CONSTRUCTIVE MATHEMATICAL ANALYSIS

Volume VI
Issue II

CMA
CONSTRUCTIVE MATHEMATICAL ANALYSIS

ISSN 2651-2939

https://dergipark.org.tr/en/pub/cma

VOLUME VI ISSUE II
ISSN 2651-2939

June 2023
https://dergipark.org.tr/en/pub/cma

CONSTRUCTIVE MATHEMATICAL ANALYSIS

CMA
CONSTRUCTIVE MATHEMATICAL ANALYSIS

Editor-in-Chief

Managing Editor

Editorial Board

Editorial Staff

Contents

CONSTRUCTIVE MATHEMATICAL ANALYSIS
6 (2023), No. 2, pp. 77-89
http://dergipark.org.tr/en/pub/cma
ISSN 2651 - 2939

Research Article

Existence and uniqueness of viscosity solutions to the infinity Laplacian relative to a class of Grushin-type vector fields

THOMAS BIESKE AND ZACHARY FORREST*

ABSTRACT. In this paper, we pose the ∞-Laplace equation as a Dirichlet Problem in a class of Grushin-type spaces whose vector fields are of the form

$$X_k(p) := \sigma_k(p)\frac{\partial}{\partial x_k}$$

and σ_k is not a polynomial for indices $m+1 \leq k \leq n$. Solutions to the ∞-Laplacian in the viscosity sense have been shown to exist and be unique in [3], when σ_k is a polynomial; we extend these results by exploiting the relationship between Grushin-type and Euclidean second-order jets and utilizing estimates on the viscosity derivatives of sub- and supersolutions in order to produce a comparison principle for semicontinuous functions.

Keywords: ∞-Laplace equation, viscosity solution, Grushin-type spaces.

2020 Mathematics Subject Classification: Primary 53C17, 35D40, 35J94; Secondary 35H20, 22E25, 17B70.

1. INTRODUCTION

In [3] the author considers the Dirichlet Problem

$$\begin{cases} \Delta_\infty w = 0 & \text{in } \Omega \\ w = g & \text{on } \partial\Omega \end{cases} \tag{1.1}$$

and establishes conditions under which a viscosity solution (see Section 3) to (1.1) exists and is unique when the problem is posed in a wide variety of Grushin-type spaces. The goal of the current paper is to extend the existence/uniqueness results of [3] to a more general class of Grushin-type spaces.

The spaces under consideration in [3] are defined by Lie Algebras consisting of vector fields of the form

$$Y_k(p) := P_k(p)\frac{\partial}{\partial x_k} \text{ for } k \leq n, \tag{1.2}$$

where P_k is a polynomial in the variables x_i ($i \leq k-1$) and $P_1 \equiv 1$. The current paper considers the situation when the vector fields are of the form

$$X_k(p) := \sigma_k(p)\frac{\partial}{\partial x_k} \text{ for } k \leq n, \tag{1.3}$$

where $\sigma_k : \mathbb{R}^n \to \mathbb{R}$ need not be a polynomial when $k > m \geq 1$. Grushin-type spaces defined by vector fields as in (1.2) are known to possess certain desirable properties – e.g. it is known

Received: 04.02.2023; Accepted: 27.04.2023; Published Online: 02.05.2023
*Corresponding author: Zachary Forrest; zachary9@usf.edu
DOI: 10.33205/cma.1245581

that the vector fields Y_j and their commutators

$$[Y_j, Y_k], [Y_j, [Y_k, Y_\ell]], [Y_j, [Y_k, [Y_\ell, Y_m]]], \ldots$$

span $\mathbb{R}^n$ and hence we may apply Chow's Theorem to conclude that points of the related Grushin-type space may be connected by appropriately smooth curves. Spaces defined by vector fields as in (1.3), however, can not be treated this way and require modified techniques.

The article will proceed as follows. In Section 2 we will define the spaces of interest and consider notions of geometry and calculus. The trappings of viscosity theory are introduced in Section 3, and a lemma relating Euclidean and Grushin second-order jets is presented. We conclude with Section 4 in which we produce results necessary to establish a comparison principle for sub- and supersolutions and existence of solutions – the culmination of these results is the theorem below.

Main Theorem. *Let $\mathbb{G}$ be a Grushin-type space whose Lie Algebra consists of vectors fields as defined in the forthcoming section. Then there exists a unique solution to the Dirichlet Problem* (1.1).

2. The Grushin-Type Environment $\mathbb{G}$

Let $n \geq 2$ and $1 \leq m < n$ be given. Fixing any $p = (x_1, \ldots, x_n) \in \mathbb{R}^n$, consider the frame $\{X_i, X_j\}$ containing the vector fields

$$X_i(p) := \frac{\partial}{\partial x_i} \ (1 \leq i \leq m) \tag{2.4}$$

and

$$X_j(p) := \sigma(p)\frac{\partial}{\partial x_j} \ (m+1 \leq j \leq n), \tag{2.5}$$

where we will assume that:

(1) $\sigma(p) = \sigma(x_1, \ldots, x_m)$. That is, $\sigma(p)$ is independent of $x_{m+1}, \ldots, x_n$.
(2) σ is Euclidean C^2 (denoted C^2_{eucl} for what follows).
(3) The set of zeroes for σ is given by $Z \times \mathbb{R}^{n-m}$, where Z is a discrete subset of $\mathbb{R}^m$.

In the case that σ is a polynomial the frame $\{X_i, X_j\}$ defines a generalized Grushin space such as the ones under consideration in [3]; otherwise $\{X_i, X_j\}$ corresponds to a member of a more general class of Grushin-type spaces.

The Lie Algebra $\mathfrak{g} := \operatorname{span}\{X_i, X_j\}$ may be endowed with an inner-product $\langle \cdot, \cdot \rangle$ which is singular on $Z \times \mathbb{R}^{n-m}$ and makes $\{X_i, X_j\}$ an orthonormal basis otherwise. Defining the space $\mathbb{G}$ to be the image of $\mathfrak{g}$ under the exponential map, note that points of $\mathbb{G}$ are also n-tuples $p = (x_1, \ldots, x_n)$ and that the tangent space to $\mathbb{G}$ at any point p is $\mathfrak{g}(p)$. One consequence of this definition is that $\mathbb{G}$ is not a group: Indeed, the dimension of the tangent space to $\mathbb{G}$ at p is $\dim \mathfrak{g}(p)$ which equals m if $p \in Z \times \mathbb{R}^{n-m}$ and otherwise equals n.

The natural metric to impose upon $\mathbb{G}$ is the *Carnot-Carathéodory* (or CC) metric

$$d_{CC}(p, q) := \inf_{\gamma \in \Gamma} \int_0^1 \|\gamma'(t)\| dt, \tag{2.6}$$

where Γ is the collection of all curves γ satisfying **(i)** $\gamma(0) = p, \gamma(1) = q$ and **(ii)** $\gamma' \in \mathfrak{g}$. Because $X_j \equiv 0$ on $Z \times \mathbb{R}^{n-m}$, Chow's Theorem (see, for example, [5]) does not apply. However, since the vector fields X_i are nonzero, points of $\mathbb{G}$ can always be connected by concatenating curves – so $\Gamma \neq \varnothing$ and $d_{CC}(\cdot, \cdot)$ is an honest metric.

We may therefore define balls in $\mathbb{G}$ by

$$B(p_0, r) := \{p \in \mathbb{G} : d_{CC}(p_0, p) < r\}$$

and consider notions of bounded domains, which we shall typically denote by $\Omega \Subset \mathbb{G}$.

Given a smooth function $u : \mathcal{O} \to \mathbb{R}$ where $\mathcal{O} \subseteq \mathbb{G}$ is open, the gradient of u in $\mathbb{G}$ is defined by

$$\nabla_{\mathbb{G}}\, u := (X_1 u, \ldots, X_n u)$$

and the second derivative matrix $\left(D^2 u\right)^\star$ is the symmetric $n \times n$ matrix whose entries are given by

$$[\left(D^2 u\right)^\star]_{k\ell} := \frac{1}{2}\left(X_\ell X_k u + X_k X_\ell u\right).$$

We also have notions of regularity.

Definition 2.1. *A function $u : \mathcal{O} \to \mathbb{R}$ is said to be $C^1_{\mathbb{G}}(\mathcal{O})$ if $X_k u$ is continuous for each $1 \le k \le n$. The function u is $C^2_{\mathbb{G}}(\mathcal{O})$ if $X_\ell X_k u$ is continuous for each $1 \le k, \ell \le n$.*

Finally, given $1 \le \mathrm{p} \le \infty$, we also may define the function spaces $L^{\mathrm{p}}(\mathcal{O}), L^{\mathrm{p}}_{\mathrm{loc}}(\mathcal{O}), W^{1,\mathrm{p}}(\mathcal{O})$ and $W^{1,\mathrm{p}}_{\mathrm{loc}}(\mathcal{O})$ in the obvious way.

3. Jets & Viscosity Solutions

With the appropriate definitions of derivatives and function spaces introduced in the previous section, we turn our attention to homogeneous PDEs of the form

$$H(p, \eta, X) = 0 \tag{3.7}$$

for $\eta \in \mathbb{R}^n$ and symmetric $n \times n$ matrices X (frequently denoted $X \in \mathcal{S}^n$). The operators H will be continuous and *proper* in the sense of [6]: That is, for $X \le Y$ we will have $H(p, \eta, Y) \le H(p, \eta, X)$. Specifically, assuming that w is smooth, we will have interest in the ∞-Laplace operator

$$\Delta_\infty\, w := -\left\langle \left(D^2 w\right)^\star \nabla_{\mathbb{G}}\, w, \nabla_{\mathbb{G}}\, w \right\rangle;$$

the related p-Laplace operators (for $1 < \mathrm{p} < \infty$)

$$\begin{aligned}\Delta_{\mathrm{p}} w &:= -\operatorname{div}\left(\|\nabla_{\mathbb{G}}\, w\|^{\mathrm{p}-2}\, \nabla_{\mathbb{G}}\, w\right)\\ &= -\|\nabla_{\mathbb{G}}\, w\|^{\mathrm{p}-2} \sum_{a=1}^{n} X_a X_a w + (\mathrm{p}-2)\|\nabla_{\mathbb{G}}\, w\|^{\mathrm{p}-4}\, \Delta_\infty\, w;\end{aligned}$$

and Jensen's Auxiliary Functions (see [7])

$$\mathcal{F}^\varepsilon(p, \nabla_{\mathbb{G}}\, w, \left(D^2 w\right)^\star) := \min\left\{\|\nabla_{\mathbb{G}}\, w\|^2 - \varepsilon^2, \Delta_\infty\, w\right\}$$

and

$$\mathcal{G}^\varepsilon(p, \nabla_{\mathbb{G}}\, w, \left(D^2 w\right)^\star) := \max\left\{\varepsilon^2 - \|\nabla_{\mathbb{G}}\, w\|^2, \Delta_\infty\, w\right\},$$

where $\varepsilon \in \mathbb{R}$ will be given. In what follows, we will use H to represent any of the four operators above.

In order to introduce the machinery of viscosity solutions to $Hw = 0$, we first must consider the following classes of test functions which "touch" the function $u : \mathcal{O} \to \mathbb{R}$. Given an open set $\mathcal{O} \subseteq \mathbb{G}$, a point $p_0 \in \mathcal{O}$, and a function $u : \mathcal{O} \to \mathbb{R}$, we have the so-called "touching above" functions

$$\mathcal{TA}(u, p_0) := \left\{\varphi \in C^2_{\mathbb{G}}(\Omega) :\ 0 = \varphi(p_0) - u(p_0) < \varphi(p) - u(p) \text{ near } p_0\right\};$$

we have also the "touching below" functions at p_0 defined by

$$\mathcal{TB}(u, p_0) := \left\{\varphi \in C^2_{\mathbb{G}}(\Omega) :\ 0 = u(p_0) - \varphi(p_0) < u(p) - \varphi(p) \text{ near } p_0\right\}.$$

Comparisons between the derivatives of smooth functions w and the touching functions φ, and between the operations $Hw, H\varphi$ then lead us to make the following definition.

Definition 3.2. *Let $\Omega \Subset \mathbb{G}$ be a domain and let $u \in \mathrm{USC}(\Omega)$. We say that u is a viscosity subsolution to (3.7) in Ω if the following is satisfied: For every $p \in \Omega$ and each $\varphi \in \mathcal{TA}(u,p)$,*

$$H(p, \nabla_{\mathbb{G}} \varphi(p), \big(D^2\varphi\big)^\star(p)) \leq 0.$$

We say that $v \in \mathrm{LSC}(\Omega)$ is a viscosity supersolution to Equation (3.7) if $-v$ is a viscosity subsolution to Equation (3.7). We say that $w \in C(\Omega)$ is a viscosity solution to Equation (3.7) if it is both a viscosity sub- and supersolution.

When convenient, we may also speak in terms of "jets" for a function u at a point p_0.

Definition 3.3. *Given $u : \mathcal{O} \to \mathbb{R}$, we define the second-order upper jet for u by*

$$J^{2,+}\, u(p_0) := \Big\{\big(\nabla_{\mathbb{G}} \varphi(p_0), \big(D^2\varphi\big)^\star(p_0)\big) \in \mathbb{R}^n \times \mathcal{S}^n : \varphi \in \mathcal{TA}(u,p_0)\Big\}$$

and the second-order lower jet for u by $J^{2,-}\, u(p_0) := -\, J^{2,+}[-u](p_0)$. We say that the ordered pair $(\eta, X) \in \mathbb{R}^n \times \mathcal{S}^n$ belongs to the closure of the upper jet, written $(\eta, X) \in \overline{J}^{2,+}\, u(p_0)$, if there exists $(p_k) \subseteq \mathcal{O}$ and jet entries $(\eta_k, X_k) \in J^{2,+}\, u(p_k)$ so that

$$(p_k, u(p_k), \eta_k, X_k) \to (p_0, u(p_0), \eta, X);$$

the definition for $\overline{J}^{2,-}\, u(p_0)$ is similar.

Remark 3.1. *Definition 3.2 above can also be stated equivalently through the lens of the jet closures: $u \in \mathrm{USC}(\Omega)$ is a viscosity subsolution if for every $p \in \Omega$*

$$H(p, \eta, X) \leq 0$$

for each $(\eta, X) \in \overline{J}^{2,+}\, u(p)$. Similar restatements can be made for viscosity supersolutions and viscosity solutions.

Remark 3.2. *If it should happen that $H = \Delta_{\mathrm{p}}$, then we will call solutions to (3.7) p-harmonic; if $H = \Delta_\infty$, then we call solutions to (3.7) infinite harmonic.*

The jets for $\mathbb{G}$ can be related to Euclidean jets via the following lemma, which is an application of [4, Corollary 3.2].

Lemma 3.1 (The $\mathbb{G}$ Twisting Lemma). *Let $\mathcal{O} \subseteq \mathbb{G}$ be open, let $u : \mathcal{O} \to \mathbb{R}$, and let $p_0 \in \mathcal{O}$. Suppose that we know $(\eta, X) \in J^{2,+}_{\text{eucl}}(u, p_0)$: Then*

$$\Big(\boldsymbol{A}(p_0) \cdot \eta, \boldsymbol{A}(p_0) \cdot X \cdot \boldsymbol{A}^{\mathrm{T}}(p_0) + \boldsymbol{M}(\eta, p_0)\Big) \in J^{2,+}(u, p_0), \tag{3.8}$$

where

$$(\boldsymbol{A}(p_0))_{k\ell} = \begin{cases} 1, & k = \ell \leq m \\ \sigma(p_0), & m+1 \leq k = \ell \leq n \\ 0, & \text{otherwise} \end{cases} \tag{3.9}$$

and

$$(\boldsymbol{M}(\eta, p_0))_{k\ell} = \begin{cases} \dfrac{1}{2} \cdot \dfrac{\partial \sigma}{\partial x_k}(p_0)\eta_\ell, & 1 \leq k \leq m < \ell \leq n \\ \dfrac{1}{2} \cdot \dfrac{\partial \sigma}{\partial x_\ell}(p_0)\eta_k, & 1 \leq \ell \leq m < k \leq n \\ 0, & \text{otherwise.} \end{cases} \tag{3.10}$$

Proof. The result in (3.8) is known (see [4, Corollary 3.2] and [1, Lemma 3]); we shall restrict our attention to verifying Equations (3.9) and (3.10). The $n \times n$ matrix $\boldsymbol{A}$ is defined by [4] as $\boldsymbol{A}(p) := (A_{k\ell}(p))$, where

$$X_k(\cdot) = \sum_{\ell=1}^{n} A_{k\ell}(\cdot)\frac{\partial}{\partial x_\ell}.$$

The definitions (2.4) and (2.5) imply:

(1) $A_{k\ell} \equiv 0$ if $k \neq \ell$;
(2) $A_{kk} \equiv 1$ if $k \leq m$ and $A_{kk} = \sigma$ if $m+1 \leq k \leq n$.

This justifies (3.9).

To verify (3.10), recall the definition of $\boldsymbol{M}(\eta, p_0)$ in [4]:

$$(\boldsymbol{M}(\eta,p_0))_{k\ell} := \begin{cases} \frac{1}{2}\sum\limits_{r=1}^{n}\sum\limits_{s=1}^{n}\left(A_{ks}(p_0)\frac{\partial A_{\ell r}}{\partial x_s}(p_0) + A_{\ell s}(p_0)\frac{\partial A_{kr}}{\partial x_s}(p_0)\right)\eta_r, & k \neq \ell \\ \sum\limits_{r=1}^{n}\sum\limits_{s=1}^{n} A_{ks}(p_0)\frac{\partial A_{kr}}{\partial x_s}(p_0)\eta_r, & k = \ell. \end{cases}$$

Because $A_{rs} \equiv 0$ whenever $r \neq s$, we may simplify the equation above:

$$\begin{aligned}(\boldsymbol{M}(\eta,p_0))_{k\ell} =& \frac{1}{2}\sum_{r=1}^{n}\left(\left(A_{kk}(p_0)\frac{\partial A_{\ell r}}{\partial x_k}(p_0) + 0\right) + \left(0 + A_{\ell\ell}(p_0)\frac{\partial A_{kr}}{\partial x_\ell}(p_0)\right)\right)\eta_r \\ =& \frac{1}{2}\left(A_{kk}(p_0)\frac{\partial A_{\ell\ell}}{\partial x_k}(p_0)\eta_\ell + A_{\ell\ell}(p_0)\frac{\partial A_{kk}}{\partial x_\ell}(p_0)\eta_k\right) \text{ if } k \neq \ell,\end{aligned} \tag{3.11}$$

and

$$(\boldsymbol{M}(\eta,p_0))_{kk} = \sum_{r=1}^{n} A_{kk}(p_0)\frac{\partial A_{kr}}{\partial x_k}(p_0)\eta_r = A_{kk}(p_0)\frac{\partial A_{kk}}{\partial x_k}(p_0)\eta_k \text{ if } k = \ell. \tag{3.12}$$

First consider Equation (3.12). If $k = 1, \ldots, m$, then $\partial A_{kk}/\partial x_k \equiv 0$. If $k = m+1, \ldots, n$ we also have $\partial A_{kk}/\partial x_k \equiv 0$ because σ is independent of the variables $x_{m+1}, \ldots, x_n$. Hence, $(\boldsymbol{M}(\eta,p_0))_{kk} = 0$ for all $k \leq n$.

Turning our attention to Equation (3.11), we reduce the expression utilizing Item 2 and the definition of σ:

- If $k, \ell \leq m$, then $A_{kk} \equiv 1 \equiv A_{\ell\ell}$ and hence
$$(\boldsymbol{M}(\eta,p_0))_{k\ell} = \frac{1}{2}\left(1 \cdot 0 \cdot \eta_\ell + 1 \cdot 0 \cdot \eta_k\right) = 0.$$
- If $k \leq m < \ell \leq n$, then $A_{kk} \equiv 1$ and $A_{\ell\ell} = \sigma$. Since σ is constant with respect to $x_{m+1}, \ldots, x_n$,
$$\begin{aligned}(\boldsymbol{M}(\eta,p_0))_{k\ell} =& \frac{1}{2}\left(1 \cdot \frac{\partial \sigma}{\partial x_k}(p_0)\eta_\ell + \sigma(p_0) \cdot 0 \cdot \eta_k\right) \\ =& \frac{1}{2} \cdot \frac{\partial \sigma}{\partial x_k}(p_0)\eta_\ell.\end{aligned}$$
- If $\ell \leq m < k \leq n$, then work similar to the above shows
$$(\boldsymbol{M}(\eta,p_0))_{k\ell} = \frac{1}{2} \cdot \frac{\partial \sigma}{\partial x_\ell}(p_0)\eta_k.$$
- If $m < k, \ell \leq n$, then $A_{kk} = \sigma = A_{\ell\ell}$ and so
$$(\boldsymbol{M}(\eta,p_0))_{k\ell} = \frac{1}{2}\left(\sigma(p_0) \cdot 0 \cdot \eta_\ell + \sigma(p_0) \cdot 0 \cdot \eta_k\right) = 0.$$

We conclude from the above that the matrix given by (3.10) is indeed $\boldsymbol{M}(\eta, p_0)$. □

4. Uniqueness of Infinite Harmonic Functions

It is standard knowledge (see, for example, [2] and [8]) that there exist solutions to the Equation (3.7), so we turn our attention to uniqueness of these solutions. This will be achieved by proving uniqueness for the operators $\mathcal{F}^\varepsilon$ and $\mathcal{G}^\varepsilon$, and will rely upon the properties of jet entries.

4.1. **Iterated Maximum Principle & Estimates on Derivatives.** The focus of this subsection is Lemma 4.4, which requires the Iterated Maximum Principle of [3]. As we shall show in Lemma 4.4, the Iterated Maximum Principle gives conditions for finding points possessing nonempty jet closures for viscosity sub- and supersolutions – this will enable us to produce necessary estimates on the "viscosity derivatives". As in [6], we will have need for a "penalty function"; specifically, we make use of the function

$$\varphi_{\tau_1,\tau_2,\tau_3,\ldots,\tau_n}(p,q) = \varphi_{\vec{\tau}}(p,q) := \frac{1}{2}\sum_{k=1}^{n} \tau_k (x_k - y_k)^2,$$

where the entries of $\vec{\tau} = (\tau_1, \tau_2, \tau_3, \ldots, \tau_n)$ are positive, real numbers. The use of n real parameters as opposed to the one employed by [6] allows us to take the set $Z \times \mathbb{R}^{n-m}$ into account.

Lemma 4.2 (The Iterated Maximum Principle). *Let $\Omega \Subset \mathbb{G}$ be a domain, $u \in$ USC(Ω), and $v \in$ LSC(Ω); assume that there exists some $p_0 \in \Omega$ so that*

$$u(p_0) - v(p_0) > 0.$$

Let $\vec{\tau} = (\tau_1, \tau_2, \tau_3, \ldots, \tau_n) \in \mathbb{R}^n$ have positive coordinates and, for each pair of points in $\mathbb{G}$ $p = (x_1, x_2, x_3, \ldots, x_n), q = (y_1, y_2, y_3, \ldots, y_n)$ define the functions

$$\begin{aligned}
\varphi_{\tau_1,\tau_2,\tau_3,\ldots,\tau_n}(p,q) &:= \frac{1}{2}\sum_{k=1}^{n} \tau_k (x_k - y_k)^2 \\
\varphi_{\tau_2,\tau_3,\ldots,\tau_n}(p,q) &:= \frac{1}{2}\sum_{k=2}^{n} \tau_k (x_k - y_k)^2 \\
\varphi_{\tau_3,\ldots,\tau_n}(p,q) &:= \frac{1}{2}\sum_{k=3}^{n} \tau_k (x_k - y_k)^2 \\
&\vdots \\
\varphi_{\tau_n}(p,q) &:= \frac{1}{2}\tau_n (x_n - y_n)^2.
\end{aligned}$$

Appealing to the compactness of $\overline{\Omega}$ and to upper semicontinuity, we may also define

$$\begin{aligned}
M_{\tau_1,\tau_2,\tau_3,\ldots,\tau_n} &:= \sup_{\overline{\Omega}\times\overline{\Omega}} \{u(p) - v(q) - \varphi_{\tau_1,\tau_2,\tau_3,\ldots,\tau_n}(p,q)\} \\
&= u(p_{\tau_1,\tau_2,\tau_3,\ldots,\tau_n}) - v(q_{\tau_1,\tau_2,\tau_3,\ldots,\tau_n}) - \varphi_{\tau_1,\tau_2,\tau_3,\ldots,\tau_n}(p_{\tau_1,\tau_2,\tau_3,\ldots,\tau_n}, q_{\tau_1,\tau_2,\tau_3,\ldots,\tau_n}) \\
M_{\tau_2,\tau_3,\ldots,\tau_n} &:= \sup_{\overline{\Omega}\times\overline{\Omega}} \{u(p) - v(q) - \varphi_{\tau_2,\tau_3,\ldots,\tau_n}(p,q) : x_1 = y_1\} \\
&= u(p_{\tau_2,\tau_3,\ldots,\tau_n}) - v(q_{\tau_2,\tau_3,\ldots,\tau_n}) - \varphi_{\tau_2,\tau_3,\ldots,\tau_n}(p_{\tau_2,\tau_3,\ldots,\tau_n}, q_{\tau_2,\tau_3,\ldots,\tau_n}) \\
M_{\tau_3,\ldots,\tau_n} &:= \sup_{\overline{\Omega}\times\overline{\Omega}} \{u(p) - v(q) - \varphi_{\tau_3,\ldots,\tau_n}(p,q) : x_k = y_k,\ k = 1,2\} \\
&= u(p_{\tau_3,\ldots,\tau_n}) - v(q_{\tau_3,\ldots,\tau_n}) - \varphi_{\tau_3,\ldots,\tau_n}(p_{\tau_3,\ldots,\tau_n}, q_{\tau_3,\ldots,\tau_n}) \\
&\vdots \\
M_{\tau_n} &:= \sup_{\overline{\Omega}\times\overline{\Omega}} \{u(p) - v(q) - \varphi_{\tau_3,\ldots,\tau_n}(p,q) : x_k = y_k,\ k = 1,\ldots,n-1\} \\
&= u(p_{\tau_n}) - v(q_{\tau_n}) - \varphi_{\tau_n}(p_{\tau_n}, q_{\tau_n}).
\end{aligned}$$

Then

$$\lim_{\tau_n\to\infty} \cdots \lim_{\tau_3\to\infty} \lim_{\tau_2\to\infty} \lim_{\tau_1\to\infty} M_{\tau_1,\tau_2,\tau_3,\ldots,\tau_n} = u(p_0) - v(p_0)$$

and

$$\lim_{\tau_n\to\infty} \cdots \lim_{\tau_3\to\infty} \lim_{\tau_2\to\infty} \lim_{\tau_1\to\infty} \varphi_{\tau_1,\tau_2,\tau_3,\ldots,\tau_n}(p_{\tau_1,\tau_2,\tau_3,\ldots,\tau_n}, q_{\tau_1,\tau_2,\tau_3,\ldots,\tau_n}) = 0.$$

Additionally, the first ℓ coordinates of $p_{\tau_{\ell+1},\ldots,\tau_n}$ and $q_{\tau_{\ell+1},\ldots,\tau_n}$ are identical – that is,

$$x_k^{\tau_{\ell+1},\ldots,\tau_n} = y_k^{\tau_{\ell+1},\ldots,\tau_n},\ k = 1,\ldots,\ell.$$

The proof of the Iterated Maximum Principle leads immediately to the following results which permit us to take the parameters $\tau_k \to \infty$ in any order, and to speak of the full limit as $\tau_{k_1}, \tau_{k_2}, \ldots, \tau_{k_n} \to \infty$.

Corollary 4.1 (cf. [3, Corollary 4.4])**.** *Under the conditions of Lemma 4.2, each iterated limit of $M_{\tau_1,\tau_2,\tau_3,\ldots,\tau_n}$ exists and is equal to $u(p_0) - v(p_0)$ – in other words,*

$$\lim_{\tau_{k_1}\to\infty} \cdots \lim_{\tau_{k_{n-2}}\to\infty} \lim_{\tau_{k_{n-1}}\to\infty} \lim_{\tau_{k_n}\to\infty} M_{\tau_1,\tau_2,\tau_3,\ldots,\tau_n} = u(p_0) - v(p_0).$$

Consequently,

$$\lim_{\tau_{k_1}\to\infty} \cdots \lim_{\tau_{k_{n-2}}\to\infty} \lim_{\tau_{k_{n-1}}\to\infty} \lim_{\tau_{k_n}\to\infty} \varphi_{\tau_1,\tau_2,\tau_3,\ldots,\tau_n}(p_{\tau_1,\tau_2,\tau_3,\ldots,\tau_n}, q_{\tau_1,\tau_2,\tau_3,\ldots,\tau_n}) = 0.$$

Lemma 4.3 (cf. [3, Lemma 4.5])**.** *Under the conditions of Lemma 4.2, the full limit of $M_{\tau_1,\tau_2,\tau_3,\ldots,\tau_n}$ exists and is equal to $u(p_0) - v(p_0)$ – more precisely,*

$$\lim_{\tau_n,\ldots,\tau_3,\tau_2,\tau_1\to\infty} M_{\tau_1,\tau_2,\tau_3,\ldots,\tau_n} = u(p_0) - v(p_0).$$

In addition,

$$\lim_{\tau_n,\ldots,\tau_3,\tau_2,\tau_1\to\infty} \varphi_{\tau_1,\tau_2,\tau_3,\ldots,\tau_n}(p_{\tau_1,\tau_2,\tau_3,\ldots,\tau_n}, q_{\tau_1,\tau_2,\tau_3,\ldots,\tau_n}) = 0.$$

Remark 4.3. *Owing to Lemma 4.3, there is no ambiguity in relabeling the intermediate points $p_{\tau_1,\tau_2,\tau_3,\ldots,\tau_n}$, $q_{\tau_1,\tau_2,\tau_3,\ldots,\tau_n}$ and function $\varphi_{\tau_1,\tau_2,\tau_3,\ldots,\tau_n}$ as $p_{\vec{\tau}}$, $q_{\vec{\tau}}$, and $\varphi_{\vec{\tau}}$. We will also denote the coordinates of $p_{\vec{\tau}}$, $q_{\vec{\tau}}$ as $x_k^{\vec{\tau}}, y_k^{\vec{\tau}}$ respectively.*

Applying the results above and [6, Theorem 3.2], we have the following estimates.

Lemma 4.4. *Let $u, v, \varphi_{\vec{\tau}}$, and $(p_{\vec{\tau}}, q_{\vec{\tau}})$ be as in Lemma 4.2 and assume additionally that at least one of u, v is locally $\mathbb{G}$-Lipschitz. Then:*

(A) There exist $(\eta^+_{\vec{\tau}}, \mathcal{X}_{\vec{\tau}}) \in \overline{J}^{2,+} u(p_{\vec{\tau}})$ and $(\eta^-_{\vec{\tau}}, \mathcal{Y}_{\vec{\tau}}) \in \overline{J}^{2,-} v(q_{\vec{\tau}})$.

(B) Define $(p\diamond q)_k$ to be the point whose k-th coordinate coincides with q and whose other coordinates coincide with p – in other words,

$$(p \diamond q)_k = (x_1, \dots, x_{k-1}, y_k, x_{k+1}, \dots, x_n).$$

Then for each index k,

$$\tau_k (x_k^{\vec{\tau}} - y_k^{\vec{\tau}})^2 \lesssim d_{CC} \left(p_{\vec{\tau}}, (p_{\vec{\tau}} \diamond q_{\vec{\tau}})_k\right). \tag{4.13}$$

For the indices $i \leq m$,

$$\tau_i \left| x_i^{\vec{\tau}} - y_i^{\vec{\tau}} \right| = O(1) \text{ as } \tau_i \to \infty. \tag{4.14}$$

(C) The vector estimate

$$\left| \left\| \eta^+_{\vec{\tau}} \right\|^2 - \left\| \eta^-_{\vec{\tau}} \right\|^2 \right| = o\left(1\right) \text{ as } \tau_k \to \infty \text{ for all } k \leq n \tag{4.15}$$

holds.

(D) The matrix estimate

$$\left\langle \mathcal{X}^{\vec{\tau}} \eta^+_{\vec{\tau}}, \eta^+_{\vec{\tau}} \right\rangle - \left\langle \mathcal{Y}^{\vec{\tau}} \eta^-_{\vec{\tau}}, \eta^-_{\vec{\tau}} \right\rangle = o\left(1\right) \text{ as } \tau_k \to \infty \text{ for all } k \leq n \tag{4.16}$$

holds.

Proof. For clarity, we split the proof between the items above.

Item (A).

[6, Theorem 3.2] guarantees the existence of elements in the Euclidean jet closures: In particular, for each fixed $\delta > 0$ we will have

$$\left(D_p \varphi_{\vec{\tau}}(p_{\vec{\tau}}, q_{\vec{\tau}}), X^{\vec{\tau}}\right) \in \overline{J}^{2,+}_{\text{eucl}} u(p_{\vec{\tau}}) \text{ and } \left(-D_q \varphi_{\vec{\tau}}(p_{\vec{\tau}}, q_{\vec{\tau}}), Y^{\vec{\tau}}\right) \in \overline{J}^{2,-}_{\text{eucl}} v(q_{\vec{\tau}}).$$

Applying the $\mathbb{G}$ Twisting Lemma (Lemma 3.1) produces the members $(\eta^+_{\vec{\tau}}, \mathcal{X}_{\vec{\tau}}) \in \overline{J}^{2,+} u(p_{\vec{\tau}})$ and $(\eta^-_{\vec{\tau}}, \mathcal{Y}_{\vec{\tau}}) \in \overline{J}^{2,-} v(q_{\vec{\tau}})$.

Item (B).

By the definition of the points $p_{\vec{\tau}}, q_{\vec{\tau}}$, for all points $p, q \in \Omega$ the inequality

$$u(p) - v(q) - \varphi_{\vec{\tau}}(p, q) \leq u(p_{\vec{\tau}}) - v(q_{\vec{\tau}}) - \varphi_{\vec{\tau}}(p_{\vec{\tau}}, q_{\vec{\tau}})$$

is satisfied. Hence assuming (without loss of generality) that u is $\mathbb{G}$-Lipschitz, decreeing $p := (p_{\vec{\tau}} \diamond q_{\vec{\tau}})_k$ and $q := q_{\vec{\tau}}$, and recollecting terms, we obtain

$$\begin{aligned} \tau_k (x_k^{\vec{\tau}} - y_k^{\vec{\tau}})^2 &= \varphi_{\vec{\tau}}(p_{\vec{\tau}}, q_{\vec{\tau}}) - \varphi_{\vec{\tau}}\left((p_{\vec{\tau}} \diamond q_{\vec{\tau}})_k, q_{\vec{\tau}}\right) \\ &\leq u(p_{\vec{\tau}}) - u\left((p_{\vec{\tau}} \diamond q_{\vec{\tau}})_k\right) \\ &\leq K\, d_{CC}\left(p_{\vec{\tau}}, (p_{\vec{\tau}} \diamond q_{\vec{\tau}})_k\right), \end{aligned} \tag{4.17}$$

where K is the Lipschitz constant for u. This is Inequality (4.13), so to complete 4.4 we turn our attention to to the expression $\tau_i \left| x_i^{\vec{\tau}} - y_i^{\vec{\tau}} \right|$ $(i \leq m)$. If $x_i^{\vec{\tau}} \neq y_i^{\vec{\tau}}$ then (4.17) shows

$$\tau_i \left| x_i^{\vec{\tau}} - y_i^{\vec{\tau}} \right| = \tau_i (x_i^{\vec{\tau}} - y_i^{\vec{\tau}})^2 \cdot \frac{1}{\left| x_i^{\vec{\tau}} - y_i^{\vec{\tau}} \right|} \leq \frac{K\, d_{CC}\left(p_{\vec{\tau}}, (p_{\vec{\tau}} \diamond q_{\vec{\tau}})_i\right)}{\left| x_i^{\vec{\tau}} - y_i^{\vec{\tau}} \right|} \text{ as } \tau_1, \cdots, \tau_n \to \infty. \tag{4.18}$$

Note that

$$d_{CC}\left(p_{\vec{\tau}}, (p_{\vec{\tau}} \diamond q_{\vec{\tau}})_i\right) \leq \left| x_i^{\vec{\tau}} - y_i^{\vec{\tau}} \right| \tag{4.19}$$

as a consequence of [5, Theorem 7.34]. Combining (4.18) and (4.19) proves Equation (4.14) and completes the proof of 4.4.

