History

The journal *Rendiconti dell'Istituto di Matematica dell'Università di Trieste* was founded in 1969 with the aim of publishing original research articles in all fields of mathematics. The first director of the journal was Arno Predonzan, subsequent directors were Graziano Gentili, Enzo Mitidieri and Bruno Zimmermann.

Rendiconti dell'Istituto di Matematica dell'Università di Trieste has been the first Italian mathematical journal to be published also on-line. The access to the electronic version of the journal is free. All articles are available on-line.

In 2008 the *Dipartimento di Matematica e Informatica*, the owner of the journal, decided to renew it. In particular, a new Editorial Board was formed, and a group of four Managing Editors was selected. The name of the journal however remained unchanged; just the subtitle *An International Journal of Mathematics* was added. Indeed, the opinion of the whole department was to maintain this name, not to give the impression, if changing it, that a further new journal was being launched.

Website Address: http://rendiconti.dmi.units.it

Rendiconti dell'Istituto di Matematica dell'Università di Trieste

An International Journal of Mathematics

Volume 44 (2012)
Dipartimento
di Matematica
e Geoscienze

ISSN 0049-4704

EUT EDIZIONI UNIVERSITÀ DI TRIESTE

ISSN 0049-4704

EUT – Edizioni Università di Trieste
via E. Weiss, 21 – 34128 Trieste
http://eut.units.it

Foreword

The first part of this volume is dedicated to our friend and colleague Fabio Zanolin, on the occasion of his sixtieth birthday. This section contains seventeen invited papers from mathematicians who have collaborated in various ways with Fabio, mainly in the fields of ordinary differential equations and topology. We thank all the authors for their contributions.

Fabio Zanolin was born on November 3, 1952 in Trieste, where he studied and obtained his university degree in mathematics in 1976, with a thesis in topology directed by Mario Dolcher. He was then appointed assistant professor at the *Istituto di Matematica* of the University of Trieste, then associate professor from 1982 to 1987, when he became full professor and moved to the University of Udine, where he still works and lives.

During his career, Fabio has had many students and collaborated with mathematicians from several countries all over the world. All those who have known Fabio have always appreciated his deep mathematical insight, as well as his kindness, generosity and modesty. Among these Alessandro Fonda and Pierpaolo Omari, who have taken care of this section, so to celebrate this special birthday.

Section 1

Rend. Istit. Mat. Univ. Trieste
Volume 44 (2012), 5–17

On the existence of forced oscillations of retarded functional motion equations on a class of topologically nontrivial manifolds

Pierluigi Benevieri, Alessandro Calamai,
Massimo Furi and Maria Patrizia Pera

Dedicated to Fabio Zanolin on the occasion of his 60th birthday

Abstract. *Using a topological approach, based on the fixed point index theory for locally compact maps on metric ANRs, we prove the existence of forced oscillations for retarded functional motion problems constrained on compact manifolds with nontrivial Euler–Poincaré characteristic, provided that the frictional coefficient is nonzero. We do not know if an analogous result holds true in the frictionless case.*

Keywords: Retarded functional differential equations, fixed point index, forced oscillations
MS Classification 2010: 34C40, 34K13, 37C25, 47H10

1. Introduction

Consider a compact boundaryless smooth manifold $M \subseteq \mathbb{R}^s$ and denote by $BU((-\infty, 0], M)$ the space of bounded and uniformly continuous maps from $(-\infty, 0]$ into M with the topology of the uniform convergence. In this paper we study a *retarded functional motion equation* on M of the type

$$x''_\pi(t) = f(t, x_t) - \varepsilon x'(t), \tag{1}$$

where

1. $x''_\pi(t)$ stands for the tangential part of the acceleration $x''(t) \in \mathbb{R}^s$ at the point $x(t) \in M$,

2. the frictional coefficient ε is a positive constant,

3. the applied force $f : \mathbb{R} \times BU((-\infty, 0], M) \to \mathbb{R}^s$ is continuous, T-periodic in the first variable and such that $f(t, \varphi) \in T_{\varphi(0)}M$ for all (t, φ), where $T_pM \subseteq \mathbb{R}^s$ stands for the tangent space of M at a point p of M.

We will call *functional field* a continuous map $f : \mathbb{R} \times BU((-\infty, 0], M) \to \mathbb{R}^s$ verifying the above tangency condition. In addition, let us recall that, given any map x, defined on a real interval J with $\inf J = -\infty$, and given $t \in J$, x_t denotes the map $\theta \mapsto x(\theta + t)$, defined on $(-\infty, 0]$.

The main result of this work, Theorem 4.1 below, shows that the equation (1) admits at least one T-periodic solution (a *forced oscillation*), provided that M has nonzero Euler-Poincaré characteristic and f is bounded and verifies a sort of Lipschitz condition.

This result provides a positive answer to a conjecture recently formulated in [4]. A key tool that allowed us to solve our conjecture is Lemma 3.1 below, proved in [10].

An existence result for a similar problem has been obtained in [1] (see also [2, 3]), with the difference that, in [1], the function f is defined and continuous on $\mathbb{R} \times C((-\infty, 0], M)$ endowed with the compact-open topology. The continuity assumption of f on $\mathbb{R} \times C((-\infty, 0], M)$ is more restrictive than the hypothesis of continuity on $\mathbb{R} \times BU((-\infty, 0], M)$, since the compact-open topology on $C((-\infty, 0], M)$ induces on $BU((-\infty, 0], M)$ a topology which is weaker than that of uniform convergence. This means that the existence of forced oscillations for (1), proved in this paper, is not a byproduct of the analogous result given in [1], whose proof, in addition, does not fit in the present context.

To get our main result we consider a first order *retarded functional differential equation* (*RFDE* for short) on the tangent bundle $TM \subseteq \mathbb{R}^{2s}$, which turns out to be equivalent to the above second order equation (1). More precisely, in the first part of the paper we study a first order RFDE of the type

$$x'(t) = g(t, x_t), \tag{2}$$

where $g \colon \mathbb{R} \times BU((-\infty, 0], N) \to \mathbb{R}^k$ is a functional field over a boundaryless smooth manifold $N \subseteq \mathbb{R}^k$.

Assuming that g is T-periodic in the first variable, we tackle the problem of the existence of T-periodic solutions of equation (2). More generally, given a closed subset X of N, we study the existence of *confined* T-periodic solutions, that is, T-periodic solutions having image in X.

The main result of the first part of the paper, Theorem 3.2 below, states that the equation (2) admits a confined T-periodic solution provided that X is a compact absolute neighborhood retract (ANR) with nonzero Euler-Poincaré characteristic, and the functional field g satisfies some additional conditions. The proof is given by applying the fixed point index theory for locally compact maps on ANRs to a sort of Poincaré T-translation operator acting in a suitable subset of the Banach space $C([-T, 0], \mathbb{R}^k)$.

For general reference on RFDEs we suggest the monograph by Hale and Verduyn Lunel [16]. For RFDEs with finite delay in Euclidean spaces we refer also to the works of Gaines and Mawhin [11], Nussbaum [22, 23] and Mallet-Paret, Nussbaum and Paraskevopoulos [19]. For RFDEs with infinite delay in Euclidean spaces we recommend the article of Hale and Kato [15] and the book by Hino, Murakami and Naito [17]. Finally, for RFDEs with finite delay on manifolds we cite the papers of Oliva [24, 25].

2. Preliminaries

Given a subset A of $\mathbb{R}^k$, we will denote by $BU((-\infty,0],A)$ the set of bounded and uniformly continuous maps from $(-\infty,0]$ into A with the topology of the uniform convergence. Clearly, $BU((-\infty,0],A)$ is a metric subspace of the Banach space $BU((-\infty,0],\mathbb{R}^k)$ and is complete if and only if A is closed. For brevity, throughout the paper we will use the notation

$$\widetilde{A} := BU((-\infty,0],A).$$

Moreover, the norm in $\mathbb{R}^k$ will be denoted by $|\cdot|$ and the norm in $\widetilde{\mathbb{R}^k}$ by $\|\cdot\|$.

