Journal of Convex Analysis
Volume 5 (1998), No. 2, 353–379

# Optimization of Positive Generalized Polynomials under $l^p$ Constraints

**L. Baratchart, M. Berthod, L. Pottier**
*Inria-Sophia Antipolis, 2004 Route des Lucioles,*
*BP 109 - 06561 Valbonne Cedex, France.*
*e-mail: Laurent.Baratchart@sophia.inria.fr*

Received October 17, 1996

The problem of maximizing a non-negative generalized polynomial of degree at most $p$ on the $l_p$-sphere is shown to be equivalent to a concave one. Arguments where the *maximum* is attained are characterized in connection with the irreducible decomposition of the polynomial, and an application to the labelling problem is presented where these results are used to select the initial guess of a continuation method.

*Keywords:* constrained optimization, convex optimization, combinatorial optimization, subhomogeneous functions, labeling

## 1. Introduction

The main purpose of this paper is to analyse when the argument of the *maximum* on the $l_p$-sphere of a non-negative polynomial is unique, in the special case where the total degree of the polynomial does not exceed $p$. The author's original incentive to study this question lies with pattern recognition and image processing, where it turns out that maximizing a polynomial under suitable constraints is an effective way to approach certain combinatorial optimization problems that would hardly be tractable otherwise. This motivation is highlighted in Section 2 where the classical labelling problem is sketched, as well as the *rationale* for replacing it by a continuous optimization scheme.

The issue just raised may be embodied into the more general problem of maximizing a non-negative generalized polynomial on the $l_p$-sphere. After some preliminaries on subhomegeneous functions and concavity in Sections 3 and 4, we give in Section 5 a systematic account of the solution when the degree is less than or equal to $p$. The approach is extremely elementary and consists in a simple change of variables that reduces the problem to concave maximization under linear constraints; determining strictly concave situations, however, involves a thorough discussion of the irreducible decomposition of polynomials which is linked to a non-linear eigenvalue problem of the Perron-Frobenius type. The results are partly carried over to $l_p$-constraints in product form in Section 6.

We finally report in Section 7 on an application to finding the Maximum a Posteriori Mode in a Markov Random Field, for which numerical algorithms are also discussed briefly.

## 2. Some motivations: a deterministic approach to the labelling problem

The labelling problem arises quite naturally in pattern recognition and image processing. Actually, many image–interpretation tasks can be cast that way and the literature on this topic is plethoric. One of the first and best known instance of this phenomenon is

ISSN 0944-6532 / $ 2.50 © Heldermann Verlag

the *Relaxation Labelling* approach that was celebrated in the computer vision community and for which we refer the reader to the seminal papers [10, 9] and [11]. A natural sequel to these developments was to rationalize the somewhat heuristic algorithms initially proposed, by defining merit functions to be optimized either locally [5] or globally [1]. An important step was taken when a well-founded probabilistic model was introduced by [7], which relied on the Hammersley-Clifford theorem [3]. As a matter of fact, all these references share a common framework that we illustrate here by describing the *Maximum a Posteriori Mode* problem (abbreviated as MAP) for a *Markov Random Field* (abbreviated as MRF).

We are given a set of units (or sites) $\mathcal{S} = S_i,\ 1 \leq i \leq N$, each of which may receive any label from 1 to $M$. A MRF on these units is defined as usual by a graph $G$, and the so-called clique potentials [7]. An edge of $G$ connecting $S_i$ and $S_j$ is denoted by $E_{ij}$ and $V_i$ is the set of vertices (or sites) connected to a given vertex $S_i$. Let $\mathcal{C}$ designate the set of all cliques of $G$, and define also $C_i = \{c \in \mathcal{C}; S_i \in c\}$. The number of sites in the clique $c$ is its degree $\deg(c)$, and we set $\deg(G) = \max_{c\in\mathcal{C}} \deg(c)$.

A global discrete labelling $L$ assigns one label $L_i$ such that $1 \leq L_i \leq M$ to each site $S_i$ in $\mathcal{S}$. The restriction of $L$ to the sites of a given clique $c$ is denoted by $L_c$. The definition of the MRF is completed by the knowledge of the clique potentials $V_{cL}$ (shorthand for $V_{cL_c}$) for every $c$ in $\mathcal{C}$ and every $L$ in $\mathcal{L}$, where $\mathcal{L}$ is the set of the $M^N$ discrete labelings.

In the MAP problem, the clique potentials stem from two sources of information: *a priori* knowledge about the restrictions that are imposed on the simultaneous labeling of connected neighboring units, and observations that were made on these units for a given occurence of the problem. The goal is to find the labeling which maximizes the *a posteriori* probability given the observations.

Following Hammersley-Clifford, the probability of a given labeling $L$ is given by:

$$P(L) \propto \prod_{c\in\mathcal{C}} \exp(-V_{cL}). \tag{2.1}$$

We assume here that the sufficient positivity condition for MRF is met *i.e.* that $P(L) > 0$ for each $L$. It follows that solving the MAP problem amounts to find

$$\max_{L\in\mathcal{L}} \sum_{c\in\mathcal{C}} W_{cL}, \tag{2.2}$$

where $W_{cL} = -V_{cL}$.

*Deterministic Pseudo Annealing* (in short: DPA) has been proposed in [2] to tackle this maximization. The idea is to first replace the combinatorial question by an equivalent continuous optimization problem, and then try to solve this continuous problem by deforming it into a convex one.

More precisely, let us define $f : \mathbb{R}^{NM} \to \mathbb{R}$ whose effect on

$$X = (x_{i,k})_{1\leq i\leq N,\ 1\leq k\leq M}$$

is given by

$$f(X) = \sum_{c\in\mathcal{C}} \sum_{l_c\in L_c} W_{cl_c} \prod_{j=1}^{\deg(c)} x_{c_j,l_{c_j}}, \tag{2.3}$$

where $c_j$ denotes the $j^{th}$ site of the clique c and $l_{c_j}$ is the label assigned to it by $l_c$. It is clear from (2.3) that $f$ is a polynomial in the $NM$ variables $x_{i,k}$'s whose degree is $\deg(G)$. Moreover, $f$ is linear in each variable separately. DPA in this case works as follows.

We define a compact subset $\mathcal{K}$ of $\mathbb{R}^{NM}$ by:

$$\forall i, k : x_{i,k} \geq 0 \quad \forall i : \sum_{k=1}^{M} x_{i,k} = 1.$$

The map $f$ may have plenty of relative *maxima* on $\mathcal{K}$. However, there is always an absolute *maximum* attained on the boundary *i.e.* at some point $X^*$ of the form:

$$\forall i, \exists k : x^\star_{ik} = 1,\ l \neq k, \Rightarrow x^\star_{il} = 0, \tag{2.4}$$

yielding naturally a discrete labelling. The difficulty is of course that standard search algorithms may typically lead to a local *maximum* and not to the absolute one. It is therefore of particular importance to find a good initial guess before applying the technique. This is precisely what DPA is designed for: we temporarily change the subset on which $f$ is maximized so as to make the problem easy to solve, and then we track the *maximum* while gradually restoring the original constraints. At each step, the projection of the former point onto the new set of contraints is used as an initial guess for the next optimization.

To be specific, we trade $\mathcal{K}$ for the set $\mathcal{K}_p$ defined by

$$\forall i, k : x_{i,k} \geq 0 \quad \forall i : \sum_{k=1}^{M} x^p_{i,k} = 1.$$

There has been numerical evidence for a while that the *maximum* is attained at a single point when $p > \deg(G)$, and further that this point lies interior to $\mathcal{K}_p$ [1]. The same holds true if $\deg(G) = p$, except for some degenerate zero-patterns of the coefficients for which the arguments of the *maximum* form a connected continuum intersecting certain coordinate axes. These facts are proved in Section 5, making the case $p \geq \deg(G)$ an easy problem. To achieve the DPA, it remains to decrease $p$ down to 1, initializing the algorithm at each step from the projection onto the new set of constraints of the solution found at the preceding step. This is the heuristic part of the procedure, as we hope to track the right solution when bifurcations do occur.

As we now see, the labelling problem raises the issue of maximizing a non-negative polynomial on the positive face of a simplex, and our contribution to the DPA approach will consist in solving the analogous problem when the simplex is replaced by an $l_p$-sphere with $p$ greater than or equal to the degree of the considered polynomial. The authors believe such a question posseses enough structure to make it worth studying, and other motivations like determining the dominant modes of certain nonlinear systems would also warrant such a study.

## 3. Preliminaries and notations.

We gather in this section a few pieces of notation and terminology which are of frequent use hereafter.

Let $\mathbb{R}^n_+$ be the non-negative cone in $\mathbb{R}^n$, that is the subset of vectors with non-negative coordinates. The interior of $\mathbb{R}^n_+$, consisting of vectors with positive coordinates, will be denoted by $\mathcal{P}_n$. For $x = (x_1 \cdots x_n) \in \mathbb{R}^n$ and $p > 0$, we denote by $\|x\|_p = (\sum_i |x_i|^p)^{1/p}$ the $l^p$ "norm" of $x$. This, actually, is a norm in the usual sense when $p \geq 1$ only (otherwise triangular inequality will fail).

Given $f : \Omega \to \mathbb{R}^m$, where $\Omega \subset \mathbb{R}^n$ is open, we denote by $\partial f/\partial i\,(x)$ the partial derivative of $f$ at $x$ with respect to the $i^{th}$ argument, and by $\partial^2 f/\partial i \partial j\,(x)$ the second partial derivative with respect to the $i^{th}$ and $j^{th}$ arguments. If $\mathcal{M}$ is a differentiable manifold and $f : \mathcal{M} \to \mathbb{R}$ a differentiable function, we say that $x \in \mathcal{M}$ is a *critical point* of $f$ if the derivative $Df(x)$ (which is defined on the tangent space $\mathcal{T}_x\mathcal{M}$ to $\mathcal{M}$ at $x$) vanishes identically. When $\mathcal{M}$ is embedded in $\mathbb{R}^n$ and $f$ extends to a differentiable function $\tilde{f}$ in a neighborhood of $x$ in $\mathbb{R}^n$, then $x$ is critical if and only if the gradient vector of $\tilde{f}$ at $x$ is normal to $\mathcal{T}_x\mathcal{M}$.

A function $f : \Omega \to \mathbb{R}$, where $\Omega \subset \mathbb{R}^n$ is open, is called *subhomogeneous of degree $h \in \mathbb{R}$ at $x \in \Omega$* if there exists $\epsilon(x) > 0$ such that

$$\forall \lambda \in [1, 1 + \epsilon(x)), \quad f(\lambda x) \leq \lambda^h f(x). \tag{3.1}$$

We say simply that $f$ is subhomogeneous of degree $h$ in $\Omega$ if it is so at each point of $\Omega$. Here, a few comments are in order. Firstly, we restrict ourselves to $\lambda \geq 1$ in the definition because if we allowed $\lambda \in ]1 - \epsilon(x), 1 + \epsilon(x)[$, $f$ would automatically be homogeneous. Secondly, the degree $h$ in the definition is by no means unique: if $f$, for instance, is positive, any $h' \geq h$ works.

Now, in the same manner as homogeneity translates into Euler's identity, subhomogeneity translates into Euler's inequality for differentiable functions: we say that a differentiable function $f : \Omega \to \mathbb{R}$ satisfies *Euler's inequality in degree $h$ at $x = (x_1 \cdots x_n)$*, for some $h \in \mathbb{R}$, if

$$\sum_{i=1}^{n} \frac{\partial f}{\partial i}(x)\, x_i \;\leq\; h\, f(x). \tag{3.2}$$

If (3.2) holds at every $x$, we simply say that $f$ satisfies Euler's inequality. The link between subhomogeneity and Euler's inequality is given in the following lemma.