Item (C).

Observe that

$$\frac{\partial}{\partial x_k}\varphi(p_{\vec{\tau}}, q_{\vec{\tau}}) = \tau_k(x_k^{\vec{\tau}} - q_k^{\vec{\tau}}) = -\frac{\partial}{\partial y_k}\varphi(p_{\vec{\tau}}, q_{\vec{\tau}});$$

consequently, referring back to the definition of the matrix $\boldsymbol{A}$, the coordinates of $\eta_{\vec{\tau}}^+$ and $\eta_{\vec{\tau}}^-$ are

$$[\eta_{\vec{\tau}}^+]_k = \begin{cases} \tau_k(x_k^{\vec{\tau}} - y_k^{\vec{\tau}}), & \text{if } k \le m \\ \tau_k(x_k^{\vec{\tau}} - y_k^{\vec{\tau}})\sigma(p_{\vec{\tau}}), & \text{if } m+1 \le k \le n \end{cases}$$

and

$$[\eta_{\vec{\tau}}^-]_k = \begin{cases} \tau_k(x_k^{\vec{\tau}} - y_k^{\vec{\tau}}), & \text{if } k \le m \\ \tau_k(x_k^{\vec{\tau}} - y_k^{\vec{\tau}})\sigma(q_{\vec{\tau}}), & \text{if } m+1 \le k \le n. \end{cases}$$

Fixing $\vec{\tau}$ for the moment, this leads to the estimate

$$\text{(4.20)} \qquad \left| \|\eta_{\vec{\tau}}^+\|^2 - \|\eta_{\vec{\tau}}^-\|^2 \right| \le \sum_{k=m+1}^{n} \left|\sigma^2(p_{\vec{\tau}}) - \sigma^2(q_{\vec{\tau}})\right| \cdot \tau_k^2 \left(x_k^{\vec{\tau}} - y_k^{\vec{\tau}}\right)^2.$$

The values τ_i for $i \le m$ are not present in Inequality (4.20). Taking the iterated limits of (4.20) as $\tau_i \to \infty$, recalling that $\sigma(p)$ depends only upon the first m coordinates of p, and applying the Iterated Maximum Principle yields

$$\lim_{\tau_m\to\infty} \cdots \lim_{\tau_1\to\infty} \left| \|\eta_{\vec{\tau}}^+\|^2 - \|\eta_{\vec{\tau}}^-\|^2 \right| = 0.$$

The above implies

$$\lim_{\tau_n\to\infty} \cdots \lim_{\tau_{m+1}\to\infty} \lim_{\tau_m\to\infty} \cdots \lim_{\tau_1\to\infty} \left| \|\eta_{\vec{\tau}}^+\|^2 - \|\eta_{\vec{\tau}}^-\|^2 \right| = 0,$$

concluding Item **(C)**.

Item (D).

[6, Theorem 3.2] and the Twisting Lemma imply

$$\langle \mathcal{X}^{\vec{\tau}}\eta_{\vec{\tau}}^+, \eta_{\vec{\tau}}^+ \rangle - \langle \mathcal{Y}^{\vec{\tau}}\eta_{\vec{\tau}}^-, \eta_{\vec{\tau}}^- \rangle = I_1 + I_2,$$

where we define

$$I_1 := \left\langle \left(\boldsymbol{A}(p_{\vec{\tau}}) \cdot X^{\vec{\tau}} \cdot \boldsymbol{A}^{\mathrm{T}}(p_{\vec{\tau}})\right) \cdot \eta_{\vec{\tau}}^+, \eta_{\vec{\tau}}^+ \right\rangle - \left\langle \left(\boldsymbol{A}(q_{\vec{\tau}}) \cdot Y^{\vec{\tau}} \cdot \boldsymbol{A}^{\mathrm{T}}(q_{\vec{\tau}})\right) \cdot \eta_{\vec{\tau}}^-, \eta_{\vec{\tau}}^- \right\rangle$$

and

$$\text{(4.21)} \qquad I_2 := \langle \boldsymbol{M}(D_p\varphi_{\vec{\tau}}(p_{\vec{\tau}}, q_{\vec{\tau}}), p_{\vec{\tau}}) \cdot \eta_{\vec{\tau}}^+, \eta_{\vec{\tau}}^+ \rangle - \langle \boldsymbol{M}(D_q\varphi_{\vec{\tau}}(p_{\vec{\tau}}, q_{\vec{\tau}}), q_{\vec{\tau}}) \cdot \eta_{\vec{\tau}}^-, \eta_{\vec{\tau}}^- \rangle.$$

Writing $\widetilde{\epsilon} := \boldsymbol{A}(p_{\vec{\tau}}) \cdot \epsilon, \widetilde{\kappa} := \boldsymbol{A}(q_{\vec{\tau}}) \cdot \kappa$ to mean the twisting of $\epsilon, \kappa \in \mathbb{R}^n$ according to the Twisting Lemma,

$$\begin{aligned} \left\langle \boldsymbol{A}(p_{\vec{\tau}}) \cdot X^{\vec{\tau}} \cdot \boldsymbol{A}^{\mathrm{T}}(p_{\vec{\tau}})\epsilon, \epsilon \right\rangle - \left\langle \boldsymbol{A}(q_{\vec{\tau}}) \cdot Y^{\vec{\tau}} \cdot \boldsymbol{A}^{\mathrm{T}}(q_{\vec{\tau}})\kappa, \kappa \right\rangle &= \langle X^{\vec{\tau}} \cdot \widetilde{\epsilon}, \widetilde{\epsilon} \rangle - \langle Y^{\vec{\tau}} \cdot \widetilde{\kappa}, \widetilde{\kappa} \rangle \\ &\le \langle \mathcal{C} \cdot \Upsilon, \Upsilon \rangle, \end{aligned}$$

where $\Upsilon := (\widetilde{\epsilon}, \widetilde{\kappa})$ and $\mathcal{C}$ is a $2n \times 2n$ block matrix resulting from [6, Theorem 3.2] of the form

$$\begin{pmatrix} B & -B \\ -B & B \end{pmatrix}$$

and

$$[B]_{ab} = \begin{cases} \tau_a + 2\delta\tau_a^2, & a = b \\ 0, & a \neq b. \end{cases}$$

(Recall that δ is a consequence of [6, Theorem 3.2].) Choosing $\epsilon := \eta^+_{\vec{\tau}}$ and $\kappa := \eta^-_{\vec{\tau}}$, the above shows

$$\begin{aligned} I_1 &\leq \left\langle B \cdot \left(\widetilde{\eta^+_{\vec{\tau}}} - \widetilde{\eta^-_{\vec{\tau}}} \right), \widetilde{\eta^+_{\vec{\tau}}} - \widetilde{\eta^-_{\vec{\tau}}} \right\rangle \\ &= \sum_{k=m+1}^{n} (\tau_k + 2\delta\tau_k^2)(\sigma^2(p_{\vec{\tau}}) - \sigma^2(q_{\vec{\tau}}))^2 \cdot \tau_k^2 (x_k^{\vec{\tau}} - y_k^{\vec{\tau}})^2. \end{aligned} \tag{4.22}$$

The right-hand side of Relation (4.22) is free of the τ_i for $i \leq m$, so proceeding as in the proof of Item **(C)**, we find

$$\lim_{\tau_m \to \infty} \cdots \lim_{\tau_1 \to \infty} I_1 = 0$$

so that

$$\lim_{\tau_n \to \infty} \cdots \lim_{\tau_{m+1} \to \infty} \lim_{\tau_m \to \infty} \cdots \lim_{\tau_1 \to \infty} I_1 = 0. \tag{4.23}$$

For the term I_2, let us begin by simplifying the notation for the matrix $\boldsymbol{M}(\cdot,\cdot)$. Appealing to Equation (3.10) in the Twisting Lemma, we see that

$$\boldsymbol{M}(D_p\varphi_{\vec{\tau}}(p_{\vec{\tau}}, q_{\vec{\tau}}), p_{\vec{\tau}}) = \begin{pmatrix} 0 & S(p_{\vec{\tau}}) \\ S(p_{\vec{\tau}})^T & 0 \end{pmatrix}$$

and

$$\boldsymbol{M}(D_q\varphi_{\vec{\tau}}(p_{\vec{\tau}}, q_{\vec{\tau}}), q_{\vec{\tau}}) = \begin{pmatrix} 0 & S(q_{\vec{\tau}}) \\ S(q_{\vec{\tau}})^T & 0 \end{pmatrix},$$

where, permitting t to represent either the point $p_{\vec{\tau}}$ or $q_{\vec{\tau}}$, the $m \times (n-m)$ matrix $S(t)$ is defined by

$$[S(t)]_{rs} := \frac{1}{2} \cdot \frac{\partial \sigma}{\partial x_r}(t) \cdot \tau_s (x_s^{\vec{\tau}} - y_s^{\vec{\tau}}).$$

Calculations with (4.21) show

$$\begin{aligned} I_2 = &\sum_{\ell=m+1}^{n} \sum_{r=1}^{m} \frac{\partial \sigma}{\partial x_r}(p_{\vec{\tau}}) \cdot \tau_r (x_r^{\vec{\tau}} - y_r^{\vec{\tau}}) \cdot \tau_\ell^2 (x_\ell^{\vec{\tau}} - y_\ell^{\vec{\tau}})^2 \sigma(p_{\vec{\tau}}) \\ - &\sum_{\ell=m+1}^{n} \sum_{r=1}^{m} \frac{\partial \sigma}{\partial x_r}(q_{\vec{\tau}}) \cdot \tau_r (x_r^{\vec{\tau}} - y_r^{\vec{\tau}}) \cdot \tau_\ell^2 (x_\ell^{\vec{\tau}} - y_\ell^{\vec{\tau}})^2 \sigma(q_{\vec{\tau}}). \end{aligned}$$

We adopt the notation

$$T_{r\ell} := \tau_r (x_r^{\vec{\tau}} - y_r^{\vec{\tau}}) \tau_\ell^2 (x_\ell^{\vec{\tau}} - y_\ell^{\vec{\tau}})^2 \left(\frac{\partial \sigma}{\partial x_r} \cdot \sigma \right)(p_{\vec{\tau}}) - \tau_r (x_r^{\vec{\tau}} - y_r^{\vec{\tau}}) \tau_\ell^2 (x_\ell^{\vec{\tau}} - y_\ell^{\vec{\tau}})^2 \left(\frac{\partial \sigma}{\partial x_r} \cdot \sigma \right)(q_{\vec{\tau}})$$

for the (r, ℓ)-term of I_2. Since the Iterated Maximum Principle implies

$$p_{\vec{\tau}} \to (x_1^0, \ldots, x_i^0, x_{i+1}^{\vec{\tau}}, \ldots, x_n^{\vec{\tau}}) \text{ and } q_{\vec{\tau}} \to (x_1^0, \ldots, x_i^0, y_{i+1}^{\vec{\tau}}, \ldots, y_n^{\vec{\tau}})$$

as $\tau_1, \dots \tau_i \to \infty$ ($i \le m$), and since $1 \le r \le m < \ell \le n$ and $\sigma \in C^2_{\text{eucl}}$, we obtain the iterated limit

$$\begin{aligned}\lim_{\tau_i\to\infty} \cdots \lim_{\tau_1\to\infty} T_{r\ell} = {} & \tau_r(x_r^{\vec{\tau}} - y_r^{\vec{\tau}})\tau_\ell^2(x_\ell^{\vec{\tau}} - y_\ell^{\vec{\tau}})^2 \left(\frac{\partial\sigma}{\partial x_r}\cdot\sigma\right)(x_1^0, \dots, x_i^0, x_{i+1}^{\vec{\tau}}, \dots, x_n^{\vec{\tau}}) \\ & - \tau_r(x_r^{\vec{\tau}} - y_r^{\vec{\tau}})\tau_\ell^2(x_\ell^{\vec{\tau}} - y_\ell^{\vec{\tau}})^2 \left(\frac{\partial\sigma}{\partial x_r}\cdot\sigma\right)(x_1^0, \dots, x_i^0, y_{i+1}^{\vec{\tau}}, \dots, y_n^{\vec{\tau}})\end{aligned}$$

if $i < r$; if $r \le i$ we may apply Item 4.4, Inequality (4.14), and arrive at

$$\begin{aligned}\lim_{\tau_i\to\infty} \cdots \lim_{\tau_1\to\infty} T_{r\ell} \approx {} & \tau_\ell^2(x_\ell^{\vec{\tau}} - y_\ell^{\vec{\tau}})^2 \left(\frac{\partial\sigma}{\partial x_r}\cdot\sigma\right)(x_1^0, \dots, x_i^0, x_{i+1}^{\vec{\tau}}, \dots, x_n^{\vec{\tau}}) \\ & - \tau_\ell^2(x_\ell^{\vec{\tau}} - y_\ell^{\vec{\tau}})^2 \left(\frac{\partial\sigma}{\partial x_r}\cdot\sigma\right)(x_1^0, \dots, x_i^0, y_{i+1}^{\vec{\tau}}, \dots, y_n^{\vec{\tau}}).\end{aligned}$$

This second limit in particular implies that

$$\begin{aligned}\lim_{\tau_m\to\infty} \cdots \lim_{\tau_1\to\infty} T_{r\ell} \approx {} & \tau_\ell^2(x_\ell^{\vec{\tau}} - y_\ell^{\vec{\tau}})^2 \left(\frac{\partial\sigma}{\partial x_r}\cdot\sigma\right)(x_1^0, \dots, x_m^0, x_{m+1}^{\vec{\tau}}, \dots, x_n^{\vec{\tau}}) \\ & - \tau_\ell^2(x_\ell^{\vec{\tau}} - y_\ell^{\vec{\tau}})^2 \left(\frac{\partial\sigma}{\partial x_r}\cdot\sigma\right)(x_1^0, \dots, x_m^0, y_{m+1}^{\vec{\tau}}, \dots, y_n^{\vec{\tau}})\end{aligned} \tag{4.24}$$

for all $r \le m$. Since $\sigma, \partial\sigma/\partial x_r$ depend only upon the first m coordinates of points p, (4.24) implies

$$\lim_{\tau_m\to\infty} \cdots \lim_{\tau_1\to\infty} I_2 = 0$$

and hence

$$\lim_{\tau_n\to\infty} \cdots \lim_{\tau_{m+1}\to\infty} \lim_{\tau_m\to\infty} \cdots \lim_{\tau_1\to\infty} I_2 = 0. \tag{4.25}$$

Equation (4.16) then follows from (4.23) and (4.25). □

4.2. **A Comparison Principle & Uniqueness.** With the completion of Lemma 4.4, we prove a comparison principle for viscosity solutions to the Dirichlet problems

$$\begin{cases} \mathcal{F}^\varepsilon\left(p, \nabla_{\mathbb{G}}\, w(p), \left(D^2 w\right)^\star(p)\right) = \min\left\{\|\, \nabla_{\mathbb{G}}\, w(p)\|^2 - \varepsilon^2, \Delta_\infty w(p)\right\} = 0, & p \in \Omega \\ w(p) = g(p), & p \in \partial\Omega \end{cases} \tag{4.26}$$

and

$$\begin{cases} \mathcal{G}^\varepsilon\left(p, \nabla_{\mathbb{G}}\, w(p), \left(D^2 w\right)^\star(p)\right) = \max\left\{\varepsilon^2 - \|\, \nabla_{\mathbb{G}}\, w(p)\|^2, \Delta_\infty w(p)\right\} = 0, & p \in \Omega \\ w(p) = g(p), & p \in \partial\Omega \end{cases} \tag{4.27}$$

in order to prove the uniqueness of solutions to

$$\begin{cases} \Delta_\infty\, w(p) = \left\langle D^2 w(p) \cdot \nabla_{\mathbb{G}}\, w(p), \nabla_{\mathbb{G}}\, w(p)\right\rangle = 0, & p \in \Omega \\ w(p) = g(p), & p \in \partial\Omega \end{cases} \tag{4.28}$$

As before, Ω is a bounded domain; we also assume $g \in C(\partial\Omega)$. In the interest of maintaining clear notation, we establish the following convention.

Definition 4.4. *A viscosity supersolution to Problems (4.26), (4.27), or (4.28) is a viscosity supersolution v to the equations $\mathcal{F}^\varepsilon = 0$, $\mathcal{G}^\varepsilon = 0$, or $\Delta_\infty = 0$ (respectively) such that $v \ge g$ on $\partial\Omega$; a viscosity supersolution u to Problems (4.26), (4.27), or (4.28) is defined similarly. A viscosity solution to any of the above three Dirichlet problems is both a viscosity sub- and supersolution to the problem in the above sense.*

Theorem 4.1. *Suppose that u, v are sub- and supersolutions to Problem* (4.26) *or Problem* (4.27) *such that at least one of the functions is locally $\mathbb{G}$-Lipschitz in Ω. Then $u \leq v$ in Ω.*

Proof. We will complete the proof for Problem (4.26) and note that the proof for Problem (4.27) is similar. Suppose, to the contrary of the theorem, that there exists some $p_0 \in \Omega$ such that

$$u(p_0) - v(p_0) = \max_{\overline{\Omega}}(u - v) > 0.$$

Appealing to Lemma 5.1 and Theorem 5.3 in [2], we may assume that v is a strict viscosity supersolution to $\mathcal{F}^\varepsilon = 0$ – that is, there exists $\mu(p) > 0$ so that

$$\mathcal{F}^\varepsilon\left(\nabla_{\mathbb{G}}\, v(p), \left(D^2 v\right)^\star(p)\right) = \mu(p) > 0$$

holds in the viscosity sense for each $p \in \Omega$. Applying Lemma 4.4 to produce the sequence of ordered pairs $(p_{\vec{\tau}}, q_{\vec{\tau}}) \in \Omega \times \Omega$, we have

$$\begin{aligned}
0 < \mu(q_{\vec{\tau}}) &\leq \mathcal{F}^\varepsilon\left(\eta_{\vec{\tau}}^-, \mathcal{Y}^{\vec{\tau}}\right) - \mathcal{F}^\varepsilon\left(\eta_{\vec{\tau}}^+, \mathcal{X}^{\vec{\tau}}\right) \\
&= \min\left\{\left\|\eta_{\vec{\tau}}^-\right\|^2 - \varepsilon^2, -\left\langle \mathcal{Y}^{\vec{\tau}}\eta_{\vec{\tau}}^-, \eta_{\vec{\tau}}^-\right\rangle\right\} - \min\left\{\left\|\eta_{\vec{\tau}}^+\right\|^2 - \varepsilon^2, -\left\langle \mathcal{X}^{\vec{\tau}}\eta_{\vec{\tau}}^+, \eta_{\vec{\tau}}^+\right\rangle\right\} \\
&\leq \max\left\{\left\|\eta_{\vec{\tau}}^-\right\|^2 - \left\|\eta_{\vec{\tau}}^+\right\|^2, \left\langle \mathcal{X}^{\vec{\tau}}\eta_{\vec{\tau}}^+, \eta_{\vec{\tau}}^+\right\rangle - \left\langle \mathcal{Y}^{\vec{\tau}}\eta_{\vec{\tau}}^-, \eta_{\vec{\tau}}^-\right\rangle\right\}.
\end{aligned} \tag{4.29}$$

[2, Lemma 5.1], [2, Theorem 5.3], Lemma 4.4, and Lemma 4.2 imply

$$\mu(q_{\vec{\tau}}) \to \mu(p_0) > 0 \tag{4.30}$$

and that

$$\max\left\{\left\|\eta_{\vec{\tau}}^-\right\|^2 - \left\|\eta_{\vec{\tau}}^+\right\|^2, \left\langle \mathcal{X}^{\vec{\tau}}\eta_{\vec{\tau}}^+, \eta_{\vec{\tau}}^+\right\rangle - \left\langle \mathcal{Y}^{\vec{\tau}}\eta_{\vec{\tau}}^-, \eta_{\vec{\tau}}^-\right\rangle\right\} \to 0 \tag{4.31}$$

as $\tau_1, \ldots, \tau_n \to \infty$; in other words, for $\tau_1, \ldots, \tau_n$ sufficiently large, we may combine (4.29), (4.30), and (4.31) and produce a contradiction. □

Because viscosity solutions are both viscosity sub- and supersolutions, Theorem 4.1 implies that solutions to (4.26) and (4.27) are unique. Observing that viscosity solutions to (4.26) are viscosity supersolutions to (4.28) and that viscosity solutions to (4.27) are viscosity subsolutions to (4.28), we may therefore conclude that solutions to (4.28) are unique by an application of the lemma below.

Lemma 4.5 (cf. [2, Lemma 5.6])**.** *Let u^ε and u_ε be solutions to the Dirichlet Problems* (4.26) *and* (4.27) *respectively. Given $\delta > 0$, there exists $\varepsilon > 0$ such that*

$$u_\varepsilon \leq u^\varepsilon \leq u_\varepsilon + \delta.$$

References

[1] F. Beatrous, T. Bieske and J. Manfredi: *The Maximum Principle for Vector Fields*, Contemp. Math., **370** (2005), Amer. Math. Soc. Providence, RI, 1–9.

[2] T. Bieske: *On Infinite Harmonic Functions on the Heisenberg Group*, Comm. in PDE, **27** (3 & 4) (2002), 727–762 .

[3] T. Bieske: *Lipschitz Extensions on Generalized Grushin Spaces*, Michigan Math. J., **53** (1) (2005), 3–31.

[4] T. Bieske: *A Sub-Riemannian Maximum Principle and its Application to the* p*-Laplacian in Carnot Groups*, Ann. Acad. Sci. Fenn., **37** (2012), 119–134 .

[5] A. Bellaïche: *The Tangent Space in Sub-Riemannian Geometry*, In *Sub-Riemannian Geometry*; Bellaïche, André., Risler, Jean-Jacques., Eds.; Progress in Mathematics; Birkhäuser: Basel, Switzerland. **144**, 1–78 (1996).

[6] M. Crandall, H. Ishii, P.-L. Lions: *User's Guide to Viscosity Solutions of Second Order Partial Differential Equations*, Bull. Amer. Math. Soc., **27** (1) (1992), 1–67.

[7] R. Jensen: *Uniqueness of Lipschitz Extensions: Minimizing the Sup Norm of the Gradient*, Arch. Rational. Mech. Anal., **123** (1993), 51–74.

[8] P. Juutinen: *Minimization Problems for Lipschitz Functions via Viscosity Solutions*, Ann. Acad. Sci. Fenn. Math. Diss., **115** (1998).

THOMAS BIESKE
UNIVERSITY OF SOUTH FLORIDA
DEPARTMENT OF MATHEMATICS AND STATISTICS
4202 E. FOWLER AVE. CMC342
TAMPA, FL 33620, USA
ORCID: 0000-0003-2029-0562
E-mail address: tbieske@usf.edu

ZACHARY FORREST
UNIVERSITY OF SOUTH FLORIDA
DEPARTMENT OF MATHEMATICS AND STATISTICS
4202 E. FOWLER AVE. CMC342
TAMPA, FL 33620, USA
ORCID: 0000-0002-6636-0047
E-mail address: zachary9@usf.edu

CONSTRUCTIVE MATHEMATICAL ANALYSIS
6 (2023), No. 2, pp. 90-101
http://dergipark.org.tr/en/pub/cma
ISSN 2651 - 2939

Research Article

King operators which preserve x^j

ZOLTÁN FINTA*

ABSTRACT. We prove the unique existence of the functions r_n $(n = 1, 2, \dots)$ on $[0, 1]$ such that the corresponding sequence of King operators approximates each continuous function on $[0, 1]$ and preserves the functions $e_0(x) = 1$ and $e_j(x) = x^j$, where $j \in \{2, 3, \dots\}$ is fixed. We establish the essential properties of r_n, and the rate of convergence of the new sequence of King operators will be estimated by the usual modulus of continuity. Finally, we show that the introduced operators are not polynomial and we obtain quantitative Voronovskaja type theorems for these operators.

Keywords: Bernstein operator, King operator, Korovkin theorem, modulus of continuity, polynomial operator.

2020 Mathematics Subject Classification: 41A10, 41A25, 41A36.

1. INTRODUCTION

Let Π_n be the space of all algebraic polynomials of degree not greater than n. The Bernstein operators $B_n : C[0, 1] \to \Pi_n$ are given by

$$(B_n f)(x) = \sum_{k=0}^{n} p_{n,k}(x) f\left(\frac{k}{n}\right), \tag{1.1}$$

where $n = 1, 2, \dots$, $x \in [0, 1]$, $f \in C[0, 1]$ and $p_{n,k}(x) = \binom{n}{k} x^k (1 - x)^{n-k}$. For $j = 0, 1, 2, \dots$, we denote by e_j the power function $e_j(x) = x^j$, $x \in [0, 1]$. It is well-known [6, p. 3] that

$$(B_n e_0)(x) = 1,\ (B_n e_1)(x) = x,\ (B_n e_2)(x) = x^2 + \frac{1}{n} x(1 - x),\ x \in [0, 1]. \tag{1.2}$$

Studying the connection between regular summability matrices and convergent positive linear operators, King [14, pp. 204-205] introduced the operators $V_n : C[0, 1] \to C[0, 1]$ defined by

$$(V_n f)(x) = \sum_{k=0}^{n} p_{n,k}(r_n^*(x)) f\left(\frac{k}{n}\right), \tag{1.3}$$

where

$$r_n^*(x) = \begin{cases} x^2, & \text{if } n = 1 \\ -\frac{1}{2(n-1)} + \sqrt{\frac{n}{n-1}x^2 + \frac{1}{4(n-1)^2}}, & \text{if } n = 2, 3, \dots \end{cases}. \tag{1.4}$$

Taking into account (1.1)-(1.3), we have $(V_n f)(x) = (B_n f)(r_n^*(x))$, $x \in [0, 1]$ and $V_n e_0 = e_0$, $V_n e_2 = e_2$. The uniform convergence $\lim_{n\to\infty} V_n f = f$ and a quantitative estimation are also discussed in [14, p. 204 and p. 206]. We mention that in [8] we obtained direct and converse

Received: 03.03.2023; Accepted: 10.05.2023; Published Online: 16.05.2023

*Corresponding author: Zoltán Finta; fzoltan@math.ubbcluj.ro

DOI: 10.33205/cma.1259505

approximation theorems for (1.3). The existence of a sequence of linear positive bounded *polynomial* operators on $C[0,1]$, possessing e_0 and e_2 as fixed points, was proved in [9]. Main results concerning certain King type modifications of the Bernstein operators and the Szász-Mirakyan operators were presented in the survey paper [1].

Replacing $f\left(\frac{k}{n}\right)$ in (1.1) with $f\left(\sqrt[j]{\frac{k(k-1)\dots(k-j+1)}{n(n-1)\dots(n-j+1)}}\right)$, $n \geq j \geq 2$, Aldaz, Kounchev and Render [3, p. 12, Proposition 11] defined a new King type operator, which preserves the functions e_0 and e_j, where $j \in \{2,3,\dots\}$ is fixed. In [10], we proved that there exist infinitely many sequences of Bernstein type operators $(L_n)_{n\geq 1}$, which approximate each continuous function on $[0,1]$ and have the functions e_0 and e_j as fixed points, where $j \in \{1,2,\dots\}$ is given and

$$(L_n f)(x) = \sum_{k=0}^{n} p_{n,k}(x)\lambda_{n,k}(f), \quad f \in C[0,1]$$

and $\lambda_{n,k} \in C^*[0,1]$ are bounded positive linear functionals. Further properties of the Bernstein type operators of Aldaz, Kounchev and Render were obtained in the papers [2], [4], [5] and [13]. In [11], among others, we studied the approximation properties of the operators $U_n : C[0,1] \to C[0,1]$ defined by

$$(U_n f)(x) = \sum_{k=0}^{n} p_{n,k}(r_n(x)) f\left(\frac{k}{n}\right), \tag{1.5}$$

where the functions $r_n \in C[0,1]$ were constructed such that U_n preserves the functions e_0 and e_{2i}, with $i \in \{1,2,\dots\}$ given. The main goal of the present paper is to prove the unique existence of the functions $r_n : [0,1] \to [0,1]$ $(n = 1,2,\dots)$ such that the corresponding King operators given by (1.5) approximate each continuous function on $[0,1]$ and satisfy the conditions $U_n e_0 = e_0$ and $U_n e_j = e_j$, where $j \in \{2,3,\dots\}$ is fixed. The essential properties of r_n $(n = 1,2,\dots)$ will be established. A necessary and sufficient condition is given for the uniform convergence of $(U_n f)_{n\geq 1}$ to $f \in C[0,1]$. The quantitative estimates for the operators (1.5) are obtained with the aid of the usual modulus of continuity. Finally, we show that U_n cannot be polynomial operator of degree n, and we obtain a quantitative Voronovskaja type theorem for the operators (1.5).

2. The construction of r_n

At first we prove the following lemma.

Lemma 2.1. *Let $f,g : [a,b] \to [\alpha,\beta]$ be strictly increasing and continuous functions such that $f(a) = \alpha = g(a)$, $f(b) = \beta = g(b)$ and $f(u) \leq g(u)$ for all $u \in [a,b]$. Then, the inverse mappings $f^{-1}, g^{-1} : [\alpha,\beta] \to [a,b]$ exist and are strictly increasing and continuous on $[\alpha,\beta]$ such that $g^{-1}(v) \leq f^{-1}(v)$ for all $v \in [\alpha,\beta]$.*

Proof. The existence of f^{-1} and g^{-1} is the consequence of the following *continuous inverse theorem*: if $\varphi : [a,b] \to \mathbb{R}$ is a strictly increasing and continuous function then the inverse mapping $\varphi^{-1} : [\varphi(a),\varphi(b)] \to [a,b]$ exists and is strictly increasing and continuous on $[\varphi(a),\varphi(b)]$. Consequently $f^{-1}, g^{-1} : [\alpha,\beta] \to [a,b]$ are strictly increasing and continuous on $[\alpha,\beta]$. Moreover, for every $v \in [\alpha,\beta]$ there exists a unique $u \in [a,b]$ such that $v = f(u)$. Then $g^{-1}(v) = g^{-1}(f(u)) \leq g^{-1}(g(u)) = u = f^{-1}(v)$, because $f(u) \leq g(u)$ and g^{-1} is strictly increasing. □

The next result contains the construction of the functions r_n $(n = 1,2,\dots)$.

Theorem 2.1. *For every $n = 1, 2, \ldots$, there exists the unique function $r_n : [0,1] \to [0,1]$ such that*

$$\sum_{k=0}^{n} p_{n,k}(r_n(x)) \left(\frac{k}{n}\right)^j = x^j \tag{2.6}$$

for all $x \in [0,1]$, being $j \in \{2, 3, \ldots\}$ fixed.

Proof. If $n = 1$ then the function $r_1(x) = x^j$, $x \in [0,1]$ satisfies the equality

$$p_{1,0}(r_1(x)) \cdot 0 + p_{1,1}(r_1(x)) \cdot 1 = x^j, \quad x \in [0,1].$$

Let $n \geq 2$ and consider the function $\phi_n : [0,1] \to \mathbb{R}$,

$$\phi_n(y) = (B_n e_j)(y) = \sum_{k=0}^{n} p_{n,k}(y) \left(\frac{k}{n}\right)^j.$$

By (1.1)-(1.2), we have $\phi_n(0) = 0$, $\phi_n(1) = 1$ and $0 \leq \phi_n(y) \leq (B_n e_0)(y) = 1$ for every $y \in [0,1]$. Because

$$(B_n f)'(y) = n \sum_{k=0}^{n-1} p_{n-1,k}(y) \left[f\left(\frac{k+1}{n}\right) - f\left(\frac{k}{n}\right) \right]$$

(see [6, p. 305, (2.2)]), we get

$$\begin{aligned}
\phi_n'(y) &= (B_n e_j)'(y) = n \sum_{k=0}^{n-1} p_{n-1,k}(y) \left[\left(\frac{k+1}{n}\right)^j - \left(\frac{k}{n}\right)^j \right] \\
&= n \left\{ (1-y)^{n-1} \left(\frac{1}{n}\right)^j + \binom{n-1}{1} y (1-y)^{n-2} \left[\left(\frac{2}{n}\right)^j - \left(\frac{1}{n}\right)^j \right] + \ldots \right. \\
&\left. + \binom{n-1}{n-2} y^{n-2}(1-y) \left[\left(\frac{n-1}{n}\right)^j - \left(\frac{n-2}{n}\right)^j \right] + y^{n-1} \left[1 - \left(\frac{n-1}{n}\right)^j \right] \right\} \\
&> 0
\end{aligned} \tag{2.7}$$

for all $y \in [0,1]$. Thus $\phi_n : [0,1] \to [0,1]$ is a strictly increasing and continuous function. But the function e_j is also strictly increasing and continuous on $[0,1]$ such that $e_j(0) = 0$ and $e_j(1) = 1$, therefore if $x \in [0,1]$ is arbitrary then the equation $\phi_n(y) = x^j$ has a unique solution $y = r_n(x)$. In view of the continuous inverse theorem, there exists the strictly increasing and continuous inverse mapping ϕ_n^{-1}. Then

$$r_n(x) = (\phi_n^{-1} \circ e_j)(x), \quad x \in [0,1] \tag{2.8}$$

and satisfies (2.6). Moreover $0 = r_n(0) \leq r_n(x) \leq r_n(1) = 1$ for all $x \in [0,1]$. □

The essential properties of r_n $(n = 1, 2, \ldots)$ are gathered in the following theorem.

Theorem 2.2. *Let $r_n : [0,1] \to [0,1]$ $(n = 1, 2, \ldots)$ be the function defined by (2.6). Then*

a) r_n is strictly increasing and continuous function on $[0,1]$;
b) $x^j \leq r_n(x) \leq r_{n+1}(x) \leq x$ for all $x \in [0,1]$;
c) $\lim_{n\to\infty} r_n(x) = x$ for all $x \in [0,1]$;
d) r_n is differentiable on $[0,1]$.

Proof. *a)* By (2.8), we have that $r_n(x) = (\phi_n^{-1} \circ e_j)(x)$, $x \in [0,1]$, where ϕ_n^{-1} is a strictly increasing and continuous function on $[0,1]$. Hence, we obtain that r_n is also a strictly increasing and continuous function on $[0,1]$.

b) In view of (1.2), we have $\sum_{k=0}^{n} p_{n,k}(r_n(x))\frac{k}{n} = r_n(x)$. Using (2.6) and Jensen's inequality on $[0,1]$ for the convex function e_j, we get

$$x^j = \sum_{k=0}^{n} p_{n,k}(r_n(x)) \left(\frac{k}{n}\right)^j \geq \left(\sum_{k=0}^{n} p_{n,k}(r_n(x))\frac{k}{n}\right)^j = (r_n(x))^j, \quad x \in [0,1].$$

Hence $r_n(x) \leq x$, $x \in [0,1]$.