A vector $v \in \mathbb{R}^k$ is said to be *inward* to A at a given point p in the closure $\overline{A}$ of A if there exist two sequences $\{\alpha_n\}$ in $[0,+\infty)$ and $\{p_n\}$ in A such that

$$p_n \to p \quad \text{and} \quad \alpha_n(p_n - p) \to v.$$

The set C_pA of the inward vectors to A at p is called the *tangent cone* of A at p (see [6]). One can easily check that the tangent cone is always closed in $\mathbb{R}^k$. The vector subspace of $\mathbb{R}^k$ spanned by C_pA is the *tangent space* T_pA of A at p, whose elements are the *tangent vectors* to A at p.

To simplify some statements and definitions we put $C_pA = T_pA = \emptyset$ whenever p does not belong to $\overline{A}$ (this can be regarded as a consequence of the definition of inward vector if one replaces the assumption $p \in \overline{A}$ with $p \in \mathbb{R}^k$).

Observe that T_pA is the trivial subspace $\{0\}$ of $\mathbb{R}^k$ if and only if p is an isolated point of A. In fact, if p is a limit point, then, given any $\{p_n\}$ in $A\backslash\{p\}$ such that $p_n \to p$, the sequence $\{\alpha_n(p_n - p)\}$, with $\alpha_n = 1/|p_n - p|$, admits a convergent subsequence whose limit is a unit vector. On the other hand, if p is an isolated point of A, the unique inward vector is the null one since the unique sequence $\{p_n\}$ in A convergent to p is the constant sequence coinciding with p.

One can show that, in the special and important case when A is a smooth differentiable manifold with (possibly empty) boundary ∂A (a *∂-manifold* for short), this definition of tangent space is equivalent to the classical one (see for instance [14, 20]). Moreover, if $p \in \partial A$, C_pA is a closed half-space in T_pA (delimited by $T_p\partial A$), while $C_pA = T_pA$ if $p \in A\backslash\partial A$.

2.1. Initial value problem

Let N be a boundaryless smooth manifold in $\mathbb{R}^k$. We say that a continuous map $g\colon \mathbb{R}\times\widetilde{N}\to\mathbb{R}^k$ is a *retarded functional tangent vector field over* N if $g(t,\varphi)\in T_{\varphi(0)}N$ for all $(t,\varphi)\in\mathbb{R}\times\widetilde{N}$. To simplify the notation, in the sequel we frequently call g a *functional field (over* N*).*

Let us consider a *retarded functional differential equation* (*RFDE* for short) of the type

$$x'(t)=g(t,x_t), \tag{3}$$

where $g\colon \mathbb{R}\times\widetilde{N}\to\mathbb{R}^k$ is a functional field over N. Here, as usual and whenever it makes sense, given $t\in\mathbb{R}$, by $x_t\in\widetilde{N}$ we mean the function $\theta\mapsto x(t+\theta)$.

A *solution* of (3) is a function $x\colon J\to N$, defined on an open real interval J with $\inf J=-\infty$, bounded and uniformly continuous on any closed half-line $(-\infty,b]\subset J$, and which verifies eventually the equality $x'(t)=g(t,x_t)$. That is, x is a solution of (3) if there exists τ, with $-\infty\le\tau<\sup J$, such that x is C^1 on the subinterval $(\tau,\sup J)$ of J, and verifies $x'(t)=g(t,x_t)$ for all $t\in(\tau,\sup J)$. Observe that the derivative of a solution x may not exist at $t=\tau$. However, the right derivative $D_+x(\tau)$ of x at τ always exists and is equal to $g(\tau,x_\tau)$. Also, notice that, since x is uniformly continuous on any closed half-line $(-\infty,b]$ of J, then $t\mapsto x_t$ is a continuous curve in $\widetilde{N}$.

A solution of (3) is said to be *maximal* if it is not a proper restriction of another solution to the same equation. As in the case of ODEs, Zorn's lemma implies that any solution is the restriction of a maximal solution.

In what follows, given $\eta\in\widetilde{N}$, we will also consider the initial value problem

$$\begin{cases} x'(t)=g(t,x_t),\\ x_0=\eta\,. \end{cases} \tag{4}$$

A solution of (4) is a solution $x\colon J\to N$ of (3) such that $\sup J>0$, $x'(t)=g(t,x_t)$ for $t>0$, and $x_0=\eta$.

Moreover, given a relatively closed subset X of N, if one takes $\eta\in\widetilde{X}$, then problem (4) will be called the *confined problem* and any X-valued solution of (4) a *confined solution*. For instance, X could be a ∂-manifold of the type $\{p\in N : F(p)\le 0\}$, where the "cutting function" $F\colon N\to\mathbb{R}$ is smooth, having $0\in\mathbb{R}$ as a regular value (this is the situation considered in Section 4). Furthermore, N could be an open subset of $\mathbb{R}^k$ and X one of its connected components.

Following [4], we say that the functional field $g\colon\mathbb{R}\times\widetilde{N}\to\mathbb{R}^k$ is *away from* N *at* $p\in X$ if either $g(t,\varphi)\notin C_p(N\backslash X)$ for all (t,φ) with $\varphi(0)=p$ or $g(t,\varphi)=0$ for all (t,φ) with $\varphi(0)=p$. We point out that this condition is obviously satisfied whenever p, which is a point of X, is not in the topological boundary of X relative to N since, in that case, $C_p(N\backslash X)=\emptyset$. Notice that

this condition is also satisfied when $X = N$, since $C_p(\emptyset) = \emptyset$. If g is away from N at any $p \in X$, we say that g is *away from N in X*.

Theorem 2.1 below is a particular case of a global existence result for the confined case (see [4, Theorem 3.9]; see also [1, Lemma 2.1]).

THEOREM 2.1 (confined global existence). *Let X be a compact subset of a boundaryless smooth manifold $N \subseteq \mathbb{R}^k$ and $g\colon \mathbb{R} \times \widetilde{N} \to \mathbb{R}^k$ a functional field away from N in X. Assume that $g(\mathbb{R} \times \widetilde{X})$ is bounded. Then, any maximal solution of the confined problem* (4) *is defined on the whole real line.*

The continuous dependence of the solutions on initial data is stated in Theorem 2.2 below and is a staightforward consequence of Theorem 4.4 of [4].

THEOREM 2.2 (continuous dependence). *Let N be a boundaryless smooth manifold and $g\colon \mathbb{R} \times \widetilde{N} \to \mathbb{R}^k$ a functional field. Assume the uniqueness of the maximal solution of problem* (4)*. Then, given $T > 0$, the set*

$$\mathcal{D} = \{\eta \in \widetilde{N} : \text{ the maximal solution of (4) is defined up to } T\}$$

is open and the map that associates to any $\eta \in \mathcal{D}$ the restriction to $[0, T]$ of the unique maximal solution of problem (4) *is continuous.*

2.2. Fixed point index

We recall that a metrizable space X is an *absolute neighborhood retract* (ANR) if, whenever it is homeomorphically embedded as a closed subset C of a metric space Y, there exists an open neighborhood V of C in Y and a retraction $r\colon V \to C$ (see e.g. [5, 13]). Polyhedra and differentiable manifolds are examples of ANRs. Let us also recall that a continuous map between topological spaces is called *locally compact* if it has the property that each point in its domain has a neighborhood whose image is contained in a compact set.

Let X be a metric ANR and consider a locally compact (continuous) X-valued map k defined on a subset $\mathcal{D}(k)$ of X. Given an open subset U of X contained in $\mathcal{D}(k)$, if the set of fixed points of k in U is compact, the pair (k, U) is called *admissible*. It is known that to any admissible pair (k, U) we can associate an integer $\mathrm{ind}_X(k, U)$ – the *fixed point index* of k in U – which satisfies properties analogous to those of the classical Leray–Schauder degree [18]. The reader can see for instance [7, 12, 21, 23] for a comprehensive presentation of the index theory for ANRs. As regards the connection with the homology theory we refer to standard algebraic topology textbooks (e.g. [8, 26]).