**Lemma 3.1.** *If $\Omega$ is open in $\mathbb{R}^n$, and $f : \Omega \to \mathbb{R}$ is a differentiable function which is subhomogeneous of degree $h$ at $x = (x_1 \cdots x_n)$, then Euler's inequality in degree $h$ holds for $f$ at $x$. Conversely, if $f$ satisfy Euler's inequality in degree $h$ at every point $\lambda x$ with $1 \leq \lambda < 1 + \epsilon(x)$ for some $\epsilon(x) > 0$, then, $f$ is subhomogeneous of degree $h$ at $x$.*

**Proof.** Suppose $f$ is subhomogeneous, and put $g_x(\lambda) = \lambda^h f(x) - f(\lambda x)$ for $x \in \Omega$. The function $g_x : [1, 1 + \epsilon(x)) \to \mathbb{R}$ is non-negative, and vanishes at 1, hence $g'_x(1) \geq 0$. Expanding the derivative yields (3.2). Conversely, assume that $f$ satisfies (3.2) at each $\lambda x$ with $1 \leq \lambda < \epsilon(x)$. For such $\lambda$'s, we have

$$\frac{d\,g_x}{d\lambda} = h\lambda^{h-1} f(x) - \sum_{i=1}^{n} \frac{\partial f}{\partial i}(\lambda x) x_i \geq \frac{h}{\lambda}\left[\lambda^h f(x) - f(\lambda x)\right] = \frac{h g_x(\lambda)}{\lambda}.$$

This means $\lambda g'_x \geq h g_x$, or equivalently $(g_x/\lambda^h)' \geq 0$. Now, since the function $g_x/\lambda^h$ vanishes at 1 and has non-negative derivative on $[1, 1+\epsilon(x))$, it is non-negative there and so is $g_x$. $\square$

For $p \neq 0$, we define throughout $\varphi_p : \mathcal{P}_n \to \mathcal{P}_n$ by putting

$$\varphi_p(x_1, ..., x_n) \;=\; (x_1^{\frac{1}{p}}, ..., x_n^{\frac{1}{p}}). \tag{3.3}$$

Clearly, $\varphi_p$ is a diffeomorphism.

Let $A = (a_{i,j})$ be a real $n \times n$ matrix. It is called *irreducible* if two distinct indices $i$ and $j$ can always be linked by a chain $i = i_1, \cdots, i_k = j$ in such a way that $a_{i_\ell, i_{\ell+1}} \neq 0$.

Let $I = \{1 \cdots n\}$ be the set of indices, and $e_i$ be the $i^{th}$ vector of the canonical basis of $\mathbb{R}^n$. If $J \subset I$, we shall denote by $E_J$ the subspace of $\mathbb{R}^n$ spanned by the $e_j$'s for $j \in J$. Obviously, $E_j$ consists of those vectors $v = (v_1 \cdots v_n)$ with $v_j = 0$ if $j \notin J$. We call $E_J$ the *coordinate subspace* of $\mathbb{R}^n$ associated with $J$. If $I_1 \cdots, I_k$ is a partition of $I$, there is an orthogonal decomposition $\mathbb{R}^n = \sum_j E_{I_j}$, and we write accordingly $v = \sum_j v_{I_j}$ for $v \in \mathbb{R}^n$.

It is a simple observation [6] that the irreducibility of a matrix $A$ is equivalent to the non-existence of a non-trivial $A$-invariant coordinate subspace (that is, distinct from $\{0\}$ and $\mathbb{R}^n$ itself).

Suppose now that $S = (s_{i,j})$ is a $n \times n$ *symmetric* matrix. It is not difficult to check that the set $I = \{1 \cdots n\}$ of indices can be partitioned into classes $I_1 \cdots I_k$ such that the submatrices $S_{I_\ell} = (s_{i,j})_{i \in I_\ell, j \in I_\ell}$ are irreducible and also $s_{i,j} = 0$ if $i$ and $j$ belong to distinct $I_\ell$'s. This partition, which we call the *irreducible partition* of $S$, is well-defined since it corresponds to the decomposition of $\mathbb{R}^n$ into minimal $S$-invariant coordinate subspaces. It follows from the definition that $S_{I_\ell}$ is the matrix of the restriction of $S$ to $E_{I_\ell}$ when the latter is endowed with the canonical basis.

## 4. A concavity property

This section is instrumental for the remaining of the paper. The main result is that a $C^2$ map $f : \mathcal{P}_n \to \mathbb{R}$ whose first partial derivatives satisfy Euler's inequality in degree $p-1$ for some $p \neq 0$, and whose second partial derivatives are non-negative, is such that $f \circ \varphi_p$ is concave. This is essentially the content of Theorem 4.2 below. We begin with a computational lemma.

Let $f : \mathcal{P}_n \to \mathbb{R}$ be $C^2$ map, and put

$$\Phi_p = f \circ \varphi_p : \mathcal{P}_n \to \mathbb{R}$$

for $p$ a nonzero real number. Denote the second derivative of $\Phi_p$ at $x$ by $D^2\Phi_p(x)$; it is a bilinear form on $\mathbb{R}^n$ that we identify with the $n \times n$ symmetric matrix whose entry $(i,j)$ is $\partial^2\Phi_p/\partial i \partial j(x)$. We also introduce another $n \times n$ matrix $M_{f,p}(x)$ whose entries at the point $x = (x_1 \cdots x_n) \in \mathcal{P}_n$ are defined by the formulae

$$[M_{f,p}(x)]_{i,i} \;=\; x_i^{1-2p}\left[x_i \frac{\partial^2 f}{\partial i \partial i}(x) - (p-1)\frac{\partial f}{\partial i}(x)\right], \tag{4.1}$$

$$[M_{f,p}(x)]_{i,j} \;=\; x_i^{1-2p}\left[x_j \frac{\partial^2 f}{\partial i \partial j}(x)\right], \quad \text{for i} \neq \text{j} \;. \tag{4.2}$$

**Lemma 4.1.** *With the above notations, $D^2\Phi_p(x)$ is conjugate to $M_{f,p}(\varphi_p(x))/p^2$ at any point $x \in \mathcal{P}_n$. More precisely, we have:*

$$B^{-1}(x)D^2\Phi_p(x)B(x) \;=\; \frac{M_{f,p}(\varphi_p(x))}{p^2}, \tag{4.3}$$

*where $B(x)$ is the diagonal matrix* $\mathrm{diag}\{x_i\}$.

**Proof.** This is a simple computation. First, we write

$$\frac{\partial \Phi_p}{\partial i}(x) \;=\; \frac{\partial f}{\partial i}(\varphi_p(x))\,\frac{1}{p}x_i^{\frac{1}{p}-1}, \tag{4.4}$$

whence

$$\frac{\partial^2 \Phi_p}{\partial i \partial i}(x) \;=\; \frac{\partial^2 f}{\partial i \partial i}(\varphi_p(x))\left(\frac{1}{p}x_i^{\frac{1}{p}-1}\right)^2 + \frac{\partial f}{\partial i}(\varphi_p(x))\,\frac{1}{p}(\frac{1}{p}-1)x_i^{\frac{1}{p}-2}. \tag{4.5}$$

This can be rearranged as

$$\frac{1}{p^2}x_i^{\frac{1}{p}-2}\left[\frac{\partial^2 f}{\partial i \partial i}(\varphi_p(x))\,x_i^{\frac{1}{p}} - (p-1)\frac{\partial f}{\partial i}(\varphi_p(x))\right]. \tag{4.6}$$

If $i \neq j$, we get similarly

$$\frac{\partial^2 \Phi_p}{\partial i \partial j}(x) \;=\; \frac{1}{p^2}x_i^{\frac{1}{p}-1}\frac{\partial^2 f}{\partial i \partial j}(\varphi_p(x))x_j^{\frac{1}{p}-1}. \tag{4.7}$$

Now, compute $B^{-1}(x)D^2\Phi_p(x)B(x)$. Since the $i^{th}$ row of $D^2\Phi_p(x)$ gets divided by $x_i$ while the $j^{th}$ column gets multiplied by $x_j$, the result is $M_{f,p}(\varphi_p(x))/p^2$. □

We are now in position to state:

**Theorem 4.2.** *Let $f : \mathcal{P}_n \to \mathbb{R}$ be a $C^2$ map. For $p \neq 0$, define $\varphi_p$ and $\Phi_p$ as before, and let $x = (x_1 \cdots x_n)$ be a point in $\mathcal{P}_n$ such that each partial derivative $\partial f/\partial i$ satisfies Euler's inequality in degree $p-1$ at $\varphi_p(x)$, while each second partial derivative $\partial^2 f/\partial i\partial j$ is non-negative at $\varphi_p(x)$. Let finally $I_1 \cdots I_k$ denote the irreducible partition of $D^2\Phi_p(x)$. Then $D^2\Phi_p(x)$ defines a non-positive quadratic form. It is negative definite unless there exists an $\ell$ such that Euler's inequality for $\partial f/\partial i$ at $\varphi_p(x)$ is in fact an equality for every $i \in I_\ell$. If we let $I' \subset \{1 \cdots k\}$ be the set of such $\ell$'s, the kernel of $D^2\Phi_p(x)$ is the subspace spanned by the $x_{I_j}$'s for $j \in I'$.*

**Proof.** From Lemma 4.1, we see that the eigenvectors of $D^2\Phi_p(x)$ are the images under $B(x)$ of those of $M_{f,p}(\varphi_p(x))/p^2$, and that the eigenvalues of the two matrices differ only by a factor $1/p^2$. In particular, since $D^2\Phi_p(x)$ is a symmetric matrix hence has real eigenvalues, so does $M_{f,p}(\varphi_p(x))$. Let us denote by $(m_{i,j})$ the entries of $M_{f,p}(\varphi_p(x))$. By assumption, we have Euler's inequality for $\partial f/\partial i$ at $\varphi_p(x)$:

$$\sum_{j=1}^{n}\frac{\partial^2 f}{\partial i \partial j}(\varphi_p(x))\,x_j^{\frac{1}{p}} \;\leq\; (p-1)\frac{\partial f}{\partial i}(\varphi_p(x)). \tag{4.8}$$

Upon multiplying by the positive quantity $x_i^{\frac{1}{p}-2}$, we get by the very definition of $(m_{i,j})$:

$$\sum_{j=1}^{n} m_{i,j} \le 0, \quad \forall i \in \{1 \cdots n\}. \tag{4.9}$$

By hypothesis, all partial second derivatives of $f$ are non-negative at $\varphi_p(x)$, so we see from the definition that $m_{i,j} \ge 0$ if $i \ne j$. Now, a well-known theorem of Gerschgorin (see e.g. [6]) tells us that every eigenvalue of $M_{f,p}(\varphi_p(x))$ belongs to a disc centered at some $m_{i,i}$ of radius $\sum_{j\ne i} |m_{i,j}| = \sum_{j \ne i} m_{i,j}$. By (4.9), all these eigenvalues lie in the left half-plane, and since they are real, they are non-positive. This shows that the eigenvalues of $D^2\Phi_p(x)$ are also non-positive, and so is the associated quadratic form.

We now compute the kernel of $D^2\Phi_p(x)$. Since the latter is symmetric and $E_{I_j}$ is $D^2\Phi_p(x)$-stable by definition, we first observe that $(D^2\Phi_p(x)\, v)_{I_j} = D^2\Phi_p(x)\, v_{I_j}$ for any $v \in \mathbb{R}^n$. Therefore, $v$ belongs to the kernel of $D^2\Phi_p(x)$ if and only if $v_{I_j}$ belongs to the kernel of the restriction of $D^2\Phi_p(x)$ to $E_{I_j}$ for all $j \in \{1, ..., k\}$. Hence, it is enough to prove that the kernel of $D^2\Phi_p(x)_{I_j}$, if non-trivial, is generated by $x_{I_j}$, and that Euler's inequality, when applied to $\partial f/\partial i$, is then an equality for every $i \in I_j$. Let $n_j$ denote the cardinality of $I_j$. Because of the relationship between $D^2\Phi_p(x)$ and $M_{f,p}(\varphi_p(x))$ asserted in Lemma 4.1, it is equivalent to show that the kernel of $(M_{f,p}(\varphi_p(x)))_{I_j}$, if nontrivial, is generated by the vector $(1 \cdots 1)$ of size $n_j$, and still Euler's inequality is an equality for $i \in I_j$.