Because $(B_n f)(y) > (B_{n+1} f)(y)$, $0 < y < 1$ for any strictly convex function f on $[0,1]$ (see [6, p. 310, Corollary 4.2]), we obtain $\phi_n(y) = (B_n e_j)(y) > (B_{n+1}e_j)(y) = \phi_{n+1}(y)$ for $y \in (0,1)$. But $\phi_n(0) = 0 = \phi_{n+1}(0)$ and $\phi_n(1) = 1 = \phi_{n+1}(1)$, therefore $\phi_n(y) \geq \phi_{n+1}(y)$, $y \in [0,1]$. Due to Lemma 2.1, we have $\phi_n^{-1}(x) \leq \phi_{n+1}^{-1}(x)$, $x \in [0,1]$. In particular $\phi_n^{-1}(x^j) \leq \phi_{n+1}^{-1}(x^j)$, $x \in [0,1]$, i.e. $r_n(x) \leq r_{n+1}(x)$, $x \in [0,1]$, because of (2.8). But $r_1(x) = x^j$, $x \in [0,1]$, thus $x^j \leq r_n(x)$, $x \in [0,1]$.

c) Because $p_{n,k}$ $(k = 0, 1, \ldots, n)$ are polynomials of degree n, we have, by Taylor's formula for $x, y \in [0,1]$ that

$$p_{n,k}(y) = p_{n,k}(x) + \frac{1}{1!}p'_{n,k}(x)(y-x) + \frac{1}{2!}p''_{n,k}(x)(y-x)^2 + \ldots + \frac{1}{n!}p^{(n)}_{n,k}(x)(y-x)^n.$$

Hence, in view of (2.6) and (1.1),

$$\begin{aligned} x^j - (B_n e_j)(x) &= \sum_{k=0}^{n} p_{n,k}(r_n(x))\left(\frac{k}{n}\right)^j - \sum_{k=0}^{n} p_{n,k}(x)\left(\frac{k}{n}\right)^j \\ &= \sum_{k=0}^{n} \left[p_{n,k}(r_n(x)) - p_{n,k}(x)\right]\left(\frac{k}{n}\right)^j \\ &= \sum_{k=0}^{n} \left\{\sum_{i=1}^{n} \frac{1}{i!}p^{(i)}_{n,k}(x)(r_n(x)-x)^i\right\}\left(\frac{k}{n}\right)^j \\ &= \sum_{i=1}^{n} \frac{1}{i!}(r_n(x)-x)^i \sum_{k=0}^{n} p^{(i)}_{n,k}(x)\left(\frac{k}{n}\right)^j = \sum_{i=1}^{n}\frac{1}{i!}(r_n(x)-x)^i (B_n e_j)^{(i)}(x). \end{aligned} \tag{2.9}$$

On the other hand the Bernstein polynomial $B_n P$ of a polynomial P of degree m is itself a polynomial of degree m, if $n \geq m$ (see [6, p. 306]). Then $(B_n e_j)^{(i)}(x) = 0$, $x \in [0,1]$ for $n \geq i > j$. By (2.9), we get for $n > j$ that

$$x^j - (B_n e_j)(x) = \sum_{i=1}^{j} \frac{1}{i!}(r_n(x)-x)^i (B_n e_j)^{(i)}(x). \tag{2.10}$$

It is known that $\lim_{n\to\infty} (B_n f)^{(i)}(x) = f^{(i)}(x)$, if $x \in [0,1]$ and $f \in C^i[0,1]$ (see [6, p. 306, Theorem 2.1]). Thus

$$\lim_{n\to\infty} (B_n e_j)^{(i)}(x) = e_j^{(i)}(x) = j(j-1)\ldots(j-i+1)x^{j-i}, \tag{2.11}$$

where $x \in [0,1]$ and $i \in \{1, 2, \ldots, j\}$. Furthermore, in view of *b*), the sequence $(r_n(x))_{n\geq 1}$ is convergent for all $x \in [0,1]$: there exists

$$\lim_{n\to\infty} r_n(x) = r(x), \quad x \in [0,1]. \tag{2.12}$$

Combining (2.10)-(2.12), we find that

$$\begin{aligned}
0 &= \lim_{n\to\infty}(x^j-(B_ne_j)(x))\\
&= \sum_{i=1}^{j}\frac{1}{i!}\lim_{n\to\infty}(r_n(x)-x)^i\lim_{n\to\infty}(B_ne_j)^{(i)}(x)\\
&= \sum_{i=1}^{j}\frac{1}{i!}(r(x)-x)^i j(j-1)\dots(j-i+1)x^{j-i} = \sum_{i=1}^{j}\binom{j}{i}(r(x)-x)^i x^{j-i}\\
&= -x^j+\sum_{i=0}^{j}\binom{j}{i}(r(x)-x)^i x^{j-i} = -x^j+(r(x)-x+x)^j = -x^j+(r(x))^j.
\end{aligned}$$

Hence $r(x)=x$, $x\in[0,1]$, thus $\lim_{n\to\infty} r_n(x)=x$.

d) Because $\phi_n'(y)>0$, $y\in[0,1]$ (see (2.7)) and $r_n(x)=\phi_n^{-1}(x^j)$, $x\in[0,1]$ (see (2.8)), it follows that r_n is a differentiable function on $[0,1]$. Moreover

$$r_n'(x)=(\phi_n^{-1})'(x^j)\cdot(x^j)'=\frac{jx^{j-1}}{(\phi_n')'(r_n(x))}=\frac{jx^{j-1}}{(B_ne_j)'(r_n(x))},\qquad x\in[0,1], \tag{2.13}$$

because $\phi_n(r_n(x))=x^j$. □

Remark 2.1. *Due to (1.4), we have for all $x\in[0,1]$ that*

$$(r_n^*)'(x)=\begin{cases}2x, & \textit{if } n=1\\ \frac{n}{n-1}x\left(\frac{n}{n-1}x^2+\frac{1}{4(n-1)^2}\right)^{-\frac12}, & \textit{if } n=2,3,\dots\end{cases}.$$

The same result can be obtained from (2.13) for $j=2$.

Indeed, by (2.7), (1.2) and (1.4), we have for $x\in[0,1]$ and $n\ge 2$ that

$$\begin{aligned}
(B_ne_2)'(r_n^*(x)) &= n\sum_{k=0}^{n-1}p_{n-1,k}(r_n^*(x))\left[\left(\frac{k+1}{n}\right)^2-\left(\frac{k}{n}\right)^2\right]\\
&= \frac{2(n-1)}{n}\sum_{k=0}^{n-1}p_{n-1,k}(r_n^*(x))\frac{k}{n-1}+\frac{1}{n}\sum_{k=0}^{n-1}p_{n-1,k}(r_n^*(x))\\
&= \frac{2(n-1)}{n}r_n^*(x)+\frac{1}{n}\\
&= \frac{2(n-1)}{n}\sqrt{\frac{n}{n-1}x^2+\frac{1}{4(n-1)^2}}.
\end{aligned}$$

Hence, by (2.13),

$$(r_n^*)'(x)=\frac{2x}{(B_ne_2)'(r_n^*(x))}=\frac{n}{n-1}x\left(\frac{n}{n-1}x^2+\frac{1}{4(n-1)^2}\right)^{-\frac12}.$$

3. The approximation properties of U_n

The operators $U_n: C[0,1]\to C[0,1]$ given by (1.5) are positive linear and $(U_nf)(0)=f(0)$ and $(U_nf)(1)=1$, because $r_n(0)=0$ and $r_n(1)=1$. Moreover, by (1.2) and (2.6), we have $U_ne_0=e_0$ and $U_ne_j=e_j$. In the following theorem, we study the convergence $U_nf\to f$ in the uniform norm defined by $\|f\|=\sup\{|f(x)|: x\in[0,1]\}$, $f\in C[0,1]$.

Theorem 3.3. *$\lim\limits_{n\to\infty}\|U_nf-f\|=0$ for each $f\in C[0,1]$ if and only if $\lim\limits_{n\to\infty}\|r_n-e_1\|=0$, where r_n $(n=1,2,\ldots)$ are defined by (2.6).*

Proof. Using (1.5), (1.1) and (1.2), we obtain

$$(U_ne_0)(x)=1,\ (U_ne_1)(x)=r_n(x),\ (U_ne_2)(x)=(r_n(x))^2+\frac{1}{n}r_n(x)(1-r_n(x)). \tag{3.14}$$

Hence

$$\|U_ne_0-e_0\|=0, \tag{3.15}$$

$$\|U_ne_1-e_1\|=\|r_n-e_1\| \tag{3.16}$$

and

$$\|U_ne_2-e_2\|\le\|r_n^2-e_1^2\|+\frac{1}{4n}\le 2\|r_n-e_1\|+\frac{1}{4n}, \tag{3.17}$$

because $r_n(x)\in[0,1]$ for $x\in[0,1]$ (see Theorem 2.1).

On the other hand, the statements a), b) and c) of Theorem 2.2, and Dini's theorem (see e.g. [15, p. 150, 7.13. Theorem]) imply that

$$\lim_{n\to\infty}\|r_n-e_1\|=0. \tag{3.18}$$

Combining (3.15)-(3.18), in view of Korovkin theorem [6, pp. 8-10], we obtain the assertion of our theorem. □

The next result contains pointwise and uniform quantitative estimates for U_n $(n=1,2,\ldots)$, using the usual modulus of continuity of $f\in C[0,1]$ given by

$$\omega(f;\delta)=\sup\{|f(u)-f(v)|:u,v\in[0,1],|u-v|<\delta\},\ \delta>0.$$

Theorem 3.4. *Let $(U_n)_{n\ge1}$ be the sequence of operators defined by (1.5). Then for every $f\in C[0,1]$, we have*

a) $|(U_nf)(x)-f(x)|\le 2\omega\left(f;\sqrt{(r_n(x)-x)^2+\frac{1}{n}r_n(x)(1-r_n(x))}\right)$, $n\ge1, x\in[0,1]$;

b)

$$|(U_nf)(x)-f(x)|\le\begin{cases}6\,\omega\left(f;\sqrt{\frac{x(1-x)}{n}}\right), & \textit{if } j=2\\ 2(1+\sqrt{C(j)})\,\omega\left(f;\frac{\sqrt{x(1-x)}}{\sqrt[2j]{n}}\right), & \textit{if } j=3,4,\ldots\end{cases},$$

where $n\ge j$, $x\in[0,1]$ and

$$C(j)=(j-1)\sqrt[j]{\frac{j(j-1)^2}{8}}+j; \tag{3.19}$$

c)

$$\|U_nf-f\|\le\begin{cases}6\,\omega\left(f;\frac{1}{\sqrt{n}}\right), & \textit{if } j=2\\ 2(1+\sqrt{C(j)})\,\omega\left(f;\frac{1}{\sqrt[2j]{n}}\right), & \textit{if } j=3,4,\ldots\end{cases},$$

where $n\ge j$ and $C(j)$ is defined by (3.19).

Proof. *a*) For any sequence $(L_n)_{n\geq 1}$ of positive linear operators on $C[a,b]$, it is known [7, p. 30] that for $f \in C[a,b]$ and $x \in [a,b]$, we have

$$|(L_n f)(x) - f(x)| \leq |f(x)| \cdot |(L_n e_0)(x) - e_0(x)| \\ + \omega(f;\delta)\left[(L_n e_0)(x) + \frac{1}{\delta}\left((L_n e_0)(x)\right)^{1/2} \cdot \left((L_n(e_1 - xe_0)^2)(x)\right)^{1/2}\right].$$

In our case $[a,b] = [0,1]$ and $U_n e_0 = e_0$ (see (3.14)), thus

$$|(U_n f)(x) - f(x)| \leq \left[1 + \delta^{-1}\left((U_n(e_1 - xe_0)^2)(x)\right)^{1/2}\right]\omega(f;\delta). \tag{3.20}$$

But, in view of (3.14), we have

$$\begin{aligned}(U_n(e_1 - xe_0)^2)(x) &= (U_n e_2)(x) - 2x(U_n e_1)(x) + x^2(U_n e_0)(x) \\ &= (r_n(x) - x)^2 + \frac{1}{n}r_n(x)(1 - r_n(x)).\end{aligned} \tag{3.21}$$

Choosing $\delta = \left((r_n(x) - x)^2 + \frac{1}{n}r_n(x)(1 - r_n(x))\right)^{1/2}$ in (3.20), we get the required estimate.
b) We will prove the following estimates below:

$$\left(U_n(e_1 - xe_0)^2\right)(x) \leq \begin{cases} \frac{4}{n}x(1-x), & \text{if } j = 2 \\ \frac{C(j)}{\sqrt[j]{n}}x(1-x), & \text{if } j \geq 3 \end{cases}, \tag{3.22}$$

where $x \in [0,1]$ is arbitrary. Hence, by (3.20) and the property $\omega(f;\lambda\delta) \leq (1+\lambda)\omega(f;\delta)$, $\lambda > 0$, we get for $\delta = \left(\left(U_n(e_1 - xe_0)^2\right)(x)\right)^{1/2}$ that

$$\begin{aligned}|(U_n f)(x) - f(x)| &\leq 2\,\omega\left(f; \left(\left(U_n(e_1 - xe_0)^2\right)(x)\right)^{1/2}\right) \\ &\leq \begin{cases} 2\,\omega\left(f; 2\sqrt{\frac{x(1-x)}{n}}\right), & \text{if } j = 2 \\ 2\,\omega\left(f; \sqrt{C(j)}\frac{\sqrt{x(1-x)}}{\sqrt[2j]{n}}\right), & \text{if } j \geq 3 \end{cases} \\ &\leq \begin{cases} 6\,\omega\left(f; \frac{\sqrt{x(1-x)}}{\sqrt{n}}\right), & \text{if } j = 2 \\ 2(1 + \sqrt{C(j)})\,\omega\left(f; \frac{\sqrt{x(1-x)}}{\sqrt[2j]{n}}\right), & \text{if } j \geq 3 \end{cases}\end{aligned}$$

which was to be proved.

Now let us prove (3.22). Using Theorem 2.2 b), we have for $x \in [0,1]$ that

$$r_n(x)(1 - r_n(x)) \leq x(1 - x^j) = x(1-x)(1 + x + \ldots + x^{j-1}) \leq jx(1-x). \tag{3.23}$$

For $j = 2$, we have in view of [8, p. 87, Lemma 1, d)] that $0 \leq x - r_n(x) \leq \frac{2}{n}(1-x)$. Hence

$$(x - r_n(x))^2 = (x - r_n(x))(x - r_n(x)) \leq \frac{2}{n}x(1-x). \tag{3.24}$$

Then (3.21), (3.24) and (3.23) imply $\left(U_n(e_1 - xe_0)^2\right)(x) \leq \frac{2}{n}x(1-x) + \frac{2}{n}x(1-x) = \frac{4}{n}x(1-x)$.

Let $j \geq 3$ and $n \geq j$. By Theorem 2.1 and [11, pp. 102-103, Lemma 1 and Lemma 2], the polynomial $\phi_n(y) \equiv P_{n,j}(y) = \sum_{k=0}^{n} p_{n,k}(y)\left(\frac{k}{n}\right)^j = a_0 y^j + a_1 y^{j-1} + \ldots + a_{j-1}y$ satisfies the following conditions:

$$P_{n,j}(r_n(x)) = x^j;$$
$$a_0 = \frac{1}{n^{j-1}}(n-1)(n-2)\dots(n-j+1);\ a_1,\dots,a_{j-1} > 0;\ a_0 + a_1 + \dots + a_{j-1} = 1;$$
$$0 \le 1 - a_0 \le \frac{j(j-1)}{2n}.$$

Hence

$$\begin{aligned}
0 \le x^j - (r_n(x))^j &= P_{n,j}(r_n(x)) - (r_n(x))^j \\
&= \sum_{k=0}^{j-1} a_k (r_n(x))^{j-k} - (r_n(x))^j \\
&= (a_0 - 1)(r_n(x))^j + \sum_{k=1}^{j-1} a_k (r_n(x))^{j-k} \\
&= -\sum_{k=1}^{j-1} a_k (r_n(x))^j + \sum_{k=1}^{j-1} a_k (r_n(x))^{j-k} \\
&= \sum_{k=1}^{j-1} a_k (r_n(x))^{j-k} \left[1 - (r_n(x))^k\right] \\
&= \sum_{k=1}^{j-1} a_k (r_n(x))^{j-k}(1 - r_n(x)) \left[1 + r_n(x) + \dots + (r_n(x))^{k-1}\right] \\
&\le r_n(x)(1 - r_n(x)) \sum_{k=1}^{j-1} k a_k \le (j-1) r_n(x)(1 - r_n(x)) \sum_{k=1}^{j-1} a_k \\
&= (j-1) r_n(x)(1 - r_n(x))(1 - a_0) \\
&\le \frac{j(j-1)^2}{2n} r_n(x)(1 - r_n(x)) \le \frac{j(j-1)^2}{8n}.
\end{aligned}$$

Using $(u-v)^{2j} \le (u^j - v^j)^2$, $u, v \in [0,1]$ (see [11, p. 103, Lemma 2, (b)]), we find that $(x - r_n(x))^{2j} \le (x^j - (r_n(x))^j)^2 \le \left(\frac{1}{8n} j(j-1)^2\right)^2$, i.e.

$$0 \le x - r_n(x) \le \sqrt[j]{\frac{1}{8n} j(j-1)^2}. \tag{3.25}$$

At the same time, due to Theorem 2.2 b), we obtain

$$0 \le x - r_n(x) \le x - x^j = x(1 - x^{j-1}) = x(1-x)(1 + x + \dots + x^{j-2}) \le (j-1)x(1-x). \tag{3.26}$$

Hence, in view of (3.21), (3.25), (3.26) and (3.23), we get

$$\begin{aligned}
\left(U_n(e_1 - xe_0)^2\right)(x) &= (x - r_n(x))(x - r_n(x)) + \frac{1}{n} r_n(x)(1 - r_n(x)) \\
&\le \sqrt[j]{\frac{1}{8n} j(j-1)^2}(j-1)x(1-x) + \frac{j}{n} x(1-x) \le \frac{C(j)}{\sqrt[j]{n}} x(1-x).
\end{aligned}$$

c) Because $x(1-x) \le 1$ for $x \in [0,1]$, the estimates formulated in c) follow from the statement of b). □

Remark 3.2. *By Theorem 2.1, we have $U_n f \equiv V_n f$ for $j = 2$. Then $V_n e_0 = e_0$ and $V_n e_2 = e_2$, thus, by (3.21), we get $(V_n(e_1 - xe_0)^2)(x) = 2x(x - r_n^*(x))$. Applying Theorem 3.4, we obtain*

$$|(V_n f)(x) - f(x)| \leq 2\,\omega\left(f; \sqrt{2x(x - r_n^*(x))}\right),\ n \geq 1,\ x \in [0,1];$$

$$|(V_n f)(x) - f(x)| \leq 6\,\omega\left(f; \sqrt{\frac{x(1-x)}{n}}\right),\ n \geq 2,\ x \in [0,1];$$

$$\|V_n f - f\| \leq 6\,\omega\left(f; \frac{1}{\sqrt{n}}\right),\ n \geq 2.$$

For the first estimate see [14, p. 206, Theorem 3.1].

Furthermore, we have the following theorem.

Theorem 3.5. *Let $U_n : C[0,1] \to C[0,1]$ $(n = 1, 2, \ldots)$ be the operators given by (1.5) with r_n defined by (2.6). Then U_n cannot be polynomial operator of degree n: there exists $f \in C[0,1]$ such that $U_n f \notin \Pi_n$.*

Proof. Let $n \geq j$ and suppose that $U_n f \in \Pi_n$ for all $f \in C[0,1]$. Then $U_n e_1 = r_n \in \Pi_n$ due to (3.14). Furthermore $B_n e_j$ is a polynomial of degree j, because $n \geq j$, and thus $(B_n e_j)(y) = a_0 y^j + a_1 y^{j-1} + \ldots + a_{j-1} y$, where $a_0 > 0$ (see [11, p. 102, Lemma 1]). Taking into account (2.6), we have

$$x^j = (U_n e_j)(x) = (B_n e_j)(r_n(x)) = a_0 (r_n(x))^j + a_1 (r_n(x))^{j-1} + \ldots + a_{j-1} r_n(x).$$

In view of $r_n \in \Pi_n$ and $a_0 > 0$, we find that r_n is a first degree polynomial. By Theorem 2.1, we have $r_n(0) = 0$ and $r_n(1) = 1$, thus $r_n(x) = x$, $x \in [0,1]$. Hence $(U_n f)(x) = (B_n f)(r_n(x)) = (B_n f)(x)$, $x \in [0,1]$. But $U_n e_j = e_j$ (see (2.6)), therefore $B_n e_j = e_j$ on $[0,1]$, contradiction, because $(B_n f)(x) > f(x)$, $0 < x < 1$ for any strictly convex function f on $[0,1]$ (see [6, p. 310, Corollary 4.2]), in particular $B_n e_j > e_j$ on $(0,1)$.

If $1 \leq n < j$ and $U_n f \in \Pi_n$ for all $f \in C[0,1]$, then $U_n e_j = e_j \in \Pi_n$ due to (2.6). Hence $j \leq n$, contradiction. □

Finally, we have the following quantitative Voronovskaja type theorem for the operators (1.5). We mention that similar result was established for the Bernstein type operators of Aldaz, Kounchev and Render in [12].

Theorem 3.6. *Let U_n $(n = 1, 2, \ldots)$ be given by (1.5). Then*

a) $\left|\, n((U_n f)(x) - f(x)) + (f'(x) - xf''(x))n(x - r_n(x)) \,\right| \leq 2(2 + \sqrt{39})x(1-x)\omega\left(f''; \frac{1}{\sqrt{n}}\right)$
for all $x \in [0,1]$, $f \in C^2[0,1]$ and $j = 2$, where

$$0 \leq \liminf_{n\to\infty} n(x - r_n(x)) \leq \limsup_{n\to\infty} n(x - r_n(x)) \leq 2;$$

b) $\left|\, \sqrt[j]{n}((U_n f)(x) - f(x)) + f'(x)\sqrt[j]{n}(U_n(xe_0 - e_1))(x) - \frac{1}{2}f''(x)\sqrt[j]{n}(U_n(xe_0 - e_1)^2)(x) \,\right|$
$\leq \sqrt{C(j)}(\sqrt{C(j)} + \sqrt{C_1(j)})x(1-x)\omega\left(f''; \frac{1}{2\sqrt[j]{n}}\right)$
for all $x \in [0,1]$, $f \in C^2[0,1]$ and $j \geq 3$, where $C(j)$ is defined by (3.19),

$$C_1(j) = \frac{3}{4}j^2 + \frac{119}{8}j + \frac{1}{4}(j-1)^2\sqrt[j]{\frac{1}{64}j^2(j-1)^4}$$

and

$$0 \leq \liminf_{n\to\infty} \sqrt[j]{n}(U_n(xe_0 - e_1))(x) \leq \limsup_{n\to\infty} \sqrt[j]{n}(U_n(xe_0 - e_1))(x) \leq \sqrt[j]{\frac{1}{8}j(j-1)^2},$$

$$0 \leq \liminf_{n\to\infty} \sqrt[j]{n}(U_n(xe_0 - e_1)^2)(x) \leq \limsup_{n\to\infty} \sqrt[j]{n}(U_n(xe_0 - e_1)^2)(x) \leq \frac{1}{4}C(j).$$

Proof. For $f \in C^2[0,1]$ and $x, t \in [0,1]$, by Taylor's formula, we have

$$f(t) = f(x) + f'(x)(t-x) + \frac{1}{2}f''(x)(t-x)^2 + \int_x^t (f''(u) - f''(x))(t-u)\,du.$$

Hence

$$\begin{aligned}(U_nf)(x) &= f(x) + f'(x)(U_n(e_1 - xe_0))(x) + \frac{1}{2}f''(x)(U_n(xe_0 - e_1)^2)(x)\\ &\quad + U_n\left(\int_x^t (f''(u) - f''(x))(t-u)\,du; x\right).\end{aligned} \tag{3.27}$$

Because

$$\begin{aligned}&\left|\int_x^t (f''(u) - f''(x))(t-u)\,du\right| \leq \left|\int_x^t |f''(u) - f''(x)||t-u|\,du\right|\\ &\leq \left|\int_x^t \omega(f''; |u-x|)\,|t-u|\,du\right| \leq \left|\int_x^t (1 + \delta^{-1}|u-x|)\,\omega(f'';\delta)\,|t-u|\,du\right|\\ &= \omega(f'';\delta)\left|\int_x^t (|t-u| + \delta^{-1}|u-x||t-u|)\,du\right| \leq \omega(f'';\delta)\left(|t-x|^2 + \delta^{-1}|t-x|^3\right),\end{aligned}$$

where $\delta > 0$, we get, by (3.27) and Hölder's inequality that

$$\begin{aligned}&\left|((U_nf)(x) - f(x)) + f'(x)(U_n(xe_0 - e_1))(x) - \frac{1}{2}f''(x)(U_n(e_1 - xe_0)^2)(x)\right|\\ &\leq \omega(f'';\delta)\left\{(U_n(e_1 - xe_0)^2)(x) + \delta^{-1}(U_n|e_1 - xe_0|^3)(x)\right\}\\ &\leq \omega(f'';\delta)\\ &\times\left\{(U_n(e_1 - xe_0)^2)(x) + \delta^{-1}\left[(U_n(e_1 - xe_0)^2)(x)\right]^{1/2}\left[(U_n(e_1 - xe_0)^4)(x)\right]^{1/2}\right\}.\end{aligned} \tag{3.28}$$

Using the first four moments of the Bernstein polynomials [6, p. 304], we have

$$\begin{aligned}&(U_n(e_1 - xe_0)^4)(x) = \sum_{k=0}^n p_{n,k}(r_n(x))\left(\frac{k}{n} - x\right)^4\\ &= \sum_{k=0}^n p_{n,k}(r_n(x))\left[\left(\frac{k}{n} - r_n(x)\right) + (r_n(x) - x)\right]^4\\ &= \sum_{k=0}^n p_{n,k}(r_n(x))\left(\frac{k}{n} - r_n(x)\right)^4 + 4(r_n(x) - x)\sum_{k=0}^n p_{n,k}(r_n(x))\left(\frac{k}{n} - r_n(x)\right)^3\\ &+ 6(r_n(x) - x)^2\sum_{k=0}^n p_{n,k}(r_n(x))\left(\frac{k}{n} - r_n(x)\right)^2\\ &+ 4(r_n(x) - x)^3\sum_{k=0}^n p_{n,k}(r_n(x))\left(\frac{k}{n} - r_n(x)\right) + (r_n(x) - x)^4\\ &= \frac{3}{n^2}(r_n(x))^2(1 - r_n(x))^2 + \frac{1}{n^3}\left[r_n(x)(1 - r_n(x)) - 6(r_n(x))^2(1 - r_n(x))^2\right]\\ &+ 4(r_n(x) - x)\frac{1}{n^2}(1 - 2r_n(x))r_n(x)(1 - r_n(x)) + 6(r_n(x) - x)^2\frac{1}{n}r_n(x)(1 - r_n(x))\\ &+ (r_n(x) - x)^4.\end{aligned} \tag{3.29}$$

a) If $j = 2$, then $r_n(x)(1 - r_n(x)) \le 2x(1 - x)$, $x \in [0,1]$, due to (3.23). Hence, by (3.29) and (3.24),

$$\begin{aligned}
&(U_n(e_1 - xe_0)^4)(x)\\
&\le \frac{12}{n^2}x^2(1-x)^2 + \frac{2}{n^2}x(1-x)(1 + 6r_n(x)(1 - r_n(x))\\
&+ \frac{8}{n^2}x(1-x)(x - r_n(x))(1 + 2r_n(x)) + \frac{12}{n}x(1-x)(x - r_n(x))^2 + (x - r_n(x))^4\\
&\le \frac{3}{n^2}x(1-x) + \frac{2}{n^2}\left(1 + \frac{3}{2}\right)x(1-x)\\
&+ \frac{24}{n^2}x(1-x) + \frac{12}{n}x(1-x)\frac{2}{n}\frac{1}{4} + \frac{4}{n^2}x(1-x)\frac{1}{4}\\
&= \frac{39}{n^2}x(1-x). \qquad (3.30)
\end{aligned}$$

Then (3.28), (3.22) and (3.30) imply that

$$\begin{aligned}
&\left| \, n((U_nf)(x) - f(x)) + f'(x)n(U_n(xe_0 - e_1))(x) - \frac{1}{2}f''(x)n(U_n(e_1 - xe_0)^2)(x) \, \right|\\
&\le \omega(f''; \delta)\left\{4x(1-x) + \delta^{-1}\frac{2\sqrt{39}}{\sqrt{n}}x(1-x)\right\}.
\end{aligned}$$

Choosing $\delta = \frac{1}{\sqrt{n}}$, and taking into account that $(U_n(xe_0 - e_1))(x) = x - r_n(x)$ and $(U_n(e_1 - xe_0)^2)(x) = 2x(x - r_n(x))$, we obtain the desired estimate.

Furthermore, in view of [8, p. 87, Lemma 1, d)], we have $0 \le x - r_n(x) \le \frac{2}{n}(1 - x) \le \frac{2}{n}$, $x \in [0, 1]$, thus $0 \le \liminf_{n\to\infty} n(x - r_n(x)) \le \limsup_{n\to\infty} n(x - r_n(x)) \le 2$.

b) If $j \ge 3$, then (3.29), (3.23), (3.25) and (3.26) imply that

$$\begin{aligned}
&(U_n(e_1 - xe_0)^4)(x)\\
&\le \frac{3}{n^2}j^2x^2(1-x)^2 + \frac{1}{n^3}jx(1-x)(1 + 6r_n(x)(1 - r_n(x))\\
&+ \frac{4}{n^2}jx(1-x)(x - r_n(x))(1 + 2r_n(x)) + \frac{6}{n}jx(1-x)(x - r_n(x))^2 + (x - r_n(x))^4\\
&\le \frac{3j^2}{4n^2}x(1-x) + \frac{5j}{2n^3}x(1-x)\\
&+ \frac{12j}{n^2}x(1-x) + \frac{3j}{8n}(j-1)^2x(1-x) + \sqrt[j]{\frac{1}{64n^2}j^2(j-1)^4}(j-1)^2\frac{1}{4}x(1-x)\\
&\le \frac{1}{\sqrt[j]{n^2}}x(1-x)\left\{\frac{3}{4}j^2 + \frac{119}{8}j + \frac{1}{4}(j-1)^2\sqrt[j]{\frac{1}{64}j^2(j-1)^4}\right\} = \frac{C_1(j)}{\sqrt[j]{n^2}}x(1-x). \qquad (3.31)
\end{aligned}$$

Using (3.28), (3.22) and (3.31), we get

$$\begin{aligned}
&\left| \, \sqrt[j]{n}((U_nf)(x) - f(x)) + f'(x)\sqrt[j]{n}(U_n(xe_0 - e_1))(x) - \frac{1}{2}f''(x))\sqrt[j]{n}(U_n(e_1 - xe_0)^2)(x) \, \right|\\
&\le \omega(f''; \delta)\left\{C(j)x(1-x) + \delta^{-1}\frac{\sqrt{C(j)}}{\sqrt[2j]{n}}\sqrt{C_1(j)x(1-x)}\right\}.
\end{aligned}$$

Choosing $\delta = \frac{1}{\sqrt[2j]{n}}$, we obtain the desired estimate.

Finally, by (3.25) and (3.22), we get

$$0 \le \liminf_{n\to\infty} \sqrt[j]{n}(U_n(xe_0 - e_1))(x) \le \limsup_{n\to\infty} \sqrt[j]{n}(U_n(xe_0 - e_1))(x) \le \sqrt[j]{\frac{1}{8}j(j-1)^2}$$

and

$$0 \le \liminf_{n\to\infty} \sqrt[j]{n}(U_n(xe_0 - e_1)^2)(x) \le \limsup_{n\to\infty} \sqrt[j]{n}(U_n(xe_0 - e_1)^2)(x) \le \frac{1}{4}C(j)$$

which completes the proof of the theorem. □

References

[1] T. Acar, M. C. Montano, P. Garrancho and V. Leonessa: *On sequences of J. P. King-type operators*, J. Funct. Spaces, **2019** (2019), Article ID 2329060, 12 pages.

[2] A. M. Acu, H. Gonska and M. Heilmann: *Remarks on a Bernstein-type operator of Aldaz, Kounchev and Render*, J. Numer. Anal. Approx. Theory, **50** (2001), 3–11.

[3] J. M. Aldaz, O. Kounchev and H. Render: *Shape preserving properties of generalized Bernstein operators on extended Chebyshev spaces*, Numer. Math., **114** (2009), 1–25.

[4] M. Birou: *A proof of a conjecture about the asymptotic formula of a Bernstein type operator*, Results Math., **72** (2017), 1129–1138.

[5] D. Cárdenas-Morales, P. Garrancho and I. Raşa: *Asymptotic Formulae via a Korovkin-Type Result*, Abstr. Appl. Anal., **2012** (2012), Article 217464, 12 pages.

[6] R. A. DeVore and G. G. Lorentz: *Constructive Approximation*, Springer, Berlin (1993).

[7] R. A. DeVore: *The Approximation of Continuous Functions by Positive Linear Operators*, Lecture Notes in Mathematics, 293, Springer, New York, (1972).

[8] Z. Finta: *Direct and converse theorems for King operators*, Acta Univ. Sapientiae, **12** (1) (2020), 85–96.

[9] Z. Finta: *Estimates for Bernstein type operators*, Math. Inequal. Appl., **15** (1) (2012), 127–135.

[10] Z. Finta: *Bernstein type operators having* 1 *and* x^j *as fixed points*, Centr.Eur. J. Math., **11** (12) (2013), 2257–2261.

[11] Z. Finta: *New properties of King's operators*, Positivity, **17** (1) (2013), 101–109.

[12] Z. Finta: *A quantitative variant of Voronovskaja's theorem for King-type operators*, Constr. Math. Anal., **2** (3) (2019), 124–129.

[13] I. Gavrea and M. Ivan: *Complete asymptotic expansions related to conjecture on a Voronovskaja-type theorem*, J. Math. Anal. Appl., **458** (2018), 452–463.

[14] J. P. King: *Positive linear operators which preserve* x^2, Acta Math. Hungar., **99** (3) (2003), 203–208.

[15] W. Rudin: *Principles of Mathematical Analysis*, Third Edition, McGraw-Hill, New York (1976).

Zoltán Finta
Babeş-Bolyai University
Department of Mathematics
1, M. Kogălniceanu st., 400084 Cluj-Napoca, Romania
ORCID: 0000-0003-2104-3483
E-mail address: fzoltan@math.ubbcluj.ro

CONSTRUCTIVE MATHEMATICAL ANALYSIS
6 (2023), No. 2, pp. 102-106
http://dergipark.org.tr/en/pub/cma
ISSN 2651 - 2939

Research Article

On an interpolation sequence for a weighted Bergman space on a Hilbert unit ball

MOHAMMED EL AIDI*

ABSTRACT. The purpose is to provide a generalization of Carleson's Theorem on interpolating sequences when dealing with a sequence in the open unit ball of a Hilbert space. Precisely, we interpolate a sequence by a function belonging to a weighted Bergman space of infinite order on a unit Hilbert ball and we furnish explicitly the upper bound corresponding to the interpolation constant.

Keywords: Analytic functions, interpolation sequences, weighted Bergman spaces, pseudohyberbolic distance, Fréchet differentiable functions.

2020 Mathematics Subject Classification: 30A99, 30H05, 32A36, 28E99, 46A04.

1. INTRODUCTION

Let us recall a known result that it has been shown that a sequence $\Gamma = (a_k)_{k\in\mathbb{N}}$ is interpolated by a function in $B^{\infty}_{\varphi^c}(\mathbb{D}^n)$, the set of holomorphic functions f on the complex unit ball $\mathbb{D}^n$ such that φf is bounded, where φ is strictly positive continuous function on $[0,1)$ satisfying a few meaningful assumptions and the power c is a strictly positive constant [3]. Precisely, it has been shown the following theorem.

