^*\|^2\leq 1$. Next, let $\tilde{\mathbf{x}}:=\boldsymbol{\ell}\,\mathbf{y}^*+r\cdot\mathbf{s}\,\mathbf{z}^*$ so that $\|\tilde{\mathbf{x}}\|^2=\|\mathbf{y}^*\|^2+r^2\,\|\mathbf{z}^*\|^2$, and choose $r$ such that $\tilde{\mathbf{x}}\in\mathbb{S}^{n-1}$. Then $\boldsymbol{\ell}^T\tilde{\mathbf{x}}=\boldsymbol{\ell}^T\mathbf{x}^*$ and therefore $\hat{h}(\tilde{\mathbf{x}})=\hat{h}(\mathbf{x}^*)=\tau$, which yields the first equality in (28). It remains to prove that $\rho=\tau$.

It is clear that $\tau\geq\rho$ as $(X,Y):=(\boldsymbol{\ell}^T\mathbf{x},\sqrt{1-\|X\|^2})\in\mathbb{S}^m$ and $\hat{f}(X,Y)=\hat{h}(\boldsymbol{\ell}^T\mathbf{x})$ whenever $\mathbf{x}\in\mathcal{E}_n$. For the converse, assume that $\rho=\rho^+$ with an optimal solution $(X^*,Y^*)\in\mathbb{S}^m$ and $Y^*\geq 0$. Let $\mathbf{x}:=\boldsymbol{\ell}\,X$ so that $\|\mathbf{x}\|=\|X\|$ as $\boldsymbol{\ell}^T\boldsymbol{\ell}=\mathrm{I}_m$. Hence $\mathbf{x}\in\mathcal{E}_n$. Moreover $\boldsymbol{\ell}^T\mathbf{x}=\boldsymbol{\ell}^T\boldsymbol{\ell}\,X=X$ and therefore by (23)-(24), $\hat{h}(\boldsymbol{\ell}^T\mathbf{x})=\hat{f}(X,Y)=\hat{f}(X,|Y|)$, which proves that $\tau\leq\rho^+$. The proof when $\rho=\rho^-$ being similar is omitted.

Of course the rationale for solving $\mathbf{Q}$ instead of $\mathbf{P}$ is based on Theorem 5, assuming that $\sum_{j=m+1}^{n}\lambda_j\,(\mathrm{E}_{\mu_n}[\nabla h\nabla h^T])$ is small. But the approximation in Theorem 5 in only in $L^2(\mathcal{E}_n)$ and not in $L^\infty(\mathcal{E}_n)$ (or equivalently in the sup-norm). This is why we have not provided an error analysis which remains to be done.

Notice that if $\lambda_j\,(\mathrm{E}_{\mu_n}[\nabla h\nabla h^T])=0$ for all $j>m$, then one retrieves the problem of Section 2. Indeed in $h(\mathbf{x})=h(\boldsymbol{\ell}\mathbf{y}+\mathbf{s}\,\mathbf{z})$ one has $\mathbf{z}=0$, and therefore in (19) and (21),

$$f(\mathbf{y}) \;=\; \mathrm{E}_{\mu_n}[h|\mathbf{y}] \;=\; h(\boldsymbol{\ell}\,\mathbf{y})\,.$$

So for instance when $\boldsymbol{\Omega}=\mathbb{S}^{n-1}$ and $h(\mathbf{x})=f(\boldsymbol{\ell}^T\mathbf{x})$ for some $\boldsymbol{\ell}\in\mathbb{R}^{n\times m}$, the sparse problem $\mathbf{Q}=\min\{f(\mathcal{L}\,\mathbf{y}):\mathbf{y}\in\mathcal{E}_m\}$ shown to be strictly equivalent to $\mathbf{P}$ in Lasserre (2021), is the limit case of $\mathbf{Q}$ in (27) when $\sum_{j=m+1}^{n}\lambda_j(\mathrm{E}_{\mu_n}[\nabla h\nabla h^T])=0$.

[2] In Constantine et al. (2014) the integral $\mathrm{E}[h|\mathbf{y}]$ has to be computed via Monte-carlo sampling with a different sample for each $\mathbf{y}$.

## REFERENCES


A. A. Ahmadi, A. Majumdar. DSOS and SDSOS optimisation: More tractable Alternatives to Sum of Squares and Semidefinite Optimization. *SIAM J. Appl. Algebra Geometry* 3(2):193–230, 2019.

V. Baldoni, N. Berline, J.A. De Loera, M. Köppe, M. Vergne. How to integrate a polynomial over a simplex. *Math. Comput.* 80:297–325, 2011.

A. Barvinok. Integration and optimization of multivariate polynomials by restriction onto a random subspace *Found. Comp. Math.* 7:229–249, 2007.

F. Camastra. Data dimensionality estimation methods: a survey: *Pattern Recognition* 362945–2954, 2003.

P.G. Constantine, E. Dow, and QiQi Wang. Active subspace methods in theory and practice: Applications to Kriging surfaces *SIAM J. Sci. Comput.* 36(4):A1500–A1524, 2014.

J.B. Lasserre. Convergent SDP-Relaxations in Polynomial Optimization with Sparsity. *SIAM J. Optim.* 17(3):218–242, 2006.

J.B. Lasserre. Optimization on the Euclidean unit sphere. *SIAM J. Optim.* to appear. *Hal-03291242*, 2021.

S. Roweis, R. Saul. Nonlinear dimensionality reduction by locally linear embedding *Science* 20:2323–2326, 2000

H. Waki, S. Kim, M. Kojima, M. Muramatsu. Sums of Squares and Semidefinite Program Relaxations for Polynomial Optimization Problems with Structured Sparsity. *SIAM J. Optim.* 17(1):822–843, 2006.

J. Wang, V. Magron, J. B. Lasserre. TSSOS: a moment-SOS hierarchy that exploits term sparsity. *SIAM J. Optim.*31(1):30–58, 2021.

J. Wang, V. Magron, J. B. Lasserre. Chordal-TSSOS: a moment-SOS hierarchy that exploits term sparsity with chordal extension. *SIAM J. Optim.* 31(1):114–141, 2021.

Y. Zheng, G. Fantuzzi, A. Papachristodoulou. Chordal and factor-width decompositions for scalable semidefinite and polynomial optimization *Annual Reviews in Control* 52:243–279, 2021