Suppose now that the kernel of $(M_{f,p}(\varphi_p(x)))_{I_j}$ does not consist of zero alone, and let $w = (w_1 \cdots w_{n_j})$ be a non-zero vector in this kernel. Choose and index $i_0 \in I_j$ such that $w_{i_0}$ is of maximum modulus. Replacing $w$ by $-w$ if necessary, we may assume that $w_{i_0} > 0$. Define $K \subset I_j$ to be the subset of indices $i$ such that $w_i = w_{i_0}$. Pick any $i_1 \in K$. If we write that the $i_1^{th}$ component of $(M_{f,p}(\varphi_p(x)))_{I_j}\, w$ is zero, we get:

$$\sum_{i \in I_j} m_{i_1,i} w_i \;=\; 0. \tag{4.10}$$

Upon multiplying (4.9) by $w_{i_1}$, and taking into account the fact that $m_{i_1,i} \;=\; 0$ if $i \notin I_j$ (since the same holds true for the matrix $D^2\Phi_p(x)$), we also have that

$$\sum_{i \in I_j} m_{i_1,i} w_{i_1} \;\le\; 0. \tag{4.11}$$

Substracting (4.10) from (4.11) yields

$$\sum_{i \in I_j,\, i \ne i_1} m_{i_1,i}(w_{i_1} - w_i) \;\le\; 0. \tag{4.12}$$

But each term in the sum is non-negative, so they all vanish, and equality holds in (4.12), hence in (4.11), and in Euler's inequality for $\partial f/\partial i_1$ as well. If $i \notin K$, then $w_{i_1} \;>\; w_i$ in (4.12) and therefore $m_{i_1,i} \;=\; 0$ whence also $m_{i,i_1} \;=\; 0$ (because they are proportional). Since $i_1$ was arbitrary in $K$, it follows that $E_K \subset E_j$ is invariant under $(M_{f,p}(\varphi_p(x)))_{I_j}$. But this matrix is irreducible since $D^2\Phi_p(x)_{I_j}$ is, by definition of $I_j$. Thus, we have $K = I_j$, and $w_i = w_{i_0}$ for every $i \in I_j$. This is precisely what we wanted to show. Conversely, it is clear that if Euler's inequality is an equality for every $i \in I_j$, the vector $(1 \cdots 1)$ lies in the kernel of $(M_{f,p}(\varphi_p(x)))_{I_j}$. ☐

To recap, Theorem 4.2 asserts that a $C^2$ function $f : \mathcal{P}_n \to \mathbb{R}$ is such that $f \circ \varphi_p$ is concave as soon as

(i) the second partial derivatives are non-negative at every $x \in \mathcal{P}_n$; this is equivalent to saying that the gradient of $f$ is non-decreasing for the usual partial ordering of $\mathcal{P}_n$, i.e. $x \leq y$ iff $y - x \in \mathcal{P}_n$.
(ii) the partial derivatives satisfy Euler's inequality in degree $p-1$ on $\mathcal{P}_n$; by Lemma 3.1, this is equivalent to the seemingly more natural property that the derivatives of $f$ are subhomogeneous of degree $p-1$ on $\mathcal{P}_n$.

## 5. $l_p$-constrained maximization of positive generalized polynomials of degree at most $p$

In this section, we apply the preceding results to the problem of maximizing a generalized polynomial with non-negative coefficients (see the definition below) on the $l^p$-sphere when $p$ is not less than the degree. This allows us to describe uniqueness and positivity properties of the solution. The approach is completely elementary and simply consists in composing the polynomial with $\varphi_p$ so as to be back to standard optimization of a concave function under linear constraints. Since we want to give complete answers on uniqueness, however, we need to analyse cases when strict concavity prevails and this requires a slightly lengthier discussion of the irreducible decomposition which makes this section somewhat reminiscent of the Perron-Frobenius theory for nonnegative matrices. In effect, at the end of the section, we use the critical point equation to derive some kind of nonlinear generalization for symmetric matrices of the Perron-Frobenius theorem.

Strictly speaking, the facts that we shall use about non-negative generalized polynomials are subhomogeneity of the derivatives and nonnegativity of the second derivatives when the exponents involved are not less than 1. *In fact, the results and the proofs can be adapted to any function sharing these properties.* In particular, everything in this section extends to *infinite* sums $\sum c_\alpha x^\alpha$ where $\alpha \in \mathbb{R}_+^n$ is bounded by $p$ in $l^1$-norm and the coefficients $c_\alpha \in \mathbb{R}_+$ decrease fast enough to ensure that the series converges absolutely when, say, $\|x\|_p < 1 + \epsilon$ for some positive $\epsilon$. Nevertheless, we shall stick to the case of generalized polynomials an application of which was described in the introduction.

A *generalized polynomial* is a function $P : \mathcal{P}_n \to \mathbb{R}$ of the form:

$$P(x) \;=\; \sum_\alpha c_\alpha x^\alpha, \tag{5.1}$$

where $\alpha$ ranges over a finite set of $\mathbb{R}_+^n$, and $x^\alpha$ stands for $x_1^{\alpha_1} \cdots x_n^{\alpha_n}$ with $\alpha = (\alpha_1 \cdots \alpha_n)$. By definition, the degree of $P$ is $h = \max_\alpha h_\alpha$, where $h_\alpha = \sum_i \alpha_i$. If $h_\alpha = h, \forall \alpha$, we call $P$ a homogeneous generalized polynomial of degree $h$. By convention, the zero polynomial is homogeneous of any degree. We say that $P$ has non-negative coefficients if $c_\alpha \geq 0$ for all $\alpha$.

Now, we need to extend the notion of irreducibility, already introduced for matrices, to generalized polynomials. If $P = \sum c_\alpha x^\alpha$ is such a polynomial, we can associate to $P$ a graph whose vertices are the variables, and an edge connects two variables $x_i$ and $x_j$ iff there exists a term $c_\alpha \neq 0$ with $\alpha_i \neq 0$ and $\alpha_j \neq 0$. The adjacency matrix of this graph is symmetric and its irreducible partition is also called the irreducible partition of

$P$. Another way to look at things is to observe that any generalized polynomial $P$ in $n$ variables can be written (in possibly many ways) as

$$P(x) = \sum_{j=1}^{k} P_{I_j}(x_{I_j}), \tag{5.2}$$

where the $I_j$'s, for $j \in \{1 \cdots k\}$, partition the set $I = \{1 \cdots n\}$; it is easy to check that there is a unique minimal such decomposition (*i.e.* one that cannot be refined), where the $PI_j{}'s$ are defined *up to a constant term* and where the $I_j$'s are nothing but the irreducible partition of $P$ already defined. If, in addition, $P$ has nonnegative coefficients, this irreducible partition is that of the second derivative $D^2P$ at any (and thus every) point of $\mathcal{P}_n$. The additive decomposition (5.2) associated to the irreducible partition is called the *irreducible decomposition* of $P$, and the polynomials $P_{I_j}$ in this decomposition are called the *irreducible components.* Such a component may well be zero; in this case, the polynomial depends on fewer variables. Note also that irreducible components are defined only up to constant terms so that *any qualification concerning them should be understood modulo a constant term.* An irreducible component which is not $P$ itself is said to be proper, and we say that $P$ is irreducible if and only if it has no proper irreducible component. Equivalently, this means that the irreducible partition has only one element, namely $I$ itself.

A family of functions to which Theorem 4.2 applies naturally is the family of generalized polynomials with non-negative coefficients; this is due to the following result.

**Proposition 5.1.** *A generalized polynomial $P$ of degree $h$ in $n$ variables with non-negative coefficients is subhomogeneous of degree $h$ in the positive cone $\mathcal{P}_n$. If $P$ is not homogeneous, it is in fact subhomogeneous of degree strictly less than $h$ at any point of $\mathcal{P}_n$.*

**Proof.** Write $P = \sum_\beta P_\beta$, where each $P_\beta$ is homogeneous of degree $\beta$. For any $x \in \mathbb{R}^n$ and $\lambda \in \mathbb{R}$, we have

$$P(\lambda x) \;=\; \sum_\beta \lambda^\beta P_\beta(x). \tag{5.3}$$

Now for $x \in \mathcal{P}_n$, each $P_\beta(x)$ is obviously non-negative, and it is clear if $\lambda \geq 1$ that $\lambda^\beta \leq \lambda^h$; therefore, $P(\lambda x) \leq \lambda^h P(x)$, showing that $P$ is subhomogeneous of degree $h$ in $\mathcal{P}_n$.

Assume now that $P$ is not homogeneous, and let $\beta_0$ be largest among those $\beta$'s such that $P_\beta \neq 0$ and $\beta < h$. Pick $\mu_0 > 0$ so small that $h - \beta_0 - \mu_0 > 0$; for $x \in \mathcal{P}_n$, $\lambda \geq 1$ and $0 < \mu < \mu_0$, put

$$g_{x,\mu}(\lambda) = \lambda^{h-\mu} P(x) - P(\lambda x). \tag{5.4}$$

We compute

$$g_{x,\mu}(\lambda) = \sum_\beta [\lambda^{h-\mu} - \lambda^\beta] P_\beta(x), \tag{5.5}$$

so that

$$\frac{dg_{x,\mu}}{d\lambda}(1) \;=\; \sum_{\beta}[h-\beta-\mu]P_\beta(x) \tag{5.6}$$

which is, by the definition of $\mu$, bounded from below by

$$(h-\beta_0-\mu)\ P_{\beta_0}(x) - \mu P_h(x). \tag{5.7}$$

Since each component of $x$ is positive, it follows that $P_{\beta_0}(x) > 0$, hence the above expression can be made positive by choosing $\mu$ sufficiently small. Then $g_{x,\mu}(\lambda)$ vanishes with positive derivative at $\lambda = 1$. Therefore, it remains non-negative on some interval $[1, 1+\epsilon(x))$, so that $P$ is subhomogeneous of degree $h-\mu < h$ at $x$. □

As an application of Proposition 5.1 and Theorem 4.2, we study for later use the concavity properties of non-negative generalized polynomials when the degree does not exceed 1.

**Theorem 5.2.** *A generalized polynomial $P$ with non negative coefficients of degree at most 1 is concave on $\mathbb{R}^n_+$. If $P(x) = \sum_{\ell=1}^k P_{I_\ell}$ is the irreducible decomposition, the kernel of $D^2P(x)$ at $x \in \mathcal{P}_n$ is the linear span of those $x_{I_\ell}$'s such that $P_{I_\ell}$ is homogeneous of degree 1. In particular, $P$ is strictly concave on $\mathcal{P}_n$ if, and only if, it has no irreducible component which is homogeneous of degree 1, in which case the second derivative is negative definite at each point of $\mathcal{P}_n$.*

**Proof.** Since $P$ is continuous on $\mathbb{R}^n_+$ (for the exponents are non-negative), it is enough to show that $P$ is concave on $\mathcal{P}_n$. Let $\alpha_m$ be the smallest non-zero exponent for a variable appearing in $P$. Set $p = \alpha_m^{-1}$ and define $f = P \circ \varphi_p^{-1} = P \circ \varphi_{1/p}$, so that $f$ is obtained from $P$ by changing each variable into its $p^{th}$ power. Then, $f$ is again a generalized polynomial with non-negative coefficients, of degree $h \le p$ (this is where we use $\deg P \le 1$), each variable of which appears at every occurence with exponent at least 1. Thus, every partial derivative $\partial f/\partial i$ of $f$ is a generalized polynomial with non-negative coefficients, of degree at most $h-1$. By Proposition 5.1 and Lemma 3.1, $\partial f/\partial i$ satisfies Euler's inequality in degree $h-1$ and so *a fortiori* in degree $p-1$ on $\mathcal{P}_n$. Since the second partial derivatives of $f$ are clearly non-negative on $\mathcal{P}_n$, Theorem 4.2 implies that $P = f \circ \varphi_p$ has a non-positive second derivative at every point of $\mathcal{P}_n$, hence is concave there.