Theorem 1.1 ([3]). *Let $\Gamma = (a_k)_{k\in\mathbb{N}}$ be a sequence in $\mathbb{D}^n$ and $\prod_{j\in\mathbb{N}\setminus\{k\}}|\psi_{a_j}(a_k)| \geq \varphi(|a_k|)$ for all $k \in \mathbb{N}$. Then Γ is interpolated by a function in $B^{\infty}_{\varphi^c}(\mathbb{D}^n)$. Furthermore, an upper bound of the interpolation constant is given explicitly and it is independent of n and φ.*

$|\psi_{a_j}(a_k)|$ is the pseudohyperbolic distance between a_j and a_k such that $\psi_{a_j}(\cdot)$ is the $\mathbb{D}^n$-valued Möbius map on $\mathbb{D}^n$. Apropos of the proof of Theorem 1.1, concisely the author sets up an interpolating function belonging in $B^{\infty}_{\varphi^c}(\mathbb{D}^n)$ given in terms of a series of functions.

The goal of the present article is to show that Theorem 1.1 remains true when we swap $\mathbb{D}^n$ by a unit Hilbert ball, so we give a positive response on a question raised in Remark 3.1 in [3]. Therefore, let $\mathbb{B}_H = \{x \in H : \|x\|_H < 1\}$ be the open unit ball in $H = (H,\langle\cdot,\cdot\rangle_H; \|\cdot\|_H)$, an infinite dimensional complex Hilbert space endowed with the inner product $\langle\cdot,\cdot\rangle_H$ and the norm $\|\cdot\|_H$. E.g., $H = L_2(X,\mu)$, the space of square-integrable measurable functions on X with respect to the measure μ such that $\langle f,g\rangle_H = \int_X f(x)\overline{g(x)}d\mu(x)$ and $\|f\|_H = \left(\int_X |f(x)|^2 d\mu(x)\right)^{\frac{1}{2}}$.

Instead to use a holomorphic function, we employ a complex-valued analytic function on $\mathbb{B}_H$, i.e., a Fréchet differentiable function at all points in $\mathbb{B}_H$.

Received: 20.01.2023; Accepted: 03.06.2023; Published Online: 04.06.2023

*Corresponding author: Mohammed El Aïdi; melaidi@unal.edu.co

DOI: 10.33205/cma.1240126

2. PRELIMINARIES AND STATEMENT OF THE MAIN THEOREM

Let $\mathcal{A}(\mathbb{B}_H)$ be the space of analytic functions on $\mathbb{B}_H$, ϕ be a strictly positive continuous function on $[0,1)$, where its inverse is logarithmically convex. Let $L^\infty_\phi(\mathbb{B}_H) = \left(L^\infty_\phi(\mathbb{B}_H), \|\cdot\|_{\infty,\phi}\right)$ be the space of complex-valued measurable functions f on $\mathbb{B}_H$ such that $\phi(\|x\|_H).f(x)$ is bounded for all $x \in \mathbb{B}_H$ and $||f||_{\infty,\phi} = \sup_{x\in\mathbb{B}_H} \phi(||x||_H)|f(x)| < \infty$.

The weighted Bergman space of infinite order on $\mathbb{B}_H$ is defined by

$$B^\infty_\phi(\mathbb{B}_H) = \left\{f \text{ complex measurable functions on } \mathbb{B}_H : f \in \mathcal{A}(\mathbb{B}_H) \cap L^\infty_\phi(\mathbb{B}_H)\right\}.$$

The space $B^\infty_\phi(\mathbb{B}_H)$ is endowed with the induced norm $\|\cdot\|_{\infty,\phi}$. We suppose that the continuous function ϕ is not identically equal to one which implies that $B^\infty_\phi(\mathbb{B}_H)$ contains strictly $H^\infty(\mathbb{B}_H)$, the Hardy space of order infinity on $\mathbb{B}_H$. We recall that interpolating a sequence by a function in $H^\infty(\mathbb{B}_H)$ has been conducted in [8].

Let $(l^\infty_\phi, ||\cdot||_{l^\infty_\phi})$ be the weighted space of bounded sequences with respect to the sequence $(x_k)_{k\in\mathbb{N}}$ in $\mathbb{B}_H$ and which is defined by

$$l^\infty_\phi = \left\{v = (v_k)_{k\in\mathbb{N}} \in \mathbb{C} \text{ such that } (\phi(\|x_k\|_H)|v_k|)_{k\in\mathbb{N}} \in l^\infty\right\}$$

such that $\|v\|_{l^\infty_\phi} = \sup_{k\in\mathbb{N}} \left(\phi(\|x_k\|_H)|v_k|\right)$. In the sequel, we need the following definition of an interpolation sequence.

Definition 2.1. *Let c be a positive constant, we say that $\Gamma = (x_k)_{k\in\mathbb{N}}$ is an interpolation sequence for $B^\infty_{\phi^c}(\mathbb{B}_H)$ if for every complex-valued sequence $v = (v_k)_{k\in\mathbb{N}} \in l^\infty_{\phi^{c-4}}$, there is $f \in B^\infty_{\phi^c}(\mathbb{B}_H)$ such that $f(a_k) = v_k$. The associated interpolation constant is the smaller constant M such that $||f||_{\infty,\phi} \leq M||v||_{l^\infty_{\phi^{c-4}}}$.*

The pseudohyperbolic distance between two points x, y belonging to $\mathbb{B}_H$ is defined by $\|\Phi_y(x)\|_H$ such that $\Phi_y(x)$ is the Möbius transformation on $\mathbb{B}_H$ defined by $\Phi_y(x) = (s_y Q_y + P_y)m_y(x)$ such that m_y is the $\mathbb{B}_H$-valued analytic map on $\mathbb{B}_H$ and defined as $m_y(x) = \frac{y-x}{1-\langle y,x\rangle_H}$, $P_y(x) = \frac{\langle y,x\rangle_H}{\|y\|^2_H} y$, $Q_y(x) = x - P_y(x)$, and $s_y = \sqrt{1-\|y\|^2_H}$. It is known (see Page 99 in [5]) that

$$\|\Phi_y(x)\|^2_H = 1 - \frac{(1-\|x\|^2_H)(1-\|y\|^2_H)}{|1-\langle x,y\rangle_H|^2}. \tag{2.1}$$

Our main result states

Theorem 2.2. *Let $\Gamma = (x_k)_{k\in\mathbb{N}}$ be a sequence in $\mathbb{B}_H$ such that $\prod_{j\in\mathbb{N}\setminus\{k\}} \|\Phi_{x_j}(x_k)\|_H \geq \phi(\|x_k\|_H)$ for all $k \in \mathbb{N}$ such that ϕ be a strictly positive continuous function on $[0,1)$ such that its inverse is logarithmically convex. Then Γ is interpolated by a function belonging to $B^\infty_{\phi^c}(\mathbb{B}_H)$ and an upper bound of the associated interpolation constant is provided explicitly and does not rely on the weight function ϕ.*

As we observe that the announcement of the main result is almost the same as the one stated in Theorem 1.1, where $\mathbb{D}_n$ is substituted by $\mathbb{B}_n$ and the complex modulus is substituted by $\|\cdot\|_H$. The novelty of the proof of Theorem 2.2 is that we use the pseudohyperbolic distance between two points in $\mathbb{B}_H$ and essentially Equality (2.1).

In the following section, we furnish the proof of Theorem 2.2 in two parts and we employ the techniques used in [2, 3, 6, 7]. The first part is on building an interpolation function, see Subsection 3.1, and the second one focuses on the interpolation constant, see Subsection 3.2.

3. Proof of the main Theorem

3.1. **On an appropriate interpolating function.** Let us consider the following series of functions on $\mathbb{B}_H$

$$G(x) = \sum_{k=1}^{\infty} v_k G_k(x) \text{ for } x \in \mathbb{B}_H, \tag{3.2}$$

where $(v_k)_{k\in\mathbb{N}} \in l^{\infty}_{\phi^{c-4}}$ such that each G_k is an analytic function on $\mathbb{B}_H$ defined as

$$G_k(x) = \left(\frac{1-\|x_k\|_H^2}{1-\langle x_k, x\rangle_H}\right)^4 \mathcal{W}(x_k,x)\mathcal{V}(x_k,x) \prod_{j\in\mathbb{N}\setminus\{k\}} \frac{\langle \Phi_{x_j}(x_k), \Phi_{x_j}(x)\rangle_H}{\|\Phi_{x_j}(x_k)\|_H^2},$$

where $x \in \mathbb{B}_H$, $(x_k)_{k\in\mathbb{N}}$ is a sequence in $\mathbb{B}_H$, $\mathcal{W}(x_k,\cdot)$ and $\mathcal{V}(x_k,\cdot)$ are two analytic functions on $\mathbb{B}_H$. Precisely,

$$\mathcal{W}(x_k,x) = \exp\left[-\sum_{m\in\mathbb{N}} (\mathfrak{f}(x)-\mathfrak{f}(x_k)) \frac{(1-\|x_m\|_H^2)(1-\|x_k\|_H^2)}{1-|\langle x_m, x_k\rangle_H|^2}\right]$$

with $\mathfrak{f}(x) = \frac{1+\langle x_m, x\rangle_H}{1-\langle x_m, x\rangle_H}$ which is well defined due the fact by using Cauchy-Schwarz inequality, we have $1-\langle x_m, x\rangle_H > 0$ and $\mathcal{V}(x_k,x) = \exp(\partial u(\widetilde{\psi}(x_k)).(\widetilde{\psi}(x)-\widetilde{\psi}(x_k)))$, where $\partial u(\widetilde{\psi}(x_k)).(\widetilde{\psi}(x)-\widetilde{\psi}(x_k))$ is the inner product in $\mathbb{C}^n$ between $\partial u(\widetilde{\psi}(x_k))$ and $\widetilde{\psi}(x)-\widetilde{\psi}(x_k)$ where u is a real-valued convex function on $\mathbb{C}^n$ and $\widetilde{\psi}$ is a $\mathbb{C}^n$-valued surjective map on $\mathbb{B}_H$. Consequently, from the definitions of $\mathcal{W}$ and $\mathcal{V}$, we have $G_k(x_k) = 1$ and for $j \neq k$ we have $G_k(x_j) = 0$ this due to the fact that $\Phi_{x_j}(x_j) = 0$, see (2.1). Whence, the sequence $(a_k)_{k\in\mathbb{N}}$ is interpolated by G and in the next subsection, we prove that $G \in B^{\infty}_{\phi^c}(\mathbb{B}_H)$ and provide explicitly an upper bound associated to the interpolation constant.

3.2. **On the interpolation constant.** By using the hypothesis of Theorem 2.2, that is, for each $k \in \mathbb{N}$, $\prod_{j\in\mathbb{N}\setminus\{k\}} \|\Phi_{x_j}(x_k)\|_H$ is bigger than $\phi(\|x_k\|_H)$, we have

$$|G_k(x)| \leq \left(\frac{1-\|x_k\|_H^2}{1-\langle x_k, x\rangle_H}\right)^4 |\mathcal{W}(x_k,x)||\mathcal{V}(x_k,x)|\phi^{-2}(\|x_k\|_H). \tag{3.3}$$

Let us look an upper bound for $|\mathcal{W}(x_k,x)|$. So, since that we work in a complex Hilbert space, we have $\Re\mathfrak{f}(x) = \frac{1-|\langle x_m, x\rangle_H|^2}{|1-\langle x_m, x\rangle_H|^2}$. Whence, we have

$$\begin{aligned} |\mathcal{W}(x_k,x)| &= \exp\left[-\sum_{m\in\mathbb{N}} \frac{1-|\langle x_m, x\rangle_H|^2}{|1-\langle x_m, x\rangle_H|^2} \frac{(1-\|x_m\|_H^2)(1-\|x_k\|_H^2)}{1-|\langle x_m, x_k\rangle_H|^2}\right] \\ &\times \exp\left[\sum_{m\in\mathbb{N}} \frac{(1-\|x_m\|_H^2)(1-\|x_k\|_H^2)}{|1-\langle x_m, x_k\rangle_H|^2}\right]. \end{aligned} \tag{3.4}$$

Let us show that the terms $\exp\left[\sum_{m\in\mathbb{N}} \frac{(1-\|x_m\|_H^2)(1-\|x_k\|_H^2)}{|1-\langle x_m, x_k\rangle_H|^2}\right]$ is upper bounded by $\exp(1)\phi^{-2}(\|x_k\|_H)$. For $x > 0$, we have $1-x \leq \exp(-x)$, thus by employing, successively, this inequality with $y_{m,k} = \frac{(1-\|x_m\|_H^2)(1-\|x_k\|_H^2)}{|1-\langle x_m, x_k\rangle_H|^2} > 0$, the square of the pseudohyperbolic distance equality (2.1), and the assumption of Theorem 2.2, we obtain

$$\begin{aligned}\exp\left[-\sum_{m\in\mathbb{N}} y_{m,k}\right] &= \prod_{m\in\mathbb{N}} \exp(-y_{m,k})\\ &= \exp(-1)\prod_{m\in\mathbb{N}\setminus\{k\}} \exp(-y_{m,k})\\ &\geq \exp(-1)\prod_{m\in\mathbb{N}\setminus\{k\}} \|\Phi_{x_m}(x_k)\|_H^2 \geq \exp(-1)\phi^2(\|x_k\|_H).\end{aligned}$$

Hence, Equality (3.4) implies

$$|\mathcal{W}(x_k,x)| \leq \frac{\exp(1)}{\phi^2(\|x_k\|_H)}\exp\left[-\sum_{m\in\mathbb{N}} A_{m,k}(x)\right] \tag{3.5}$$

such that $A_{m,k}(x) = \frac{1-|\langle x_m,x\rangle_H|^2}{|1-\langle x_m,x\rangle_H|^2}\frac{(1-\|x_m\|_H^2)(1-\|x_k\|_H^2)}{1-|\langle x_m,x_k\rangle_H|^2}$.

Let us reorder the sequence $(x_k)_{k\in\mathbb{N}}$, for obtaining an increasing sequence $(\|x_k\|_H)_{k\in\mathbb{N}}$, then by using the fact that $\frac{1-|\langle x_m,x\rangle_H|^2}{1-|\langle x_m,x_k\rangle_H|^2} \geq \frac{1-\|x_m\|_H^2}{8(1-|\langle x_k,x\rangle_H|^2)}$ whenever $\|x_m\|_H \geq \|x_k\|_H$, for the proof see Lemmas 3.8 and 3.9 in [8], and Inequality (3.5) becomes

$$|\mathcal{W}(x_k,x)| \leq \frac{\exp(1)}{\phi^2(\|x_k\|_H)}\exp\left[-\frac{\mathfrak{X}\mathfrak{T}_k}{8}\right] \tag{3.6}$$

such that $\mathfrak{X} = \frac{1-\|x_k\|_H^2}{1-|\langle x_k,x\rangle_H|^2}$ and $\mathfrak{T}_k = \sum_{m\geq k}\left(\frac{1-\|x_m\|_H^2}{|1-\langle x_m,x\rangle_H|}\right)^2$.

Let $b_m(x) = \left(\frac{1-\|x_m\|_H^2}{|1-\langle x_m,x\rangle_H|}\right)^2$, then thanks to the triangle inequality, we have $b_k(x) \leq 4\mathfrak{X}^2$, and we observe that the function $g_{\mathfrak{X}}(\tau) = \mathfrak{X}^2\exp\left(-\frac{\mathfrak{X}\tau}{8}\right)$ for $\tau > 0$, is at most equal $h(\tau) = \min\left(1, \frac{256}{\exp(2)\tau^2}\right)$. Accordingly, Inequality (3.6) becomes

$$\begin{aligned}b_k(x)\,|\mathcal{W}(x_k,x)| &\leq \frac{4\exp(1)\mathfrak{X}^2}{\phi^2(\|x_k\|_H)}\exp\left(-\frac{\mathfrak{X}\mathfrak{T}_k}{8}\right)\\ &\leq \frac{4\exp(1)}{\phi^2(\|x_k\|_H)}h\left(\mathfrak{T}_k\right).\end{aligned} \tag{3.7}$$

Now, from the definition of $\mathcal{V}$ and the use the properties of the convex function u, we have $|\mathcal{V}(x_k,x)| \leq \exp(u(\widetilde{\psi}(x)) - u(\widetilde{\psi}(x_k)))$. Furthermore, since that the inverse of ϕ is logarithmically convex, let us choose $u(\widetilde{\psi}(x)) = -c\log(\phi(\|x\|_H))$ and we have

$$|\mathcal{V}(x_k,x)| \leq \phi^c(\|x_k\|_H)\phi^{-c}(\|x\|_H). \tag{3.8}$$

We recall that G_k satisfies

$$|G_k(x)| \leq \left(\frac{1-\|x_k\|_H^2}{1-\langle x_k,x\rangle_H}\right)^4 |\mathcal{W}(x_k,x)||\mathcal{V}(x_k,x)|\phi^{-2}(\|x_k\|_H). \tag{3.9}$$

Whence, by using Inequalities (3.7)-(3.9) we obtain

$$\phi^c(\|x\|_H)\phi^{4-c}(\|x_k\|_H)|G_k(x)| \leq 4\exp(1)b_k(x)h\left(\mathfrak{T}_k\right). \tag{3.10}$$

The function $h(\tau)$ decreases on $[\mathfrak{T}_{k+1},\mathfrak{T}_k]$, then by using Inequality (3.10), we have

$$\phi^c(\|x\|_H)\phi^{4-c}(\|x_k\|_H)|G_k(x)| \leq 4\exp(1)\int_{\mathfrak{T}_{k+1}}^{\mathfrak{T}_k} h(\tau)d\tau. \tag{3.11}$$

Therefore, by using the definition of $h(\tau)$ and Inequality (3.11), we have

$$\begin{aligned}\sum_{k\in\mathbb{N}} \phi^c(\|x\|_H)\phi^{4-c}(\|x_k\|_H)|G_k(x)| &\leq 4\exp(1)\sum_{k\in\mathbb{N}}\int_{\mathfrak{T}_{k+1}}^{\mathfrak{T}_k} h(\tau)d\tau\\ &\leq 4\exp(1)\int_0^\infty h(\tau)d\tau\\ &= 47.0886. \qquad (3.12)\end{aligned}$$

We recall that $G(x) = \sum_{k=1}^\infty v_k G_k(x)$, then from (3.12), we have

$$\begin{aligned}|G(x)| \leq \sum_{k=1}^\infty |v_k||G_k(x)| &\leq \|v\|_{l^\infty_{\phi^c-4}}\sum_{k=1}^\infty \phi^{4-c}(|x_k|)|G_k(x)|\\ &\leq 47.0886\|v\|_{l^\infty_{\phi^c-4}}\phi^{-c}(\|x\|_H).\end{aligned}$$

Thus, $\|G\|_{\infty,\phi} = \sup_{x\in\mathbb{B}_H}\phi^c(\|x\|_H)|G(x)| \leq 47.0886\|v\|_{l^\infty_{\phi^c-4}} < \infty$, i.e., $G \in B^\infty_{\phi^c}(\mathbb{B}_H)$, consequently the sequence Γ is interpolated by the function G, furthermore an upper bound of the interpolation constant is equal to 47.0886. The proof of Theorem 2.2 is complete.

On an extension

We are asking whether it possible to state an analogue result of Theorem 2.2, for a proper subspace of a suitable weighted Bergman space of infinite order on $\mathbb{B}_H$ and containing a proper subspace of $H^\infty(\mathbb{B}_H)$. E.g., interpolating sequences for a proper space of $H^\infty(\mathbb{D})$ has been conducted by Dyakonov [1]. Also, we are asking whether our result remains true for a function belonging to a Bloch-type space on $\mathbb{B}_H$, see, e.g., [9].

References

[1] K. M. Dyakonov: *A free interpolation problem for a subspace of H^∞*, Bull Lond Math Soc., **50** (2018), 477–486.

[2] B. Berndtsson: *Interpolating sequences for H^∞ in the ball*, Nederl Akad Wetensch Indag Math., **47** (1) (1985), 1–10.

[3] M. El Aïdi: *On the interpolation constant for weighted Bergman spaces of infinite order*, Complex Var. Elliptic Equ., **64** (6) (2019), 1043–1049.

[4] P. Galindo, A. Miralles: *Interpolating sequences for bounded analytic functions*, Proc Amer Math Soc., **135** (10) (2007), 3225–3231.

[5] K. Goebel, S. Reich: *Uniform convexity, hyperbolic geometry, and nonexpansive mappings*, Marcel Dekker, Inc., New York and Basel, 1984.

[6] P. Jones: *L^∞-estimates for the δ problem in a half-plane*, Acta Math., **150** (1983) 137–152.

[7] X. Massaneda: *Interpolation by holomorphic functions in the unit ball with polynomial growth* Ann Fac Sci Toulouse Math., **6** (2) (1997), 277–296.

[8] A. Miralles: *Interpolating sequences for $H^\infty(B_H)$*, Quaest Math., **39** (6) (2016), 785–795.

[9] T. T. Quang: *Banach-valued Bloch-type functions on the unit ball of a Hilbert space and weak spaces of Bloch-type*, Constr. Math. Anal., **6** (1) (2023), 6–21.

Mohammed El Aïdi
Universidad Nacional De Colombia, Sede Bogotá
Facultad De Ciencias
Departamento De Matemáticas
Edificio 404, Bogotá , D.C. Colombia
ORCID: 0000-0002-3032-0879
E-mail address: melaidi@unal.edu.co

CONSTRUCTIVE MATHEMATICAL ANALYSIS
6 (2023), No. 2, pp. 107-127
http://dergipark.org.tr/en/pub/cma
ISSN 2651 - 2939

Research Article

Principal eigenvalues of elliptic problems with singular potential and bounded weight function

TOMAS GODOY*

ABSTRACT. Let Ω be a bounded domain in $\mathbb{R}^n$ with $C^{0,1}$ boundary, and let $d_\Omega : \Omega \to \mathbb{R}$ be the distance function $d_\Omega(x) := dist(x, \partial\Omega)$. Our aim in this paper is to study the existence and properties of principal eigenvalues of self-adjoint elliptic operators with weight function and singular potential, whose model problem is $-\Delta u + bu = \lambda mu$ in Ω, $u = 0$ on $\partial\Omega$, $u > 0$ in Ω, where $b : \Omega \to \mathbb{R}$ is a nonnegative function such that $d_\Omega^2 b \in L^\infty(\Omega)$, $m : \Omega \to \mathbb{R}$ is a nonidentically zero function in $L^\infty(\Omega)$ that may change sign, and the solutions are understood in weak sense.

Keywords: Weighted principal eigenvalue problems, second order elliptic operators, singular potentials.

2020 Mathematics Subject Classification: 35J20, 35J25.

1. INTRODUCTION

Let Ω be a bounded domain in $\mathbb{R}^n$ with $C^{1,1}$ boundary if $n > 1$, let m be a real valued function defined on Ω, let $\lambda \in \mathbb{R}$, and let $\mathcal{L}$ be a second order elliptic linear operator on Ω. We recall that λ is said a principal eigenvalue of the operator $\mathcal{L}$ with weight function m and Dirichlet boundary condition, if there exists a solution u to the problem

$$(1.1) \qquad \begin{cases} \mathcal{L}u = \lambda mu \text{ in } \Omega, \\ u = 0 \text{ on } \partial\Omega, \\ u \geq 0 \text{ in } \Omega \text{ and } u \not\equiv 0 \text{ in } \Omega. \end{cases}$$

These problems have received a lot of attention in the literature, in part because they appear naturally when one studies semilinear bifurcation problems via the implicit function theorem (for details see e.g., [8], Chapter 5, Section 5.3). Let us recall some works related to problem (1.1).

Manes and Micheletti in [15] studied the problem (with the solutions understood in weak sense and belonging to $H_0^1(\Omega) \cap C(\overline{\Omega})$)

$$(1.2) \qquad \begin{cases} -\operatorname{div}(A\nabla u) = \lambda mu \text{ in } \Omega, \\ u = 0 \text{ on } \partial\Omega, \\ u > 0 \text{ in } \Omega \end{cases}$$

in the case when $m \in L^r(\Omega)$ for some $r > \frac{n}{2}$ and $A = (a_{ij}(x))$ is a symmetric uniformly elliptic $n \times n$ whose coefficients belong to $C^{0,1}(\overline{\Omega})$. They proved, by variational methods, the following facts:

Received: 28.03.2023; Accepted: 01.06.2023; Published Online: 05.06.2023
*Corresponding author: Tomas Godoy; godoy@famaf.unc.edu.ar
DOI: 10.33205/cma.1272110

a) If $m \geq 0$, then problem (1.2) has a principal eigenvalue $\lambda_1 (m)$, which is positive and simple, and that it is the first positive eigenvalue of the problem

$$\text{(1.3)} \qquad \begin{cases} -\operatorname{div}(A\nabla u) = \lambda m u \text{ in } \Omega, \\ u = 0 \text{ on } \partial\Omega, \end{cases}$$

that is, if λ is any other eigenvalue λ of (1.3), then $\lambda > \lambda_1 (m)$.

b) If $m \leq 0$, then problem (1.2) has a principal eigenvalue $\lambda_{-1} (m)$, which is negative and simple, and satisfies that $\lambda < \lambda_{-1} (m)$ for any other eigenvalue λ of problem (1.3).

c) If $m^+ \not\equiv 0$ and $m^- \not\equiv 0$, then problem (1.2) has two principal eigenvalues $\lambda_1 (m)$ and $\lambda_{-1} (m)$, with $\lambda_1 (m) > 0$ and $\lambda_{-1} (m) < 0$; both of them are simple eigenvalues, and $\lambda \notin (\lambda_{-1} (m), \lambda_1 (m))$ for any eigenvalue λ of problem (1.3).

They proved also a maximum principle with weight, which reads as: If $h \in L^q (\Omega)$ for some $q > n$ and $0 \leq h \not\equiv 0$, and if either $m^+ \not\equiv 0, m^- \not\equiv 0$ and $\lambda_{-1}(m) < \lambda < \lambda_1(m)$, or $m \geq 0$ and $\lambda < \lambda_1(m)$, or $m \leq 0$ and $\lambda > \lambda_{-1}(m)$, then the problem

$$\text{(1.4)} \qquad \begin{cases} -\operatorname{div}(A\nabla u) = \lambda m u + h \text{ in } \Omega, \\ u = 0 \text{ on } \partial\Omega \end{cases}$$

has a unique solution, and it is positive in Ω.

On the other hand, motivated by problems of genetic population dynamics, Brown and Lin in [4] studied the existence and properties of principal eigenvalues for problem (1.2) in the case of the Laplace operator with homogeneous Neumann boundary condition, Hess and Kato in [13] investigated principal eigenvalue problems with weight for a general uniformly elliptic second order linear operator

$$\mathcal{L}u := -\sum_{1\leq i,j\leq n} a_{ij} (x) \frac{\partial^2 u}{\partial x_i \partial x_j} + \sum_{1\leq i\leq n} a_i (x) \frac{\partial u}{\partial x_i} + a_0 (x) u.$$

Indeed, they studied the problem

$$\text{(1.5)} \qquad \begin{cases} \mathcal{L}u = \lambda m u \text{ in } \Omega, \\ u = 0 \text{ on } \partial\Omega, \\ u > 0 \text{ in } \Omega, \end{cases}$$

where the weight m may change sign and belongs to $C^\gamma \left(\overline{\Omega}\right)$ for some $\gamma \in (0, 1)$, and with the solutions understood in classical sense (i.e., $u \in C^2 (\Omega) \cap C \left(\overline{\Omega}\right)$). Under standard regularity assumptions on the coefficients of $\mathcal{L}$ (among them that $a_0 \in C^\gamma \left(\overline{\Omega}\right)$ for some $\gamma \in (0, 1)$), they proved, by using the Krein Rutman theorem, that if $a_0 \geq 0$ in Ω, and $m^+ \not\equiv 0$ (respectively $m^- \not\equiv 0$), then problem (1.5) admits a unique positive (resp. negative) principal eigenvalue $\lambda_1 (m)$ (resp $\lambda_{-1} (m)$) which is simple. They also showed that the solutions u of (1.5) belong to $C^1 \left(\overline{\Omega}\right)$ and satisfy, for some positive constants c_1 and c_2,

$$c_1 d_\Omega \leq u \leq c_2 d_\Omega \text{ in } \Omega.$$

They proved also the following maximum principle with weight: If $a_0 \geq 0$ in Ω, and if $m^+ \not\equiv 0$ (respectively $m^- \not\equiv 0$) and if $0 \leq \lambda < \lambda_1 (m)$ (resp. $\lambda_{-1} (m) < \lambda \leq 0$) then, for any nonidentically zero h such that $0 \leq h \in C^\gamma \left(\overline{\Omega}\right)$, the problem

$$\text{(1.6)} \qquad \begin{cases} \mathcal{L}u = \lambda m u + h \text{ in } \Omega, \\ u = 0 \text{ on } \partial\Omega \end{cases}$$

has a unique (classical) solution u and it is positive in Ω.

Hess and Senn in [18] studied problem (1.5) with the Dirichlet replaced by the Neumann boundary condition.

Lopez-Gomez in [14] addressed problem (1.5) in the case when a_0 is not necessarily nonnegative and, by using arguments relying on the maximum principle, they stated sufficient conditions for the existence and the nonexistence of principal eigenvalues.

Hernandez, Mancebo and Vega (see [10], Section 2), studied problem (1.5) in situations where some coefficients of $\mathcal{L}$ and the weight m are allowed to have a certain kind of singularity along $\partial\Omega$. They assumed that:

1) Ω is a bounded domain in $\mathbb{R}^n$ with $C^{3+\gamma}$ boundary for some $\gamma \in (0,1)$,
2) $A(x) = (a_{i,j}(x))$ is a symmetric $n \times n$ matrix, uniformly and strongly elliptic in $\overline{\Omega}$, and for each i, j, $a_{ij} \in C^3(\Omega) \cap C(\overline{\Omega})$,
3) $a_i \in C^2(\Omega)$ and there exists a constant K and $\alpha \in (-1,1)$ such that $\left|\frac{\partial a_{ij}}{\partial x_k}\right| + |a_i| \leq K\left(1 + d_\Omega^\alpha\right)$ and $\left|\frac{\partial^2 a_{ij}}{\partial x_i \partial x_j}\right| + \left|\frac{\partial a_i}{\partial x_j}\right| \leq K d_\Omega^{\alpha-1}$ for all $x \in \Omega$ and $1 \leq i, j \leq n$; and their assumptions on the functions a_0 and m were:
4) $a_0 \in C^1(\Omega)$ and, for all $k = 1, 2, ..., n$, $d_\Omega^{2-\alpha}\left|\frac{\partial a_0}{\partial x_k}\right| \in L^\infty(\Omega)$, with α as in 3),
5) m is strictly positive in Ω and satisfies the conditions in 4).

Under the hypothesis 1)-5), they proved (see [10, Theorem 2.6]), that there exists a unique real eigenvalue λ with an associated eigenfunction u in the interior of the positive cone of $C^1(\overline{\Omega})$ (i.e., such that $u > 0$ in Ω and $\frac{\partial u}{\partial \nu} < 0$ on $\partial\Omega$, where ν denotes the unit outward normal to $\partial\Omega$), and that such a λ is a simple eigenvalue of problem (1.5).

Let us mention also that Berestycki, Varadhan an Nirenberg in [2] studied, in a generalized sense, problem (1.5) in the case where each $a_{ij} \in C(\Omega)$, $a_0 \in L^\infty(\Omega)$, and $a_i \in L^\infty(\Omega)$ for $i = 1, 2, ..., n$. Additional results and more references concerning principal eigenvalues for elliptic problems can be found in [6].

Principal eigenvalue problems for periodic parabolic operators with Dirichlet boundary condition were studied by Beltramo and Hess in [1], and applications to semilinear periodic parabolic problems were given in [11]. A very good exposition of these results, including problems with either Neumann or Robin boundary conditions and its nonlinear applications, as well as additional references, can be found in the book [12].

Problems of the form

$$\begin{cases} -\Delta u + bu = \lambda m u \text{ in } \Omega, \\ u = 0 \text{ on } \partial\Omega, \\ u > 0 \text{ in } \Omega, \end{cases} \tag{1.7}$$

were studied in [9] in the case when m is a nonnegative and nonidentically zero function belonging to $L^\infty(\Omega)$, and b is a singular potential of the form $b = av^{-\alpha-1}$, where:

1′) $0 < \alpha < 3$,
2′) $a \in L^\infty(\Omega)$ and there exists $\delta > 0$ such that $ess \inf_{A_\delta} a > 0,$, with $A_\delta := \{x \in \Omega : d_\Omega(x) \leq \delta\}$,
3′) $v \in D_\alpha := \left\{v \in H_0^1(\Omega) : \vartheta_\alpha^{-1} v \in L^\infty(\Omega) \text{ and } ess \inf_\Omega \vartheta_\alpha^{-1} v > 0\right\}$, where $\vartheta_\alpha := d_\Omega$ if $0 < \alpha < 1$, $\vartheta_1 := d_\Omega\left(\log\left(\frac{\omega}{d_\Omega}\right)\right)^{\frac{1}{2}}$, where ω is an arbitrary constant greater than the diameter of Ω, and $\vartheta_\alpha := d_\Omega^{\frac{2}{1+\alpha}}$ if $1 < \alpha < 3$.

Under these assumptions, Lemmas 4.3 and 4.4 in [9] state the existence of a positive principal eigenvalue for problem (1.7), and a maximum principle with weight.

Our aim in this paper is to study principal eigenvalue problems with singular potential and bounded weight function of the form

$$\text{(1.8)}\qquad \begin{cases} -\operatorname{div}(A\nabla u) + bu = \lambda m u \text{ in } \Omega, \\ u = 0 \text{ on } \partial\Omega, \\ u > 0 \text{ in } \Omega, \end{cases}$$

where the solution u is understood in weak sense (see Definition 1.1 below), and Ω, A, b and m satisfy the following assumptions:

H1) Ω is a bounded domain in $\mathbb{R}^n$, with $C^{1,1}$ boundary if $n > 1$.

H2) $A : \Omega \to M_n(\mathbb{R})$, with $A = (a_{ij}(x))$ uniformly elliptic (i.e., there exists a constant $\gamma > 0$ such that $\langle A(x)\xi, \xi\rangle \geq \gamma |\xi|^2$ for any $x \in \overline{\Omega}$ and $\xi \in \mathbb{R}^n$) and such that $a_{ij} \in C^{0,1}(\overline{\Omega})$, $a_{ij} = a_{ji}$ for $1 \leq i, j \leq n$.

H3) The potential $b : \Omega \to \mathbb{R}$ is nonnegative and $bd_\Omega^2 \in L^\infty(\Omega)$, where $d_\Omega : \Omega \to \mathbb{R}$ denotes the distance function given by

$$\text{(1.9)}\qquad d_\Omega(x) := dist(x, \partial\Omega).$$

H4) $m \in L^\infty(\Omega)$ and $m \not\equiv 0$ in Ω, i.e., $|\{x \in \Omega : m(x) \neq 0\}| > 0$.

Observe that H3) allows b to be singular along $\partial\Omega$ and H4) allows m to change sign in Ω. The notion of weak solution we use is the usual one, given by the following:

Definition 1.1. *Let $f : \Omega \to \mathbb{R}$ be such that $f\varphi \in L^1(\Omega)$ for any $\varphi \in H_0^1(\Omega)$, and let $u : \Omega \to \mathbb{R}$. We say that u is a weak solution of the problem*

$$\begin{cases} -\operatorname{div}(A\nabla u) = f \textit{ in } \Omega, \\ u = 0 \textit{ on } \partial\Omega \end{cases}$$

if $u \in H_0^1(\Omega)$ and $\int_\Omega \langle A\nabla u, \nabla\varphi\rangle = \int_\Omega f\varphi$ for any $\varphi \in H_0^1(\Omega)$.