We summarize below the main properties of the fixed point index.

i) (*Existence*) If $\mathrm{ind}_X(k, U) \neq 0$, then k admits at least one fixed point in U.

ii) (*Normalization*) If X is compact, then $\mathrm{ind}_X(k, X) = \Lambda(k)$, where $\Lambda(k)$ denotes the Lefschetz number of k.

iii) (*Additivity*) Given two disjoint open subsets U_1, U_2 of U such that any fixed point of k in U is contained in $U_1 \cup U_2$, then

$$\mathrm{ind}_X(k, U) = \mathrm{ind}_X(k, U_1) + \mathrm{ind}_X(k, U_2).$$

iv) (*Excision*) Given an open subset U_1 of U such that k has no fixed points in $U \backslash U_1$, then $\mathrm{ind}_X(k, U) = \mathrm{ind}_X(k, U_1)$.

v) (*Commutativity*) Let X and Y be metric ANRs. Suppose that U and V are open subsets of X and Y respectively and that $k\colon U \to Y$ and $h\colon V \to X$ are locally compact maps. Assume that one of the sets of fixed points of hk in $k^{-1}(V)$ or kh in $h^{-1}(U)$ is compact. Then the other set is compact as well and $\mathrm{ind}_X(hk, k^{-1}(V)) = \mathrm{ind}_Y(kh, h^{-1}(U))$.

vi) (*Homotopy invariance*) Let $H\colon U \times [0,1] \to X$ be a locally compact map such that the set $\{(x, \lambda) \in U \times [0,1] : H(x, \lambda) = x\}$ is compact. Then $\mathrm{ind}_X(H(\cdot, \lambda), U)$ is independent of λ.

3. Existence of periodic solutions

Let $N \subseteq \mathbb{R}^k$ be a boundaryless differentiable manifold and $X \subseteq N$ a compact ANR. Given $T > 0$, denote by $\widehat{X} := C([-T, 0], X)$ the metric subspace of $C([-T, 0], \mathbb{R}^k)$ of the X-valued continuous function on $[-T, 0]$ and by $\widehat{X}_0$ the set $\big\{\psi \in \widehat{X} : \psi(-T) = \psi(0)\big\}$. Observe that $\widehat{X}$ is complete since X is closed. Moreover, it is not difficult to show that $\widehat{X}$ is itself an ANR.

Let $g\colon \mathbb{R} \times \widetilde{N} \to \mathbb{R}^k$ be a functional field. Given $T > 0$, assume that g is T-periodic in the first variable. We are interested in proving the existence of X-valued T-periodic solutions of equation (3). To this end, let us consider the family of RFDE

$$x'(t) = \lambda\, g(t, x_t) \tag{5}$$

depending on the parameter $\lambda \in [0,1]$. Our aim is to define a parametrized Poincaré-type T-translation operator whose fixed points are the restrictions to the interval $[-T, 0]$ of the T-periodic solutions of (5). For this purpose, we need to introduce a suitable backward extension of the elements of $\widehat{X}$. The properties of such an extension are contained in Lemma 3.1 below, obtained in [10]. In what follows, by a T-periodic map defined on $(-\infty, 0]$ (or on $(-\infty, -T]$) we mean the restriction of a T-periodic map on $\mathbb{R}$.

LEMMA 3.1. *There exist an open neighborhood U of $\widehat{X}_0$ in $\widehat{X}$ and a continuous map from U to $\widetilde{X}$, $\psi \mapsto \widetilde{\psi}$, with the following properties:*

1) $\widetilde{\psi}$ is an extension of ψ;

2) $\widetilde{\psi}$ is T-periodic on $(-\infty, -T]$;

3) $\widetilde{\psi}$ is T-periodic on $(-\infty, 0]$, whenever $\psi \in \widehat{X}_0$.

Let us now state our existence result.

THEOREM 3.2. *Let $N \subseteq \mathbb{R}^k$ be a boundaryless smooth manifold and $g\colon \mathbb{R}\times\widetilde{N} \to \mathbb{R}^k$ a T-periodic functional field. Let $X \subseteq N$ be a compact ANR with Euler-Poincaré characteristic $\chi(X) \neq 0$. Assume that g is away from N in X and that $g(\mathbb{R} \times \widetilde{X})$ is bounded. Also assume that, for any $\eta \in \widetilde{X}$, the maximal solution of problem (4) is unique. Then, the equation $x'(t) = g(t, x_t)$ has a T-periodic solution in X.*

Proof. Given $\eta \in \widetilde{X}$ and $\lambda \in [0,1]$, let $x(\eta, \lambda, \cdot)$ be the X-valued maximal solution of the parametrized confined problem

$$\begin{cases} x'(t) = \lambda\, g(t, x_t), \\ x_0 = \eta, \end{cases} \tag{6}$$

whose global existence is ensured by Theorem 2.1 (observe that $\lambda\, g$ is still away from N in X even for $\lambda = 0$). Let now U be an open neighborhood of $\widehat{X}_0$ in $\widehat{X}$ as in Lemma 3.1 and consider the homotopy $P\colon U \times [0,1] \to \widehat{X}$ defined by $P(\psi, \lambda)(\theta) = x(\widetilde{\psi}, \lambda, T + \theta)$, where $\widetilde{\psi} \in \widetilde{X}$ is the continuous extension of ψ as in Lemma 3.1.

By an argument similar to that used in [2, Proposition 3.2], we get that $\psi \in U$ is a fixed point of $P(\cdot, \lambda)$, $\lambda \in [0,1]$, if and only if it is the restriction to $[-T, 0]$ of a T-periodic solution of (5).

Let us show that P is admissible for the fixed point index.

P is continuous. Consider the problem

$$\begin{cases} x'(t) = \mu\, g(t, x_t), \\ \mu'(t) = 0, \\ x_0 = \eta, \\ \mu(0) = \lambda. \end{cases} \tag{7}$$

The continuity of P follows immediately by Lemma 3.1 and by applying Theorem 2.2 to the auxiliary problem (7).

The image of P is contained in a compact subset of $\widehat{X}$. By assumption, there exists $c > 0$ such that $|g(t, \varphi)| \leq c$ for any $(t, \varphi) \in \mathbb{R} \times \widetilde{X}$. Hence, $P(U \times [0,1])$ is contained in the set $K = \{y \in \widehat{X} : |y'(t)| \leq c\}$ which is compact by Ascoli's theorem, since X is bounded and $\widehat{X}$ complete.

The set $\{(\psi,\lambda)\in U\times[0,1] : P(\psi,\lambda)=\psi\}$ *is compact.* Observe that, for any $\lambda\in[0,1]$, the set $\{\psi\in U : P(\psi,\lambda)=\psi\}$ is contained in $K\cap\widehat{X}_0$ that is clearly a compact subset of U.

The three steps proved above imply that P is an admissible homotopy in U. Consequently, by the homotopy invariance of the fixed point index, we get

$$\mathrm{ind}_{\widehat{X}}(P(\cdot,1),U)=\mathrm{ind}_{\widehat{X}}(P(\cdot,0),U).$$

Now, observe that $P(\cdot,0)$ sends U onto the subset of $\widehat{X}_0\subseteq U$ of the constant X-valued functions, which will be identified with X itself. According to this identification, the restriction $P(\cdot,0)|_X$ coincides with the identity I_X of X. Therefore, by the commutativity and normalization properties of the fixed point index, we get

$$\mathrm{ind}_{\widehat{X}}(P(\cdot,0),U)=\mathrm{ind}_X(P(\cdot,0)|_X,X)=\Lambda(I_X).$$

As well-known, the Lefschetz number $\Lambda(I_X)$ coincides with the Euler-Poincaré characteristic $\chi(X)$ of X that, by assumption, is nonzero. Hence,

$$\mathrm{ind}_{\widehat{X}}(P(\cdot,1),U)=\chi(X)\neq 0,$$

which implies that $P(\cdot,1)$ has a fixed point in U. Thus, as previously observed, this is equivalent to the existence of a T-periodic solution of equation (3), as claimed. □

Remark 3.3. *We believe that the above existence result is still valid without the uniqueness assumption on the solutions of the initial value problem.*