Let us now determine the kernel of the second derivative. Since $P$ is nonnegative, its irreducible partition is also that of $D^2P(x)$ at any $x \in \mathcal{P}_n$ and we deduce from Theorem 4.2 again that the kernel of this matrix is generated by those $x_{I_\ell}$'s such that $\partial f/\partial i$ satisfies Euler's *equality* in degree $p-1$ at $\varphi_p(x)$ for each $i \in I_\ell$. By construction, the irreducible partitions of $P$ and $f$ are identical, so we can write

$$f = \sum_{\ell=1}^{k} f_{I_\ell},$$

and it is clear that $f_{I_\ell}$ is homogeneous of degree $p$ if, and only if, $P_{I_\ell}$ is homogeneous of degree 1. Hence, it remains for us to show that $\partial f_{I_\ell}/\partial i$ satisfies Euler's identity in degree $p-1$ at $\varphi_p(x)$ for each $i \in I_\ell$ if, and only if, $f_{I_\ell}$ is homogeneous of degree $p$. Sufficiency is obvious. By Proposition 5.1, necessity amounts to prove that a generalized polynomial which is not homogeneous of degree $p$ (up to a constant term) cannot have partial derivatives each of which is homogeneous of degree $p-1$. This is easy: argueing by contradiction, suppose that $Q(y_1, \cdots, y_s)$ is such a polynomial and consider the homogeneous generalized polynomial $\partial Q/\partial 1$ of degree $p-1$. By elementary integration, we get

$$Q(y_1, \cdots, y_s) = Q_1(y_1, \cdots, y_s) + Q_2(y_2, \cdots, y_s),$$

where $Q_1$ is homogeneous of degree $p$ and $Q_2$ does not depend on $y_1$. Now, for $2 \leq i \leq s$,

$$\frac{\partial Q_2}{\partial i} = \frac{\partial Q}{\partial i} - \frac{\partial Q_1}{\partial i}$$

is homogeneous of degree $p-1$, so we can iterate the process and write $Q_2 = Q_1^1 + Q_2^1$ where $Q_1^1$ is homogeneous of degree $p$, and $Q_2^1$ depends on $y_3, \cdots, y_s$ only, while still having homogeneous derivatives of degree $p-1$. By induction, we conclude that

$$Q = Q_1 + Q_1^1 + \cdots + Q_1^{s-1} + \text{constant}$$

is homogeneous of degree $p$ up to a constant, contradicting the hypothesis.

Thus, if $P$ has no homogeneous irreducible component of degree 1, we conclude that $D^2P$ is negative definite on $\mathcal{P}_n$, so that $P$ is strictly concave there; on the contrary, if it has some irreducible homogeneous component of degree 1, say, $P_{I_\ell}$, observe that if we fix all the other variables and if we evaluate $P_{I_\ell}$ on the diagonal, we get an affine function. This achieves the proof. □

Now, we turn to optimization. For $p > 0$, let $S_p^n = \{x \in \mathbb{R}^n; \sum_j |x_j|^p = 1\}$ denote the $l^p$ unit sphere in $\mathbb{R}^n$ and set $S_{p,+}^n = \mathbb{R}_+^n \cap S_p^n$. Our goal is to study the following problem:

*Given $p > 0$ and a non-constant generalized polynomial with non-negative coefficients $P$, of degree $h$ with $h \leq p$, characterize the argument(s) of*

$$\max_{x \in S_{p,+}^n} P(x). \tag{5.8}$$

It should be observed in the first place, since a generalized polynomial is continuous on $\mathbb{R}_+^n$, that the max in Problem (5.8) is indeed attained by compactness. We shall further investigate uniqueness and positivity properties of the argument of the max. In the second place, it is perhaps appropriate to take a look at the limiting case $p = \infty$ which was tacitly excluded in the statement of the problem. Then, it is clear that the maximum is attained by setting to 1 each variable which *actually appears* in $P$ so that the problem is, in some sense, totally decoupled. For finite $p$, our first result concerns positivity:

**Proposition 5.3.** *Assume a solution $x^* = (x_1^* \cdots x_n^*)$ to Problem (5.8) satisfies $x_i^* = 0$ for $i \in I_1$ and $x_i^* > 0$ for $i \in I_2$, where $I_1$ and $I_2$ partition the set of indices. Then one can decompose $P$ as*

$$P(x) = P_1(x_{I_1}) + P_2(x_{I_2}), \tag{5.9}$$

*where $P_1$ is some (possibly zero) homogeneous generalized polynomial with non-negative coefficients of degree $p$.*

**Proof.** Assume without loss of generality that $I_1 = \{1 \cdots k\}$ and $I_2 = \{k+1 \cdots n\}$. Then $k < n$ for $x^* \neq 0$. If a decomposition of the form (5.9) is impossible, this means that there exists a non-zero monomial in $P$ whose degree with respect to the variables $x_1, \cdots, x_k$ is positive but less than $p$. Define $\Phi_p = P \circ \varphi_p$. This is a nonnegative generalized polynomial of degree $h/p \leq 1$ deduced from $P$ by dividing each exponent by $p$, say

$$\Phi_p(y) = \sum c_\beta y^\beta, \tag{5.10}$$

and it attains its maximum on $S^n_{1,+}$ at

$$y^* = \varphi_p^{-1}(x^*) = (0, \cdots, 0, y^*_{k+1}, \cdots, y^*_n).$$

By a previous observation, one of the monomials, say, $c_\gamma y^\gamma$ has non-zero degree less than 1 in the variables $y_1 \cdots y_k$. Thus, if we write

$$y^\gamma = y_1^{\gamma_1} \cdots y_n^{\gamma_n},$$

we have $0 < \sum_{\ell=1}^k \gamma_\ell < 1$. In particular, there is an index $i \leq k$ such that $0 < \gamma_i$ and

$$1 - \gamma_i > \sum_{1 \leq \ell \leq k, \ell \neq i} \gamma_\ell \stackrel{\text{def}}{=} \delta_i. \tag{5.11}$$

Select any $j > k$ so that $y_j^* > 0$ by definition of $k$. For $0 \leq t \leq t_0 < y_j^*/k$, the point

$$Y_t = (t, \cdots, t, y^*_{k+1}, \cdots, y^*_{j-1}, y^*_j - kt, y^*_{j+1}, \cdots, y^*_n)$$

belongs to $S^n_{1,+}$ and we can define $G : [0, t_0] \to \mathbb{R}$ by the formula

$$G(t) = \Phi_p(Y_t).$$

This function is $C^\infty$ for $0 < t < t_0$. Since every quantity involved is nonnegative, we have $\partial\Phi_p/\partial\ell(Y_t) \geq 0$ for $1 \leq \ell \leq n$, and for $\ell = i$ the stronger inequality:

$$\frac{\partial \Phi_p}{\partial i}(Y_t) \geq c_\gamma \gamma_i \frac{t^{\delta_i}}{t^{1-\gamma_i}} (y^*_{k+1})^{\gamma_{k+1}} \cdots (y^*_{j-1})^{\gamma_{j-1}} (y^*_j - kt)^{\gamma_j} (y^*_{j+1})^{\gamma_{j+1}} \cdots (y^*_n)^{\gamma_n}. \tag{5.12}$$

Now, we evaluate

$$\frac{dG}{dt} = \sum_{\ell=1}^{k} \frac{\partial \Phi_p}{\partial \ell}(Y_t) \; - \; k \frac{\partial \Phi_p}{\partial j}(Y_t),$$

and we observe that the only negative contribution comes from the last term wich is bounded for $0 \leq t \leq t_0$ (since the only quantities appearing in the denominators of $\partial\Phi_p/\partial j(Y_t)$ are of the form $(y_j^* - kt)^{1-\beta_j}$), whereas the term corresponding to $\ell = i$ is arbitrarily large when $t$ is small enough by (5.12) and (5.11). Therefore, if $t_0$ is small enough, we have $dG/dt > 0$ hence (notice that the integral converges by continuity)

$$G(t_0) - G(0) = \int_0^{t_0} \frac{dG}{dt}(t)\, dt > 0,$$

contradicting the fact that $y^*$ is a maximum. □

With the aid of Proposition 5.3, we can now treat uniqueness of the solution to (5.8) in an important special case.

**Theorem 5.4.** *If $P$ has no proper homogeneous irreducible component of degree $p$ (in particular if $P$ is irreducible), then Problem* (5.8) *has a unique solution $x^*$. Moreover, $x^* \in \mathcal{P}_n$ in this case, and is the unique critical point of $P$ on $S^n_{p,+} \cap \mathcal{P}_n$.*

**Proof.** By Proposition 5.3, any solution $x^*$ belongs to $\mathcal{P}_n$. Define again $\Phi_p = P \circ \varphi_p$, so that the *maxima* of $\Phi_p$ on $S^n_{1,+}$ are the images under $\varphi_p^{-1}$ of those of $P$ on $S^n_{p,+}$. Similarly, the critical points of $\Phi_p$ on $S^n_{1,+} \cap \mathcal{P}_n$ are the images under $D\varphi_p^{-1}$ of those of $P$ on $S^n_{p,+} \cap \mathcal{P}_n$. Since $\Phi_p$ is a generalized polynomial of degree at most 1, it is concave on $\mathbb{R}^n_+$ by Theorem 5.2 and so is its restriction to the *linear* manifold $S^n_{1,+}$. Consequently its *maxima* on the latter form a convex set. Also, by concavity, these *maxima* coincide with the critical points of $\Phi_p$ on $S^n_{1,+} \cap \mathcal{P}_n$, and the second derivative at such a point $y^*$ is just the restriction of $D^2\Phi_p(y^*)$ to the tangent space

$$\mathcal{T}_{y^*}\, S^n_{1,+} = \{x \in \mathbb{R}^n;\ \sum x_i = 0\}.$$

If $P$ is reducible or if $P$ is not homogeneous of degree $p$, the hypothesis strengthens to: "$P$ has no irreducible homogeneous component of degree $p$" so that $\Phi_p$ has then no irreducible homogeneous component of degree 1; otherwise, $P$ is irreducible homogeneous in degree $p$ and so is $\Phi_p$ in degree 1. According to each possibility, Theorem 5.2 tells us that the kernel of $D^2\Phi_p(y^*)$ is either zero or one dimensional generated by $y^*$. In any case, this kernel intersects $\mathcal{T}_{y^*}\, S^n_{1,+}$ at zero only. Therefore, the critical points of $\Phi_p$ on $S^n_{1,+} \cap \mathcal{P}_n$ are isolated, while at the same time forming a connected set since it is convex. Hence, there is a unique such point. $\square$

To complete our study of Problem (5.8), we still have to examine what happens if $P$ does have homogeneous irreducible components of degree $p$. To this effect, it will be convenient to generalize Problem (5.8) slightly and to consider

$$\max_{\substack{x \in \mathbb{R}^n_+ \\ \|x\|_p = r}} P(x) \tag{5.13}$$

for $r$ a non-negative real number. When $r = 1$, this is just Problem (5.8). Conversely, (5.13) reduces to (5.8) upon scaling each variable by $r$, so that the results established so far transpose immediately to Problem (5.13). In particular, if $P$ has *no* proper homogeneous component of degree $p$, there is a unique argument for the max in (5.13) that we denote by $x^*(r)$. We have of course $x^*(0) = 0$. If $r > 0$, we know from Theorem 5.4 that $x^*(r)$ is the unique critical point of $P$ on $\mathcal{P}_n \cap \{\|x\|_p = r\}$; this means that there exists a *Lagrange multiplier* $\lambda^*(r)$ such that

$$\lambda^*(r)\frac{(x_i^*(r))^{p-1}}{r^{1-1/p}} = \frac{\partial P}{\partial i}(x^*(r)) \quad \forall i \in \{1, \cdots, n\},$$

as this equation merely expresses that the gradient of $P$ is proportional to the gradient of $\|x\|_p$ at $x^*(r)$. Clearly, $\lambda^*(r)$ is positive for $x^*(r) \in \mathcal{P}_n$ and $P$ is not constant. Introducing the *Lagrangian* function

$$L_r(x, \lambda) = P(x) + \lambda(r - \|x\|_p),$$

this may be capsulized by saying that for $r>0$, then $(x^*(r),\lambda^*(r))$ is the unique critical point of $L_r(x,\lambda)$ on $\mathcal{P}_n\times\mathbb{R}$. We shall need a few differential properties of $x^*(r)$ and $\lambda^*(r)$ as functions of $r$:

**Lemma 5.5.** *If $P$ has no proper homogeneous component of degree $p$, then, with the above notations, $x^*(r)$ and $\lambda^*(r)$ are $C^\infty$ functions of $r$ on $(0,\infty)$. If, in addition, $P$ is not homogeneous of degree $p$, then*

$$\frac{d\left[r^{1/p-1}\lambda^*(r)\right]}{dr}\neq 0 \tag{5.14}$$

*at every point of* $(0,\infty)$.