The paper is organized as follows: In Section 2, we present some general facts need later. In Section 3, following the approach of [13] we study, for each $\lambda \in \mathbb{R}$ and under the assumptions H1)-H4), the principal eigenvalue problem without weight (i.e., with weight $\mathbf{1}$)

$$\text{(1.10)}\qquad \begin{cases} -\operatorname{div}(A\nabla u) + bu = \lambda m u + \mu u \text{ in } \Omega, \\ u = 0 \text{ on } \partial\Omega, \\ u > 0 \text{ in } \Omega. \end{cases}$$

We prove that, for each $\lambda \in \mathbb{R}$, problem (1.10) has a unique principal eigenvalue $\mu = \mu_{m,b}(\lambda)$, which has the variational characterization

$$\text{(1.11)}\qquad \mu_{m,b}(\lambda) := \inf_{w \in H_0^1(\Omega)\setminus\{0\}} \frac{\int_\Omega \langle A\nabla w, \nabla w\rangle + \int_\Omega (b - \lambda m) w^2}{\int_\Omega w^2}.$$

We prove also that the eigenspace $V_{\mu_{m,b}(\lambda)}$ corresponding to $\mu_{m,b}(\lambda)$ is one dimensional, and that if $0 \not\equiv u \in V_{\mu_{m,b}(\lambda)}$ then $u \in H_0^1(\Omega) \cap C^1(\Omega)$ and either $u \equiv 0$ in Ω, or $u > 0$ in Ω, or $u < 0$ in Ω. In addition, we show that $\mu_{m,b}$ is a concave function which satisfies $\mu_{m,b}(0) > 0$, $\lim_{\lambda\to\infty} \mu_{m,b}(\lambda) = -\infty$ if $m^+ \not\equiv 0$, and $\lim_{\lambda\to-\infty} \mu_{m,b}(\lambda) = -\infty$ if $m^- \not\equiv 0$. We show also that if $m \geq 0$ in Ω then $\mu_{m,b}(\lambda) > 0$ for any $\lambda \leq 0$, and that if $m \leq 0$ in Ω then $\mu_{m,b}(\lambda) > 0$ for any $\lambda \geq 0$. From these facts, it follows that if m changes sign in Ω then the equation $\mu_{m,b}(\lambda) = 0$ has exactly two roots, $\lambda = \lambda_{-1}(m,b) < 0$ and $\lambda = \lambda_1(m,b) > 0$, whereas if $m \geq 0$ (respectively $m \leq 0$) the same equation has a unique solution $\lambda = \lambda_1(m,b) > 0$ (resp. $\lambda = \lambda_{-1}(m,b) < 0$). From these facts, and since the principal eigenvalues of problem (1.8) are exactly the roots of the equation $\mu_{m,b}(\lambda) = 0$, we state, in Section 4 (see Theorem 4.1) the corresponding results for the principal eigenvalues of (1.8). A maximum principle with weight is given in Theorem 4.2, the variational formula for the principal eigenvalues of problem (1.8) is given in Theorem 4.3. In Theorem 4.4 we prove that the eigenfunctions corresponding to these eigenvalues belong to $H_0^1(\Omega) \cap C^1(\Omega) \cap C(\overline{\Omega})$ and we give lower and upper estimates for them (in terms of powers of d_Ω), and in Theorem 4.5 we study the continuity of the maps $(m, b) \to \lambda_1(m, b)$ and $(m, b) \to \Phi_{m,b}$, where $\Phi_{m,b}$ is the positive eigenfunction associated to $\lambda_1(m, b)$ and normalized by $\|\Phi_{m,b}\|_{L^2(\Omega)} = 1$.

2. Preliminaries

For $1 \leq p \leq \infty$, we will write p' for the Hölder conjugate exponent defined by $\frac{1}{p} + \frac{1}{p'} = 1$ (with the convention that $\frac{1}{\infty} = 0$); and p^* will denote the Sobolev critical exponent defined by $\frac{1}{p^*} = \frac{1}{p} - \frac{1}{n}$ if $p < n$ and by $p^* := \infty$ otherwise.
For a measurable function $v : \Omega \to \mathbb{R}$ such that $v\varphi \in L^1(\Omega)$ for any $\varphi \in H_0^1(\Omega)$, we will write S_v to denote the functional $S_v : H_0^1(\Omega) \to \mathbb{R}$ defined by $S_v(\varphi) := \int_\Omega v\varphi$; and we will say $v \in \left(H_0^1(\Omega)\right)'$ to mean that $S_v \in \left(H_0^1(\Omega)\right)'$ and, in this case, if no confusion arises, we will write sometimes v instead of S_v. We will denote by d_Ω the distance to the boundary function $d_\Omega : \Omega \to \mathbb{R}$ defined by

$$d_\Omega(x) = dist(x, \partial\Omega).$$

From now on, $\mathcal{L}_0$ will denote the operator $\mathcal{L}_0 : H_0^1(\Omega) \to \left(H_0^1(\Omega)\right)'$ defined by $\mathcal{L}_0 u := -\operatorname{div}(A\nabla u)$ and, for $\zeta \in \left(H_0^1(\Omega)\right)'$, $\mathcal{L}_0^{-1}(\zeta)$ will denote the unique weak solution $u \in H_0^1(\Omega)$ (given by the Riesz theorem) to the problem $\mathcal{L}_0 u = \zeta$ in Ω, $u = 0$ on $\partial\Omega$.

Remark 2.1. *Let us recall the following well known facts:*

i) (Poincaré's inequality, see e.g., [16], Proposition 1.9.6) If $n > 2$ then there exists a positive constant c such that $\|\varphi\|_{L^{2^*}(\Omega)} \leq c\|\nabla\varphi\|_{L^2(\Omega)}$ for all $\varphi \in H_0^1(\Omega)$ and, if $n = 2$ then for each $q \in [1, \infty)$ there exists a positive constant c_q such that $\|\varphi\|_{L^q(\Omega)} \leq c_q \|\nabla\varphi\|_{L^2(\Omega)}$ for all $\varphi \in H_0^1(\Omega)$.

ii) (Hardy's inequality, see e.g., [3], p. 313) There exists a positive constant c such that $\left\|\frac{\varphi}{d_\Omega}\right\|_{L^2(\Omega)} \leq c\|\nabla\varphi\|_{L^2(\Omega)}$ for all $\varphi \in H_0^1(\Omega)$.

iii) (weak maximum principle, see e.g., [8], Theorem 1.3.7) If $g : \Omega \to \mathbb{R}$ is nonnegative and belongs to $\left(H_0^1(\Omega)\right)'$, then $\mathcal{L}_0^{-1} g \geq 0$.

iv) (weak comparison principle) If $g : \Omega \to \mathbb{R}$ and $h : \Omega \to \mathbb{R}$ belong to $\left(H_0^1(\Omega)\right)'$ and $g \leq h$ in Ω, then $\mathcal{L}_0^{-1} g \leq \mathcal{L}_0^{-1} h$.

Remark 2.2. *Let $v : \Omega \to \mathbb{R}$. From the Poincaré's and Hardy's inequalities of Remark 2.1, it follows immediately that if either $v \in L^{(2^*)'}(\Omega)$ or $d_\Omega v \in L^2(\Omega)$, then:*

i) The functional $S_v : H_0^1(\Omega) \to \mathbb{R}$ is well defined, belongs to $\left(H_0^1(\Omega)\right)'$, and there exists a positive constant c, independent of v, such that: If $v \in L^{(2^*)'}(\Omega)$ then $\|S_v\| \leq c\|v\|_{(2^*)'}$, and if $d_\Omega v \in L^2(\Omega)$ then $\|S_v\| \leq c\|d_\Omega v\|_2$.

ii) The problem $\mathcal{L}_0 z = v$ in Ω, $z = 0$ on $\partial\Omega$, has a unique weak solution $z \in H_0^1(\Omega)$, and it satisfies, for some positive constant c independent of v, $\|z\|_{H_0^1(\Omega)} \leq c\|v\|_{(2^*)'}$ when $v \in L^{(2^*)'}(\Omega)$, and $\|z\|_{H_0^1(\Omega)} \leq c\|d_\Omega v\|_2$ when $d_\Omega v \in L^2(\Omega)$.

Remark 2.3. *If $v:\Omega\rightarrow\mathbb{R}$ be a measurable function such that $v\varphi\in L^{1}\left(\Omega\right)$ for any $\varphi\in H_{0}^{1}\left(\Omega\right)$ and if $S_{v}\in\left(H_{0}^{1}\left(\Omega\right)\right)^{\prime}$, then, by the Riesz theorem, the problem*

$$\mathcal{L}_{0}z=v \textit{ in } \Omega,\quad z=0 \textit{ on } \partial\Omega$$

has a unique weak solution $z\in H_{0}^{1}\left(\Omega\right)$, and it satisfies $\left\Vert z\right\Vert_{H_{0}^{1}\left(\Omega\right)}=\left\Vert S_{v}\right\Vert_{\left(H_{0}^{1}\left(\Omega\right)\right)^{\prime}}$.

If g and h are real functions defined $a.e.$ in Ω, we will write sometimes $f\approx g$ to mean that there exist positive constants c_{1} and c_{2} such that $c_{1}f\leq g\leq c_{2}f$ $a.e.$ in Ω. We will write also $f\lesssim g$ to mean that there exists a positive constant c such that $f\leq cg$ $a.e.$ in Ω.
For $\delta>0$, we set $\Omega_{\delta}:=\left\{x\in\Omega:d_{\Omega}\left(x\right)>\delta\right\}$.

Lemma 2.1. *If w and φ belong to $H_{0}^{1}\left(\Omega\right)$, then $d_{\Omega}^{-2}w\varphi\in L^{1}\left(\Omega\right)$ and there exists a positive constant, independent of w and φ, such that*

$$\left\Vert d_{\Omega}^{-2}w\varphi\right\Vert_{1}\leq c\left\Vert w\right\Vert_{H_{0}^{1}\left(\Omega\right)}\left\Vert\varphi\right\Vert_{H_{0}^{1}\left(\Omega\right)}. \tag{2.12}$$

Proof. The lemma follows immediately from the Hardy's inequality. □

Lemma 2.2. *Let $b:\Omega\rightarrow\mathbb{R}$ be a nonnegative function such that $d_{\Omega}^{2}b\in L^{\infty}\left(\Omega\right)$, and let $h:\Omega\rightarrow\mathbb{R}$ be such that $h\in\left(H_{0}^{1}\left(\Omega\right)\right)^{\prime}$. Then:*

i) There exists a unique weak solution $u\in H_{0}^{1}\left(\Omega\right)$ to the problem

$$\begin{cases}\mathcal{L}_{0}u+bu=h \textit{ in } \Omega,\\ u=0 \textit{ on } \partial\Omega.\end{cases} \tag{2.13}$$

ii) If $h\geq0$, and if u is the weak solution of (2.13), then $u\geq0$ $a.e$ in Ω.

iii) If $h\geq0$ and $h\not\equiv0$, and if u is the weak solution of (2.13), then, for any $\delta>0$ such that $\Omega_{\delta}\neq\varnothing$, there exists a positive constant c such that $u\geq cd_{\Omega_{\delta}}$ $a.e$ in Ω_{δ}. In particular, $u>0$ $a.e.$ in Ω.

Proof. Let $B:H_{0}^{1}\left(\Omega\right)\times H_{0}^{1}\left(\Omega\right)\rightarrow\mathbb{R}$ be defined by

$$B\left(\varphi,\psi\right):=\int_{\Omega}\left(\left\langle A\nabla\varphi,\nabla\psi\right\rangle+b\varphi\psi\right).$$

By Lemma 2.1, B is a continuous bilinear form on $H_{0}^{1}\left(\Omega\right)\times H_{0}^{1}\left(\Omega\right)$ and, since $b\geq0$, B is also coercive. Then $i)$ follows from the Lax Milgram theorem. Suppose now $h\geq0$. By taking $-u^{-}$ as a test function in (2.13), we get

$$\int_{\Omega}\left(\left\langle A\nabla u^{-},\nabla u^{-}\right\rangle+b\left(u^{-}\right)^{2}\right)=\int_{\Omega}\left(\left\langle A\nabla u,-\nabla u^{-}\right\rangle+bu\left(-u^{-}\right)\right)=-\int_{\Omega}hu^{-}\leq0,$$

which gives $u^{-}=0$ $a.e.$ in Ω. Thus $ii)$ holds.

To prove $iii)$, observe that if $h\geq0$ $a.e$ in Ω and $h\not\equiv0$ in Ω, then, for δ positive and small enough, there exist $\varepsilon>0$ and a measurable set $E\subset\Omega_{\delta}$ such that $|E|>0$ and $h\geq\varepsilon\chi_{E}$ in Ω_{δ}. For such a δ, let $\Omega^{\prime}$ be a regular domain such that $\Omega_{\delta}\subset\subset\Omega^{\prime}\subset\subset\Omega$, and consider the problem

$$\begin{cases}-\mathcal{L}_{0}z+bz=\varepsilon\chi_{E} \text{ in } \Omega^{\prime},\\ z=0 \text{ on } \partial\Omega^{\prime}.\end{cases}$$

Since $0\leq b_{|\Omega^{\prime}}\in L^{\infty}\left(\Omega^{\prime}\right)$ and $\varepsilon\chi_{E}\in L^{\infty}\left(\Omega^{\prime}\right)$, by the inner elliptic estimates in ([7], Theorem 9.11), we have $z\in W^{2,q}\left(\Omega^{\prime}\right)\cap W_{0}^{1,q}\left(\Omega^{\prime}\right)$ for any $q\in[1,\infty)$ and so $z\in C^{1}\left(\overline{\Omega^{\prime}}\right)$. By the maximum principle (as stated e.g., in [7, Theorem 9.1]) we have $z\left(x\right)>0$ for any $x\in\Omega^{\prime}$, and by the Hopf's boundary lemma (as stated e.g., in [17, Theorem 1.1]), we have also $\frac{\partial z}{\partial\nu}<0$ on $\partial\Omega^{\prime}$ and from these two facts it follows that z belongs to the interior of the positive cone of $C^{1}\left(\overline{\Omega^{\prime}}\right)$, and so there exists a constant $c > 0$ (which may depend on Ω') such that $z \geq c d_{\Omega'}$ in Ω'. Therefore, since $d_{\Omega'} \geq d_{\Omega_\delta}$ in Ω_δ, we have $z \geq c d_{\Omega_\delta}$ in Ω_δ. Now,

$$\begin{cases} \mathcal{L}_0 (u - z) + b (u - z) = h - \varepsilon \chi_E \geq 0 \text{ in } D' (\Omega'), \\ u - z \geq 0 \text{ on } \partial\Omega', \end{cases}$$

with the inequality on $\partial\Omega'$ understood in the sense of the trace. Thus, by the maximum principle (as stated, e.g., in [7, Theorem 9.1]), $u \geq z$ in Ω' and then $u \geq c d_{\Omega_\delta}$ *a.e.* in Ω_δ. Thus *iii*) holds for δ positive and small enough, and so *iii*) holds also for any $\delta > 0$ such that $\Omega_\delta \neq \varnothing$ (because if $0 < \delta_1 < \delta_2$ and $\Omega_{\delta_2} \neq \varnothing$ then $d_{\Omega_1} \leq d_{\Omega_2}$ in Ω_{δ_2}). □

Remark 2.4. *Let $b : \Omega \to \mathbb{R}$ be a nonnegative function such that $d_\Omega^2 b \in L^\infty (\Omega)$, and let $(\mathcal{L}_0 + b)^{-1} : L^2 (\Omega) \to H_0^1 (\Omega)$ be the solution operator of problem (2.13), i.e., the operator defined by $(\mathcal{L}_0 + b)^{-1} h = u$, where u is the weak solution of (2.13). Then $(\mathcal{L}_0 + b)^{-1} : L^2 (\Omega) \to H_0^1 (\Omega)$ is continuous and $(\mathcal{L}_0 + b)^{-1} : L^2 (\Omega) \to L^2 (\Omega)$ is a compact operator. Indeed, for $h \in L^2 (\Omega)$ and $u = (\mathcal{L}_0 + b)^{-1} h$, we have*

$$c \|u\|_{H_0^1(\Omega)}^2 \leq \int_\Omega \langle A\nabla u, \nabla u\rangle + \int_\Omega b u^2 = \int_\Omega h u \leq c_P \|h\|_2 \|u\|_{H_0^1(\Omega)},$$

where c is the ellipticity constant of A and c_P is the constant of the Poincaré's inequality, and so, if $u \not\equiv 0$, then $\|u\|_{H_0^1(\Omega)} \leq c^{-1} c_P \|h\|_2$. Since clearly this inequality holds also when $u \equiv 0$, it follows that $(\mathcal{L}_0 + b)^{-1} : L^2 (\Omega) \to H_0^1 (\Omega)$ is continuous. Then, since $H_0^1 (\Omega)$ has compact inclusion in $L^2 (\Omega)$, we conclude that $(\mathcal{L}_0 + b)^{-1} : L^2 (\Omega) \to L^2 (\Omega)$ is a compact operator.

3. A one parameter eigenvalue problem with singular potential

From now on, b and m will denote, respectively, a nonnegative function $b : \Omega \to \mathbb{R}$ such that $d_\Omega^2 b \in L^\infty (\Omega)$, and a nonidentically zero function $m \in L^\infty (\Omega)$, which (except if otherwise is explicitly stated) may change sign.

Definition 3.2. *For $\lambda \in \mathbb{R}$, let*

(3.14) $$\mu_{m,b} (\lambda) := \inf_{w \in H_0^1(\Omega) \setminus \{0\}} \frac{\int_\Omega \langle A\nabla w, \nabla w\rangle + \int_\Omega (b - \lambda m) w^2}{\int_\Omega w^2}.$$

Notice that, by the Hardy's inequality,

(3.15) $$0 \leq \int_\Omega b w^2 = \int_\Omega d_\Omega^2 b \frac{w^2}{d_\Omega^2} \leq \left\| d_\Omega^2 b \right\|_\infty \|w\|_{H_0^1(\Omega)}^2 \leq c \|w\|_{H_0^1(\Omega)}^2$$

for any $w \in H_0^1 (\Omega)$, where c is a positive constant independent of w. Also,

$$\begin{aligned} &\frac{\int_\Omega \langle A\nabla w, \nabla w\rangle + \int_\Omega (b - \lambda m) w^2}{\int_\Omega w^2} \\ &\geq \frac{\int_\Omega \langle A\nabla w, \nabla w\rangle + \int_\Omega b w^2}{\int_\Omega w^2} - \|m\|_\infty |\lambda| \geq - \|m\|_\infty |\lambda|, \end{aligned}$$

and then $\mu_{m,b} (\lambda)$ is well defined and finite for any $\lambda \in \mathbb{R}$.

Proposition 3.1. *For any $\lambda \in \mathbb{R}$, we have:*

i) If $\mu \in \mathbb{R}$ and if u is a weak solution of the problem

(3.16) $$\begin{cases} -\operatorname{div} (A\nabla u) + b u = \lambda m u + \mu u \text{ in } \Omega, \\ u = 0 \text{ on } \partial\Omega \end{cases}$$

then $u \in C^1(\Omega)$ and $\mu_{m,b}(\lambda) \leq \mu$.

ii) The infimum in (3.14) is achieved at some nonnegative and nonidentically zero $u \in H_0^1(\Omega)$.

Proof. To prove $i)$, it is enough to see that: if u is a weak solution of (3.16), and if Ω' is an arbitrary regular domain such that $\Omega' \subset\subset \Omega$, then $u \in C^1(\Omega')$. We consider first the case $n = 2$. For Ω' as above, let U_0 be a regular domain such that $\Omega \supset\supset U_0 \supset\supset \Omega'$. Since $n = 2$, we have $u \in H_0^1(\Omega) \subset L^q(\Omega)$ for any $q \in [1, \infty)$, and so $u \in L^q(U_0)$, $\lambda mu + \mu u \in L^q(U_0)$ for some $q > 2$. Also, $b \in L^\infty(U_0)$. Then, taking into account (3.16), and the inner elliptic estimates in ([7], Theorem 9.11), we get $u \in W^{2,q}(\Omega') \subset C^1(\Omega')$. Suppose now $n > 2$, and let $\{\Omega_j\}_{j \in \mathbb{N} \cup \{0\}}$ and $\{U_j\}_{j \in \mathbb{N} \cup \{0\}}$ be two sequences of regular domains such that $\Omega_0 = \Omega$ and $\Omega_j \supset\supset U_j \supset\supset \Omega_{j+1} \supset\supset \Omega'$ for all $j \in \mathbb{N} \cup \{0\}$. For $j \in \mathbb{N} \cup \{0\}$, let q_j be inductively defined by $q_0 = 2$, and by $q_{j+1} = q_j^*$ (with $q_j^* := \infty$ if $q_j \geq n$). Let $j_0 = \max\left\{j \in \mathbb{N} \cup \{0\} : q_j^* < \infty\right\}$. Thus $q_{j_0} < n$ and $q_{j_0}^* \geq n$. Let us show, inductively, that

$$u \in W^{2,q_j}(\Omega_{j+1}) \text{ for } j = 0, 1, ..., j_0. \tag{3.17}$$

Since $u \in L^2(\Omega)$, we have $u \in L^2(U_0)$, $\lambda mu + \mu u \in L^2(U_0)$. Also, $b \in L^\infty(U_0)$ and thus, by (3.16) and ([7], Theorem 9.11), $u \in W^{2,2}(\Omega_1) = W^{2,q_0}(\Omega_1)$. Then (3.17) holds for $j = 0$. Suppose now that (3.17) holds for some $j \in \{0, 1, ..., j_0 - 1\}$. Then $u \in L^{q_j^*}(U_{j+1})$, $\lambda mu + \mu u \in L^{q_j^*}(U_{j+1})$, and also $b \in L^\infty(U_{j+1})$, and so, again now from (3.16) and ([7], Theorem 9.11), $u \in W^{2,q_j^*}(\Omega_{j*2}) = W^{2,q_{j+1}}(\Omega_{j*2})$, which completes the inductive proof of (3.17). Then $u \in W^{2,q_{j_0}}(\Omega_{j_0+1})$ and so, by using again now the above argument, $u \in W^{2,q_{j_0}^*}(\Omega_{j_0+2})$. If $q_{j_0}^* > n$ then $W^{2,q_{j_0}^*}(\Omega_{j_0+2}) \subset C^1(\Omega_{j_0+2}) \subset C^1(\Omega')$ and we are done. If $q_{j_0}^* = n$ then $W^{2,q_{j_0}^*}(\Omega_{j_0+2}) \subset L^r(\Omega_{j_0+2})$ for any $r \in [1, \infty)$. We take $r > n$ to obtain, proceeding as above, $u \in W^{2,r}(\Omega_{j_0+3}) \subset C^1(\Omega_{j_0+3}) \subset C^1(\Omega')$. Thus the first assertion of $i)$ holds.

On the other hand, from (3.16),

$$\int_\Omega \left(\langle A\nabla u, \nabla u\rangle + (b - \lambda m)u^2\right) = \mu \int_\Omega u^2$$

and so $\mu = \left(\int_\Omega u^2\right)^{-1} \int_\Omega \left(\langle A\nabla u, \nabla u\rangle + (b - \lambda m)u^2\right) \geq \mu_{m,b}(\lambda)$, the last inequality by (3.14), which completes the proof of $i)$. To prove $ii)$ consider a minimizing sequence $\{w_j\}_{j \in \mathbb{N}}$ for (3.14). After normalizing it, and by replacing, if necessary, w_j by $|w_j|$ we can assume that $w_j \geq 0$ and $\|w_j\|_2 = 1$ for each j. From (3.14), we have

$$\mu_{m,b}(\lambda) = \lim_{j \to \infty} \left(\int_\Omega \langle A\nabla w_j, \nabla w_j\rangle + \int_\Omega (b - \lambda m) w_l^2\right) \tag{3.18}$$

$$\geq \liminf_{j \to \infty} \int_\Omega \langle A\nabla w_j, \nabla w_j\rangle - |\lambda| \, \|m\|_\infty \tag{3.19}$$

and so, after pass to a further subsequence if necessary, we can assume that $\{w_j\}_{j \in \mathbb{N}}$ is bounded in $H_0^1(\Omega)$. Thus there exist $u \in H_0^1(\Omega)$ and a subsequence, still denoted by $\{w_j\}_{j \in \mathbb{N}}$, such that $\{\nabla w_j\}_{j \in \mathbb{N}}$ converges weakly in $L^2(\Omega, \mathbb{R}^n)$ to ∇u and $\{w_j\}_{j \in \mathbb{N}}$ converges strongly in $L^2(\Omega)$ to u. Thus $\|u\|_2 = 1$. After pass to a further subsequence if necessary, we can assume also that $\{w_j\}_{j \in \mathbb{N}}$ converges to u $a.e.$in Ω and so, since each w_j is nonnegative, we have $u \geq 0$. Let $k \in \mathbb{R}$

such that $b-\lambda m+k\geq 0$. From the equality in (3.18) and since $\|w_j\|_2=1$, we have

$$\begin{aligned}\mu_{m,b}(\lambda)+k&=\lim_{j\to\infty}\left(\int_\Omega\langle A\nabla w_j,\nabla w_j\rangle+\int_\Omega(b-\lambda m+k)\,w_l^2\right)\\&\geq\liminf_{j\to\infty}\int_\Omega\langle A\nabla w_j,\nabla w_j\rangle+\liminf_{j\to\infty}\int_\Omega(b-\lambda m+k)\,w_l^2\\&\geq\int_\Omega\langle A\nabla u,\nabla u\rangle+\int_\Omega(b-\lambda m+k)\,u^2\\&=\int_\Omega\langle A\nabla u,\nabla u\rangle+\int_\Omega(b-\lambda m)\,u^2+k,\end{aligned}$$

where in the last inequality it was used the Fatou's Lemma and the fact that $\|\langle A\nabla u,\nabla u\rangle\|_2\leq\liminf_{j\to\infty}\|\langle A\nabla w_j,\nabla w_j\rangle\|_2$. Then $\mu_{m,b}(\lambda)\geq\int_\Omega\langle A\nabla u,\nabla u\rangle+\int_\Omega(b-\lambda m)\,u^2$. On the other hand, from the definition of $\mu_{m,b}(\lambda)$, we get the opposite inequality. Then $\mu_{m,b}(\lambda)=\int_\Omega\langle A\nabla u,\nabla u\rangle+\int_\Omega(b-\lambda m)\,u^2$ and so $ii)$ holds. □

Proposition 3.2. *For any $\lambda\in\mathbb{R}$, we have:*

i) If u is a minimizer of (3.14), then u is a weak solution of the problem

$$\left\{\begin{array}{c}-\operatorname{div}(A\nabla u)+bu=\lambda mu+\mu_{m,b}(\lambda)\,u\text{ in }\Omega,\\u=0\text{ on }\partial\Omega.\end{array}\right.\tag{3.20}$$

ii) For $\mu\in\mathbb{R}$, if u is a nonidentically zero weak solution of the problem

$$\left\{\begin{array}{c}-\operatorname{div}(A\nabla u)+bu=\lambda mu+\mu u\text{ in }\Omega,\\u=0\text{ on }\partial\Omega\end{array}\right.\tag{3.21}$$

such that $u\geq 0$ in Ω, then $\mu=\mu_{m,b}(\lambda)$ and u is a minimizer of (3.14).

Proof. To prove $i)$, consider a minimizer w of (3.14). Thus

$$\mu_{m,b}(\lambda)=\frac{\int_\Omega\left(\langle A\nabla w,\nabla w\rangle+(b-\lambda m)\,w^2\right)}{\int_\Omega w^2}.\tag{3.22}$$

Let $\psi\in H_0^1(\Omega)$. Then there exists $\varepsilon_0>0$ such that $w+t\psi\in H_0^1(\Omega)\setminus\{0\}$ for any $t\in(-\varepsilon_0,\varepsilon_0)$. Then, for such a t,

$$\mu_{m,b}(\lambda)\leq\frac{\int_\Omega\left(\langle A\nabla(w+t\psi),\nabla(w+t\psi)\rangle+(b-\lambda m)\,(w+t\psi)^2\right)}{\int_\Omega(w+t\psi)^2}.\tag{3.23}$$

From (3.23), a computation using gives that, for $t\in(0,\varepsilon_0)$,

$$\begin{aligned}&\mu_{m,b}(\lambda)\left(\int_\Omega w\psi+\frac{t}{2}\int_\Omega\psi^2\right)\\&\leq\int_\Omega\left(\langle A\nabla w,\nabla\psi\rangle+\frac{t}{2}\langle A\nabla\psi,\nabla\psi\rangle+(b-\lambda m)\left(w\psi+\frac{t}{2}w^2\right)\right),\end{aligned}$$

and so, by taking $\lim_{t\to 0^+}$ we get $\mu_{m,b}(\lambda)\int_\Omega w\psi\leq\int_\Omega\left(\langle A\nabla w,\nabla\psi\rangle+(b-\lambda m)\,w\psi\right)$. By replacing ψ by $-\psi$, the reversed inequality is obtained, and thus $i)$ holds.

To prove $ii)$, suppose that $u\in H_0^1(\Omega)$ is a nonidentically zero weak solution of (3.16) such that $u\geq 0$ in Ω. Let $w\in C_c^\infty(\Omega)$ and let $\varepsilon>0$. Then $\frac{w^2}{u+\varepsilon}\in H_0^1(\Omega)$. We take $\frac{w^2}{u+\varepsilon}$ as a test function in (3.16) to obtain

$$\int_\Omega \left\langle A\nabla u, \frac{(u+\varepsilon)\,2w\nabla w - w^2\nabla u}{(u+\varepsilon)^2} \right\rangle + \int_\Omega bw^2 \frac{u}{u+\varepsilon}$$
$$= \lambda \int_\Omega mw^2 \frac{u}{u+\varepsilon} + \int_\Omega w^2 \frac{\mu u}{u+\varepsilon},$$

that is

$$\int_\Omega \left\langle A\nabla u, \frac{2w\nabla w}{u+\varepsilon} \right\rangle - \int_\Omega \left\langle A\nabla u, \frac{w^2\nabla u}{(u+\varepsilon)^2} \right\rangle + \int_\Omega bw^2 \frac{u}{u+\varepsilon}$$
$$= \lambda \int_\Omega mw^2 \frac{u}{u+\varepsilon} + \int_\Omega w^2 \frac{\mu u}{u+\varepsilon},$$

i.e.,

$$\int_\Omega 2\left\langle Aw\nabla \ln(u+\varepsilon), \nabla w\right\rangle - \int_\Omega \left\langle Aw\nabla \ln(u+\varepsilon), w\nabla \ln(u+\varepsilon)\right\rangle + \int_\Omega bw^2 \frac{u}{u+\varepsilon}$$
$$= \lambda \int_\Omega mw^2 \frac{u}{u+\varepsilon} + \int_\Omega w^2 \frac{\mu u}{u+\varepsilon},$$

that is

$$-\int_\Omega \left\langle A\left(w\nabla \ln(u+\varepsilon) - \nabla w\right), w\nabla \ln(u+\varepsilon) - \nabla w\right\rangle + \int_\Omega \left\langle A\nabla w, \nabla w\right\rangle + \int_\Omega bw^2 \frac{u}{u+\varepsilon}$$
$$= \lambda \int_\Omega mw^2 \frac{u}{u+\varepsilon} + \mu \int_\Omega w^2 \frac{u}{u+\varepsilon},$$

and so

$$\int_\Omega w^2 \frac{\mu u}{u+\varepsilon} \le \int_\Omega \left\langle A\nabla w, \nabla w\right\rangle + \int_\Omega bw^2 \frac{u}{u+\varepsilon} - \lambda \int_\Omega mw^2 \frac{u}{u+\varepsilon}. \tag{3.24}$$

From (3.24), by taking $\lim_{\varepsilon\to 0^+}$ and using the Lebesgue's dominated convergence theorem, we get

$$\mu \int_\Omega w^2 \le \int_\Omega \left\langle A\nabla w, \nabla w\right\rangle + \int_\Omega bw^2 - \lambda \int_\Omega mw^2. \tag{3.25}$$

Since this holds for any $w \in C_c^\infty(\Omega)$, and taking into account Lemma 2.1, a density argument gives that (3.25) holds also for any $w \in H_0^1(\Omega)$. Therefore,

$$\mu \le \frac{\int_\Omega \left\langle A\nabla w, \nabla w\right\rangle + \int_\Omega bw^2 - \lambda \int_\Omega mw^2}{\int_\Omega w^2} \tag{3.26}$$

for any $w \in H_0^1(\Omega) \setminus \{0\}$. On the other hand, by taking $w = u$ as a test function in (3.14), we get $\mu = \left(\int_\Omega u^2\right) \int_\Omega \left(\left\langle A\nabla u, \nabla u\right\rangle + bu^2 - \lambda m u^2\right)$. Thus, from this fact and (3.26), $\mu = \mu_{m,b}(\lambda)$. Then $ii)$ holds. □

Proposition 3.3. *For any $\lambda \in \mathbb{R}$, we have:*

- *i) If u is a nonidentically zero weak solution of problem (3.20), then either $u > 0$ in Ω or $u < 0$ in Ω.*
- *ii) The space of the weak solutions u of (3.20) is one dimensional.*

Proof. To prove $i)$ we follow, partly, [15] (see also [5, Theorem 1.13]). We proceed by the way of contradiction. Suppose that $u \in H_0^1(\Omega) \setminus \{0\}$ is a weak solution of (3.20), and that $u^+ \not\equiv 0$ and $u^- \not\equiv 0$. Let

$$\begin{aligned}
\alpha &:= \int_\Omega \left(\langle A\nabla u, \nabla u\rangle + (b - \lambda m)\, u^2\right), & \beta &:= \int_\Omega u^2,\\
\alpha_1 &:= \int_\Omega \left(\left\langle A\nabla u^+, \nabla u^+\right\rangle + (b - \lambda m)\left(u^+\right)^2\right), & \beta_1 &:= \int_\Omega \left(u^+\right)^2,\\
\alpha_2 &:= \int_\Omega \left(\left\langle A\nabla u^-, \nabla u^-\right\rangle + (b - \lambda m)\left(u^-\right)^2\right), & \beta_2 &:= \int_\Omega \left(u^-\right)^2.
\end{aligned}$$

Thus $\alpha = \alpha_1 + \alpha_2$ and $\beta = \beta_1 + \beta_2$. Now,

$$\mu_{m,b}(\lambda) = \frac{\alpha_1 + \alpha_2}{\beta_1 + \beta_2},$$

and so, since u^+ and u^- belong to $H_0^1(\Omega) \setminus \{0\}$,

$$\frac{\alpha_1 + \alpha_2}{\beta_1 + \beta_2} \le \frac{\alpha_1}{\beta_1} \text{ and } \frac{\alpha_1 + \alpha_2}{\beta_1 + \beta_2} \le \frac{\alpha_2}{\beta_2},$$

that is

$$\begin{aligned}
\alpha_1\beta_1 + \alpha_2\beta_1 &\le \beta_1\alpha_1 + \beta_2\alpha_1,\\
\alpha_1\beta_2 + \alpha_2\beta_2 &\le \beta_1\alpha_2 + \beta_2\alpha_2,
\end{aligned} \tag{3.27}$$

i.e., $\frac{\alpha_1}{\beta_1} \ge \frac{\alpha_2}{\beta_2}$ and $\frac{\alpha_1}{\beta_1} \le \frac{\alpha_2}{\beta_2}$. Then $\frac{\alpha_1}{\beta_1} = \frac{\alpha_2}{\beta_2}$ and so $\frac{\alpha_1+\alpha_2}{\beta_1+\beta_2} = \frac{\alpha_1}{\beta_1} = \frac{\alpha_2}{\beta_2}$. Thus $\mu_{m,b}(\lambda) = \frac{\alpha_1}{\beta_1} = \frac{\alpha_2}{\beta_2}$. Therefore u^+ and u^- are nonnegative minimizers of (3.14) and then, by Proposition 3.1 $ii)$, they are nonnegative and nonidentically zero weak solutions of (3.20) and so, for $q \in \mathbb{R}$ such that $b - \lambda m + q \ge 0$ and $\mu_{m,b}(\lambda) + q > 0$ we have, in weak sense,

$$\begin{cases} -\operatorname{div}\left(A\nabla u^+\right) + (b - \lambda m + q)\, u^+ = \left(\mu_{m,b}(\lambda) + q\right) u^+ \text{ in } \Omega,\\ \qquad\qquad\qquad\qquad\qquad\qquad u^+ = 0 \text{ on } \partial\Omega. \end{cases} \tag{3.28}$$

Thus, from Lemma 2.2 (used with b replaced by $b - \lambda m + q$ and with h replaced by $\left(\mu_{m,b}(\lambda) + q\right) u$), we get that, for any $\delta > 0$ such that $\Omega_\delta \ne \varnothing$, there exists a positive constant c such that $u^+ \ge c d_{\Omega_\delta}$ in Ω_δ. In particular, $u^+ > 0$ in Ω, and so $u^- \equiv 0$ in Ω, which contradicts our assumptions. Then $i)$ holds.