Remark 3.4. *A functional field* $g\colon\mathbb{R}\times\widetilde{N}\to\mathbb{R}^k$ *is said to be* compactly Lipschitz *(for short,* c-Lipschitz*) if, given any compact subset* Q *of* $\mathbb{R}\times\widetilde{N}$*, there exists* $L\geq 0$ *such that*

$$|g(t,\varphi)-g(t,\psi)|\leq L\|\varphi-\psi\|$$

for all $(t,\varphi),(t,\psi)\in Q$*. Moreover, we will say that* g *is* locally c-Lipschitz *if for any* $(\tau,\eta)\in\mathbb{R}\times\widetilde{N}$ *there exists an open neighborhood of* (τ,η) *in which* g *is c-Lipschitz. In spite of the fact that a locally Lipschitz map is not necessarily (globally) Lipschitz, one could actually show that if* g *is locally c-Lipschitz, then it is also (globally) c-Lipschitz. As a consequence, if* g *is* C^1 *or, more generally, locally Lipschitz in the second variable, then it is additionally c-Lipschitz. In [4] we proved that if* g *is a c-Lipschitz functional field, then problem* (4) *has a unique maximal solution for any* $\eta\in\widetilde{N}$*. For a characterisation of compact subsets of* $\widetilde{N}$ *see e.g. [9, Part 1, IV.6.5].*

4. Retarded functional motion equations

Let $M \subseteq \mathbb{R}^s$ be a boundaryless smooth manifold and let

$$TM = \big\{(q,v) \in \mathbb{R}^s \times \mathbb{R}^s : \ q \in M, \ v \in T_qM\big\}$$

be the tangent bundle of M. Given $q \in M$, let $(T_qM)^\perp \subseteq \mathbb{R}^s$ denote the normal space of M at q. Since $\mathbb{R}^s = T_qM \oplus (T_qM)^\perp$, any vector $u \in \mathbb{R}^s$ can be uniquely decomposed into the sum of the parallel (or tangential) component $u_\pi \in T_qM$ of u at q and the normal component $u_\nu \in (T_qM)^\perp$ of u at q.

Consider the retarded functional motion equation on the constraint M

$$x''_\pi(t) = f(t, x_t) - \varepsilon x'(t), \tag{8}$$

where $x''_\pi(t)$ stands for the parallel component of the acceleration $x''(t) \in \mathbb{R}^s$ at the point $x(t)$, the parameter $\varepsilon > 0$ is the frictional coefficient, and the map $f\colon \mathbb{R} \times \widetilde{M} \to \mathbb{R}^s$ is a functional field, T-periodic in the first variable. Any T-periodic solution of (8) is called a *forced oscillation*.

Theorem 4.1 below gives a positive answer to the conjecture presented by the authors in [4].

THEOREM 4.1. *Let M be a compact boundaryless smooth manifold with nonzero Euler-Poincaré characteristic, and let $f\colon \mathbb{R} \times \widetilde{M} \to \mathbb{R}^k$ be a T-periodic functional field on M. Assume that f is locally Lipschitz in the second variable and has bounded image. Then, the equation (8) has a forced oscillation.*

Proof. Let us observe first that the equation (8) can be equivalently written as

$$x''(t) = r(x(t), x'(t)) + f(t, x_t) - \varepsilon x'(t), \tag{9}$$

where $r\colon TM \to \mathbb{R}^s$ is a smooth map (the so-called reactive force or inertial reaction) satisfying the following properties:

(a) $r(q,v) \in (T_qM)^\perp$ for any $(q,v) \in TM$;

(b) r is quadratic in the second variable;

(c) given $(q,v) \in TM$, $r(q,v)$ is the unique vector such that $(v, r(q,v))$ belongs to $T_{(q,v)}(TM)$;

(d) any C^2 curve $\gamma\colon (a,b) \to M$ verifies the condition $\gamma''_\nu(t) = r(\gamma(t), \gamma'(t))$ for any $t \in (a,b)$, i.e. for each $t \in (a,b)$, the normal component $\gamma''_\nu(t)$ of $\gamma''(t)$ at $\gamma(t)$ equals $r(\gamma(t), \gamma'(t))$.

Now, let us transform the second order equation (9) into the first order system

$$\begin{cases} x'(t) = y(t), \\ y'(t) = r(x(t), y(t)) + f(t, x_t) - \varepsilon y(t). \end{cases} \tag{10}$$

System (10) is actually a first order RFDE on the noncompact manifold TM, since it can be written as

$$(x'(t), y'(t)) = G(t, (x_t, y_t)),$$

where the map $G\colon \mathbb{R} \times \widetilde{TM} \to \mathbb{R}^s \times \mathbb{R}^s$ is the T-periodic functional field over TM given by

$$G(t, (\varphi, \psi)) = (\psi(0), r(\varphi(0), \psi(0)) + f(t, \varphi) - \varepsilon\psi(0)).$$

It is easy to see that equation (9) and system (10) are equivalent in the sense that a function $x\colon J \to M$ is a solution of (9) if and only if the pair $(x, x')\colon J \to TM$ is a solution of (10).

Given $c > 0$, consider the closed subset

$$X_c = \big\{(q, v) \in TM : |v| \le c\big\}$$

of TM. It is not difficult to show that X_c is a ∂-manifold in $\mathbb{R}^s \times \mathbb{R}^s$ with boundary

$$\partial X_c = \big\{(q, v) \in X_c : |v| = c\big\}.$$

Moreover, since M is a deformation retract of X_c, then the two spaces are homotopically equivalent. Thus, $\chi(X_c) = \chi(M)$, so that $\chi(X_c) \neq 0$.

Observe now that $G(\mathbb{R} \times \widetilde{X}_c)$ is a bounded subset of $\mathbb{R}^s \times \mathbb{R}^s$, since f is bounded by assumption and X_c is compact.

Let us prove that if c is sufficiently large, then G is away from TM in X_c. To this end, write X_c by means of the inner product $\langle \cdot, \cdot \rangle$ in $\mathbb{R}^s$, as $\big\{(q, v) \in TM : \langle v, v\rangle \le c^2\big\}$ and observe first that the tangent cone of X_c at $(q, v) \in \partial X_c$ is the half subspace of $T_{(q,v)}X_c$ given by

$$C_{(q,v)}X_c = \big\{(\dot q, \dot v) \in T_{(q,v)}(TM) : \langle v, \dot v\rangle \le 0\big\}.$$

Analogously,

$$C_{(q,v)}(TM \backslash X_c) = \big\{(\dot q, \dot v) \in T_{(q,v)}(TM) : \langle v, \dot v\rangle \ge 0\big\}.$$

Take any $t \in \mathbb{R}$ and any pair $(\varphi, \psi) \in \widetilde{X}_c$ with $|\psi(0)| = c$ and consider the inner product

$$\begin{aligned} &\langle \psi(0), r(\varphi(0), \psi(0)) + f(t, \varphi) - \varepsilon\psi(0)\rangle \\ &\quad = \langle \psi(0), r(\varphi(0), \psi(0))\rangle + \langle \psi(0), f(t, \varphi)\rangle - \varepsilon\langle \psi(0), \psi(0)\rangle. \end{aligned}$$

Now,

$$\langle \psi(0), r(\varphi(0), \psi(0)) \rangle = 0,$$

since $r(\varphi(0), \psi(0))$ belongs to $(T_{\varphi(0)}M)^{\perp}$. Moreover,

$$\langle \psi(0), f(t, \varphi) \rangle \leq |\psi(0)|\, |f(t, \varphi)| \leq K|\psi(0)|,$$

where K is such that $|f(t, \varphi)| \leq K$ for all $(t, \varphi) \in \mathbb{R} \times \widetilde{M}$. Finally,

$$\langle \psi(0), \psi(0) \rangle = c^2,$$

since $(\varphi(0), \psi(0)) \in \partial X_c$. Therefore, by choosing $c > K/\varepsilon$, we get

$$\big\langle \psi(0), r(\varphi(0), \psi(0)) + f(t, \varphi) - \varepsilon\psi(0) \big\rangle \leq Kc - \varepsilon c^2 < 0.$$

Thus, $G(t, (\varphi, \psi)) \notin C_{(q,v)}(TM \backslash X_c)$ for all $(t, (\varphi, \psi))$ with $(\varphi(0), \psi(0)) = (q, v) \in \partial X_c$. This shows that G is away from TM in X_c, as claimed.