**Proof.** Put $\Phi_p=P\circ\varphi_p$ and note that

$$(y^*(r),\mu^*(r))=\left(\varphi_p^{-1}(x^*(r)),\ \frac{r^{1/p-1}\lambda^*(r)}{p}\right)$$

is the unique critical point over $\mathcal{P}_n\times\mathbb{R}$ of the modified Lagrangian

$$L_r^1(y,\mu)=\Phi_p(y)+\mu(r^p-\sum_i y_i).$$

As in the proof of Theorem 5.4, we are now back to the elementary problem of maximizing a concave functional under some linear constraint, and the lemma is a standard application of the implicit function theorem granted Theorem 5.2 which guarantees nondegeneracy of the second derivative on the tangent space to the constraint. This computation we redo for the ease of the reader; by the implicit function theorem, we will obtain the desired smoothness if we show that the second derivative $D^2L_r^1$ is nonsingular at $(y^*(r),\mu^*(r))$ because $y^*$ and $\mu^*$ will then be smooth functions of $r$ and the same will obviously be true of $x^*$ and $\lambda^*$. Compute this second derivative as

$$D^2L_r^1(y^*(r),\mu^*(r))=\begin{pmatrix} D^2\Phi_p(y^*(r)) & -\mathbf{1}\\ -\mathbf{1}^T & 0\end{pmatrix}, \tag{5.15}$$

where $-\mathbf{1}$ stands for the vector in $\mathbb{R}^n$ all components of which are -1's and where the superscript "$^T$" means "transpose". Assume $(v,\nu)\in\mathbb{R}^n\times\mathbb{R}$ is in the kernel of this matrix. From the last row, we get $\sum v_i=0$ so that $v$ belongs to the tangent space of $S^n_{1,+}\cap\mathcal{P}_n$. Then, multiplying (5.15) on the right by $(v,\nu)$ and on the left by $(v,\nu)^T$, we deduce that $v^T\,D^2\Phi_p(y^*(r))\,v=0$. But since $\Phi_p$ has no proper irreducible homogeneous component of degree 1, we deduce from Theorem 5.2 (as in the proof of Theorem 5.4) that its second derivative restricted to the tangent space of $S^n_{1,+}\cap\mathcal{P}_n$ is negative definite. Therefore, we have $v=0$ hence $\nu=0$ so that $D^2L_r^1(y^*(r),\mu^*(r))$ is non-singular. Now, equation (5.14) is equivalent to $d\mu^*/dr\neq 0$. Still from the implicit function theorem, and denoting by 0 the zero vector in $\mathbb{R}^n$, we have that

$$\frac{d\mu}{dr}(y^*(r),\mu^*(r)=(\mathbf{0}^T\ 1)\ \left[D^2L_r^1(y^*(r),\mu^*(r))\right]^{-1}\begin{pmatrix}\mathbf{0}\\ pr^{p-1}\end{pmatrix},$$

and this quantity cannot vanish if $P$ has no homogeneous component of degree $p$, because it would mean that $D^2L_r^1(y^*(r),\mu^*(r))$ applied to some vector of the form $(v,0)^T$ yields $(\mathbf{0},pr^{p-1})^T$ so that $D^2\Phi_p(y^*(r))\,v$ should in turn vanish and $v$ itself should vanish since $D^2\Phi_p$ is non-singular by Theorem 5.4. This is absurd for $r>0$. $\square$

For our purposes, we need to know a bit more about the behaviour of the objective function in Problem (5.13), that we define as

$$M_{P,p}(r) = \max_{\substack{x\in\mathbb{R}_+^n\\ \|x\|_p=r}} P(x). \tag{5.16}$$

For instance, when $P$ is homogeneous of degree $h$, then $M_{P,p}(r)$ is just $r^h\,M_{P,p}(1)$. When $P$ is not homogeneous, things get more complicated but the properties of $M_{P,p}$ that we will use are gathered in the following lemma.

**Lemma 5.6.** *If $P$ has no homogeneous component of degree $p$, then $M_{P,p}$ is a continuous function on $\mathbb{R}_+$ which is $C^\infty$ on $(0,\infty)$. Moreover, $M_{P,p}(t^{1/p})$ is a strictly concave function on $(0,\infty)$ whose first derivative is positive and whose second derivative is negative there. When $p\le 1$, the same is therefore true of the function $M_{P,p}(r)$ itself.*

**Proof.** Smoothness of $M_{P,p}$ follows from Lemma 5.5 and from the relation

$$M_{P,p} = L_r(x^*(r),\lambda^*(r)). \tag{5.17}$$

Continuity of $M_{P,p}$ at $0^+$ is obvious. Note also that

$$\frac{d\,M_{P,p}}{dr}(r) = \lambda^*(r), \tag{5.18}$$

which is an ultraclassical result in optimization asserting that the Lagrange multiplier can be interpreted as the sensitivity of the *optimal* value to the constraint level; (5.18) drops out immediately from (5.17) and from the fact that $(x^*(r),\lambda^*(r))$ is critical for $L_r$. From (5.18), we see that $dM_{P_p}/dr>0$ on $(0,\infty)$, hence also $dM_{P,p}(t^{1/p})/dt>0$ by the chain rule. Setting as usual $\Phi_p = P\circ\varphi_p$, we readily observe that

$$M_{P,p}(t^{1/p}) = M_{\Phi_p,1}(t), \tag{5.19}$$

and it follows then from (5.18) and (5.14) that $d^2M_{\Phi_p,1}/dr^2$ is never zero. It will in fact be negative for the function is concave; indeed, let $x^*$ and $y^*$ have $l^1$ norm $r_1$ and $r_2$ respectively, and be such that $\Phi_p(x^*) = M_{\Phi_p,1}(r_1)$ and $\Phi_p(y^*) = M_{\Phi_p,1}(r_2)$. Since $\Phi_p$ has degree at most 1, it is concave on $\mathbb{R}_+^n$ by Theorem 5.2. Hence, we get for $\mu_1\ge 0$ and $\mu_2\ge 0$ satisfying $\mu_1+\mu_2=1$:

$$\mu_1 M_{\Phi_p,1}(r_1) + \mu_2 M_{\Phi_p,1}(r_2) = \mu_1\Phi_p(x^*) + \mu_2\Phi_p(y^*) \le \Phi_p(\mu_1x^*+\mu_2y^*).$$

As $\|\mu_1x^*+\mu_2y^*\|_1 \le \mu_1r_1+\mu_2r_2$, it follows that

$$\Phi_p(\mu_1x^*+\mu_2y^*) \le M_{\Phi_p,1}(\mu_1r_1+\mu_2r_2),$$

for $M_{\Phi_p,1}$ is an increasing function. Finally, when $p\le 1$, it is immediate from the chain rule that $dM_{P,p}(t^{1/p})/dt>0$ and $d^2M_{P,p}(t^{1/p})/dt^2<0$ together imply $d^2M_{P,p}(t)/dt^2<0$. This achieves the proof. $\square$

We can now assess positivity and uniqueness in the general case for Problem (5.8). *For definiteness, we shall assume that $P$ has no constant coefficient.* This does not change the arguments of the *maximum* in Problem (5.8) and consequently does not affect our results. But it is to the effect that the irreducible components of $P$ will in turn be free of constant coefficients, hence homogeneous components will really be homogeneous, and not only up to an additive constant. So, assume that

$$P(x) = \sum_{\ell=1}^{k} P_{I_\ell}(x_{I_\ell}) \tag{5.20}$$

is the irreducible decomposition of $P$, each $I_\ell$ being of cardinality $n_\ell$, and, say $P_{I_1}, \cdots, P_{I_s}$ are homogeneous of degree $p$. Define $H = \cup_{\ell=1}^{s} I_\ell$, $K = \cup_{\ell=s+1}^{k} I_\ell$. Reordering the indices is necessary, we may suppose that $\{1, \cdots, m\}$ are those indices $\ell \le s$ for which $M_{P_{I_\ell},p}(1)$ is *maximum*, and we denote this *maximum* common value by $M_h$. This gives rise to a partition $H = H_1 \cup H_2$ of $H$, with

$$H_1 = \cup_{\ell=1}^{m} I_\ell \ \text{ and } \ H_2 = \cup_{\ell=m+1}^{s} I_\ell.$$

We set accordingly

$$P_{H_1}(x_{H_1}) = \sum_{\ell=1}^{m} P_{I_\ell}(x_{I_\ell}), \qquad P_{H_2}(x_{H_2}) = \sum_{\ell=m+1}^{s} P_{I_\ell}(x_{I_\ell}),$$

$$\text{and } P_K(x_K) = \sum_{\ell=s+1}^{k} P_{I_\ell}(x_{I_\ell}),$$

which depend on

$$n_{H_1} = \sum_{\ell=1}^{m} n_\ell, \qquad n_{H_2} = \sum_{\ell=m+1}^{s} n_\ell, \qquad \text{and } n_K = \sum_{\ell=s+1}^{k} n_\ell$$

variables respectively. These are generalized polynomial with non-negative coefficients, the first two being homogeneous of degree $p$ while the third is of degree at most $p$ and has *no* homogeneous irreducible component of degree $p$. For $1 \le \ell \le m$, we shall denote by $z^*_{I_\ell} \in \mathcal{P}_{n_\ell} \cap S^{n_\ell}_{p,+}$ the maximizing vector such that $P_{I_\ell}(z^*_{I_\ell}) = M_h$, whose existence and uniqueness is asserted in Theorem 5.4. If $s < k$, that is, if $P$ is not homogeneous of degree $p$, we further define $z^*_K(r) \in \mathcal{P}_{n_K} \cap \{\|x\|_p = r\}$ to be the solution to Problem (5.13) for $P_K$, whose existence and uniqueness again follows from Theorem 5.4, and by

$$\lambda^*_K(r) = r^{1-1/p}\, (z^*_K)_i^{1-p}(r)\, \frac{\partial P_K}{\partial i}(z^*_K(r))$$

the associated Lagrange multiplier as introduced in Lemma 5.5 (whose value does not depend on $i \in \{1, \cdots, n_K\}$). We simply set $z^*_K = z^*_K(1)$ and $\lambda^*_K = \lambda^*(1)$ for the pair associated with Problem (5.8). When $s = k$, then $K$ is of course empty so we should forget about $P_K$, $x^*_K$, $z^*_K$, and $\lambda^*_K$ in the statement of the next theorem.

**Theorem 5.7.** *With the above notations, the set of solutions to Problem* (5.8) *consists of those* $x^*$ *satisfying*

$$
\begin{aligned}
&x^*_{I_\ell} = \mu_\ell z^*_{I_\ell} \quad \text{for } 1 \le \ell \le m \quad \text{and } \mu_\ell \in \mathbb{R}^+ \quad \text{subject to } \sum_{\ell=1}^{m} \mu_\ell^p = 1 - r_0^p, \\
&x^*_{H_2} = 0, \qquad x^*_K = z^*_K(r_0),
\end{aligned}
\tag{5.21}
$$

*where* $r_0$ *is the unique maximum on* $[0,1]$ *of the function*

$$g(r) = (1 - r^p)M_h + M_{P_K,p}(r).$$

*We have* $0 < r_0 < 1$ *unless* $s = k$, *in which case* $r_0 = 0$, *or else* $s < k$ *and* $\lambda^*_K \ge pM_h$, *in which case* $r_0 = 1$ *and* $x^*_{H_1} = 0$. *In particular, the solution* $x^*$ *is unique if, and only if, either* $m = 1$ *or* $s < k$ *and* $\lambda^*_K \ge pM_h$ *(these two cases are not exclusive). There exists a solution in* $\mathcal{P}_n$ *if and only if* $m = s$ *(i. e.* $H_2$ *is void). Every solution lies in* $\mathcal{P}_n$ *if, and only if,* $m = s = 1$ *(hence the solution is unique) and* $\lambda^* < pM_h$ *in case* $s < k$*; then, the solution* $x^*$ *is the unique critical point of* $P$ *on* $S^n_{p,+} \cap \mathcal{P}_n$.