To prove $ii)$, suppose that v and w are two linearly independent solutions of (3.20) and let $x_0 \in \Omega$. Taking into account $i)$ and Proposition 3.1, we can assume (by replacing, if necessary, v and/or w by $-v$ and/or $-w$ respectively) that $v(x_0) > 0$ and $w(x_0) > 0$. Let $t_0 = (v(x_0))^{-1} w(x_0)$ and let $z := t_0 v - w$. Then $t_0 > 0$ and z is a solution of (3.20) such that $z(x_0) = 0$. Thus, by $i)$, z is identically zero on Ω, which contradicts the assumed linear independence of v and w. □

Proposition 3.4. *Let $b : \Omega \to \mathbb{R}$ be a nonnegative function such that $d_\Omega^2 b \in L^\infty(\Omega)$, let $m \in L^\infty(\Omega)$ be a nonidentically zero function, and, for $\lambda \in \mathbb{R}$, let $\mu_{m,b}(\lambda)$ be defined by (3.14). Then:*

i) The map $\lambda \to \mu_{m,b}(\lambda)$ is concave and $\mu_{m,b}(0) > 0$.

ii) If $m^+ \not\equiv 0$ then $\lim_{\lambda\to\infty} \mu_{m,b}(\lambda) = -\infty$; and there exists a unique $\lambda > 0$ such that $\mu_{m,b}(\lambda) = 0$. If, in addition, $m \ge 0$ in Ω, then $\mu_{m,b}(\lambda) > 0$ for any $\lambda \le 0$.

iii) If $m^- \not\equiv 0$ then $\lim_{\lambda\to-\infty} \mu_{m,b}(\lambda) = -\infty$; and there exists a unique $\lambda < 0$ such that $\mu_{m,b}(\lambda) = 0$. If, in addition, $m \le 0$ in Ω, then $\mu_{m,b}(\lambda) > 0$ for any $\lambda \ge 0$.

Proof. The first assertion of $i)$ follows from the facts that $\mu_{m,b}(\lambda)$ is finite for any $\lambda \in \mathbb{R}$, and that $\lambda \to \left(\int_\Omega w^2\right)^{-1}\left(\int_\Omega \langle A\nabla w, \nabla w\rangle + \int_\Omega (b-\lambda m)\, w^2\right)$ is an affine function for any $w \in H_0^1(\Omega)\backslash\{0\}$. Observe also that, from Proposition 3.1 ii), Proposition 3.2 $i)$ and Proposition 3.3 $i)$, all of them used with $\lambda = 0$, the problem

$$\begin{cases} -\operatorname{div}(A\nabla u) + bu = \mu_m(0)\, u \text{ in } \Omega, \\ u = 0 \text{ on } \partial\Omega, \\ u > 0 \text{ in } \Omega \end{cases} \tag{3.29}$$

has a weak solution u. By taking u as a test function in (3.29), we get

$$\int_\Omega \langle A\nabla u, \nabla u\rangle + \int_\Omega bu^2 = \mu_m(0)\int_\Omega u^2$$

which gives $\mu_m(0) > 0$. Thus $i)$ holds.

To see $ii)$, suppose $m^+ \not\equiv 0$ and let $w_0 \in H_0^1(\Omega) \setminus \{0\}$ such that $\int_\Omega m w_0^2 > 0$. By normalizing w_0, if necessary, we can assume that $\int_\Omega m w_0^2 = 1$. Then, for any $\lambda \in \mathbb{R}$, $\mu_{m,b}(\lambda) \leq \int_\Omega \langle A\nabla w_0, \nabla w_0\rangle + \int_\Omega b w_0^2 - \lambda \int_\Omega m w_0^2$. From this fact, and since μ_m is concave and $\mu_m(0) > 0$, it follows that $\lim_{\lambda\to\infty}\mu_{m,b}(\lambda) = -\infty$; and that there exists a unique $\lambda > 0$ such that $\mu_{m,b}(\lambda) = 0$. On the other hand, if $m \geq 0$ in Ω and $\lambda \leq 0$, and if u is a positive solution of the problem

$$\begin{cases} -\operatorname{div}(A\nabla u) + bu = \lambda m u + \mu_{m,b}(\lambda)\, u \text{ in } \Omega, \\ u = 0 \text{ on } \partial\Omega, \\ u > 0 \text{ in } \Omega, \end{cases}$$

then, by taking u as a test function, we get

$$\int_\Omega \left(\langle A\nabla u, \nabla u\rangle + bu^2\right) = \lambda\int_\Omega mu^2 + \mu_{m,b}(\lambda)\int_\Omega u^2$$

and so $\int_\Omega u^2 > 0$, which implies $\mu_{m,b}(\lambda) > 0$. Thus $ii)$ holds.

Finally, $iii)$ follows from $ii)$ by using that, by (3.14), $\mu_{m,b}(\lambda) = \mu_{-m}(-\lambda)$. □

4. Principal eigenvalues problems with singular potential and bounded weight

Definition 4.3. *Let $b : \Omega \to \mathbb{R}$ be a nonnegative function such that $d_\Omega^2 b \in L^\infty(\Omega)$ and let $m \in L^\infty(\Omega) \setminus \{0\}$. We say that $\lambda \in \mathbb{R}$ is a principal eigenvalue of the operator $\mathcal{L}_0 + b$ on Ω, with weight function m and homogeneous Dirichlet boundary condition, if the problem*

$$\begin{cases} -\operatorname{div}(A\nabla \phi) + b\phi = \lambda m\phi \textit{ in } \Omega, \\ \phi = 0 \textit{ on } \partial\Omega \end{cases} \tag{4.30}$$

has a weak solution $\phi \in H_0^1(\Omega)$ such that $\phi \geq 0$ a.e. in Ω and $\phi \not\equiv 0$ in Ω. In such a case, any nonidentically zero solution of (4.30) will be called a principal eigenfunction associated to the principal eigenvalue λ.

Theorem 4.1. *Let $b : \Omega \to \mathbb{R}$ be a nonnegative function such that $d_\Omega^2 b \in L^\infty(\Omega)$ and let $m \in L^\infty(\Omega)$ be such that $m \not\equiv 0$. Then:*

i) $\lambda \in \mathbb{R}$ is a principal eigenvalue for problem (4.30) if, and only if, $\mu_{m,b}(\lambda) = 0$.
ii) If $m^+ \not\equiv 0$ (respectively if $m^- \not\equiv 0$) there exists a unique positive (resp. a unique negative) principal eigenvalue for problem (4.30), which will be denoted by $\lambda_1(m,b)$ (resp. by $\lambda_{-1}(m)$).
iii) If $m \geq 0$ (respectively if $m \leq 0$), then $\lambda_1(m,b)$ (resp. $\lambda_{-1}(m)$) is the unique principal eigenvalue for problem (4.30).

iv) *If $\lambda \in \mathbb{R}$ is a principal eigenvalue for problem (4.30), and if u is an associated eigengunction, then $u \in H_0^1(\Omega) \cap C^1(\Omega)$. Moreover, if $u \in H_0^1(\Omega)$ nonidentically zero then either $u > 0$ in Ω or $u < 0$ in Ω.*

v) *The space of solutions of (4.30) is one dimensional.*

Proof. The proposition follows directly from Propositions 3.1, 3.2, 3.3, and 3.4. □

The following form of the maximum principle for problems with singular potential and weight function holds:

Theorem 4.2. *Let $b : \Omega \to \mathbb{R}$ be a nonnegative function such that $d_\Omega^2 b \in L^\infty(\Omega)$ and let $m \in L^\infty(\Omega)$ be a nonidentically zero function. For $\lambda \in \mathbb{R}$, let $\mu_{m,b}(\lambda)$ be defined by (3.14) and let $h : \Omega \to \mathbb{R}$ be such that $h \in \left(H_0^1(\Omega)\right)'$. Then:*

i) *If $\mu_{m,b}(\lambda) > 0$, the problem*

$$\begin{cases} -\operatorname{div}(A\nabla u) + bu = \lambda mu + h \text{ in } \Omega, \\ u = 0 \text{ on } \partial\Omega \end{cases} \tag{4.31}$$

has a unique weak solution.

ii) *If $\mu_{m,b}(\lambda) > 0$ and $0 \not\equiv h \geq 0$, then the solution u of (4.31) is positive a.e in Ω.*

iii) *If $0 \not\equiv h \geq 0$ and if (4.31) has a nonnegative solution, then $\mu_{m,b}(\lambda) > 0$.*

iv) *If $0 \not\equiv h \geq 0$ and $\mu_{m,b}(\lambda) = 0$, then (4.31) has no weak solutions.*

Proof. To prove $i)$, suppose $\mu_{m,b}(\lambda) > 0$ and let $k \in [0, \infty)$ be such that $b - \lambda m + k \geq 0$. Let $T : L^2(\Omega) \to L^2(\Omega)$ be defined by $T := (\mathcal{L}_0 + b - \lambda m + k)^{-1}$. Thus T is a continuous, compact, linear and it is self-adjoint operator on $L^2(\Omega)$. Notice that ρ is an eigenvalue of T if and only if $\rho = \frac{1}{k+\mu}$ with μ an eigenvalue of $\mathcal{L}_0 + b + k - \lambda m$ with (homogeneous Dirichlet boundary condition). By Proposition 3.1 $i)$, we have $\mu \geq \mu_{m,b+k}(\lambda) = \mu_{m,b}(\lambda) + k > 0$, and so $\rho < \frac{1}{k}$. Thus, by the Fredholm alternative theorem, $\frac{1}{k}I - T : L^2(\Omega) \to L^2(\Omega)$ is bijective, and so the problem $\frac{1}{k}u - Tu = \frac{1}{k}Th$ has a unique weak solution $u \in H_0^1(\Omega)$, that is, the problem

$$\begin{cases} \dfrac{1}{k}(\mathcal{L}_0 + b - \lambda m + k)\, u - u = \dfrac{1}{k}h \text{ in } \Omega, \\ u = 0 \text{ on } \partial\Omega \end{cases}$$

has a unique weak solution u. Then $i)$ holds.

To see $ii)$ observe that if $\mu_{m,b}(\lambda) > 0$ and if $u \in H_0^1(\Omega)$ is a weak solution of

$$\begin{cases} -\operatorname{div}(A\nabla u) + (b - \lambda m)\, u = h \text{ in } \Omega, \\ u = 0 \text{ on } \partial\Omega \end{cases}$$

then, by taking $-u^-$ as a test function,

$$\mu_{m,b}(\lambda) \int_\Omega \left(u^-\right)^2 \leq \int_\Omega \left(\left\langle A\nabla u^-, \nabla u^-\right\rangle + (b - \lambda m)\left(u^-\right)^2\right) = -\int_\Omega hu^- \leq 0$$

and so $u^- = 0$. Thus $u \geq 0$. In addition, since $-\operatorname{div}(A\nabla u) + (b - \lambda m + k)\, u = h + ku$ and $0 \not\equiv h + ku \geq 0$, Lemma 2.2 gives $u > 0$ in Ω. Thus $ii)$ holds.

To see $iii)$ suppose that $0 \not\equiv h \geq 0$ and that u is a nonnegative solution of (4.31). Take k as in the proof of $i)$, to get

$$\begin{cases} -\operatorname{div}(A\nabla u) + (b - \lambda m + k)\, u = h + ku \text{ in } \Omega, \\ u = 0 \text{ on } \partial\Omega \end{cases}$$

Then, by Lemma 2.2 $iii)$, $u>0$ $a.e.$ in Ω. Now we can repeat, line by line, the first part of the proof of Lemma 3.2 $ii)$, replacing there, in each appearance, μu by h, to obtain, instead of (3.24), that for any $w\in C_c^\infty(\Omega)$ and $\varepsilon>0$,

$$\int_\Omega w^2\frac{h}{u+\varepsilon}\le\int_\Omega\langle A\nabla w,\nabla w\rangle+\int_\Omega bw^2\frac{u}{u+\varepsilon}-\lambda\int_\Omega mw^2\frac{u}{u+\varepsilon}$$

and so, by taking $\liminf_{\varepsilon\to0^+}$

$$0\le\int_\Omega\langle A\nabla w,\nabla w\rangle+\liminf_{\varepsilon\to0^+}\left(\int_\Omega bw^2\frac{u}{u+\varepsilon}-\lambda\int_\Omega mw^2\frac{u}{u+\varepsilon}\right).\tag{4.32}$$

Notice that $u>0$ $a.e.$ in Ω, $\lim_{\varepsilon\to0^+}bw^2\frac{u}{u+\varepsilon}=bw^2$ $a.e.$ in Ω, and $\lim_{\varepsilon\to0^+}mw^2\frac{u}{u+\varepsilon}=mw^2$ $a.e.$ in Ω. Also, $bw^2\frac{u}{u+\varepsilon}\le bw^2$ and $mw^2\frac{u}{u+\varepsilon}\le mw^2$. Observe also that, by Lemma 2.1 and that, from our assumption on b, $bw^2\in L^1(\Omega)$. Also, clearly $mw^2\in L^1(\Omega)$. Thus, from (4.32) and the Lebesgue's dominated convergence theorem,

$$0\le\int_\Omega\langle A\nabla w,\nabla w\rangle+\int_\Omega bw^2-\lambda\int_\Omega mw^2$$

and so

$$\frac{\int_\Omega\left(\langle A\nabla w,\nabla w\rangle+bw^2-\lambda mw^2\right)}{\int_\Omega w^2}\ge0$$

and thus, since $w\to\int_\Omega bw^2$ and $w\to\int_\Omega mw^2$ are continuous on $H_0^1(\Omega)$, the same inequality holds for any $w\in H_0^1(\Omega)\setminus\{0\}$. Thus $\mu_{m,b}(\lambda)\ge0$. If $\mu_{m,b}(\lambda)=0$, then there exists $\phi\in H_0^1(\Omega)$ such that

$$\begin{cases}-\operatorname{div}(A\nabla\phi)+b\phi=\lambda m\phi\text{ in }\Omega,\\ \phi=0\text{ on }\partial\Omega,\\ \phi>0\text{ in }\Omega.\end{cases}\tag{4.33}$$

Then , $\int_\Omega(\langle A\nabla\phi,\nabla u\rangle+b\phi u)=\lambda\int_\Omega m\phi u$ and also $\int_\Omega(\langle A\nabla u,\nabla\phi\rangle+bu\phi)=\lambda\int_\Omega m\phi u+\int_\Omega h\phi$. Then $\int_\Omega h\phi=0$, which is impossible. □

Remark 4.5. *From Proposition 3.4, it follows immediately that:*

i) If $m\ge0$ in Ω, then $\{\lambda\in\mathbb{R}:\mu_{m,b}(\lambda)>0\}=(-\infty,\lambda_1(m,b))$.
ii) If $m\le0$ in Ω, then $\{\lambda\in\mathbb{R}:\mu_{m,b}(\lambda)>0\}=(\lambda_{-1}(m),\infty)$.
iii) $m^+\not\equiv0$ and $m^-\not\equiv0$, then $\{\lambda\in\mathbb{R}:\mu_{m,b}(\lambda)>0\}=(\lambda_{-1}(m),\lambda_1(m,b))$.

Theorem 4.3. *If $m^+\not\equiv0$, then*

$$\lambda_1(m,b)=\inf_{\left\{w\in H_0^1(\Omega):\int_\Omega mw^2>0\right\}}\frac{\int_\Omega\left(\langle A\nabla w,\nabla w\rangle+bw^2\right)}{\int_\Omega mw^2}\tag{4.34}$$

or, equivalently,

$$\lambda_1(m,b)=\inf_{w\in W_m}\int_\Omega\left(\langle A\nabla w,\nabla w\rangle+bw^2\right),\tag{4.35}$$

where $W_m:=\left\{w\in H_0^1(\Omega):\int_\Omega mw^2=1\right\}$.

Proof. For $\lambda>0$, from (3.14), we have $\mu_{m,b}(\lambda)=0$ if and only if

$$\inf_{\left\{w\in H_0^1(\Omega):\int_\Omega mw^2>0\right\}}\frac{\int_\Omega\left(\langle A\nabla w,\nabla w\rangle+(b-\lambda m)w^2\right)}{\int_\Omega mw^2}=0,$$

i.e., if and only if (4.34) holds. □

Remark 4.6. *From proposition 4.3, it is clear that the following three facts follow:*

i) Let $b_i : \Omega \to \mathbb{R}$, $i = 1, 2$, be nonnegative functions such that $d_\Omega^2 b_i \in L^\infty(\Omega)$, $i = 1, 2$ and let $m \in L^\infty(\Omega) \setminus \{0\}$ be such that $m^+ \not\equiv 0$. If $b_1 \le b_2$ in Ω, then $\lambda_1(m, b_1) \le \lambda_1(m, b_2)$.

ii) Let $b : \Omega \to \mathbb{R}$, $i = 1, 2$, be a nonnegative function such that $d_\Omega^2 b \in L^\infty(\Omega)$, and let $m_i : \Omega \to \mathbb{R}$, $i = 1, 2$, be functions in $L^\infty(\Omega)$ such that $m_1^+ \not\equiv 0$. If $m_1 \le m_2$ in Ω, then $\lambda_1(m_1, b) \ge \lambda_1(m_2, b)$.

iii) Let Ω_1, Ω_2 be bounded domains in $\mathbb{R}^n$ such that $\Omega_1 \subset \Omega_2$, let $m \in L^\infty(\Omega_2)$ be such that $m^+ \not\equiv 0$ in Ω_1 and let $b : \Omega_2 \to \mathbb{R}$ be a nonnegative function such that $d_{\Omega_2}^2 b \in L^\infty(\Omega_2)$. Let $\lambda_1(m, b, \Omega_i)$, $i = 1, 2$, be the positive principal eigenvalue of the operator $\mathcal{L}_0 + b$ on Ω_i with weight function m. Then $\left\{w \in H_0^1(\Omega_1) : \int_{\Omega_1} m w^2 = 1\right\} \subset \left\{w \in H_0^1(\Omega_2) : \int_{\Omega_2} m w^2 = 1\right\}$ and so $\lambda_1(m, b, \Omega_2) \le \lambda_1(m, b, \Omega_1)$.

For $\delta > 0$, we set $A_\delta := \{x \in \Omega : dist(x, \partial\Omega) < \delta\}$.

Remark 4.7. *Let $b : \Omega \to \mathbb{R}$ be a nonnegative function such that $d_\Omega^2 b \in L^\infty(\Omega)$, and let $\delta > 0$ be such that $\Omega_\delta \ne \varnothing$. If $v \in H^1(\Omega) \cap C(\overline{\Omega})$ and $\mathcal{L}_0 v + bv \ge 0$ in $D'(A_\delta)$, $v \ge 0$ on ∂A_δ then $v \ge 0$ in A_δ. Indeed, we have $v^- \in H^1(A_\delta) \cap C(\overline{\Omega})$ and $v^- = 0$ on ∂A_δ, and so $v^- \in H_0^1(A_\delta)$. Let $\{\varphi_j\}_{j\in\mathbb{N}}$ be a sequence in $C_c^\infty(A_\delta)$ such that $\{\varphi_j\}_{j\in\mathbb{N}}$ converges to v^- in $H_0^1(A_\delta)$. By replacing $\{\varphi_j\}_{j\in\mathbb{N}}$ by $\left\{\sqrt{\varphi_j^2 + \frac{1}{j^2}} - \frac{1}{j}\right\}_{j\in\mathbb{N}}$ if necessary, we can assume that each φ_j is nonnegative. Then*

$$\int_{A_\delta} \left(\langle A\nabla v^-, \nabla v^-\rangle + b\left(v^-\right)^2\right) = \lim_{j\to\infty} \int_{A_\delta} \left(\langle A\nabla v^-, \nabla\varphi_j\rangle + bv^-\varphi_j\right)$$
$$= -\lim_{j\to\infty} \int_\Omega \left(\langle A\nabla v, \nabla\varphi_j\rangle + bv\varphi_j\right) \le 0$$

and so $v^- = 0$ on A_δ.

In the case when $0 \le b \in L^\infty(\Omega)$ (and m such that $m \in L^\infty(\Omega)$ and $m^+ \not\equiv 0$), it is well known that any positive eigenfunction u associated to $\lambda_1(b, m)$ satisfies $u \approx d_\Omega$ in Ω (because $u \in C^1(\overline{\Omega})$ and $\frac{\partial u}{\partial \nu} < 0$ on $\partial\Omega$, see e.g., [5], Proposition 1.6 and the Remark immediately before it). Let us mention that, if we require only that $b \ge 0$ and $d_\Omega^2 b \in L^\infty(\Omega)$, the assertion that $u \approx d_\Omega$ in Ω may not hold, as the following example shows:

Example 4.1. *Let $\gamma_1 > 1$ and let φ_1 be a principal eigenfunction for the problem without weight $-\Delta\varphi_1 = \lambda_1\varphi_1$ in Ω, $\varphi_1 = 0$ on $\partial\Omega$, $\varphi_1 > 0$ in Ω. A computation shows that $-\Delta(\varphi_1^\gamma) = \gamma\lambda_1\varphi_1^\gamma - \gamma(\gamma-1)\varphi_1^{\gamma-2}|\nabla\varphi_1|^2$, i.e., $-\Delta(\varphi_1^\gamma) + b\varphi_1^\gamma = \gamma\lambda_1\varphi_1^\gamma$ in Ω, where $b := \gamma(\gamma-1)\varphi_1^{-2}|\nabla\varphi_1|^2$, and, since $\varphi_1 \approx d_\Omega$ in Ω and $|\nabla\varphi_1| \in L^\infty(\Omega)$, we have $b \ge 0$ and $d_\Omega^2 b \in L^\infty(\Omega)$. It is easy to see that $\varphi_1^\gamma \in H_0^1(\Omega)$ and that φ_1^γ satisfies, in weak sense, $-\Delta(\varphi_1^\gamma) + b\varphi_1^\gamma = \gamma\lambda_1\varphi_1^\gamma$ in Ω, $\varphi_1^\gamma = 0$ on $\partial\Omega$, and so φ_1^γ is a principal eigenfunction corresponding to the potential b and the weight $m = 1$, and clearly $\varphi_1^\gamma \not\approx d_\Omega$ in Ω.*

In order to prove the next theorem, we need the following elementary lemma:

Lemma 4.3. *For $\delta > 0$ such that $\Omega_\delta \ne \varnothing$, we have*

$$\{x \in \Omega : dist(x, \partial\Omega) = \delta\} \subset \overline{\Omega_{\frac{\delta}{2}}}. \tag{4.36}$$

Proof. If $x \in \Omega$ and $dist(x, \partial\Omega) = \delta$, then $dist\left(z, \partial\Omega_{\frac{\delta}{2}}\right) = \frac{\delta}{2}$ for any $z \in \partial\Omega_{\frac{\delta}{2}}$, and so there exists $p_z \in \partial\Omega$ such that $|z - p_z| = \frac{\delta}{2}$. Now,

$$|x - z| = |x - p_z - (z - p_z)| \ge |x - p_z| - |z - p_z| = |x - p_z| - \frac{\delta}{2} \ge \delta - \frac{\delta}{2} = \frac{\delta}{2}$$

then, since $z \in \partial\Omega_{\frac{\delta}{2}}$ was arbitrary, we conclude that $dist\left(x, \partial\Omega_{\frac{\delta}{2}}\right) \geq \frac{\delta}{2}$. Thus (4.36) holds. □

Theorem 4.4. *Let $b : \Omega \to \mathbb{R}$ be a nonnegative function such that $d_\Omega^2 b \in L^\infty(\Omega)$, let $m \in L^\infty(\Omega)$ such that $m \not\equiv 0$ in Ω, and let $\lambda \in \mathbb{R}$. If $u \in H_0^1(\Omega)$ is a weak solution of the problem*

$$\left\{ \begin{aligned} -\operatorname{div}(A\nabla u) + bu = \lambda mu \text{ in } \Omega, \\ u = 0 \text{ on } \partial\Omega, \\ u > 0 \text{ in } \Omega, \end{aligned} \right. \tag{4.37}$$

then:

i) There exists a positive constant c_1 such that $u \leq c_1 d_\Omega$ in Ω.
ii) $u \in C\left(\overline{\Omega}\right)$.
iii) If, in addition, $d_\Omega^\beta b \in L^\infty(\Omega)$ for some $\beta < 2$, then for any $\gamma > 1$ there exists a positive constant c_2 such that $u \geq c_2 d_\Omega^\gamma$ in Ω.

Proof. Since $\lambda m = -\lambda(-m)$ it is enough to consider the case when $\lambda > 0$. Notice that, for $k > 0$, the equation $\mathcal{L}_0 u + bu = \lambda mu$ can be written as $\mathcal{L}_0 u + (b + \lambda k) u = \lambda (m + k) u$ and that $b + \lambda k$ satisfies the condition on b assumed in the statements of the lemma. Therefore, by taking k positive and large enough, we can assume that $m \geq 1$.

We first prove $i)$ and $ii)$. For $\delta > 0$ such that $\Omega_\delta \neq \varnothing$ let $b_\delta := b\chi_{\Omega_\delta}$. Then $0 \leq b_\delta \in L^\infty(\Omega)$ and, in weak sense, $\mathcal{L}_0 u + b_\delta u \leq \mathcal{L}_0 u + bu = \lambda mu$ in Ω. Thus

$$0 < u \leq (\mathcal{L}_0 + b_\delta)^{-1}(\lambda mu) \text{ in } \Omega. \tag{4.38}$$

If $2^* = \infty$ (i.e., if $n = 1, 2$) then $(\mathcal{L}_0 + b_\delta)^{-1}(\lambda mu) \in L^r(\Omega)$ for any $r \in [1, \infty)$ (because $\lambda mu \in L^2(\Omega)$) and thus, by (4.38), $u \in L^r(\Omega)$ for any $r \in [1, \infty)$. In particular, $\lambda mu \in L^r(\Omega)$ for some $r > n$ which implies $(\mathcal{L}_0 + b_\delta)^{-1}(\lambda mu) \in C^1\left(\overline{\Omega}\right)$. Then, by (4.38), u is continuous at $\partial\Omega$ and, since by Proposition 3.1 $i)$, $u \in C(\Omega)$ we conclude that $u \in C\left(\overline{\Omega}\right)$. Also, since $(\mathcal{L}_0 + b_\delta)^{-1}(\lambda mu) \in C^1\left(\overline{\Omega}\right)$ and $(\mathcal{L}_0 + b_\delta)^{-1}(\lambda mu) = 0$ on $\partial\Omega$, there exists a positive constant c such that $(\mathcal{L}_0 + b_\delta)^{-1}(\lambda mu) \leq c d_\Omega$ in Ω, and then, by (4.38), $u \leq c d_\Omega$ in Ω.

In the case when $2^* < \infty$, since $u \in H_0^1(\Omega)$ we have $u \in L^{2^*}(\Omega)$. Thus $\lambda mu \in L^{2^*}(\Omega)$ and then $(\mathcal{L}_0 + b_\delta)^{-1}(\lambda mu) \in L^{2^{**}}(\Omega)$ (when $2^{**} < \infty$) and thus, from (4.38), $u \in L^{2^{**}}(\Omega)$ and so $\lambda mu \in L^{2^{**}}(\Omega)$. By iterating this procedure, we get that $\lambda mu \in L^r(\Omega)$ for some $r > n$. Then $(\mathcal{L}_0 + b_\delta)^{-1}(\lambda mu) \in C^1\left(\overline{\Omega}\right)$ and thus, as above, we get that $u \in C\left(\overline{\Omega}\right)$ and that there exists a positive constant c such that $u \leq c d_\Omega$ in Ω. Thus $i)$ and $ii)$ hold.

To prove $iii)$, assume that $d_\Omega^\beta b \in L^\infty(\Omega)$ for some $\beta < 2$. Notice that if $\gamma > r$ then (since Ω is bounded) there exists a constant $c_{r,s}$ such that $d_\Omega^\gamma \leq c_{s,r} d_\Omega^r$ in Ω. Therefore it is enough to prove $iii)$ when $1 < \gamma < 2$. Consider the solution $\psi \in \cap_{1 \leq q < \infty} W^{2,q}(\Omega) \cap W_0^{1,q}(\Omega)$ of the problem

$$\left\{ \begin{aligned} \mathcal{L}_0 \psi =& 1 \text{ in } \Omega, \\ \psi =& 0 \text{ on } \partial\Omega. \end{aligned} \right.$$

The regularity of ψ and the Hopf's boundary lemma give that there exist $\delta > 0$ and a constant $c_3 > 0$ such that

$$\langle A\nabla\psi, \nabla\psi \rangle \geq c_3^2 \text{ in } A_\delta. \tag{4.39}$$

From this fact, the strong maximum principle and the fact that $\psi \in C^1\left(\overline{\Omega}\right)$, it follows that, for some positive constants c_4 and c_5,

$$c_4 d_\Omega < \psi \leq c_5 d_\Omega \text{ in } \Omega. \tag{4.40}$$

Let $c_6 \in (0, \infty)$ be such that $d_\Omega^\beta b < c_6$ in Ω. A computation shows that

$$\mathcal{L}_0 \left(\psi^\gamma\right) + b\psi^\gamma = \gamma \psi^{\gamma-1} - \gamma\left(\gamma - 1\right) \psi^{\gamma-2} \left\langle A\nabla\psi, \nabla\psi \right\rangle + b\psi^\gamma \text{ in } \Omega,$$

and so, for δ as above,

$$\begin{aligned} \mathcal{L}_0 \left(\psi^\gamma\right) + b\psi^\gamma &\leq \gamma c_5^{\gamma-1} d_\Omega^{\gamma-1} - \gamma\left(\gamma-1\right) c_3^{\gamma-2} c_5^2 d_\Omega^{\gamma-2} + c_6 c_5^\gamma d_\Omega^{-\beta+\gamma} \\ &= d_\Omega^{\gamma-2} \left(-\gamma\left(\gamma-1\right) c_3^{\gamma-2} c_5^2 + \gamma c_3^{\gamma-1} d_\Omega + c_6 c_5^\gamma d_\Omega^{-\beta+2}\right) \end{aligned}$$

and thus, by diminishing δ if necessary,

$$\mathcal{L}_0 \left(\psi^\gamma\right) + b\psi^\gamma \leq 0 \text{ in } A_\delta.$$

Then, for any $\varepsilon > 0$,

$$\{\mathcal{L}_0 \left(u - \varepsilon\psi^\gamma\right) + b\left(u - \varepsilon\psi^\gamma\right) \geq 0 \text{ in } D'\left(A_\delta\right).$$

Let us show that, for ε small enough, $u - \varepsilon\psi^\gamma \geq 0$ on ∂A_δ. Indeed, clearly $u - \varepsilon\psi^\gamma = 0$ on $\partial\Omega$. Also, by Lemma 2.2 iii), there exists a positive constant c_7 such that

$$u \geq c_7 d_{\Omega_{\frac{\delta}{2}}} \text{ in } \Omega_{\frac{\delta}{2}}. \tag{4.41}$$

Thus, since $u \in C\left(\overline{\Omega}\right)$ we have

$$u \geq c_7 \frac{\delta}{2} \text{ in } \overline{\Omega_{\frac{\delta}{2}}}. \tag{4.42}$$

Then, by (4.42), (4.36) and (4.40), for ε small enough (perhaps depending on δ) we have

$$\begin{aligned} u - \varepsilon\psi^\gamma &\geq c_6 \frac{\delta}{2} - \varepsilon c_5^\gamma d_\Omega^\gamma \geq c_6 \frac{\delta}{2} - \varepsilon c_5^\gamma \delta^\gamma \\ &= \delta\left(\frac{c_6}{2} - \varepsilon c_5^\gamma \delta^{\gamma-1}\right) > 0 \text{ in } \left\{x \in \Omega : dist\left(x, \partial\Omega\right) = \delta\right\}. \end{aligned}$$

Then, by Remark 4.7,

$$u - \varepsilon\psi^\gamma \geq 0 \text{ in } A_\delta.$$

On the other hand, since $\psi \leq M := c_5 diam\left(\Omega\right)$ in Ω, by diminishing ε if necessary we have $u - \varepsilon\psi^\gamma \geq c_6 \frac{\delta}{2} - \varepsilon M^\gamma > 0$ in $\Omega_{\frac{\delta}{2}}$ and so $u - \varepsilon\psi^\gamma > 0$ in $\overline{\Omega_\delta}$). Then $u - \varepsilon\psi^\gamma \geq 0$ in Ω and the Proposition follows from (4.40). □

Let us to introduce some convenient notation. We set

$$\mathcal{B} := \left\{b : \Omega \to \mathbb{R} : d_\Omega^2 b \in L^\infty\left(\Omega\right)\right\}$$

and for $b \in \mathcal{B}$, we set $\|b\|_\mathcal{B} := \left\|d_\Omega^2 b\right\|_\infty$ and $\mathcal{B}^+ := \{b \in \mathcal{B} : b \geq 0\}$. Thus $(\mathcal{B}, \|b\|_\mathcal{B})$ is a Banach space and $\mathcal{B}^+$ is its positive cone. We set also $\mathcal{P} := \{m \in L^\infty\left(\Omega\right) : m^+ \not\equiv 0\}$.

For $m \in \mathcal{P}$ and $b \in \mathcal{B}^+$, we will write $\lambda_1\left(m, b\right)$ for the (unique) positive principal eigenvalue of problem (4.33), and we will denote by $\phi_{m,b}$ the (unique) associated positive principal eigenfunction, normalized by $\|\phi_{m,b}\|_2 = 1$.

Lemma 4.4. *Let $(m, b) \in \mathcal{P}\times\mathcal{B}^+$ and let $\{(m_j, b_j)\}_{j\in\mathbb{N}}$ be a sequence in $\mathcal{P}\times\mathcal{B}^+$ such that $\{(m_j, b_j)\}_{j\in\mathbb{N}}$ converges to (m, b) in $\mathcal{P}\times\mathcal{B}$ (with $\mathcal{P}$ endowed with the topology of the norm of $L^\infty\left(\Omega\right)$ and $\mathcal{B}^+$ endowed with the topology of the norm $\|.\|_\mathcal{B}$). Then:*

i) $\{\lambda_1\left(m_j, b_j\right)\}_{j\in\mathbb{N}}$ is bounded.

ii) $\left\{\phi_{m_j,b_j}\right\}_{j\in\mathbb{N}}$ is bounded in $H_0^1\left(\Omega\right)$.