Consequently, we are reduced to the context of Theorem 3.2 with $\mathbb{R}^k = \mathbb{R}^s \times \mathbb{R}^s$, $N = TM$, $g = G$ and the confining set X given by the compact ∂-manifold X_c.

Moreover, since f is locally Lipschitz in the second variable and r is smooth, then G is locally Lipschitz as well. Therefore, taking into account Remark 3.4, we get that the initial value problem

$$\begin{cases} (x'(t), y'(t)) = G(t, (x_t, y_t)), \\ (x_0, y_0) = (\varphi, \psi) \end{cases} \tag{11}$$

has a unique maximal solution for any $(\varphi, \psi) \in \widetilde{TM}$.

Thus, we can apply Theorem 3.2 to the first order equation $(x'(t), y'(t)) = G(t, (x_t, y_t))$, obtaining that system (10) has a T-periodic solution and, equivalently, that the motion equation (8) has a forced oscillation. □

Remark 4.2. *We believe that the assertion of Theorem 4.1 still holds without the Lipschitz assumption.*

Remark 4.3. *In the frictionless case (i.e. $\varepsilon = 0$) we do not know whether or not the equation*

$$x''_{\pi}(t) = f(t, x_t) \tag{12}$$

has a forced oscillation. As far as we know, the problem of the existence of forced oscillations of (12) is still open, even in the undelayed situation. In the particular case of the spherical pendulum, i.e. $X = S^2$, or, more generally, in the case of the even dimensional pendulum (i.e. $X = S^{2n}$), the existence of forced oscillations for equation (12) *has been proved by the authors in [3], assuming the stronger hypothesis of the continuity of the functional field f on $\mathbb{R} \times C((-\infty, 0], X)$.*

References

[1] P. Benevieri, A. Calamai, M. Furi and M.P. Pera, *Retarded functional differential equations on manifolds and applications to motion problems for forced constrained systems*, Adv. Nonlinear Stud. **9** (2009), 199–214.

[2] P. Benevieri, A. Calamai, M. Furi and M.P. Pera, *A continuation result for forced oscillations of constrained motion problems with infinite delay*, to appear in Adv. Nonlinear Stud.

[3] P. Benevieri, A. Calamai, M. Furi and M.P. Pera, *On the existence of forced oscillations for the spherical pendulum acted on by a retarded periodic force*, J. Dynam. Differential Equations **23** (2011), 541–549.

[4] P. Benevieri, A. Calamai, M. Furi and M.P. Pera, *On general properties of retarded functional differential equations on manifolds*, Discrete Contin. Dyn. Syst. **33** (2013), 27–46.

[5] K. Borsuk, *Theory of Retracts*, Polish Sci. Publ., Warsaw, 1967.

[6] G. Bouligand, *Introduction à la Géométrie Infinitésimale Directe*, Gauthier-Villard, Paris, 1932.

[7] R.F. Brown, *The Lefschetz Fixed Point Theorem*, Scott, Foresman and Co., Glenview, Ill.-London, 1971.

[8] A. Dold, *Lectures on Algebraic Topology*, Springer-Verlag, Berlin, 1972.

[9] N. Dunford and J.T. Schwartz, *Linear Operators*, Wiley & Sons, Inc., New York, 1957.

[10] M. Furi, M.P. Pera and M. Spadini, *Periodic solutions of functional differential perturbations of autonomous differential equations*, Commun. Appl. Anal. **15** (2011), 381–394.

[11] R. Gaines and J. Mawhin, *Coincidence Degree and Nonlinear Differential Equations*, Lecture Notes in Math., **568**, Springer Verlag, Berlin, 1977.

[12] A. Granas, *The Leray-Schauder index and the fixed point theory for arbitrary ANRs*, Bull. Soc. Math. France **100** (1972), 209–228.

[13] A. Granas and J. Dugundji, *Fixed Point Theory*, Springer-Verlag, New York, 2003.

[14] V. Guillemin and A. Pollack, *Differential Topology*, Prentice-Hall Inc., Englewood Cliffs, New Jersey, 1974.

[15] J.K. Hale and J. Kato, *Phase Space for Retarded Equations with Infinite Delay*, Funkc. Ekvac. **21** (1978), 11–41.

[16] J.K. Hale and S.M. Verduyn Lunel, *Introduction to Functional Differential Equations*, Springer Verlag, New York, 1993.

[17] Y. Hino, S. Murakami and T. Naito, *Functional-differential Equations with Infinite Delay*, Lecture Notes in Math., **1473**, Springer Verlag, Berlin, 1991.

[18] J. Leray and J. Schauder, *Topologie et équations fonctionnelles*, Ann. Sci. École Norm. Sup. **51** (1934), 45–78.

[19] J. Mallet-Paret, R.D. Nussbaum and P. Paraskevopoulos, *Periodic solutions for functional-differential equations with multiple state-dependent time lags*, Topol. Methods Nonlinear Anal. **3** (1994), 101–162.

[20] J.M. Milnor, *Topology from the Differentiable Viewpoint*, Univ. Press of Virginia, Charlottesville, 1965.

[21] R.D. Nussbaum, *The fixed point index for local condensing maps*, Ann. Mat. Pura Appl. **89** (1971), 217–258.
[22] R.D. Nussbaum, *Periodic solutions of some nonlinear autonomous functional differential equations*, Ann. Mat. Pura Appl. **101** (1974), 263–306.
[23] R.D. Nussbaum, *The fixed point index and fixed point theorems*, Topological methods for ordinary differential equations (Montecatini Terme, 1991), Lecture Notes in Math., **1537**, Springer, Berlin, 1993, 143–205.
[24] W.M. Oliva, *Functional differential equations on compact manifolds and an approximation theorem*, J. Differential Equations **5** (1969), 483–496.
[25] W.M. Oliva, *Functional differential equations–generic theory*. Dynamical systems (Proc. Internat. Sympos., Brown Univ., Providence, R.I., 1974), Vol. I, Academic Press, New York, 1976, 195–209.
[26] E. Spanier, *Algebraic Topology*, Mc Graw-Hill Series in High Math., New York, 1966.

Authors' addresses:

Pierluigi Benevieri
Dipartimento di Sistemi e Informatica
Università degli Studi di Firenze
Via S. Marta 3, 50139 Firenze, Italy
and
Instituto de Matemática e Estatística
Universidade de São Paulo
Rua do Matão 1010, São Paulo, 05508-090, Brasil
E-mail: pierluigi.benevieri@unifi.it

Alessandro Calamai
Dipartimento di Ingegneria Industriale e Scienze Matematiche
Università Politecnica delle Marche
Via Brecce Bianche, 60131 Ancona, Italy
E-mail: calamai@dipmat.univpm.it

Massimo Furi
Dipartimento di Sistemi e Informatica
Università degli Studi di Firenze
Via S. Marta 3, 50139 Firenze, Italy
E-mail: massimo.furi@unifi.it

Maria Patrizia Pera
Dipartimento di Sistemi e Informatica
Università degli Studi di Firenze
Via S. Marta 3, 50139 Firenze, Italy
E-mail: mpatrizia.pera@unifi.it

Received February 20, 2012
Revised April 7, 2012

Rend. Istit. Mat. Univ. Trieste
Volume 44 (2012), 19–32

Stability criteria for impulsive Kolmogorov-type systems of nonautonomous differential equations

Shair Ahmad and Ivanka Stamova

Dedicated to Fabio Zanolin on the occasion of his sixtieth birthday

Abstract. *In this paper we consider a class of impulsive Kolmogorov-type systems. The problems of uniform stability and uniform asymptotic stability of the solutions are studied. We establish stability criteria by employing piecewise continuous Lyapunov functions. Examples are given to demonstrate the effectiveness of the obtained results. We show, also, that the role of impulses in changing the behavior of impulsive models is very important.*

Keywords: stability, Kolmogorov-type models, Lyapunov functions, impulses
MS Classification 2010: 34D20, 34A37, 92D25

1. Introduction

The studies for Kolmogorov systems has long been and will continue to be one of the dominant themes in both ecology and mathematical ecology due to its theoretical and practical significance. Many authors established a series of criteria on the boundedness, persistence, permanence, global asymptotic stability and the existence of positive periodic solutions [8, 9, 12, 14, 16, 18]. Some interesting work on this topic of interest has been done by Zanolin and his co-authors [6, 19, 20].