**Proof.** Let $x^*$ be a solution; then the optimal value is equal to

$$M_{P,p}(1) = M_{P_{H_1},p}(\|x^*_{H_1}\|_p) + M_{P_{H_2},p}(\|x^*_{H_2}\|_p) + M_{P_K,p}(\|x^*_K\|_p),$$

so, by homogeneity of $P_{H_1}$ and $P_{H_2}$,

$$M_{P,p}(1) = M_h\|x^*_{H_1}\|_p^p + \sum_{\ell=m+1}^{s} M_{P_{I_\ell},p}(1)\|x^*_{I_\ell}\|_p^p + +M_{P_K,p}(\|x^*_K\|_p). \tag{5.22}$$

As $M_h > M_{P_{I_\ell},p}(1)$ for $m+1 \le \ell \le s$, and since we are maximizing under the constraint

$$\sum_{\ell=1}^{k} \|x_{I_\ell}\|_p^p = 1,$$

it is clear that $x^*_{H_2}$ must be zero and that $x^*_{I_\ell} = \mu_\ell z^*_{I_\ell}$ for $1 \le \ell \le m$, where the $\mu_\ell$'s should satisfy $\sum_{\ell=1}^m \mu_\ell^p = 1 - \|x^*_K\|_p^p$ but otherwise produce the same value for $M_{P,p}(1)$. If $s = k$, then $K$ is void, $M_{P_K,p}$ is not present in (5.22), and $\|x^*_K\|_p$ should be interpreted as zero. Since also $r_0 = 0$ in this case, (5.21) then holds true. If $s < k$, (5.22) reads now:

$$M_{P,p}(1) = M_h(1 - \|x^*_K\|_p^p) + M_{P_K,p}(\|x^*_K\|_p) = g(\|x^*_K\|_p).$$

Setting $t = \|x^*_K\|_p^p$, we get

$$g(\|x^*_K\|_p) = M_h(1-t) + M_{P_K,p}(t^{1/p}),$$

and, since $P_K$ has no homogeneous components of degree $h$, we deduce from Lemma 5.6 that the above is a smooth strictly concave function of $t$ on $[0,\infty)$. Therefore, it has a unique *maximum* on $[0,1]$, which is necessarily attained at $t_0 = \|x^*_K\|_p^p$ by the optimality of $x^*$. From Proposition 5.3, we also see that $t_0 \ne 0$ for otherwise $P$ would reach a *maximum* on $S^n_{p,+}$ at the point $x^*$ satisfying $x^*_K = 0$, whereas $P_K$ is a sum of components, none of which is homogeneous of degree $p$. This shows in particular that $x_K^* = z_K^*(r_0)$, where $r_0 = t_0^{1/p}$ is indeed the *maximum* of $g(r)$ on $(0,1)$. Now, this *maximum* is attained on $(0,1]$, and $x_{H_1}$ will be zero if, and only if, it is attained at 1, that is, if, and only if,

$$\frac{g(t^{1/p})}{dt}(1) \geq 0.$$

Expanding the derivative using (5.18) yields then $\lambda_K^* \geq pM_h$. The remaining assertions are now obvious. □

**Remark 5.8.** 1) In principle, Theorem 5.7 reduces the general case of Problem (5.8) to a sequence of situations covered by Theorem 5.4 where the *optimum* can be computed by almost any method in optimization since, composing with $\varphi_p$, we are back to maximizing a smooth strictly concave function over a convex open subset of a linear space and we know the *maximum* is attained. In particular, the problem of deciding which variables are zero at an *optimum* is equivalent to determining the homogeneous irreducible components of degree $p$ –a combinatorial step– and then comparing the optimal values they achieve on Problem (5.8) –an analytical step– while these values can be computed rather easily as we just mentioned.

2) As we noticed already, a point $x^c \in S_{p,+}^n \cap \mathcal{P}_n$ is critical for $P$ if and only if there exists a Lagrange multiplier $\lambda$ such that

$$\lambda\,(x_i^c)^{p-1} = \frac{\partial P}{\partial i}(x^c) \quad \forall i \in \{1, \cdots, n\}. \tag{5.23}$$

Now, as in the proof of Theorem 5.2, let $\alpha_m$ be the smallest non-zero exponent for a variable appearing in $P$; changing $P$ into $P \circ \varphi_{\alpha_m}$ if necessary, we may assume in Problem (5.8) that $\alpha_m \geq 1$ implying $p \geq 1$ also. In this case, equation (5.23) makes sense for any point in $S_{p,+}^n$, that is even if some components of $x^c$ are equal to zero, so that we could *define* a critical point of $P$ on $S_{p,+}^n$ in this way . If $\alpha_m = p = 1$, then $P$ is a linear polynomial and the *maxima* of $P$ are generally not critical points. But if $p > 1$, they are critical because if $x^*$ is such a point and $I_1$, $I_2$ partition the set of indices in such a way that $x_{I_1}^* = 0$ while $x_i^* > 0$ for $i \in I_2$, then we know from Proposition 6.1 that $P$ decomposes as $P_1(x_{I_1}) + P_2(x_{I_2})$ where $P_1$ is homogeneous of degree $p$ and those equations in (5.23) that correspond to null components of $x^*$ can be read $0 = 0$ so that they are automatically satisfied. It is natural to ask for the converse, namely is a critical point necessarily a *maximum*? The answer is no: for $p \geq 3$, if we denote by $(x_1^c, x_2^c)$ the *maximum* of $x_1x_2$ on $S_{p,+}^2$, the point $(x_1^c, x_2^c, 0, 0)$ is critical for $P(x) = x_1x_2 + x_2x_3x_4$ on $S_{p,+}^4$ but is not a *maximum* for it does not belong to $\mathcal{P}_4$ though $P$ is irreducible.

*To recap, assuming* $\alpha_m \geq 1$, *we have in Problem* (5.8) *that a critical point lying in* $\mathcal{P}_n$ *is necessarily a solution whereas a critical point with some zero components may not be a solution.*

As a byproduct of the preceding discussion, we may also point out a kind of non-linear generalization of the Perron-Frobenius theorem [6] in case $A$ is symmetric.

**Corollary 5.9.** *Let* $A$ *be a real symmetric* $n \times n$ *matrix with non-negative entries. For any* $\alpha \geq 1$, *there exists a nonzero* $x^* \in \mathbb{R}_+^n$ *such that*

$$Ax^* = \lambda^* \varphi_{1/\alpha}(x^*), \qquad \textit{for some } \lambda^* \geq 0. \tag{5.24}$$

*If $A$ is irreducible, then $x^*$ is unique up to a multiplicative constant and belongs to $\mathcal{P}_n$. If $A$ is reducible, the solution is no longer unique. If $\alpha$ is an integer (so that equation (5.24) makes sense for negative $x$'s as well), the largest possible value for $\lambda^*$ is also the largest value of $|\lambda|$ for which*

$$Ax = \lambda \varphi_{1/\alpha}(x) \tag{5.25}$$

*is solvable with respect to $x \in \mathbb{R}^n - \{0\}$ for some (possibly negative) $\lambda$. When $\alpha > 1$ (the non-linear case), the vector $x^*$ associated with this largest value belongs to $\mathcal{P}_n$.*

**Proof.** If $A = 0$, there is nothing to prove. Otherwise, set $P(x) = x^T Ax$, which is homogeneous of degree 2. Consider Problem (5.8) with $p = \alpha + 1$. Let $I_1 \cdots I_k$ denote the irreducible partition of $A$, so that equation (5.24) splits into $k$ subequations in the $x_{I_j}$'s (because $A$ is symmetric). Now, the irreducible decomposition of $P$ writes

$$P(x) = \sum_{\ell=1}^{k} P_{I_\ell}(x_{I_\ell}), \tag{5.26}$$

each $P_{I_\ell}$ being again homogeneous of degree 2. Let $x^*$ be a solution to (5.8). By Proposition 5.3, the indices of the null coordinates of $x^*$ range over a union $\cup_{j \in J} I_j$, for some proper subset $J$ of $\{1 \cdots k\}$, and this union can be nonempty only when $p = 2$, that is, when $\alpha = 1$. In this case, the subequations of (5.24) corresponding to the $x_{I_j}$'s for $j \in J$ read $0 = 0$ and are automatically satisfied; the problem then reduces to a similar one in fewer variables. Altogether, we may assume $x^* \in \mathcal{P}_n$. Then, writing that $x^*$ is a critical point of $P$ on $S^n_{p,+} \cap \mathcal{P}_n$ and observing that the vector $\varphi_{1/(p-1)}(x^*)$ is normal to $S^n_{p,+}$ at $x^*$, we get $Ax^* = \lambda^* \varphi_{1/(p-1)}(x^*)$ for some $\lambda^*$ which is obviously nonnegative, that is, (5.24) is satisfied. If $A$ is irreducible, every non-zero solution to (5.24) belongs to $\mathcal{P}_n$ (easy checking) and Theorem 5.4 tells us that there is a unique critical point. This means that a solution of unit $l^p$ norm to (5.24) is unique in this case. If $A$ is reducible, a solution is no longer unique as it is clear from what precedes that we may set $x_{I_\ell}$ to zero for $\ell$ ranging over a strict subset $\{1, \cdots, k\}$, and still find a nonzero solution in terms of the remaining variables (of course this will not, in general, lead to a solution of (5.24) which is at the same time *a maximum* of $P$ on $S^n_{p,+}$). Finally, if $\alpha$ is an integer and $x \in \mathbb{R}^n$ any non-zero solution to (5.25) of unit $l^p$ norm, we have $x^T Ax = \lambda$ and, since $A$ is non-negative, this number cannot exceed the *maximum* of $P$ on $S^n_{p,+}$. □

**Remark 5.10.** (i) Extending a remark of [8] concerning this theorem, we may notice that *existence* of a solution to (5.24) would also follow from Brouwer's fixed-point theorem as applied to the map $x \to \varphi_\alpha(Ax)/\|\varphi_\alpha(Ax)\|_p$ from $S^n_{2,+}$ into itself.

(ii) Corollary 5.9 would remain valid for matrices depending on $x$, provided $Ax$ is the gradient of some non-negative generalized polynomial. This entails algebraic conditions that we shall not analyse here.

## 6. A generalization to product-type constraints

Some of the previous results extend easily to the case where the set of indices $I = \{1, \cdots, n\}$ is partitioned into $d$ blocks $J_1, \cdots, J_d$ of respective sizes $\nu_1, \cdots, \nu_d$, that is

$$J_i = \{x_k;\quad \sum_{\ell=1}^{i-1} \nu_\ell < k \le \sum_{\ell=1}^{i} \nu_\ell\}$$

(the empty sum occuring if $i = 1$ has of course to be interpreted as zero), with $\sum_i \nu_i = n$ and the constraint is now $\|x_{J_i}\|_p = 1$ for each $i \in \{1, \cdots, d\}$. We shall not analyse the solution to this generalized problem as completely as we did for Problem (5.8). However, since this generalization is relevant to the applications presented in the introduction, we shall proceed in this section with those results that can be stated in a fairly general manner and at the same time warrant the search for critical points in practice. Letting $\nu$ stand for the $d$-tuple $(\nu_1, \cdots, \nu_d)$, the set over which we optimize becomes

$$\mathcal{S}^{\nu}_{p,+} \stackrel{\text{def}}{=} S^{\nu_1}_{p,+} \times S^{\nu_2}_{p,+} \times \cdots S^{\nu_d}_{p,+},$$

and we state the problem formally as:

*Given $p > 0$ and a non-constant generalized polynomial $P$ with non-negative coefficients of degree $h$ with $h \leq p$, characterize the argument(s) of*

$$\max_{x \in \mathcal{S}^{\nu}_{p,+}} P(x). \tag{6.1}$$

Proposition 5.3 carries over *mutatis mutandis* to Problem (6.1):

**Proposition 6.1.** *Assume a solution $x^*$ to Problem (6.1) satisfies $x^*_i = 0$ for $i \in I_1$ and $x^*_i > 0$ for $i \in I_2$, where $I_1$ and $I_2$ partition the set of indices. Then one can decompose $P$ as*

$$P(x) = P_1(x_{I_1}) + P_2(x_{I_2}), \tag{6.2}$$

*where $P_1$ is some (possibly zero) homogeneous generalized polynomial with non-negative coefficients of degree $p$.*

**Proof.** The proof is similar to that of Proposition 5.3 except for two facts:

1) we cannot assume that $I_1 = \{1, \cdots, k\}$, because this time the indices have been fixed by the way we formulated the constraints. This creates only notational inconvenience.