Proof. To see i), consider an arbitrarily chosen function $z \in H_0^1(\Omega) \cap L^\infty(\Omega)$ such that $z > 0$ $a.e.$ in Ω. Since $\{b_j\}_{j\in\mathbb{N}}$ converges to b in $\mathcal{B}$, there exists a positive constant c such that $b_j \leq c d_\Omega^{-2}$ $a.e.$ in Ω for any $j \in \mathbb{N}$ and, by Lemma 2.1, $\int_\Omega d_\Omega^{-2} z^2 < \infty$. Then, for $j \in \mathbb{N}$,

$$\int_\Omega b_j z^2 \leq c'' \tag{4.43}$$

with c'' a positive constant independent of j. Also, taking into account that $\{m_j\}_{j\in\mathbb{N}}$ converges to m in $L^\infty(\Omega)$ and that $z^2 \in L^1(\Omega)$, the Lebesgue's dominated convergence gives $\lim_{j\to\infty} \int_\Omega m_j z^2 = \int_\Omega m z^2 > 0$. Then there exists a positive constant c''' such that, for any $j \in \mathbb{N}$,

$$\int_\Omega m_j z^2 \geq c''' \tag{4.44}$$

then i) follows from (4.43), (4.44) and from the fact that

$$\lambda_1(m_j, b_j) \leq \frac{\int_\Omega \left[|\nabla z|^2 + b_j z^2 \right]}{\int_\Omega m_j z^2}.$$

To prove ii), observe that

$$\int_\Omega \left|\nabla \phi_{m_j,b_j}\right|^2 = \lambda_1(m_j, b_j) \int_\Omega m_j \phi_{m_j,b_j}^2 - \int_\Omega b_j \phi_{m_j,b_j}^2 \leq \lambda_1(m_j, b_j) \int_\Omega m_j \phi_{m_j,b_j}^2,$$

and so, since $\{m_j\}_{j\in\mathbb{N}}$ is bounded in $L^\infty(\Omega)$, ii) follows from i). □

Theorem 4.5. *i) The map $(m, b) \to \lambda_1(m, b)$ is continuous from $\mathcal{P} \times \mathcal{B}_+$ into $\mathbb{R}$.*
ii) The map $(m, b) \to \phi_{m,b}$ is continuous from $\mathcal{P} \times \mathcal{B}_+$ into $H_0^1(\Omega)$.

Proof. To prove the lemma, it is enough to see that if $(m, b) \in \mathcal{P} \times \mathcal{B}_+$ and if $\{(m_j, b_j)\}_{j\in\mathbb{N}}$ is a sequence in $\mathcal{P} \times \mathcal{B}_+$ which converges to (m, b) in $\mathcal{P} \times \mathcal{B}$, then there exists a subsequence $\{(m_{j_k}, b_{j_k})\}_{k\in\mathbb{N}}$ such that $\lim_{k\to\infty} \lambda_1(m_{j_k}, b_{j_k}) = \lambda_1(m, b)$ and $\lim_{k\to\infty} \left\|\phi_{m_{j_k},b_{j_k}} - \phi_{m,b}\right\|_{H_0^1(\Omega)} = 0$. To do it, consider a pair $(m, b) \in \mathcal{P} \times \mathcal{B}_+$ and a sequence $\{(m_j, b_j)\}_{j\in\mathbb{N}} \subset \mathcal{P} \times \mathcal{B}_+$ such that $\lim_{j\to\infty}(m_j,_j) = (m, b)$ with convergence in $\mathcal{P} \times \mathcal{B}$. From Lemma 4.4 i) and ii), after pass to a subsequence if necessary (still denoted by $\{(m_j, b_j)\}_{j\in\mathbb{N}}$, we can assume that $\{\lambda_1(m_j, b_j)\}_{j\in\mathbb{N}}$ converges to some $\mu \in [0, \infty)$, and that there exists $\phi \in H_0^1(\Omega)$ such that $\left\{\phi_{m_j,b_j}\right\}_{j\in\mathbb{N}}$ converges to ϕ strongly in $L^2(\Omega)$ and $a.e.$ in Ω, and $\left\{\nabla\phi_{m_j,b_j}\right\}_{j\in\mathbb{N}}$ converges weakly to $\nabla\phi$ in $L^2(\Omega, \mathbb{R}^n)$. In particular, this implies $\|\phi\|_2 = 1$, and then ϕ is nonnegative (because each ϕ_{m_j,u_j} is positive) and nonidentically zero in Ω.

Let us show that $\left\{\phi_{m_j,b_j}\right\}_{j\in\mathbb{N}}$ converges to ϕ strongly in $H_0^1(\Omega)$. For $j, k \in \mathbb{N}$ we have, in weak sense,

$$\begin{aligned} \mathcal{L}_0\left(\phi_{m_j,b_j} - \phi_{m_k,b_k}\right) &= -\left(b_j \phi_{m_j,b_j} - b_k \phi_{m_k,b_k}\right) \\ &\quad + \lambda_1(m_j, b_j) m_j \phi_{m_j,b_j} - \lambda_1(m_k, b_k) m_k \phi_{m_k,b_k} \text{ in } \Omega, \\ \phi_{m_j,b_j} - \phi_{m_k,b_k} &= 0 \text{ on } \partial\Omega, \end{aligned} \tag{4.45}$$

and so, by taking $\phi_{m_j,b_j} - \phi_{m_k,b_k}$ as a test function in (4.45), we get

$$\int_\Omega \left\langle A\nabla\left(\phi_{m_j,b_j} - \phi_{m_k,b_k}\right), \left(\phi_{m_j,b_j} - \phi_{m_k,b_k}\right)\right\rangle = I_{j,k} + II_{j,k},$$

where

$$I_{j,k} := -\int_\Omega \left(b_j\phi_{m_j,b_j} - b_k\phi_{m_k,b_k}\right)\left(\phi_{m_j,b_j} - \phi_{m_k,b_k}\right),$$
$$II_{j,k} := \int_\Omega \left(\lambda_1(m_j,b_j)\,m_j\phi_{m_j,b_j} - \lambda_1(m_k,b_k)\,m_k\phi_{m_k,b_k}\right)\left(\phi_{m_j,b_j} - \phi_{m_k,b_k}\right).$$

Now, $b_j = \beta_j d_{\Omega j}^{-2}$ in Ω, with $\beta_j \in L^\infty(\Omega)$ such that, for some positive constant c and for all $j \in \mathbb{N}$, $\|\beta_j\|_\infty \le c$. Thus

$$\begin{aligned}(4.46)\qquad I_{j,k} &= -\int_\Omega (b_j - b_k)\,\phi_{m_j,b_j}\left(\phi_{m_j,b_j} - \phi_{m_k,b_k}\right) - \int_\Omega b_k\left(\phi_{m_j,b_j} - \phi_{m_k,b_k}\right)^2\\ &\le \int_\Omega \phi_{m_j,b_j}\,|b_j - b_k|\,\left|\phi_{m_j,b_j} - \phi_{m_k,b_k}\right|\\ &= \int_\Omega \frac{\phi_{m_j,b_j}}{d_\Omega} d_\Omega^2\,|b_j - b_k|\left|\frac{\phi_{m_j,b_j} - \phi_{m_k,b_k}}{d_\Omega}\right|\\ &= \int_\Omega \frac{\phi_{m_j,b_j}}{d_\Omega}\,|\beta_j - \beta_k|\left|\frac{\phi_{m_j,u_j} - \phi_{m_k,u_k}}{d_\Omega}\right|.\end{aligned}$$

Then, by the Hardy's inequality,

$$\begin{aligned}I_{j,k} &\le c\,\|\beta_j - \beta_k\|_\infty \left\|\frac{\phi_{m_j,b_j} - \phi_{m_k,b_k}}{d_\Omega}\right\|_2 \left\|\frac{\phi_{m_j,b_j}}{d_\Omega}\right\|_2\\ &\le c'\,\|\beta_j - \beta_k\|_\infty \left\|\phi_{m_j,b_j} - \phi_{m_k,b_k}\right\|_{H_0^1(\Omega)} \left\|\phi_{m_j,b_j}\right\|_{H_0^1(\Omega)}\\ &\le c''\varepsilon(j,k)\left\|\phi_{m_j,b_j} - \phi_{m_k,b_k}\right\|_{H_0^1(\Omega)},\end{aligned}$$

where $\varepsilon(j,k) := \|\beta_j - \beta_k\|_\infty$ and where c, c' and c'' are positive constants independent of j and k. Therefore

$$(4.47)\qquad I_{j,k} \le c''\varepsilon(j,k)\left\|\phi_{m_j,b_j} - \phi_{m_k,b_k}\right\|_{H_0^1(\Omega)}.$$

On the other hand,

$$\begin{aligned}(4.48)\qquad II_{j,k} &\le \int_\Omega \left|\left(\lambda_1(m_j,b_j) - \lambda_1(m_k,b_k)\right)m_j\phi_{m_j,b_j}\left(\phi_{m_j,b_j} - \phi_{m_k,b_k}\right)\right|\\ &+ \int_\Omega \left|\lambda_1(m_k,b_k)\,(m_j - m_k)\,\phi_{m_j,b_j}\left(\phi_{m_j,b_j} - \phi_{m_k,b_k}\right)\right|\\ &+ \int_\Omega \lambda_1(m_k,b_k)\,m_k\left(\phi_{m_j,b_j} - \phi_{m_k,b_k}\right)\left(\phi_{m_j,b_j} - \phi_{m_k,b_k}\right)\\ &\le c'\delta(j,k)\left\|\phi_{m_j,b_j} - \phi_{m_k,b_k}\right\|_{H_0^1(\Omega)},\end{aligned}$$

where c' is a positive constant independent of j and k and

$$\begin{aligned}\delta(j,k) &:= \left\|\left(\lambda_1(m_j,b_j) - \lambda_1(m_k,b_k)\right)m_j\phi_{m_j,b_j}\right\|_2\\ &+ \left\|\lambda_1(m_k,b_k)\,(m_j - m_k)\,\phi_{m_j,b_j}\right\|_2 + \left\|\lambda_1(m_k,b_k)\,m_k\left(\phi_{m_j,b_j} - \phi_{m_k,b_k}\right)\right\|_2.\end{aligned}$$

Now, $\lim_{j,k\to\infty}\left(\lambda_1(m_j,b_j) - \lambda_1(m_k,b_k)\right) = 0$, $\{m_j\}_{j\in\mathbb{N}}$ is bounded in $L^\infty(\Omega)$, and $\left\{\phi_{m_j,b_j}\right\}_{j\in\mathbb{N}}$ converges to ϕ in $L^2(\Omega)$. Then

$$\lim_{j,k\to\infty}\left\|\left(\lambda_1(m_j,b_j) - \lambda_1(m_k,b_k)\right)m_j\phi_{m_j,u_j}\right\|_2 = 0.$$

Also, $\{\lambda_1(m_k, b_k)\}_{k\in\mathbb{N}}$ is bounded, $\lim_{j\to\infty} m_j = m$ with convergence in $L^\infty(\Omega)$, and $\{\phi_{m_j,u_j}\}_{j\in\mathbb{N}}$ is bounded in $L^2(\Omega)$. Thus

$$\lim_{j,k\to\infty} \left\|\lambda_1(m_k, b_k)(m_j - m_k)\phi_{m_j,b_j}\right\|_2 = 0,$$

and, since $\{\lambda_1(m_k, b_k)\}_{k\in\mathbb{N}}$ and $\{m_k\}_{k\in\mathbb{N}}$ are bounded in $\mathbb{R}$ and $L^\infty(\Omega)$ respectively, and $\{\phi_{m_j,b_j}\}_{j\in\mathbb{N}}$ converges to ϕ in $L^2(\Omega)$, we have

$$\lim_{j,k\to\infty} \left\|\lambda_1(m_k, b_k)\, m_k\left(\phi_{m_j,b_j} - \phi_{m_k,b_k}\right)\right\|_2 = 0.$$

Then $\lim_{j,k\to\infty}\delta(j,k) = 0$ and, since $\{b_j\}_{j\in\mathbb{N}}$ converges to b in $\mathcal{B}$, we have also that $\lim_{j,k\to\infty}\varepsilon(j,k) = 0$. Now,

$$\begin{aligned}
&\left\|\phi_{m_j,b_j} - \phi_{m_k,b_k}\right\|^2_{H_0^1(\Omega)}\\
&= I_{j,k} + II_{j,k}\\
&\leq c\varepsilon_{j,k}\left\|\phi_{m_j,b_j} - \phi_{m_k,b_k}\right\|_{H_0^1(\Omega)} + c'\delta_{j,k}\left\|\phi_{m_j,b_j} - \phi_{m_k,b_k}\right\|_{H_0^1(\Omega)}
\end{aligned}$$

and so

$$\lim_{j,k\to\infty}\left\|\phi_{m_j,b_j} - \phi_{m_k,b_k}\right\|_{H_0^1(\Omega)} = 0.$$

Thus $\{\phi_{m_j,b_j}\}_{j\in\mathbb{N}}$ converges in $H_0^1(\Omega)$ to some $\widetilde{\phi}$. Since ϕ_{m_j,b_j} converges $a.e.$ in Ω to ϕ, we conclude that $\widetilde{\phi} = \phi$. Therefore,

$$\{\phi_{m_j,b_j}\}_{j\in\mathbb{N}} \text{ converges to } \phi \text{ in } H_0^1(\Omega). \tag{4.49}$$

To complete the proof of the lemma, it only remains to see that $\mu = \lambda_1(m, b)$ and $\phi = \phi_{m,b}$. For $\varphi \in H_0^1(\Omega)$ and $j \in \mathbb{N}$, we have

$$\int_\Omega \left(\left\langle A\nabla\phi_{m_j,b_j}, \nabla\varphi\right\rangle + b_j\phi_{m_j,b_j}\varphi\right) = \lambda_1(m_j, b_j)\int_\Omega m_j\phi_{m_j,b_j}\varphi, \tag{4.50}$$

and, by (4.49), $\lim_{j\to\infty}\int_\Omega\left\langle\nabla\phi_{m_j,b_j}, \nabla\varphi\right\rangle = \int_\Omega\langle\nabla\phi, \nabla\varphi\rangle$. Also, $b_j\phi_{m_j,b_j}\varphi$ converges to $b\phi\varphi$ $a.e.$ in Ω and, by Lemma 4.4 i), we have

$$|b_j\phi\varphi| \leq cd_\Omega^{-2}\phi\,|\varphi|$$

with c a positive constant independent of j and, by Lemma 2.1, $d_\Omega^{-2}\phi\,|\varphi| \in L^1(\Omega)$. Thus, by the Lebesgue's dominated convergence theorem,

$$\lim_{j\to\infty}\int_\Omega b_j\phi_{m_j,b_j}\varphi = \int_\Omega b\phi\varphi.$$

Also, since $\lim_{j\to\infty}\lambda_1(m_j, b_j) = \mu$, $\lim_{j\to\infty} m_j = m$ with convergence in $L^\infty(\Omega)$, and $\lim_{j\to\infty}\phi_{m_j,b_j} = \phi$ with convergence in $H_0^1(\Omega)$, we have

$$\lim_{j\to\infty}\lambda_1(m_j, b_j)\int_\Omega m_j\phi_{m_j,b_j}\varphi = \mu\int_\Omega m\phi\varphi.$$

Then, from (4.50),

$$\int_\Omega\left(\langle A\nabla\phi, \nabla\varphi\rangle + b\phi\varphi\right) = \mu\int_\Omega m\phi\varphi$$

and so $\mu = \lambda_1(m, b)$ and $\phi = \phi_{m,b}$. □

References

[1] A. Beltramo, P. Hess: *On the principal eigenvalue of a periodic-parabolic operator*, Comm. Partial Differential Equations **9** (9) (1984), 919–941.

[2] H. Berestycki, S. R. S. Varadhan and L. Nirenberg: *The principal eigenvalue and maximum principle for second-order elliptic operators in general domains*, Comm. Pure Appl. Math. **47** (1) (1994), 47–92.

[3] H. Brezis: *Functional Analysis, Sobolev Spaces and Partial Differential Equations*, Springer, 2011.

[4] K. Brown, S. Lin: *On the existence of positive eigenfunctions for an eigenvalue problem with indefinite weight function*, J. Math. Anal. Appl., **75** (1980), 112–120.

[5] D. G. De Figueiredo: *Positive solutions of semilinear elliptic equations*, Lect. Notes Math., Springer, **957** (1982), 34–87.

[6] J. Fleckinger, J. Hernandez and F. de Thelin: *Existence of multiple eigenvalues for some indefinite linear eigenvalue problems*, Bolletino U.M.I., **7** (2004), 159-188.

[7] D. Gilbarg, N. S. Trudinger: *Elliptic Partial Differential Equations of Second Order,* Springer-Verlag, Berlin Heidelberg New York, 2001.

[8] M. Ghergu, V. D. Rădulescu: *Singular Elliptic Problems: Bifurcation and Asymptotic Analysis*, Oxford Lecture Series in Mathematics and Its Applications, Oxford University Press, No 37, 2008.

[9] T. Godoy, A. Guerin: *Regularity of the lower positive branch for singular elliptic bifurcqation problems*, Electron. J. Differential Equations, **2019** (49) (2019), 1–32.

[10] J. Hernández, F. J. Mancebo and J. M. Vega: *On the linearization of some singular nonlinear elliptic problems and applications*, Ann. Inst. H. Poincaré C Anal. Non Linéaire, **19** (6) (2002), 777–813.

[11] P. Hess: *On positive solutions of semilinear periodic-parabolic problems*, Infinite-dimensional systems (Retzhof, 1983), 101–114, Lecture Notes in Math. **1076**, Springer, Berlin, 1984.

[12] P. Hess: *Periodic parabolic problems and positivity*, Pitman Research Notes, 1991.

[13] P. Hess, T. Kato: *On some linear and nonlinear eigenvalue problems with an indefinite weight function*, Comm. Partial Differential Equations, **5** (1980), 999–1030.

[14] J. Lopez-Gomez: *The maximum principle and the existence of principal eigenvalues for some linear weighted boundary value problems*, J. Differential Equations **127** (1) (1996), 263–294.

[15] A. Manes, A.M. Micheletti: *Un'estensione della teoria variazionale classica degli autovalori per operatori ellitici del secondo ordine*, Bollettino U.M.I., **7** (1973), 285–301.

[16] N. S. Papageorgiou, V. D. Rădulescu and D. D. Repovš: *Nonlinear Analysis – Theory and Methods*, Springer Monographs in Mathematics, Springer Nature Switzerland, 2019.

[17] J. Sabina de Lis: *Hopf maximum principle revisited*, Electron. J. Differential Equations, **2015** (115) (2015), 1–9.

[18] S. Senn, P. Hess: *On positive solutions of a linear elliptic eigenvalue problem with Neumann boundary conditions*, Math. Ann., **258** (1982), 459–470.

Tomas Godoy
Universidad Nacional de Córdoba
Facultad de Matemática, Astronomía, Física y Compùtación
Av. Medina Allende s/n , Ciudad Universitaria, CP:X5000HUA Córdoba, Argentina
ORCID: 0000-0002-8804-9137
E-mail address: godoy@famaf.unc.edu.ar

CONSTRUCTIVE MATHEMATICAL ANALYSIS
6 (2023), No. 2, pp. 128-141
`http://dergipark.org.tr/en/pub/cma`
ISSN 2651 - 2939

Research Article

Beyond Descartes' rule of signs

VLADIMIR PETROV KOSTOV*

ABSTRACT. We consider real univariate polynomials with all roots real. Such a polynomial with c sign changes and p sign preservations in the sequence of its coefficients has c positive and p negative roots counted with multiplicity. Suppose that all moduli of roots are distinct; we consider them as ordered on the positive half-axis. We ask the question: If the positions of the sign changes are known, what can the positions of the moduli of negative roots be? We prove several new results which show how far from trivial the answer to this question is.

Keywords: Real polynomial in one variable, hyperbolic polynomial, sign pattern, Descartes' rule of signs.

2020 Mathematics Subject Classification: 26C10.

1. INTRODUCTION

In the present paper, we study a problem related to a generalization of Descartes' rule of signs formulated in [5]. About this rule see [1], [2], [3], [4], [7], [9], [10], [16] or [17]. For its tropical analog see [6]. A related problem concerning polynomials in one variable is considered in [15]. A degree d real polynomial $Q := \sum_{j=0}^{d} a_j x^j$ is *hyperbolic* if all its roots are real. Suppose that all coefficients a_j are non-zero. For such a polynomial, Descartes' rule of signs implies that it has c positive and p negative roots (counted with multiplicity, so $c + p = d$), where c is the number of sign changes and p the number of sign preservations in the sequence of coefficients of Q. The signs of these coefficients define the *sign pattern* $(\operatorname{sgn}(a_d), \operatorname{sgn}(a_{d-1}), \ldots, \operatorname{sgn}(a_0))$. We deal mainly with monic polynomials in which case sign patterns begin with a $+$. In this case, we can use instead of and equivalently to a sign pattern the corresponding *change-preservation pattern* which is a d-vector and (by some abuse of notation) whose jth component equals c if $a_{d-j+1}a_{d-j} < 0$ and p if $a_{d-j+1}a_{d-j} > 0$. One can consider also the moduli of the roots of a hyperbolic polynomial defining a given sign pattern. We study the generic case when all moduli are distinct. A natural question to ask is:

Question 1.1. *When these moduli are ordered on the real positive half-axis, at which positions can the moduli of the negative roots be?*

Descartes' rule of signs provides no hint for the answer to this question. In the present paper, we recall known and we introduce new results in this direction which show how far from trivial the situation is.

Notation 1.1. *(1) We denote by $0 < \alpha_1 < \cdots < \alpha_c$ the positive and by $0 < \gamma_1 < \cdots < \gamma_p$ the moduli of the negative roots of a hyperbolic polynomial. We explain the notation of the order of*

Received: 17.02.2023; Accepted: 02.06.2023; Published Online: 06.06.2023

*Corresponding author: Vladimir Petrov Kostov; vladimir.kostov@unice.fr

DOI: 10.33205/cma.1252639

these moduli on the positive half-axis by an example. Suppose that $d=6$, $c=2$, $p=4$ *and*

$$\alpha_1 < \gamma_1 < \gamma_2 < \alpha_2 < \gamma_3 < \gamma_4 \ .$$

Then for the order of moduli we write $PNNPNN$, *i. e. the letters* P *and* N *denote the relative positions of the moduli of the positive and negative roots.*

(2) *A sign pattern beginning with* i_1 *signs* $+$ *followed by* i_2 *signs* $-$ *followed by* i_3 *signs* $+$ *etc. is denoted by* $\Sigma_{i_1,i_2,i_3,\ldots}$.

In what follows, we consider for each given degree d couples of the form (change-preservation pattern, order of moduli) (called *couples* for short). Such a couple is *compatible* with Descartes' rule of signs if the number of components c (resp. p) of the change-preservation pattern is equal to the number of components P (resp. N) of the order of moduli. A couple is called *realizable* if there exists a polynomial defining the change-preservation pattern of the couple and whose moduli of roots define the given order.

Remark 1.1. *For fixed* d *and* c, *there are* $\binom{d}{c}$ *change-preservation patterns and* $\binom{d}{c}$ *orders of moduli hence* $\binom{d}{c}^2$ *compatible couples. Thus for a given degree* d, *the total number of compatible couples is*

$$\chi(d) := \sum_{c=0}^{d} \binom{d}{c}^2 = \sum_{c=0}^{d} \binom{d}{c}\binom{d}{d-c} = \binom{2d}{d} \ . \tag{1.1}$$

This is the coefficient of x^d *in the polynomial* $(x+1)^d(x+1)^d = (x+1)^{2d}$. *Using Stirling's formula* $n! \sim \sqrt{2\pi n}(n/e)^n$, *one concludes that* $\chi(d) \sim 2^{2d}/\sqrt{\pi d}$.

Example 1.1. (1) *For* $d=1$, *the only compatible couples are* $(c,\ P)$ *and* $(p,\ N)$. *They are realizable respectively by the polynomials* $x-1$ *and* $x+1$.

(2) *For* $d=2$, *there are* $\binom{4}{2}=6$ *compatible couples. Out of these, the couples* $(cp,\ PN)$ *and* $(pc,\ NP)$ *are not realizable. Indeed, for a hyperbolic polynomial* x^2-ux-v *(resp.* x^2+ux-v*),* $u>0$, $v>0$, *one has the order of moduli* NP *(resp.* PN*). The remaining* 4 *couples are realizable. To see this one can consider the family of polynomials* $x^2+a_1x+a_0$. *In the plane of the variables* (a_1,a_0) *the domain of hyperbolic polynomials is the one below the parabola* $\mathcal{P} : a_0 = a_1^2/4$. *We list the realizable couples and the open domains in which they are realizable:*

$$(cc,\ PP) \quad \{a_1<0,\ 0<a_0<a_1^2/4\}\ , \qquad (pp,\ NN) \quad \{a_1>0,\ 0<a_0<a_1^2/4\}\ ,$$

$$(cp,\ NP) \quad \{a_1<0,\ a_0<0\}\ , \qquad (pc,\ PN) \quad \{a_1>0,\ a_0<0\}\ .$$

We can make Question 1.1 more precise:

Question 1.2. *For a given degree* d, *which compatible couples are realizable?*

The above example answers this question for $d=1$ and 2. For $d=3$, 4 and 5, the exhaustive answer is given in Section 3.

Remark 1.2. *There exist two commuting involutions acting on the set of degree* d *polynomials with non-vanishing coefficients. These are*

$$i_m \ : \ Q(x) \mapsto (-1)^d Q(-x) \quad \text{and} \quad i_r \ : \ Q(x) \mapsto x^d Q(1/x)/Q(0)\ .$$

The role of the factors $(-1)^d$ *and* $1/Q(0)$ *is to preserve the set of monic polynomials. When acting on a couple, the involution* i_m *changes the components* c *to* p, P *to* N *and vice versa while the involution* i_r *reads the vectors of a given couple from the right. A given couple is realizable or not simultaneously with all other couples from its orbit under the action of* i_m *and* i_r. *An orbit consists of four or two couples.*

Notation 1.2. *For a sign pattern σ, we denote by $k^*(\sigma)$ the number of orders of moduli with which σ is realizable. For an order of moduli Ω, we denote by $l_*(\Omega)$ the number of sign patterns realizable with Ω. For a given d, we denote by $\tilde{r}^*(d)$ the ratio between the numbers of realizable and of all compatible couples.*

Example 1.2. *(1) For the sign pattern $\Sigma_{3,3,1}$ one has $k^*(\Sigma_{3,3,1}) = 6$. Indeed, consider the polynomial*

$$\begin{aligned}&(x-1)(x+1)^4(x-b)\\ =&x^6+(3-b)x^5+(2-3b)x^4+(-2b-2)x^3+(2b-3)x^2+(3b-1)x+b\,.\end{aligned}$$

For $b > 0$ sufficiently small, it defines the sign pattern $\Sigma_{3,3,1}$. One can perturb its 4-fold root at -1 to obtain polynomials with the same sign pattern and with exactly k moduli of negative roots which are > 1 and $4 - k$ moduli which are < 1, where $k = 0, 1, \ldots, 4$; these moduli are close to 1. On the other hand, the only other realizable order with this sign pattern is

$$\gamma_1 < \alpha_1 < \alpha_2 < \gamma_2 < \gamma_3 < \gamma_4\ , \quad \text{i. e.} \quad NPPNNN\ ,$$

see [11, Theorems 3 and 4], *which makes a total of* 6 *orders of moduli realizable with $\Sigma_{3,3,1}$.*

(2) For $m \geq 1$, $n \geq 1$, one has $k^*(\Sigma_{m,n}) = 2\min(m,n) - 1$, see [11, Theorem 1 and Corollary 1].

Our first result is the following theorem:

Theorem 1.1. *(1) For $d \geq 1$, the only orders realizable with all compatible change-preservation patterns are $PP\ldots P$ and $NN\ldots N$. The corresponding change-preservation patterns are $cc\ldots c$ and $pp\ldots p$.*

(2) For any $d \geq 1$, there exist sign patterns realizable with all compatible orders. For $d \geq 5$, there exist sign patterns with $c = 2$ which are realizable with all $\binom{d}{2}$ compatible orders.

(3) There exists no sign pattern σ such that $k^*(\sigma) = 2$.

(4) The only sign patterns σ with $k^*(\sigma) = 3$ are the ones of the form $\Sigma_{2,d-1}$, $i_r(\Sigma_{2,d-1})$, $i_m(\Sigma_{2,d-1})$ and $i_r i_m(\Sigma_{2,d-1})$.

(5) For any $\ell \in \mathbb{N}^*$, there exist a degree d and an order Ω such that $l_*(\Omega) = \ell$.

The theorem is proved in Section 4. In Section 2, we recall some notions and known results and we continue the formulation of the new ones. In particular, for each of the 6 classes of non-realizable couples introduced in Section 2, we compare the number of couples which it contains with the number of all compatible couples, see (1.1). In all 6 cases, the limit of their ratio as $d \to \infty$ is 0 (see part (2) of Remarks 2.3, part (2) of Remarks 2.4, Remark 2.5, Remark 2.6, Remark 2.7 and part (4) of Theorem 2.3). On the other hand, when considering the cases $d = 3$, 4 and 5 in Section 3, we arrive to the conclusion that it is plausible to have $\lim_{d\to\infty} \tilde{r}^*(d) = 0$ (see Notation 1.2). This however cannot be explained by the presence of the 6 classes of non-realizable couples, so for the moment it is not evident what the exhaustive answer to Question 1.2 should be.

We finish this section by a result of geometric nature. Consider the space of coefficients $Oa_{d-1}\cdots a_0 \cong \mathbb{R}^d$. The *hyperbolicity domain* is the set of values of $(a_{d-1}, \ldots, a_0)$ for which the corresponding monic polynomial Q is hyperbolic. The resultant $R := \text{Res}(Q(x), (-1)^d Q(-x), x)$ vanishes exactly when Q has two opposite roots or a root at 0. When the coefficients a_j are real, the polynomials $Q(x)$ and $Q(-x)$ have a root in common either when $Q(0) = 0$ or when Q has two opposite real non-zero roots or when Q has a pair of purely imaginary roots.

Example 1.3. *For $d = 1$, 2 and 3, one obtains $R = -2a_0$, $R = 4a_0a_1^2$ and $R = -8a_0(a_2a_1 - a_0)^2$, respectively.*

We denote by [.] the integer part and we set

$$Q^1 := x^{[d/2]} + a_{d-2}x^{[d/2]-1} + a_{d-4}x^{[d/2]-2} + \cdots \ ,$$

$$Q^2 := a_{d-1}x^{[(d-1)/2]} + a_{d-3}x^{[(d-1)/2]-1} + a_{d-5}x^{[(d-1)/2]-2} + \cdots \quad \text{and}$$

$$R_0 := \mathrm{Res}(Q^1(x), Q^2(x), x)) \ .$$

Theorem 1.2. *(1) One has $R = (-1)^{[d/2]+1}2^{d-[(d+1)/2]+1}a_0R_0^2$.*
(2) The quantity R_0 is an irreducible polynomial in the variables a_j.

The theorem is proved in Section 5. Properties of the set $\{R_0 = 0\}$ and its pictures for $d \leq 4$ can be found in [8].

2. Canonical sign patterns, rigid orders of moduli and further results

Definition 2.1. *For a given change-preservation pattern, the corresponding canonical order is obtained by reading the pattern from the right and by replacing each component c (resp. p) by P (resp. by N). E. g., the canonical order corresponding to the pattern $ccpcp$ is $NPNPP$. This definition allows to define the canonical order corresponding to each given sign pattern beginning with $+$.*

Each sign or change-preservation pattern is realizable with its canonical order, see [12, Proposition 1].

Definition 2.2. *(1) A sign pattern (or equivalently a change-preservation pattern) realizable only with its corresponding canonical order is called* canonical.
(2) If all monic hyperbolic polynomials having a given order of moduli define one and the same sign pattern, then the order is called rigid.

Remark 2.3. *(1) It is shown in [13] that canonical are exactly these sign patterns which have no four consecutive signs equal to*

$$(+,+,-,-,) \ , \quad (-,-,+,+) \ , \quad (+,-,-,+) \quad \text{or} \quad (-,+,+,-) \ .$$

Hence canonical are these change-preservation patterns having no isolated sign changes and no isolated sign preservations, i. e. having no three consecutive components cpc or pcp.
(2) In the proof of Proposition 10 in [13], the set of all *canonical change-preservation patterns is represented as union of four subsets, namely of patterns beginning with a single p or c, patterns ending by a single p or c, patterns both beginning and ending by a single p or c and patterns whose two first letters are equal and whose last two letters are also equal. For $d \geq 100$, the number of patterns in each of these sets can be majorized by $2 \cdot [d/2] \cdot 2^{d-[0.26d]-1}$. Hence the number of* all *canonical sign-preservation patterns is $\leq \tau(d) := 8 \cdot [d/2] \cdot 2^{d-[0.26d]-1}$ and for large d, the number of all non-realizable couples with canonical sign-preservation patterns is*

$$\leq \tau(d) \sum_{c=0}^{d} \binom{d}{c} = 8 \cdot [d/2] \cdot 2^{2d-[0.26d]-1} < 2^{2d}/\sqrt{\pi d} \sim \chi(d) \ ,$$

see Remark 1.1; we majorize one of the factors $\binom{d}{c}$ in (1.1) by $\tau(d)$.

Remark 2.4. *(1) It is proved in [14] that rigid are the orders of moduli $PP\ldots P$, $NN\ldots N$ (defining the change-preservation patterns $cc\ldots c$ and $pp\ldots p$, the two corresponding couples are realizable by any polynomials having distinct positive or distinct negative roots) and also*

$$P_N := PNPNPN\ldots \ , \qquad N_P := NPNPNP\ldots \ . \tag{2.2}$$

Each of the latter two orders (we call them standard*) defines, depending on the parity of d, one of the sign patterns*

$$\sigma_+ := (+,+,-,-,+,+,-,-,\ldots) \quad \text{or} \quad \sigma_- := (+,-,-,+,+,-,-,+,+,\ldots)\,. \tag{2.3}$$

(2) *For each fixed degree d, there are $\binom{d}{[d/2]}$ compatible couples with the order P_N and $\binom{d}{[d/2]}$ with the order N_P, see (2.2). Hence there are $2\binom{d}{[d/2]} - 2$ compatible couples in which the order of moduli is rigid (more exactly standard) and which are not realizable, and one has $\lim_{d\to\infty}(2\binom{d}{[d/2]} - 2)/\chi(d) = 0$, see (1.1) and use Stirling's formula.*

Definition 2.3. *We call superposition of two standard orders of moduli Ω_1 and Ω_2 any order obtained as follows. One inserts the components of Ω_2 at any places between the components of Ω_1 or in front of the first or after the last component of Ω_1 by preserving their relative order. Example: the order*

$$P\bar{N}NP\bar{P}N\bar{N}\bar{P}\bar{N} \quad \text{is superposition of} \quad PNPN \quad \text{and} \quad NPNPN$$

(we overline in this superposition the moduli coming from Ω_2; in this example there is more than one way to attribute the moduli of roots in the superposition as coming from Ω_1 or Ω_2; the superposition of two standard orders is not uniquely defined).

The following proposition explains how one can obtain new examples of non-realizable couples on the basis of standard orders.

Proposition 2.1. *Each superposition of two standard orders is realizable only with sign patterns of the form*

$$(+,+,?,-,?,+,?,-,\ldots)\,, \quad (+,?,-,?,+,?,-,\ldots) \quad \text{or} \quad (+,-,?,+,?,-,?,+,\ldots)$$

which are the "products" of sign patterns $\sigma_+\sigma_+$, $\sigma_+\sigma_-$ and $\sigma_-\sigma_-$.