On the other hand, impulsive effect likewise exists in a wide variety of evolutionary processes in which states are changed abruptly at certain moments of time, involving such fields as medicine and biology, economics, mechanics, electronics, telecommunications, etc. Since time perturbations occur so often in nature, a number of models in ecology can be formulated as systems of impulsive differential equations [2, 3, 4, 5, 13, 15, 21]. One of the most important problems for these types of systems is to analyze the effect of impulsive time perturbations on the dynamic activity patterns in the systems. Impulses can make unstable systems stable; so they have been widely used as a control [17].

Recently, some qualitative properties of populations, which undergo impulsive effects at fixed times between interval of continuous evolutions, have been investigated for impulsive classes of Kolmogorov systems [5, 15, 21]. However, in all of these papers so far, authors mostly focused on the existence of periodic solutions and permanence.

In our previous papers [2] and [3] we studied stability properties of some special cases of impulsive Kolmogorov systems with or without delays.

In the present paper, we consider the uniform stability and uniform asymptotic stability of the solutions for a class of impulsive Kolmogorov-type systems of nonautonomous differential equations. For this purpose piecewise continuous auxiliary functions are used which are an analogue of Lyapunov functions. Examples are given to demonstrate the effectiveness of the obtained results. We show, also, that the role of impulses in changing the behavior of impulsive models is very important.

2. Preliminaries

Let R^n be the n-dimensional Euclidean space with norm $||x|| = \sum_{i=1}^{n} |x_i|$. Let $R_+ = [0, \infty)$, $t_0 \in R_+$ and $t_0 < t_1 < t_2 < ...$, $\lim_{k\to\infty} t_k = \infty$.

Consider the following n- dimensional impulsive Kolmogorov-type system

$$\begin{aligned} &\dot{x}_i(t) = x_i(t) f_i(t, x(t)),\ t \neq t_k, \\ &\Delta x_i(t_k) = P_{ik}(x_i(t_k)),\ k = 1, 2, ..., \end{aligned} \tag{1}$$

$i = 1, 2, ..., n$, where n corresponds to the number of units in the system, $x_i(t)$ corresponds to the state of the ith unit at time t, $f_i : [t_0, \infty) \times R_+^n \to R$, $f = col(f_1, f_2, ..., f_n)$, $f \in C[[t_0, \infty) \times R_+^n, R^n]$, $\Delta x_i(t) = x_i(t + 0) - x_i(t - 0)$, t_k, $k = 1, 2, ...$ are the moments of impulsive perturbations and $P_{ik}(x_i(t_k))$ represents the abrupt change of the state $x_i(t)$ at the impulsive moment t_k, $P_k = col(P_{1k}, P_{2k}, ..., P_{nk})$, $P_k \in C[R_+^n, R^n]$.

Let $x_0 = col(x_{10}, x_{20}, ..., x_{n0})$ and $x_{i0} \geq 0$, $i = 1, 2, ..., n$. Denote by $x(t) = x(t; t_0, x_0) = col(x_1(t), x_2(t), ..., x_n(t))$ the solution of system (1), satisfying the initial condition

$$x(t_0 + 0; t_0, x_0) = x_0. \tag{2}$$

We suppose that the existence, uniqueness, and continuous dependence of solutions of system (1) hold. For the efficient sufficient conditions which guarantee the existence, uniqueness, and continuous dependence of solutions of system (1) (see [11]).

The solutions $x(t)$ of system (1) are piecewise continuous functions with points of discontinuity of the first kind t_k at which they are left continuous; i.e.

the following relations are satisfied:

$$x_i(t_k - 0) = x_i(t_k),\ x_i(t_k + 0) = x_i(t_k) + P_{ik}(x_i(t_k)),$$
$$i = 1, 2, ..., n,\ k = 1, 2, ...$$

We also assume that solutions of (1) with initial conditions (2) are nonnegative, and if $x_{i0} > 0$ for some i, then $x_i(t) > 0$ for all $t \geq t_0$, If, moreover, $(t_k, x_i) \in (t_0, \infty) \times (0, \infty)$, then $x_i(t_k) + P_{ik}(x_i(t_k)) > 0$ for all $i = 1, 2, ..., n$ and $k = 1, 2, ...$ Note that these assumptions are natural from the applicability point of view.

Let $x(t) = x(t; t_0, x_0) = col(x_1(t), x_2(t), ..., x_n(t))$ and $x^*(t) = x^*(t; t_0, x_0^*) = col(x_1^*(t), x_2^*(t), ..., x_n^*(t))$ be any two solutions of (1) with initial conditions

$$x(t_0 + 0; t_0, x_0) = x_0,$$

$$x^*(t_0 + 0; t_0, x_0^*) = x_0^*,$$

where $x_0^* = col(x_{10}^*, x_{20}^*, ..., x_{n0}^*)$ and $x_{i0}^* \geq 0$, $i = 1, 2, ..., n$.

We will use the following definitions of some stability properties of the solutions of (1).

Definition 2.1. *The solution $x^*(t)$ of system (1) is said to be:*

(a) stable, if for all $t_0 \in R_+$ and for all $\varepsilon > 0$ there exists $\delta = \delta(t_0, \varepsilon) > 0$ such that if $x_0, x_0^* \in R_+^n$, with $||x_0 - x_0^*|| < \delta$, then for all $t \geq t_0$:

$$||x(t; t_0, x_0) - x^*(t; t_0, x_0^*)|| < \varepsilon;$$

(b) uniformly stable, if the number δ in (a) is independent of $t_0 \in R_+$;

(c) uniformly attractive, if there exists $\lambda > 0$ such that for all $\varepsilon > 0$ there exists $\gamma = \gamma(\varepsilon) > 0$ such that if $t_0 \in R_+$ and $x_0, x_0^* \in R_+^n$, with $||x_0 - x_0^*|| < \lambda$, then for all $t \geq t_0 + \gamma$:

$$||x(t; t_0, x_0) - x^*(t; t_0, x_0^*)|| < \varepsilon;$$

(d) uniformly asymptotically stable, if it is uniformly stable and uniformly attractive.

Introduce the sets

$$G_k = \Big\{(t, x, x^*) \in [t_0, \infty) \times R_+^n \times R_+^n :\ t_{k-1} < t < t_k\Big\},\quad k = 1, 2, ...,$$
$$G = \bigcup_{k=1}^{\infty} G_k\,.$$

DEFINITION 2.2. *A function $V : [t_0, \infty) \times R_+^n \times R_+^n \to R_+$ belongs to class V_0, if:*

1. *V is continuous in G and locally Lipschitz continuous with respect to its second and third arguments on each of the sets G_k, $k = 1, 2, ...$ and*

$$V(t, x^*, x^*) = 0, \ t \in [t_0, \infty).$$

2. *For each $k = 1, 2, ...$ there exist the finite limits*

$$V(t_k - 0, x, x^*) = \lim_{\substack{t \to t_k \\ t < t_k}} V(t, x, x^*), \ V(t_k + 0, x, x^*) = \lim_{\substack{t \to t_k \\ t > t_k}} V(t, x, x^*)$$

and the equality $V(t_k - 0, x, x^*) = V(t_k, x, x^*)$ holds.