2) the variable $y_j$ such that $y^*_j > 0$ has to be replaced by a collection $y_\alpha$ with $\alpha \in \Lambda$, where $y^*_\alpha > 0$ and $\Lambda$ contains exactly one element in each intersection $J_i \cap I_2$, for which $J_i \cap I_1 \neq \emptyset$ as $i$ ranges over $\{1, \cdots, d\}$. For $\alpha \in J_i \cap I_2$, we then define $n_\alpha > 0$ to be the cardinality of $J_i \cap I_1$.

Letting now $Y_t$ be the vector such that each component of $(Y_t)_{I_1}$ is $t$ and $(Y_t)_\alpha = y^*_\alpha - n_\alpha t$ while all other components of $Y_t$ are equal to those of $y^*$, we leave it to the reader to check that the proof of Proposition 5.3 carries over with obvious changes. □

We now obtain a straightforward generalization of Theorem 5.4:

**Theorem 6.2.** *The solutions to Problem (6.1) form a nonempty, connected, and closed subset of $\mathcal{S}^{\nu}_{p,+}$. The set of critical points and the set of maxima of $P$ coincide on $\mathcal{S}^{\nu}_{p,+} \cap \mathcal{P}_n$. If $P$ has no proper irreducible homogeneous components of degree $p$ (in particular if $P$ is irreducible), then there is a unique solution to Problem (6.1) and it lies on $\mathcal{S}^{\nu}_{p,+} \cap \mathcal{P}_n$ where it is thus the unique critical point of $P$. More generally, if $x^*$ denotes any solution to Problem (6.1), the components of $x^*$ that are not involved in a proper irreducible homogeneous component of degree $p$ are uniquely determined.*

**Proof.** We mimic the proof of Theorem 5.4. The set of solutions is nonempty and closed by the compactness of $\mathcal{S}^{\nu}_{p,+}$ and the continuity of $P$. Put $\Phi_p = P \circ \varphi_p$. Since $\Phi_p$ has degree at most 1, it is concave on $\mathbb{R}^n_+$ by Theorem 5.2 and so is its restriction to the convex subset $\mathcal{S}^{\nu}_{1,+}$ of the affine subspace

$$\{y \in \mathbb{R}^n;\ \sum_{j \in J_i} y_j = 1 \text{ for } 1 \leq i \leq d\};$$

consequently, the *maxima* of $\Phi_p$ on $\mathcal{S}^{\nu}_{1,+}$ form a convex set which is therefore connected an since they are the images under $\varphi_p^{-1}$ of the *maxima* of $P$ on $\mathcal{S}^{\nu}_{p,+}$, the latter must form a connected set as well. Further, we get by concavity that any critical point of $\Phi_p$ on the linear manifold $S^{\nu}_{1,+} \cap \mathcal{P}_n$ is a *maximum* of $\Phi_p$ with respect to $\mathcal{S}^{\nu}_{1,+}$, and conversely any *maximum* lying in $\mathcal{P}_n$ is a critical point. Since the critical points of $\Phi_p$ on $\mathcal{S}^{\nu}_{1,+} \cap \mathcal{P}_n$ are the images under $D\varphi_p^{-1}$ of those of $P$ on $\mathcal{S}^{\nu}_{p,+} \cap \mathcal{P}_n$, we see that critical points and *maxima* of $P$ coincide on $\mathcal{S}^{\nu}_{p,+} \cap \mathcal{P}_n$.

If $P$ has no proper irreducible component which is homogeneous of degree $p$, Proposition 6.1 tells us that the solutions to Problem (6.1) belong to $\mathcal{P}_n$ and the same then holds for any *maximum*, say $y^*$, of $\Phi_p$ on $\mathcal{S}^{\nu}_{1,+}$. We get, as in the proof of Theorem 5.4, that that the kernel of $D^2\Phi_p(y^*)$ is either zero or one dimensional generated by $y^*$, and therefore intersects

$$\mathcal{T}_{y^*}\, \mathcal{S}^{\nu}_{1,+} = \{y \in \mathbb{R}^n;\ \sum_{j \in J_i} y_j = 0 \text{ for } 1 \leq i \leq d\} \tag{6.3}$$

at zero only. Therefore, the critical points of $\Phi_p$ on $\mathcal{S}^{\nu}_{1,+} \cap \mathcal{P}_n$ are isolated, and the same is true of those of $P$ on $\mathcal{S}^{\nu}_{p,+} \cap \mathcal{P}_n$. Because we just proved they form a connected set, there is only one such point. More generally, if we assume that $\Phi_p$ assumes a *maximum* both at $y^*$ and $z^*$, then $\Phi_p(ty^* + (1-t)z^*)$ is also *maximum* for $0 \leq t \leq 1$ so the vector $y^* - z^*$ lies in the kernel of $D^2\Phi_p(y^*)$, and this implies by Theorem 5.2 that the coordinates of $y^* - z^*$ whose index is not involved in some homogeneous component of degree 1 do vanish. $\square$

**Remark 6.3.** The second remark we made after Theorem 5.7 remains of course valid: if the smallest exponent for a variable in $P$ is not less than 1, we can define critical points on $\mathcal{S}^{\nu}_{p,+}$. However, such a point need not be a *maximum* if it does not belong to $\mathcal{P}_n$.

At this point, it would be possible to derive an analog of Theorem 5.7 for Problem (6.1) by comparing the optimal values of the homogeneous and the remaining parts of $P$. However, it seems hardly worthwhile to build such a general statement because it would be rather intricated. Indeed, the function $M_{P,p}(r)$ defined in (5.16) should be replaced by some $M_{P,p}(r_1, \cdots, r_d)$ whose behaviour is more complex even if $P$ is homogeneous because it need not be homogeneous with respect to each $x_{J_i}$ separately. Leaving it to the interested reader to analyse further specific cases, we shall rather illustrate how multiple constraints may interact with the irreducible decomposition by giving a simple triangular criterion for uniqueness.

**Proposition 6.4.** *Let $P$ have $s \geq 1$ homogeneous irreducible components of degree p and let $I_1, \cdots, I_s$ be the corresponding elements of the irreducible partition. If the ordering $1, \cdots, s$ can be arranged so that for each $I_j$ there is a $i(j) \in \{1, \cdots, d\}$ with the property that $J_{i(j)} \cap I_j \neq \emptyset$ but $J_{i(j)} \cap I_k = \emptyset$ for $k < j$, then the solution to Problem* (6.1) *is unique.*

**Remark 6.5.** When $P$ is irreducible, this is nothing new in view of Theorem 6.2. We can even assert in this case that the solution is the unique critical point on $\mathcal{S}^{\nu}_{p,+} \cap \mathcal{P}_n$.

**Proof.** Let $x^*$ be a solution and assume first that $x^* \in \mathcal{P}_n$. Define as usual $\Phi_p = P \circ \varphi_p$, so that $y^* = \varphi_p^{-1}(x^*)$ is a critical point of $\Phi_p$ on $\mathcal{S}^{\nu}_{1,+}$. It follows from Theorem 5.2 that the kernel of $D^2\Phi_p(y^*)$ consists of vectors of the form $\sum_j \mu_j y^*_{I_j}$ with $\mu_j \in \mathbb{R}$. By (6.3), such a vector belongs to the tangent space of $\mathcal{S}^{\nu}_{1,+}$ if, and only if,

$$\sum_{j=1}^{s} \mu_j \sum_{\ell \in J_i \cap I_j} y^*_\ell = 0 \text{ for } 1 \leq i \leq d\}.$$

This means that the vector $\mu = (\mu_1, \cdots, \mu_s)^T$ lies in the kernel of the $d \times s$ matrix whose $(i,j)$ entry is $\sum_{\ell \in J_i \cap I_j} y^*_\ell$. By our hypothesis, this matrix has a nonsingular triangular submatrix of size $s$ so that $\mu = 0$ and $D^2\Phi_p(y^*)$ restricted to $\mathcal{T}_{y^*}\, \mathcal{S}^{\nu}_{1,+}$ is nonsingular. Consequently, $y^*$ is isolated and so is $x^*$. The latter is therefore unique by Theorem 6.2.

Assume now that $x^*$, hence also $y^*$, has some zero components whose indices then range over a union $\cup I_\alpha$ by Proposition 6.1. If $z^* \neq y^*$ is another point at which $\Phi_p$ attains a *maximum* on $\mathcal{S}^{\nu}_{1,+}$, the zero components of $z^*$ also have indices ranging over a union $\cup I_\beta$. By concavity, $\Phi_p$ attains a *maximum* at every point of the form $ty^* + (1-t)z^*$ with0 $< t < 1$, and the zero components of these points have indices ranging over $\cup I_\alpha \cap \cup I_\beta$ which is again a union, say, $\cup I_\gamma$. Setting $y_{I_\gamma} = 0$ for each $\gamma$, we get from $\Phi_p(y)$ a new generalized polynomial in fewer variables which still meets the assumptions (because we have merely suppressed a few homogeneous components) but attains a *maximum* at infinitely many points with positive components. This contradicts the first part of the proof and establishes the proposition. □

## 7. Finding the maximal mode of a Potts model

In this section, we study in more details the behaviour of DPA on the optimization of randomly weighted *square* Markov Random Fields. The idea is to generate reasonably large graphs, so that it is possible to make an exhaustive search of the optimal solution, and so perform a more objective evaluation of the results of DPA.

We build such an MRF in the following way. The sites are the pixels on an image, and the cliques are determined by the maximum distance of neighbouring pixels. This way, we successively study cliques of order 2, generated by 4-connectivity, then cliques of order 4, generated by 8-connectivity. We also vary the number of labels, or states, for each site, from 2 (corresponding to the Ising model), to 4 (an instance of a Potts model). Once the type of connectivity, the number of labels and the size of the graph (i.e. the image) have been selected, then the values of the clique potentials are generated randomly (typically with uniform distribution between 0 and 1, but the results do not change much if another distribution is used). In each case, 10 and 20 such random MRF's are generated. The value of the optimal configuration can be determined exactly by dynamic programming, following Derin & Elliott [4]. If the image has width $w$ and height $h$, then the complexity of the search is $hw^M$. Images of width up to 10, and height up to 20 can thus easily be searched on a workstation. The experiments related here were made on $5*5$ to $16*32$ images.

Applying DPA to this problem is straightforward. The maximization of $f$ on $\mathcal{K}_p$ can be performed by any method as long as $p \geq \deg(G)$, for we know that the maximum is unique in this case. As suggested in [1], we used a heuristic generalization of the iterative power method for finding an eigenvector corresponding to a maximum eigenvalue: we select some $X^{(0)}$ and subsequently compute

$$\forall i : X_i^{(n+1)} = \alpha_i^{(n+1)} \left( D_{X_i} f(X_i^{(n)}) \right)^{\frac{1}{p-1}}, \tag{7.1}$$

where $X_i$ denotes the vector $(x_{i,1} \cdots x_{i,M})$ and $D_{X_i} f$ the partial gradient with respect to these variables, while the $\alpha_i^{(n+1)}$'s are adjusted so that each $X_i$ has unit $l^p$-norm. This simply means that we select at each iteration the point on the pseudo-sphere of degree $p$ where the normal is parallel to the gradient of $f$, and the unique fixed point is necessarily the *maximum* we are looking for. Though this attractive procedure performed very well in our practice, the sequence of values for the criterion that are generated in this way need not be monotonically increasing, so one may have to fall back on a classical descent algorithm. Here are example of non–monotonicity:

**Example 7.1 (polynomial with integer exponents).** Let $n = 3$, $a$ be a positive real parameter, $f$ be the polynomial $xy + az^2$, and $p$ a real number strictly greater than 2.