Proof. Indeed, suppose that in the superposition of standard orders, the roots coming from the order Ω_i are roots of a polynomial T_i, $i = 1, 2$. Then in the product T_1T_2 every second coefficient, the leading coefficient and the constant term are sums of products of a coefficient of T_1 and a coefficient of T_2 either all with opposite or all with same signs, so the corresponding components of the "products" of sign patterns are well-defined. □

Remark 2.5. *The number of letters N in a standard order is equal to the number of letters P or differs from the latter by 1. Hence in the superposition of two standard orders the modulus of this difference is majorized by 2. Besides, not more than $[d/2]$ of the signs of coefficients are not determined by the order of moduli, so the number of non-realizable couples corresponding to superpositions of standard orders is less than*

$$2\left(\binom{d}{[d/2]} + \binom{d}{[d/2]-1} + \binom{d}{[d/2]-2}\right)\cdot 2^{[d/2]} < 6\binom{d}{[d/2]}\cdot 2^{(d+1)/2}$$

which is $\sim 12\cdot 2^{3d/2}/\sqrt{\pi d}$ (we use Stirling's formula here). At the same time $\chi(d) \sim 2^{2d}/\sqrt{\pi d}$ (see Remark 1.1).

There exist other situations in which the order of moduli defines the signs of part of the coefficients of the polynomial.

Example 2.4. *Consider for $d = 8k+2$, $k \in \mathbb{N}^*$, and for $c = 2$ the order of moduli*

$$\Omega \;:\; \gamma_1 < \cdots < \gamma_{4k} < \alpha_1 < \alpha_2 < \gamma_{4k+1} < \cdots < \gamma_{8k}\,.$$

It is realizable only with sign patterns having two sign changes. Denote by U_1 and U_2 monic hyperbolic degree $4k+1$ polynomials with roots

$$-\gamma_1\,, \quad -\gamma_2\,,\ldots\,, \quad -\gamma_{2k}\,, \quad -\gamma_{4k+1}\,, \quad -\gamma_{4k+2}\,,\;\ldots\;, \quad -\gamma_{6k}\,, \quad \alpha_1$$

and

$$-\gamma_{2k+1}\ ,\quad -\gamma_{2k+2}\ ,\ \ldots\ ,\quad -\gamma_{4k}\ ,\quad -\gamma_{6k+1}\ ,\quad -\gamma_{6k+2}\ ,\ \ldots\ ,\quad -\gamma_{8k}\ ,\quad \alpha_2$$

respectively. Hence they define sign patterns of the form Σ_{m_i,n_i}, $i = 1, 2$. *According to* [11, Theorem 1], *if* $n_i < m_i$, *then the polynomial* U_i *has* $\leq 2n_i - 2$ *moduli of negative roots which are* $\leq \alpha_i$*; if* $n_i > m_i$, *then it has* $\leq 2m_i - 2$ *moduli of negative roots which are* $\geq \alpha_i$. *Hence one has* $n_i \geq k+1$ *and* $m_i \geq k+1$. *This implies that the first* $k+1$ *and the last* $k+1$ *coefficients of the product* U_1U_2 *are positive, i. e. the order of moduli* Ω *is not realizable with sign patterns* Σ_{j_1,j_2,j_3} *which do not satisfy the conditions* $j_1 \geq k+1$ *and* $j_3 \geq k+1$.

Remark 2.6. *There are* $\binom{d}{2}^2$ *compatible couples with* $c = 2$ *hence less than* $\binom{d}{2}^2$ *non-realizable couples concerned by Example 2.4. Using the involution* i_m *(see Remark 1.2), one can give as many such examples with* $c = d-2$. *One has* $\lim_{d\to\infty}\binom{d}{2}^2/\chi(d) = 0$, *see (1.1).*

The proposition and theorem that follow describe other situations in which certain compatible couples are not realizable.

Proposition 2.2. *Suppose that* d *is even, that the leading monomial and the constant term are positive (hence* c *is even), that all coefficients of odd powers are negative and that* $c < d$. *Then there is no modulus of a negative root in any of the intervals* $(0, \alpha_1)$, (α_2, α_3), $\ldots$, $(\alpha_{c-2}, \alpha_{c-1})$, (α_c, ∞).

Proof. Indeed, for a monic hyperbolic polynomial Q satisfying these conditions one has $Q(t) > 0$, if t belongs to any of the mentioned intervals. As all odd monomials are with negative coefficients, one has also $Q(-t) > Q(t)$ from which the proposition follows. □

Remark 2.7. For d even, the number of sign patterns as defined in Proposition 2.2 is $\leq 2^{d/2}$ (half of the signs of coefficients are fixed), so if d is large, then the number of such non-realizable couples is

$$\leq 2^{d/2}\sum_{c=0}^{d}\binom{d}{c} = 2^{3d/2} < \chi(d) \sim 2^{2d}/\sqrt{\pi d}\,,$$

see Remark 1.1.

Theorem 2.3. *(1) Suppose that*

$$c \leq p \quad \text{and} \quad \alpha_c < \gamma_p, \quad \alpha_{c-1} < \gamma_{p-1}\ ,\ \ldots\ ,\quad \alpha_1 < \gamma_{p-c+1}\ . \tag{2.4}$$

Then $a_{d-1} > 0$. *Hence a couple with* $a_{d-1} < 0$ *and order satisfying conditions (2.4) is not realizable.*

(2) For fixed d*, the number of orders of moduli satisfying conditions (2.4) is*

$$T_d^c := \binom{d}{c} - C_0\binom{d-1}{c-1} - C_1\binom{d-3}{c-2} - C_2\binom{d-5}{c-3} - C_3\binom{d-7}{c-4} - \cdots\ , \tag{2.5}$$

where $C_k := \binom{2k}{k}/(k+1)$ *is the* $k-$*th Catalan number.*

(3) One has

$$T_d^c = \binom{d}{c}\left(1 - \frac{c}{d-c+1}\right) = \binom{d}{c}\frac{d-2c+1}{d-c+1}\ . \tag{2.6}$$

(4) For the number $\nu(d)$ *of non-realizable couples satisfying condition (2.4) and with* $a_{d-1} < 0$ *one has* $\lim_{d\to\infty}\nu(d)/\chi(d) = 0$, *see (1.1).*

Remark 2.8. *The quantity $T_d^c\binom{d-1}{c}$ (resp. $\binom{d}{c}\binom{d-1}{c}$) is the number of couples in which the change-preservation pattern begins with p and the order satisfies condition (2.4) (resp. of all compatible couples in which the change-preservation pattern begins with p). For c fixed, one has $\lim_{d\to\infty} T_d^c/\binom{d}{c} = 1$. Indeed, this is the ratio of two degree c polynomials in d whose leading coefficients equal $1/c!$.*

Proof of Theorem 2.3. Part (1). Indeed, $a_{d-1} = \gamma_1 + \cdots + \gamma_p - \alpha_1 - \cdots - \alpha_c > 0$.

Part (2). The first term in the right-hand side of (2.5) is the number of all orders with c components equal to P. The second term is the number of orders beginning with P; they do not satisfy conditions (2.4). The third (resp. the fourth) term is the number of orders beginning with NPP (resp. with $NPNPP$ or $NNPPP$). The fifth term is the number of orders beginning with $NPNPNPP$, $NNPPNPP$, $NPNNPPP$, $NNPNPPP$ or $NNNPPPP$ etc.

That is, for $k \geq 2$, the kth term is the number of orders among whose first $2k-1$ components there are k letters P and which are not included in one of the previous terms (excluding the initial $\binom{d}{c}$). In an equivalent way, the kth term contains orders among whose $2k-2$ first components there are exactly $k-1$ letters P and for $s \leq 2k-2$, among their s first letters there are not less letters N than letters P. Hence this is the number of lattice paths in the plane with possible steps $(1,1)$ and $(1,-1)$ going from $(0,0)$ to $(2k-2,0)$ which do not descend below the abscissa-axis. The number of such paths is C_{k-1}.

Part (3). Formula (2.6) can be proved by induction on d. For $d=1$ and 2 and for $c \leq d$, it is to be checked directly. Suppose that it is true for $d \leq d_0$. Then for $d = d_0+1$, one applies to any binomial coefficient in the formula the well-known equality $\binom{n}{k} = \binom{n-1}{k-1} + \binom{n-1}{k}$. Thus

$$\begin{aligned} T_d^c = T_{d-1}^c + T_{d-1}^{c-1} &= \binom{d-1}{c}\left(1-\tfrac{c}{d-c}\right) + \binom{d-1}{c-1}\left(1-\tfrac{c-1}{d-c+1}\right) \\ &= \binom{d}{c}\left(1-\tfrac{c}{d-c+1}\right) , \end{aligned}$$

where the rightmost equality is to be checked straightforwardly.

Part (4). Suppose that $d=2k$, $k \in \mathbb{N}^*$. Set

$$h_{k,m} := \frac{k(k-1)\cdots(k-m+1)}{(k+1)(k+2)\cdots(k+m)} , \quad \text{so} \quad \binom{2k}{k-m} = \binom{2k}{k} h_{k,m} .$$

For k fixed, the sequence $h_{k,m}$ is decreasing in m; one has $h_{k,0} = 1$. The sum $\sum_{c=0}^{d}\binom{d}{c}^2$ of all compatible couples equals $\tilde{b} := \binom{2k}{k}^2(1+2\sum_{m=1}^{k} h_{k,m}^2)$. The number $\nu(d) = \nu(2k)$ is bounded by

$$\sum_{c=0}^{k}\binom{2k}{c}T_{2k}^c = \sum_{m=0}^{k}\binom{2k}{k-m}T_{2k}^{k-m} = \binom{2k}{k}^2\sum_{m=0}^{k}\frac{2m+1}{k+m+1}h_{k,m}^2$$

(we remind that the orders satisfying condition (2.4) are defined under the assumption that $c \leq p$). Fix $s \in (0,1)$. Then

$$g_1 := \sum_{m=0}^{[sk]}\frac{2m+1}{k+m+1}h_{k,m}^2 \leq \frac{2[sk]+1}{k+[sk]+1}\sum_{m=0}^{[sk]} h_{k,m}^2 .$$

It is clear that $g_1 < \frac{2[sk]+1}{k+[sk]+1}\sum_{m=0}^{k} h_{k,m}^2$, so

$$\binom{2k}{k}^2 g_1 < \frac{2[sk]+1}{k+[sk]+1}\tilde{b} . \tag{2.7}$$

For large values of k and for $m \geq [sk]+1$, the quantity $h_{k,m}$ is majorized by

$$\frac{(k-[sk/2])\cdots(k-m+1)}{(k+[sk/2]+1)\cdots(k+m)} \leq \left(\frac{k-[sk/2]}{k+[sk/2]+1}\right)^{[sk]-[sk/2]} \left(\frac{k-[sk]+1}{k+[sk]}\right)^{m-[sk]-1} .$$

Set $u := \frac{k-[sk/2]}{k+[sk/2]+1}$ and $v := \frac{k-[sk]+1}{k+[sk]}$. Hence

$$\begin{array}{rclcl} g_2 & := & \sum_{m=[sk]+1}^{k} h_{k,m}^2 & < & u^{[sk]-[sk/2]} \sum_{m=[sk]+1}^{\infty} v^{m-[sk]-1} \\ & = & \frac{u^{[sk]-[sk/2]}}{1-v} & = & u^{[sk]-[sk/2]} \frac{k+[sk]}{2[sk]+1} . \end{array}$$

The latter quantity tends to 0 as $k \to \infty$, therefore

$$\lim_{k\to\infty} \binom{2k}{k}^2 g_2/\tilde{b} = 0.$$

As

$$g_3 := \sum_{m=[sk]+1}^{k} \frac{2m+1}{k+m+1} h_{k,m}^2 < g_2,$$

one obtains

$$\lim_{k\to\infty} \binom{2k}{k}^2 g_3/\tilde{b} = 0 . \tag{2.8}$$

One has $\nu(d) \leq \binom{2k}{k}^2(g_1+g_3)$. The coefficient of $\tilde{b}$ in (2.7) can be made smaller than any positive number by choosing s small enough. Therefore inequality (2.7) and equality (2.8) imply part (4) of Theorem 2.3 for d even.

If $d = 2k+1$, $k \in \mathbb{N}^*$, then one can prove part (4) in much the same way, so we point out only some technical differences. One sets

$$h_{k,m} := \frac{k(k-1)\cdots(k-m+1)}{(k+2)(k+3)\cdots(k+m+1)} , \quad \text{so} \quad \binom{2k+1}{k-m} = \binom{2k+1}{k} h_{k,m} ,$$

and $\tilde{b} = 2\binom{2k+1}{k}^2(1+\sum_{m=1}^{k} h_{k,m}^2)$. The definitions of the quantities g_1, g_2 and g_3 are the same, but with respect to the new formula for $h_{k,m}$. One sets $u := \frac{k-[sk/2]}{k+[sk/2]+2}$ and $v := \frac{k-[sk]+1}{k+[sk]+1}$. Inequality (2.7) and equality (2.8) remain the same. □

3. Realizable couples for $d = 3$, 4 and 5

We give the exhaustive answer to Question 1.2 for $d = 3, 4$ and 5; for $d = 1$ and 2, this answer is given by Example 1.1; one finds that $\tilde{r}^*(1) = 1$ and $\tilde{r}^*(2) = 2/3$, see Notation 1.2. It is clear from part (1) of Theorem 1.1 that $\tilde{r}^*(1) < 1$ for $d > 1$. We make use of the involution i_m, see Remark 1.2, to consider only the cases with $a_{d-1} > 0$. For $d = 3$, we give the list of sign patterns and (non)-realizable orders in the following table:

sign pattern	realizable orders	non − realizable orders
$(+,+,+,-)$	PNN	NPN , NNP
$(+,+,-,-)$	PNN , NPN , NNP	
$(+,+,+,+)$	NNN	
$(+,+,-,+)$	PPN	NPP , PNP .

Thus $\tilde{r}^*(3) = 3/5$. The (non)-realizability of these cases can be justified using the results in [11]. For $d = 4$, we list the sign patterns by the value of c:

c	sign pattern	realizable orders	non − realizable orders
0	$(+,+,+,+,+)$	$NNNN$	
1	$(+,+,+,+,-)$	$PNNN$	$NPNN$, $NNPN$, $NNNP$
	$(+,+,+,-,-)$	$PNNN$, $NPNN$, $NNPN$	$NNNP$
	$(+,+,-,-,-)$	$NPNN$, $NNPN$, $NNNP$	$PNNN$
2	$(+,+,-,+,+)$	$NPPN$	$NNPP$, $NPNP$, $PNNP$ $PNPN$, $PPNN$
	$(+,+,-,-,+)$	$PNPN$, $NPPN$, $PPNN$, $PNNP$	$NPNP$, $NNPP$
	$(+,+,+,-,+)$	$PPNN$	$PNPN$, $NPPN$, $NPNP$ $PNNP$, $NNPP$
3	$(+,+,-,+,-)$	$PPPN$	$NPPP$, $PNPP$, $PPNP$

Hence $\tilde{r}^*(4) = 3/7$. The (non)-realizability of the cases can be proved using the results in [11]. The involution i_m transforms the sign pattern with $c = 3$ into $(+,-,-,-,-)$. We illustrate the realizability of the cases with the sign pattern $(+,+,-,-,+)$ by examples:

$$\begin{array}{ll} PNPN & (x+1.3)(x-1.2)(x+1.1)(x-1) = \\ & x^4+0.2x^3-2.65x^2-0.266x+1.716 \\ \\ NPPN & (x+2)(x-1)(x-0.9)(x+0.8) = \\ & x^4+0.9x^3-2.82x^2-0.52x+1.44 \\ \\ PPNN & (x+2)(x+1.1)(x-1)(x-0.1) = \\ & x^4+2x^3-1.11x^2-2.11x+0.22 \\ \\ PNNP & (x-2)(x+1.9)(x+1)(x-0.8) = \\ & x^4+0.1x^3-4.62x^2-0.68x+3.04\ . \end{array}$$

For $d = 5$, we show for each sign pattern only the number of realizable and the total number of orders compatible with the sign pattern and in some cases the realizable orders. To justify the tables below, one can use the results in [11] and [13]. There are the following canonical sign patterns:

$c = 0$	$(+,+,+,+,+,+)$	$1/1$	$c = 1$	$(+,+,+,+,+,-)$	$1/5$
$c = 2$	$(+,+,-,+,+,+)$	$1/10$	$c = 3$	$(+,+,-,+,-,-)$	$1/10$
	$(+,+,+,-,+,+)$	$1/10$		$(+,+,+,-,+,-)$	$1/10$
	$(+,+,+,+,-,+)$	$1/10$			
$c = 4$	$(+,+,-,+,-,+)$	$1/5$.			

The remaining sign patterns are:

$c = 1$	$(+,+,+,+,-,-)$	$PNNNN$, $NPNNN$, $NNPNN$	$3/5$
	$(+,+,+,-,-,-)$		$5/5$
	$(+,+,-,-,-,-)$	$NNPNN$, $NNNPN$, $NNNNP$	$3/5$
$c = 2$	$(+,+,-,-,-,+)$	$PPNNN$, $PNPNN$, $PNNPN$, $PNNNP$, $NPPNN$	$5/10$
	$(+,+,+,-,-,+)$	$PPNNN$, $PNPNN$, $PNNPN$, $NPPNN$	$4/10$
	$(+,+,-,-,+,+)$		$10/10$
$c = 3$	$(+,+,-,+,+,-)$		$5/10$
	$(+,+,-,-,+,-)$		$4/10$.

Therefore $\tilde{r}^*(5) = 47/126$. The two latter sign patterns (with $c = 3$) are obtained from two of the sign patterns with $c = 2$ via the involution $i_m i_r$.

The realizability of the sign pattern $(+,+,-,-,+,+)$ with all possible orders results from

$$(x+1)^3(x-1)^2 = x^5 + x^4 - 2x^3 - 2x^2 + x + 1\ .$$

Indeed, by perturbing the triple root at -1 and the double root at 1, one obtains polynomials with the same sign pattern and with any order of the moduli of the roots, see the proof of part (2) of Theorem 1.1.

Remark 3.9. *We obtained the following sequence for the values of the quantity $\tilde{r}^*(d)$: 1, 2/3, 3/5, 3/7, 47/126, ... One could conjecture that the sequence is decreasing. For the sequence of the ratios of two consecutive terms, one gets*

$$2/3 = 0.66\ldots\ ,\quad 9/10 = 0.9\ ,\quad 5/7 = 0.71\ldots\ ,\quad 47/54 = 0.87\ldots\ .$$

It seems that the even and the odd terms form two adjacent sequences and that $\lim_{d\to\infty} \tilde{r}^*(d) = 0^+$.

4. Proof of Theorem 1.1

Part (1). As already mentioned, for the orders $PP\ldots P$ and $NN\ldots N$, the only change-preservation patterns compatible with them are $cc\ldots c$ and $pp\ldots p$ respectively and the corresponding couples are realizable.

Suppose that for given $c > 0$ and $p > 0$, the order of moduli Ω is realizable with all compatible change-preservation patterns. Then, in particular, it is realizable with the sign patterns σ' and σ'', where σ' has all its c sign changes at the beginning followed by its p sign preservations and vice-versa for σ''. However, the sign patterns σ' and σ'' are canonical hence realizable only with their respective canonical orders Ω' and Ω'', see Definition 2.2. As $\Omega' \neq \Omega''$, the order Ω is not realizable with both σ' and σ''.

Part (2). For $d \geq 1$, the all-pluses sign pattern is realizable with its only compatible order $N \dots N$. To prove the rest of part (2) for $d \geq 5$, we construct sign patterns with $c = 2$ which are realizable with all compatible orders. Consider the polynomial

$$\begin{aligned}(x+1)^{d-2}(x-1)^2 &= \left(\sum_{k=0}^{d-2} \binom{d-2}{k} x^k\right)(x^2 - 2x + 1) \\ &= \sum_{k=0}^{d} h_k x^k \, , \quad h_k := \binom{d-2}{k} - 2\binom{d-2}{k-1} + \binom{d-2}{k-2} \, .\end{aligned}$$

It has two sign changes (so its sign pattern is of the form Σ_{i_1,i_2,i_3}). To understand in which positions they are, one observes that

$$h_k = \frac{(d-2)!}{k!(d-k)!}(4k^2 - 4dk + d(d-1)) \, ,$$

so $h_k = 0$ if and only if $k = k_\pm := (d \pm \sqrt{d})/2$. If d is not an exact square, then the sign changes occur between the powers $x^{s_\pm}$ and $x^{s_\pm+1}$, where $s_\pm < k_\pm < s_\pm + 1$. If d is an exact square, then the coefficients of $x^{k_\pm}$ are 0.

Suppose that d is not an exact square. One can perturb the roots of the polynomial by keeping the sign pattern the same. If d is an exact square, then one can perturb them so that all coefficients become non-zero. One can choose such a perturbation for any possible order of the moduli of roots which proves part (2). One can observe that as $k_+ - k_- = \sqrt{d}$, for $d \geq 5$, there are at least two consecutive negative coefficients (i. e. $i_2 \geq 2$) and the sign pattern is not canonical.

We prove part (3) of the theorem by induction on d. For $d = 1$, 2 and 3, the claim is to be checked straightforwardly, see Example 1.1 and Section 3. Suppose that $d \geq 4$ and that σ is not canonical. Represent σ in the form $(\sigma_d, \sigma^\dagger, \sigma_0)$, where σ_d and σ_0 are its first and last components. Then at least one of the sign patterns $(\sigma_d, \sigma^\dagger)$ and $(\sigma^\dagger, \sigma_0)$ contains an isolated sign change or an isolated sign preservation. Suppose that this is $(\sigma_d, \sigma^\dagger)$. Then $(\sigma_d, \sigma^\dagger)$ is not canonical and hence is realizable by at least three orders by polynomials P_j. This means that σ is also realizable by at least three orders defined by the roots of the polynomials $P_j(x)(x \pm \varepsilon)$, where $\varepsilon > 0$ is small enough and the sign is $+$ (resp. $-$) if the last two components of σ are equal (resp. are different).

Part (4) is also proved by induction on d. For $d \leq 4$, it is to be checked directly. Suppose that $d \geq 5$. If neither of the sign patterns $(\sigma_d, \sigma^\dagger)$ and $(\sigma^\dagger, \sigma_0)$ contains an isolated sign change or sign preservation, then this is the case of σ as well, so σ is canonical and $k^*(\sigma) = 1$ – a contradiction. Hence at least one of these sign patterns is not canonical. Without loss of generality, we suppose that this is $(\sigma_d, \sigma^\dagger)$ (otherwise we apply the involution i_r). Hence $k^*((\sigma_d, \sigma^\dagger)) \geq 3$, so $k^*((\sigma_d, \sigma^\dagger)) = 3$, otherwise similarly to the proof of part (3) we obtain that $k^*(\sigma) > 3$. Applying if necessary the involution i_m, we assume that $(\sigma_d, \sigma^\dagger) = \Sigma_{2,d-2}$ or $\Sigma_{d-2,2}$. In the first case, one has $\sigma = \Sigma_{2,d-1}$. Indeed, if $\sigma = \Sigma_{2,d-2,1}$, then $k^*(\sigma) > 3$, see [11, Theorems 3 and 4]. In the second case, either $\sigma = \Sigma_{d-2,3}$ and $k^*(\sigma) = 5$ (see [11, Theorem 1]) or $\sigma = \Sigma_{d-2,2,1}$ and $k^*(\sigma) = 4$ (see [11, Theorems 3 and 4]).

Part (5). For d even, the order $\Omega := PNN \dots N$ is realizable exactly with the sign patterns $\Sigma_{m,n}$, $m + n = d + 1$, $n < m$, see [11, Theorem 1], so $\ell_*(\Omega) = d/2$.

5. Proof of Theorem 1.2

Proof of part (1). A) For a vector-row v of length $2d$, we denote by v_ℓ the vector-row obtained from v by shifting v by ℓ positions to the right (the rightmost ℓ positions are then lost and the leftmost ℓ positions are filled with zeros). We represent R as determinant of the Sylvester $2d \times 2d$-matrix of the polynomials $Q(x)$ and $(-1)^d Q(-x)$ whose first and $(d+1)$st row equal respectively

$$u := (\, 1 \quad a_{d-1} \quad a_{d-2} \quad a_{d-3} \quad a_{d-4} \quad \ldots \quad a_1 \quad a_0 \quad 0 \quad \ldots \quad 0 \,)$$

and

$$w := (\, 1 \quad -a_{d-1} \quad a_{d-2} \quad -a_{d-3} \quad a_{d-4} \quad \ldots \quad (-1)^{d-1}a_1 \quad (-1)^d a_0 \quad 0 \quad \ldots \quad 0 \,)\,;$$

its second and $(d+2)$nd rows equal u_1 and w_1, its third and $(d+3)$rd rows equal u_2 and w_2 etc. For $d = 2$ and $d = 3$, we obtain the determinants

$$\begin{vmatrix} 1 & a_1 & a_0 & 0 \\ 0 & 1 & a_1 & a_0 \\ 1 & -a_1 & a_0 & 0 \\ 0 & 1 & -a_1 & a_0 \end{vmatrix} \quad \text{and} \quad \begin{vmatrix} 1 & a_2 & a_1 & a_0 & 0 & 0 \\ 0 & 1 & a_2 & a_1 & a_0 & 0 \\ 0 & 0 & 1 & a_2 & a_1 & a_0 \\ 1 & -a_2 & a_1 & -a_0 & 0 & 0 \\ 0 & 1 & -a_2 & a_1 & -a_0 & 0 \\ 0 & 0 & 1 & -a_2 & a_1 & -a_0 \end{vmatrix} .$$

B) For $j = 1, \ldots, d$, we add the $(j+d)$th row to the jth row. Hence the first row of the determinant is now

$$g := (\, 2 \quad 0 \quad 2a_{d-2} \quad 0 \quad 2a_{d-4} \quad \ldots \quad 2a_{d-2[d/2]} \quad 0 \quad 0 \quad \ldots \quad 0 \,)$$

and the next $d-1$ rows equal g_j, $j = 1, \ldots, d-1$. After this one subtracts the kth row multiplied by $1/2$ from the $(d+k)$th one, $k = 1, \ldots, d$. Hence, the $(d+1)$st row equals

$$h := (\, 0 \quad -a_{d-1} \quad 0 \quad -a_{d-3} \quad 0 \quad \ldots \quad -a_{d-2[(d+1)/2]+1} \quad 0 \quad 0 \quad \ldots \quad 0 \,)$$

and the next $d-1$ rows are of the form h_j, $j = 1, \ldots, d-1$. For $d = 2$ and $d = 3$, this gives

$$\begin{vmatrix} 2 & 0 & 2a_0 & 0 \\ 0 & 2 & 0 & a_0 \\ 0 & -a_1 & 0 & 0 \\ 0 & 0 & -a_1 & 0 \end{vmatrix} \quad \text{and} \quad \begin{vmatrix} 2 & 0 & 2a_1 & 0 & 0 & 0 \\ 0 & 2 & 0 & 2a_1 & 0 & 0 \\ 0 & 0 & 2 & 0 & 2a_1 & 0 \\ 0 & -a_2 & 0 & -a_0 & 0 & 0 \\ 0 & 0 & -a_2 & 0 & -a_0 & 0 \\ 0 & 0 & 0 & -a_2 & 0 & -a_0 \end{vmatrix} .$$

C) We permute the rows of the determinant (which does not change the determinant up to a sign). In the first $d - [d/2]$ positions we place the first, third, fifth etc. rows, in the next $[d/2]$ positions the $(d+2)$nd, $(d+4)$th, $(d+6)$th etc. rows, in the next $[d/2]$ positions the second, fourth, sixth etc. rows and in the last $d - [d/2]$ positions the $(d+1)$st, $(d+3)$rd, $(d+5)$th etc. rows. After this permutation the first d rows have non-zero entries only in the odd and the last d rows have non-zero entries only in the even columns.

Then we permute the columns of the determinant placing the odd columns in the first d positions and the even columns in the last d positions by preserving the relative order of the even and odd columns. For $d = 2$ and $d = 3$, the result is

$$\begin{vmatrix} 2 & 2a_0 & 0 & 0 \\ 0 & -a_1 & 0 & 0 \\ 0 & 0 & 2 & 2a_0 \\ 0 & 0 & -a_1 & 0 \end{vmatrix} \quad \text{and} \quad \begin{vmatrix} 2 & 2a_1 & 0 & 0 & 0 & 0 \\ 0 & 2 & 2a_1 & 0 & 0 & 0 \\ 0 & -a_2 & -a_0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 2 & 2a_1 & 0 \\ 0 & 0 & 0 & -a_2 & -a_0 & 0 \\ 0 & 0 & 0 & 0 & -a_2 & -a_0 \end{vmatrix} .$$

For any $d \geq 2$, the determinant is now block-diagonal, with two diagonal blocks $d \times d$. For $d = 4$, these blocks are

$$\begin{vmatrix} 2 & 2a_2 & 2a_0 & 0 \\ 0 & 2 & 2a_2 & 2a_0 \\ 0 & -a_3 & -a_1 & 0 \\ 0 & 0 & -a_3 & -a_1 \end{vmatrix} \quad \text{and} \quad \begin{vmatrix} 2 & 2a_2 & 2a_0 & 0 \\ 0 & 2 & 2a_2 & 2a_0 \\ -a_3 & -a_1 & 0 & 0 \\ 0 & -a_3 & -a_1 & 0 \end{vmatrix} .$$

The first and the $(d + 1)$st rows equal respectively

$$\tilde{g} := (\, 2 \quad 2a_{d-2} \quad 2a_{d-4} \quad \ldots \quad 2a_{d-2[d/2]} \quad 0 \quad 0 \quad \ldots \quad 0 \,)$$

and $\tilde{g}_d$. The first $d - [d/2]$ rows equal $\tilde{g}$, $\tilde{g}_1$, $\tilde{g}_2$, ..., $\tilde{g}_{d-[d/2]-1}$ while the rows with indices $d + 1$, $d + 2$, ..., $d + [d/2]$ are $\tilde{g}_d$, $\tilde{g}_{d+1}$, ..., $\tilde{g}_{d+[d/2]-1}$. The $(d - [d/2] + 1)$st row equals

$$\tilde{h} := (\, 0 \quad -a_{d-1} \quad -a_{d-3} \quad -a_{d-5} \quad \ldots \quad -a_{d-2[(d+1)/2]+1} \quad 0 \quad 0 \quad \ldots \quad 0 \,) .$$

The next $[d/2] - 1$ rows are $\tilde{h}_j$, $j = 1, \ldots, [d/2] - 1$. The last $d - [d/2]$ rows equal $\tilde{h}_k$, $k = d - 1$, ..., $2d - [d/2] - 2$.

The total number of transpositions of rows and columns is even, so the sign of the determinant does not change.

D) One develops the determinant thus obtained w.r.t. its first and then w.r.t. its last column. For d even (resp. for d odd), this yields $-4a_0\Delta$ (resp. $-2a_0\Delta$), where the $(2d - 2) \times (2d - 2)$-determinant Δ is block-diagonal, with two diagonal blocks $(d - 1) \times (d - 1)$ each of which is the Sylvester matrix of the polynomials $2Q^1$ and $-Q^2$. This implies part (1) of the theorem. □

Proof of part (2). One can assign quasi-homogeneous weights to the variables a_j as follows: 0 to a_{d-1}, 1 to a_{d-2} and a_{d-3}, 2 to a_{d-4} and a_{d-5}, 3 to a_{d-6} and a_{d-7} etc., in accordance with the fact that a_{d-2}, a_{d-4}, ... and a_{d-3}/a_{d-1}, a_{d-5}/a_{d-1}, ... are up to a sign elementary symmetric polynomials of the roots of Q^1 and Q^2. Hence R_0 is a quasi-homogeneous polynomial of weight $d_0 := [(d - 1)/2][d/2]$. For d even (resp. for d odd), it contains monomials $\alpha a_0^{[(d-1)/2]} a_{d-1}^{[d/2]}$ and $\beta a_1^{[d/2]}$, $\alpha \neq 0 \neq \beta$ (resp. $\gamma a_1^{[(d-1)/2]} a_{d-1}^{[d/2]}$ and $\delta a_0^{[d/2]}$, $\gamma \neq 0 \neq \delta$), all other monomials containing factors a_0^k and a_1^s only with $k < [(d-1)/2]$ and $s < [d/2]$ (resp. with $k < [d/2]$ and $s < [(d-1)/2]$). Hence R_0 cannot be the product of two quasi-homogeneous polynomials of weights b_1 and b_2, $0 < b_1, b_2 < d_0$. □

References

[1] F. Cajori: *A history of the arithmetical methods of approximation to the roots of numerical equations of one unknown quantity*. Colo. Coll. Publ. Sci. Ser., **12** (7) (1910), 171–215 .

[2] D. R. Curtiss: *Recent extensions of Descartes' rule of signs*, Ann. of Math., **19** (4) (1918), 251–278.

[3] J.-P. de Gua de Malves: *Démonstrations de la Règle de Descartes*, Pour connoître le nombre des Racines positives & négatives dans les Équations qui n'ont point de Racines imaginaires, Memoires de Mathématique et de Physique tirés des registres de l'Académie Royale des Sciences, (1741), 72–96.

[4] The Geometry of René Descartes with a facsimile of the first edition, translated by D. E. Smith and M.L. Latham, New York, Dover Publications, 1954.

[5] J. Forsgård, V. P. Kostov and B. Shapiro: *Could René Descartes have known this?*, Exp. Math., **24** (4) (2015), 438–448.

[6] J. Forsgård, D. Novikov and B. Shapiro: *A tropical analog of Descartes' rule of signs*, Int. Math. Res. Not. IMRN, **2017** (12), 3726–3750.

[7] J. Fourier: Sur l'usage du théorème de Descartes dans la recherche des limites des racines. Bulletin des sciences par la Société philomatique de Paris, (1820) 156–165, 181–187; œuvres 2, 291–309, Gauthier-Villars, 1890.

[8] Y. Gati, V. P. Kostov and M. C. Tarchi: *Sign patterns and rigid moduli orders*, Grad. J. Math., **6** (1) (2021), 60–72.

[9] C. F. Gauss: *Beweis eines algebraischen Lehrsatzes*. J. Reine Angew. Math., **3** (1828), 1-4; Werke 3, 67–70, Göttingen, 1866.

[10] J. L. W. Jensen: *Recherches sur la théorie des équations*, Acta Math., **36** (1913), 181–195 .

[11] V. P. Kostov: *Descartes' rule of signs and moduli of roots*, Publ. Math. Debrecen, **96** (1-2) (2020) 161–184,

[12] V. P. Kostov: *Hyperbolic polynomials and canonical sign patterns*, Serdica Math. J., **46** (2020) 135–150.

[13] V. P. Kostov: *Which Sign Patterns are Canonical*, Results Math., **77** (6) (2022), 235.

[14] V. P. Kostov: *Hyperbolic polynomials and rigid moduli orders*, Publ. Math. Debrecen, **100** (1-2) (2022), 119–128,

[15] V. P. Kostov: *The disconnectedness of certain sets defined after uni-variate polynomials*, Constr. Math. Anal., **5** (3) (2022), 119–133.

[16] E. Laguerre: *Sur la théorie des équations numériques*, Journal de Mathématiques pures et appliquées, s. 3, t. 9, 99–146 (1883); œuvres 1, Paris, 1898, Chelsea, New-York, 1972, pp. 3–47.

[17] B. E. Meserve: *Fundamental Concepts of Algebra*, New York, Dover Publications, 1982.

VLADIMIR PETROV KOSTOV
UNIVERSITÉ CÔTE D'AZUR
DÉPARTEMENT DE MATHÉMATIQUES
CNRS, LJAD, FRANCE
ORCID: 0000-0001-5836-2678
E-mail address: vladimir.kostov@unice.fr