3. *For each $k = 1, 2, ...$ and $x, x^* \in R_+^n$ the following inequality holds:*

$$V(t_k + 0, x + P_k(x), x^* + P_k(x^*)) \leq V(t, x, x^*). \tag{3}$$

Let $V \in V_0$. For $(t, x, x^*) \in G$ we set

$$\dot{V}_{(1)}(t, x, x^*) = \lim_{h \to 0^+} \sup \frac{1}{h}[V(t+h, x+hxf(t, x), x^*+hx^*f(t, x^*)) - V(t, x, x^*)].$$

Note that if $x = x(t)$ and $x^* = x^*(t)$ are solutions of system (1), then $D^+_{(1)}V(t, x(t), x^*(t)) = \dot{V}_{(1)}(t, x, x^*)$, $t \geq t_0$, $t \neq t_k$, where

$$D^+_{(1)}V(t, x(t), x^*(t)) = \lim_{h \to 0^+} \sup \frac{1}{h}[V(t+h, x(t+h), x^*(t+h)) - V(t, x(t), x^*(t))]$$

is the upper right Dini derivative of the function $V(t, x(t), x^*(t))$ (with respect to the system (1)).

We shall use the following class of functions:

$$K = \{a \in C[R_+, R_+] : \ a(r) \text{ is strictly increasing and } a(0) = 0\} \ .$$

3. Main results

In the proofs of our main theorems in this section we shall use piecewise continuous Lyapunov functions $V \in V_0$. Similar results for systems with delays are discussed in [13].

THEOREM 3.1. *Assume that there exist functions $V \in V_0$ and $a, b \in K$ such that*

$$a(||x - x^*||) \leq V(t, x, x^*) \leq b(||x - x^*||),\ t \in [t_0, \infty),\ x, x^* \in R_+^n, \tag{4}$$

$$\dot{V}_{(1)}(t, x, x^*) \leq 0,\ (t, x, x^*) \in G. \tag{5}$$

Then the solution $x^*(t)$ of system (1) is uniformly stable.

Proof. Let $\varepsilon > 0$ be chosen. Choose $\delta = \delta(\varepsilon) > 0$ so that $b(\delta) < a(\varepsilon)$. Let $t_0 \in R_+$, $x_0, x_0^* \in R_+^n$, with $||x_0 - x_0^*|| < \delta$, and $x(t) = x(t; t_0, x_0) = col(x_1(t), x_2(t), ..., x_n(t))$, $x^*(t) = x^*(t; t_0, x_0^*) = col(x_1^*(t), x_2^*(t), ..., x_n^*(t))$ be the solutions of (1).

From the properties of the function V and conditions (4), (5), we get to the inequalities

$$\begin{aligned} a(||x(t; t_0, x_0) - x^*(t; t_0, x_0^*)||) &\leq V(t, x(t; t_0, x_0), x^*(t; t_0, x_0^*)) \\ &\leq V(t_0 + 0, x_0, x_0^*) \\ &\leq b(||x_0 - x_0^*||) < b(\delta) < a(\varepsilon)\,, \end{aligned}$$

from which it follows that $||x(t; t_0, x_0) - x^*(t; t_0, x_0^*)|| < \varepsilon$ for $t \geq t_0$. This proves the uniform stability of the solution $x^*(t)$ of system (1). □

THEOREM 3.2. *Let the condition (4) of Theorem 3.1 be fulfilled and let a function $c \in K$ exist such that for $x, x^* \in R_+^n$ the inequality*

$$\dot{V}_{(1)}(t, x, x^*) \leq -c(||x - x^*||),\ t \in [t_0, \infty), t \neq t_k,\ k = 1, 2, ... \tag{6}$$

holds.

Then the solution $x^*(t)$ of system (1) is uniformly asymptotically stable.

Proof. From Theorem 3.1 it follows that the solution $x^*(t)$ of system (1) is uniformly stable. Hence, for any ε, $\varepsilon > 0$, there exists $\delta > 0$, such that if $t_0 \in R_+$, $x_0, x_0^* \in R_+^n$, with $||x_0 - x_0^*|| < \delta$, then

$$||x(t; t_0, x_0) - x^*(t; t_0, x_0^*)|| < \varepsilon$$

for $t \geq t_0$.

Now, we shall prove that the solution $x^*(t)$ of system (1) is uniformly attractive.

1. Let $\alpha = const > 0$ be so small, that $\{x \in R^n : ||x - x^*(t)|| \leq \alpha\} \subset R_+^n$. For any $t \geq t_0$ denote

$$V_{t,\alpha}^{-1} = \left\{x \in R_+^n : V(t + 0, x, x^*) \leq a(\alpha)\right\}.$$

From (4) we deduce

$$V_{t,\alpha}^{-1} \subset \{x \in R^n : ||x - x^*|| \leq \alpha\}.$$

From conditions of Theorem 3.2 it follows that for any $t_0 \in R_+$ and any $x_0 \in R_+^n : x_0 \in V_{t_0,\alpha}^{-1}$ we have $x(t;t_0,x_0) \in V_{t,\alpha}^{-1}$, $t \geq t_0$. Choose $\eta = \eta(\varepsilon)$ so that $b(\eta) < a(\varepsilon)$ and let $\gamma = \gamma(\varepsilon) > \frac{b(\alpha)}{c(\eta)}$. If we assume that for each $t \in [t_0, t_0 + \gamma]$ the inequality $||x(t;t_0,x_0) - x^*(t;t_0,x_0^*)|| \geq \eta$ is valid, then from (3) and (6) we deduce the inequalities

$$\begin{aligned} V(t_0+\gamma,\, & x(t_0+\gamma;t_0,x_0), x^*(t_0+\gamma;t_0,x_0^*)) \\ &\leq V(t_0+0,x_0,x_0^*) - \int_{t_0}^{t_0+\gamma} c(||x(s;t_0,x_0) - x^*(s;t_0,x_0^*)||)\, ds \\ &\leq b(\alpha) - c(\eta)\gamma < 0\,, \end{aligned}$$

which contradicts (4). The contradiction obtained shows that there exists $t^* \in [t_0, t_0 + \gamma]$ such that $||x(t^*;t_0,x_0) - x^*(t^*;t_0,x_0^*)|| < \eta$. Then for $t \geq t^*$ (hence for any $t \geq t_0 + \gamma$) the following inequalities hold:

$$\begin{aligned} a(||x(t) - x^*(t)||) &\leq V(t;x(t),x^*(t)) \\ &\leq V(t^*,x(t^*),x^*(t^*)) \\ &\leq b(||x(t^*;t_0,x_0) - x^*(t^*;t_0,x_0^*)||) \\ &< b(\eta) < a(\varepsilon)\,. \end{aligned}$$

Therefore $||x(t;t_0,x_0) - x^*(t;t_0,x_0^*)|| < \varepsilon$ for $t \geq t_0 + \gamma$.

2. Let $\lambda = const > 0$ be such that $b(\lambda) \leq a(\alpha)$. Then if $x_0 \in R_+^n : ||x_0 - x_0^*|| < \lambda$, (4) implies

$$V(t_0 + 0, x_0, x_0^*) \leq b(||x_0 - x_0^*||) < b(\lambda) \leq a(\alpha),$$

which shows that for $x_0 \in V_{t_0,\alpha}^{-1}$. From what we proved in item 1 it follows that the solution $x^*(t)$ of system (1) is uniformly attractive.

Therefore, the solution $x^*(t)$ of system (1) is uniformly asymptotically stable. □

Corollary 3.3. *If in Theorem 3.2 condition* (6) *is replaced by the condition*

$$\dot{V}_{(1)}(t,x,x^*) \leq -cV(t,x,x^*),\ t \neq t_k,\ k = 1,2,...,\ x, x^* \in R_+^n, \qquad (7)$$

where $c = const > 0$, then the solution $x^*(t)$ of system (1) *is uniformly asymptotically stable.*

Proof. The proof of Corollary 3.3 is analogous to the proof of Theorem 3.2. It uses the fact that

$$V(t, x(t; t_0, x_0), x^*(t; t_0, x_0^*)) \leq V(t_0 + 0, x_0, x_0^*) \exp[-c(t - t_0)]$$

for $t \geq t_0$, which is obtained from (7) and (3).