Let $x_0$ be the point on $S^3_{p,+}$ with coordinates $((1 - \gamma_0^p)^{1/p}, 0, \gamma_0)$, where $0 < \gamma_0 < 1$. The gradient $g_0$ of $f$ at $x_0$ is the vector $(0, (1 - \gamma_0^p)^{1/p}, 2a\gamma_0)$. The unique point $x_1$ of $S^3_{p,+}$ where the normal is parallel to $g_0$ is

$$(0, (1 - \gamma_1^p)^{1/p}, \gamma_1),$$

where

$$\gamma_1 = \left( 1 + \frac{(1 - \gamma_0^p)^{\frac{1}{p-1}}}{(2a\gamma_0)^{\frac{p}{p-1}}} \right)^{\frac{-1}{p}}.$$

We get $f(x_0) = a\gamma_0^2$ and $f(x_1) = a\gamma_1^2$. As $\gamma_1$ vanishes when $a$ vanishes, we have that $f(x_1) < f(x_0)$ for $a$ small enough. That one coordinate is zero is unimportant: by perturbation we get an example with positive coordinates as well. The fact that $f$ has two irreducible components is not relevant either: just perturb by adding a monomial $xyz$ with a small coefficient to obtain, by continuity, an irreducible polynomial.

The example even shows that the sequence of values can be monotonically *decreasing*: since the polynomial $f$ is symmetric in $x$ and $y$, then we can obtain points $x_2, \ldots$ through the formula for $\gamma_1$; taking $p = 3, \gamma_0 = 0.5$, and $a = 10^{-4}$ generates a strictly decreasing sequence $f(x_n)$. In this case the sequence $x_n$ has two alternating points of accumulation yielding the same value for $f$:

$$y_0 := \left[ \left(1 - 8\,\frac{a^3}{8\,a^3+1}\right)^{1/3},\, 0\,,\, \frac{8^{1/3}\left(\frac{a^6}{8\,a^3+1}\right)^{1/3}}{a} \right]$$

and

$$y_1 := \left[ 0\,,\, \left(1 - 8\,\frac{a^3}{8\,a^3+1}\right)^{1/3},\, \frac{8^{1/3}\left(\frac{a^6}{8\,a^3+1}\right)^{1/3}}{a} \right].$$

**Example 7.2 (monomial with rational exponents).** This example shows that monotonicity may fail even for a monomial. Let $n = 2$, $p = 3$, and $f = x^{1/2}y^{3/2}$. Let $x_0$ be the point of $S^2_{3,+}$ with coordinates $(a, (1-a^3)^{1/3})$, where $0 < a < 1$.

Then the point $x_1$ of $S^2_{3,+}$ where the normal is parallel to the gradient of $f$ at $x_0$ has coordinates:

$$x_1 = \left( \frac{1}{2}\, \frac{\sqrt{2}\sqrt{\frac{\sqrt{1-a^3}}{\sqrt{a}}}}{\left(\frac{1}{4}\sqrt{2}\left(\frac{\sqrt{1-a^3}}{\sqrt{a}}\right)^{3/2} + \frac{3}{4}\sqrt{3}\sqrt{2}\,\left(\sqrt{a}\,(1-a^3)^{1/6}\right)^{3/2}\right)^{1/3}}, \right.$$
$$\left. \frac{1}{2}\, \frac{\sqrt{3}\sqrt{2}\sqrt{\sqrt{a}\,(1-a^3)^{1/6}}}{\left(\frac{1}{4}\sqrt{2}\left(\frac{\sqrt{1-a^3}}{\sqrt{a}}\right)^{3/2} + \frac{3}{4}\sqrt{3}\sqrt{2}\,\left(\sqrt{a}\,(1-a^3)^{1/6}\right)^{3/2}\right)^{1/3}} \right).$$

When $a$ vanishes, the limit of $f(x_1)/f(x_0)$ is zero because it is easily checked that $f(x_1)/f(x_0) = ka^{1/4} + o(a^{1/4})$ for some constant $k$. Thus, for sufficiently small $a$, $f(x_1) < f(x_0)$.

In order to account for the experimental monotonicity that was observed in practice by the authors, it would be interesting to know whether there always exist a neighbourhood of the maximum each point of which generates an increasing sequence of iterates.

Back to the labelling issue, and having initialized our continuation method using what precedes, we continue to lower the value of $p$ so that a bifurcation typically occurs for some unknown $p_b < \deg(C)$, and the maximum is no longer unique. Here comes the heuristic part of the procedure: we simply ignored the bifurcation, and we did carry on with the iterative power algorithm even when $p < p_b$. More precisely, the complete heuristic goes as follows:

(1) set $X^{(0)} = [1, \cdots, 1], k = 1$,

(2) while $p > 1$ do
- $Y^{(k)} \leftarrow$ the projection of $X^{(k-1)}$ on $\mathcal{K}_p$,
- find $X^{(k)} = \max_{\mathcal{K}_p} f(X)$ using the iterative power method, starting from $Y^{(k)}$,
- $k \leftarrow k+1$, $p \leftarrow p - \beta$
- od

Here, projecting a non-negative $Y$ on $\mathcal{K}_p$ is performed by scaling separately for each site, i.e. $X_i = \alpha_i Y_i, \forall i$, just as shown above.

For $p < p_b$, it may theoretically, happen that $Y^{(k)}$ is a critical point, possibly a local minimum, or a saddle-point, in which case the iterative power method gets stuck. This never occurred in our practice.

We have run comparisons with a Gibbs sampler, setting the number of iterations for it to the total number of iterations (summed over the $\beta$'s) set for DPA. The results are displayed on Tables 7.3 (for 4-connectivity and cliques of order 2) and 7.4 (for 8-connectivity and cliques of order 4). Here, $E_G$ is the average number of errors for Gibbs, and $V_G$ the average value of the criterion. The last column ($DPA/Gibbs$) displays the percentage of trials for which the value reached by DPA was better than the value reached by Gibbs.

The results on Table 7.3 show that, for 2 labels and cliques of order 2 (4-connectivity), DPA is definitely better, and quite close to optimal. For 3 or more labels, DPA is still better than Gibbs, but farther from optimal. For higher-order cliques (Table 7.4), DPA degrades faster than Gibbs.

We have also applied on larger graphs a search by dynamic programming with pruning of the current hypotheses, thus implementing a variation of the Viterbi algorithm. On an image 8-pixel wide, and 2 labels, an exhaustive search implies to maintain $2^9$ current hypotheses. We have found, as shown on Table 7.5, that the results are quite poor as soon as pruning (by discarding the worst hypothese) excedes 50%. This makes this last method just as intractable as optimal search.

| $N_\beta$ | $E_{DPA}$ | $V_{DPA}$ | $V_{opt}$ |
|---|---|---|---|
| 2 | 2.2 | 128.40 | 128.74 |
| 3 | 2.6 | 128.43 | 128.74 |
| 5 | 2.6 | 128.45 | 128.74 |

Table 7.1: Size $8 \times 8$, 2 labels, 4-connectivity, $N_{it} = 50$, $Th = 10^{-5}$

| $Th$ | $E_{DPA}$ | $V_{DPA}$ | $V_{opt}$ |
|---|---|---|---|
| $10^{-7}$ | 3.6 | 128.25 | 128.74 |
| 0.3 | 3.6 | 128.19 | 128.74 |

Table 7.2: Size $8 \times 8$, 2 labels, 4-connectivity, $N_\beta = 3$, $N_{it} = 5$

## 8. Conclusion

Having analysed in detail the $l_p$-constrained maximization problem for generalized polynomials of degree at most $p$, it is natural to ask about the best algorithmic approach to it. In the first place, our experimental success with the power iteration method, as mentioned

| $N_{lab}$ | $E_{DPA}$ | $E_G$ | $V_{DPA}$ | $V_G$ | $V_{opt}$ | $DPA/Gibbs$ |
|---|---|---|---|---|---|---|
| 2 | 12.1 | 29.4 | 211 | 206 | 214.2 | 100% |
| 3 | 20.0 | 27.7 | 131.3 | 128.5 | 136.3 | 90% |
| 4 | 13.7 | 20.3 | 73.2 | 70.9 | 77.7 | 100% |

Table 7.3: Size $10 \times 10$, 4-connectivity, $N_{it} = 2$, $N_\beta = 3$ $N_{itGibbs} = 6$

| $N_{lab}$ | $E_{DPA}$ | $E_G$ | $V_{DPA}$ | $V_G$ | $V_{opt}$ | $DPA/Gibbs$ |
|---|---|---|---|---|---|---|
| 2 | 14.6 | 15.9 | 288.9 | 292.6 | 298.0 | 30% |
| 3 | 15.2 | 18.2 | 155.5 | 158.1 | 164.3 | 20% |

Table 7.4: Size $8 \times 8$ or $6 \times 6$, 8-connectivity, $N_{it} = 3$, $N_\beta = 4$ $N_{itGibbs} = 12$

| $N_{hyp}$ | $E_{prun}$ | $V_{prun}$ | $V_{opt}$ |
|---|---|---|---|
| 256 | 11.2 | 128 | 129.3 |
| 16 | 22.7 | 123 | 129.3 |

Table 7.5: Size $8 \times 8$, 2 labels $\Rightarrow$ 512 running hypotheses.

in the previous section, is somewhat intriguing. Secondly, the results we proved are an invitation to interior point methods, as one may hope that such methods can be provided by perturbing the polynomial to be maximized so as to make it irreducible.

Finally, in connection with the original motivation presented in Section 2, more ambitious questions arise about what happens if the constraint on the degree is relaxed. Deterministic Pseudo-Annealing has been used for a variety of applications, and proven to be an efficient parallel and deterministic substitute to stochastic methods like Simulated Annealing. While uniqueness of the solution to the deformed problem was experimentally conjectured a while ago, the present paper contributes the establishment of the method by making sure that no pathological situation may arise at this early stage. However, when performing subsequent steps, the rate of decreasing of the degree while restoring the original constraints may change the solution we reach, and no theoretical foundations are presently available about such continuation methods.

## References

[1] M. Berthod: Definition of a consistent labeling as a global extremum, International Conference on Pattern Recognition, Munich (1982) 339–341.

[2] M. Berthod, S. Liu-Yu, J. P. Stromboni: Deterministic pseudo-annealing: a new optimization scheme applied to texture segmentation, International Conference on Pattern Recognition, vol. 2, The Hague (1992) 533–536.

[3] J. E. Besag: Spatial interaction and the statistical analysis of lattice systems (with discussion), Journal of Royal Statis. Society B. 36 (1974) 192–236.

[4] H. Derin, H. Elliott, R. Cristi, D. Geman: Bayes smoothing algorithms for segmentation of binary images modeled by markov random fields, IEEE Transactions on Pattern Analysis and Machine Intelligence 6 (1984) 107–120.

[5] O. D. Faugeras, M. Berthod: Improving consistency and reducing ambiguity in stochastic

labeling: an optimization approach, IEEE Transactions on Pattern Analysis and Machine Intelligence 4 (1981) 412–423.

[6] F. R. Gantmacher: The Theory of Matrices, vol. I,II, Chelsea, 1959.

[7] S. Geman, D. Geman: Stochastic relaxation, Gibbs distributions and the Bayesian restoration of images, IEEE Transactions on Pattern Analysis and Machine Intelligence 6 (1984) 721–741.

[8] V. Guillemin, A. Pollack: Differential Topology, Prentice-Hall, 1974.

[9] R. Hummel, S. Zucker: On the foundations of relaxation labeling, IEEE Transactions on Pattern Analysis and Machine Intelligence 5(3) (1983) 267–287.

[10] A. Rosenfeld, R. A. Hummel, S. W. Zucker: Scene labeling by relaxation operations, IEEE Transactions on Systems, Man and Cybernetics 6 (1976) 420–433.

[11] L. G. Shapiro, R. Haralick: Structural description and inexact matching, IEEE Transactions on Pattern Analysis and Machine Intelligence 3(5) (1981) 504–519.