In fact, let $\alpha = const > 0 : \{x \in R^n : ||x - x^*(t)|| \leq \alpha\} \subset R_+^n$. Choose $\lambda > 0$ so that $b(\lambda) < a(\alpha)$. Let $\varepsilon > 0$ and $\gamma \geq \frac{1}{c} \ln \frac{a(\alpha)}{a(\varepsilon)}$. Then for $t_0 \in R_+$, $x_0, x_0^* \in R_+^n$, with $||x_0 - x_0^*|| < \lambda$ and $t \geq t_0 + \gamma$ the following inequalities hold

$$V(t, x(t; t_0, x_0), x^*(t; t_0, x_0^*)) \leq V(t_0 + 0, x_0, x_0^*) \exp[-c(t - t_0)] < a(\varepsilon),$$

whence, in view of (4), we deduce that the solution $x^*(t)$ of system (1) is uniformly attractive. □

4. Applications

The results obtained can be applied in the investigation of the stability of any solution which is of interest. One of the solutions which is an object of investigations for the systems of type (1) is the *equilibrium* state, i.e. the constant solution $x^* = col(x_1^*, x_2^*, ..., x_n^*)$ such that

$$\begin{aligned} &\dot{x}_i^*(t) = 0,\ t \neq t_k, \\ &\Delta x_i^*(t_k) = 0,\ k = 1, 2, ..., i = 1, 2, ..., n. \end{aligned}$$

In the applications, uniform stability and uniform asymptotic stability of the equilibria will be discussed for a special case of impulsive Kolmogorov-type models.

Consider the following n-species Lotka-Volterra type impulsive system

$$\begin{cases} \dot{x}_i(t) = x_i(t) \left[b_i(t) - a_{ii}(t) x_i(t) - \sum\limits_{\substack{j=1 \\ j \neq i}}^{n} a_{ij}(t) x_j(t) \right], \quad t \neq t_k, \\ x_i(t_k + 0) = x_i(t_k) + P_{ik}(x_i(t_k)), \quad i = 1, ..., n,\ k = 1, 2, ..., \end{cases} \tag{8}$$

where $n \geq 2$, $t \geq 0$, $a_{ij} \in C[R_+, R_+]$, $b_i \in C[R_+, R]$, $P_{ik} : R_+ \to R$, $i, j = 1, ..., n$, $k = 1, 2, ...$, $0 < t_1 < t_2 < ... < t_k < ...$ are fixed impulsive points and $\lim\limits_{k \to \infty} t_k = \infty$. In mathematical ecology, the system (8) denotes a model of the dynamics of an n-species system in which each individual competes with all others of the system for a common resource and at the fixed moments of time t_k, $k = 1, 2, ...$, the system is subject to short-term perturbations. The

numbers $x_i(t_k)$ and $x_i(t_k + 0)$ are, respectively, the population densities of species i before and after impulse perturbation at the moment t_k and P_{ik} are functions which characterize the magnitude of the impulse effect on the species i at the moments t_k.

Let $x_0 = col(x_{10}, x_{20}, ..., x_{n0})$ and $x_{i0} \geq 0$, $i = 1, 2, ..., n$. Denote by $x(t) = x(t; t_0, x_0) = col(x_1(t), x_2(t), ..., x_n(t))$ the solution of system (8), satisfying the initial condition

$$x(t_0 + 0; t_0, x_0) = x_0. \tag{9}$$

Given a continuous function $g(t)$ which is defined on J, $J \subseteq R$, we set

$$g^L = \inf_{t \in J} g(t), \quad g^M = \sup_{t \in J} g(t).$$

For $0 \leq \tau_1 < \tau_2$, we define the following notation:

$$A[g, \tau_1, \tau_2] = \frac{1}{\tau_2 - \tau_1} \int_{\tau_1}^{\tau_2} g(s)ds.$$

The lower and upper averages of $g(t)$, denoted by $m[g]$ and $M[g]$ are defined by

$$m[g] = \lim_{s \to \infty} \inf \{A[g, \tau_1, \tau_2] \,|\, \tau_2 - \tau_1 \geq s\},$$
$$M[g] = \lim_{s \to \infty} \sup \{A[g, \tau_1, \tau_2] \,|\, \tau_2 - \tau_1 \geq s\}.$$

In our subsequent analysis, we shall assume that the functions b_i and a_{ij}, $i, j = 1, 2, ..., n$, are continuous on R_+, $a_{ij} \geq 0$, $a_{ij}^M < \infty$, $b_i^M < \infty$, $b_i^L > 0$, and $a_{ii}^L > 0$ for $i = 1, 2, ..., n$.

Furthermore, in order to restrict our attention only to those solutions which evolve in the phase space $\{x \in R_+^n : \ x_i > 0, \ i = 1, 2, ..., n\}$, we also shall assume that the functions P_{ik} are continuous on R_+, and $x_i + P_{ik}(x_i) > 0$ for $x_i > 0$, $i = 1, 2, ..., n$, $k = 1, 2, ...$ This restriction prevents the instantaneous extinction of any population x_i at an impulse time t_k. We point out that efficient sufficient conditions which guarantee the positivity of the solutions of such systems are given in [2].

Ahmad and Lazer [1] proved that, if for $i = 1, ..., n$,

$$m[b_i] > \sum_{\substack{i=1 \\ i \neq j}}^{n} \frac{a_{ij}^M}{a_{jj}^L} M[b_j], \tag{A}$$

then for any solution $x(t) = col(x_1(t), ..., x_n(t))$ of the corresponding system to system (8) without impulses (i.e. with $x_i(t_k + 0) = x_i(t_k)$, $i = 1, ..., n$, $k = 1, 2, ...$) if $x_i(0) > 0$, $i = 1, ..., n$, then:

$$0 < \inf_{t \geq 0} x_i(t) < \sup_{t \geq 0} x_i(t) < \infty.$$

LEMMA 4.1. *Assume that the condition (A) is satisfied and the functions P_{ik} are such that*

$$-x_i \le P_{ik}(x_i) \le 0 \quad for \quad x_i \in R_+, \ \ i = 1, 2, ..., n, \ \ k = 1, 2, ...$$

Then there exist positive constants r and R such that

$$r \le x_i(t) \le R, \quad t \in [0, \infty). \tag{10}$$

Proof. From corresponding theorem for the continuous case ([1]), it follows that for all $t \in [0, t_1] \cup (t_k, t_{k+1}]$, $k = 1, 2, ...$ and $1 \le i \le n$ there exist positive constants r_i^* and R_i^* such that the following inequalities hold:

$$r_i^* \le x_i(t) \le R_i^*.$$

Using the positivity of the solutions and the condition of Lemma 4.1, we obtain

$$0 < x_i(t_k + 0) = x_i(t_k) + P_{ik}(x_i(t_k)) \le x_i(t_k) \le R_i^*.$$

Therefore, there exist positive constants r and R such that the inequalities (10) are valid. □

Next, we will give sufficient conditions for the uniform stability and uniform asymptotic stability of the equilibrium states of (8). The problems of existence and uniqueness of equilibria of Lotka-Volterra systems with or without impulses have been investigated by many authors. Some sufficient conditions for impulsive models are given in [2, 3, 13].

THEOREM 4.2. *Assume that:*

1. *The assumptions of Lemma 4.1 holds.*
2. $r \le x_i + P_{ik}(x_i) \le x_i \le R \ \ for \ \ r \le x_i \le R, \ \ i = 1, 2, ..., n, \ \ k = 1, 2, ...$
3. *The following inequalities are valid*

$$a_{jj}(t) \ge \sum_{\substack{i=1 \\ i \ne j}}^{n} a_{ij}(t), \quad t \ne t_k, \quad k = 1, 2, ...$$

Then the equilibrium x^* of system (8) is uniformly stable.

Proof. Define a Lyapunov function

$$V(t, x, x^*) = \sum_{i=1}^{n} \left| \ln \frac{x_i}{x_i^*} \right|. \tag{11}$$

By Mean Value Theorem and by (10), it follows that for any closed interval contained in $[0, t_1] \cup (t_k, t_{k+1}]$, $k = 1, 2, ...$ and for all $i = 1, 2, ...$

$$\frac{1}{R}|x_i(t) - x_i^*| \leq |\ln x_i(t) - \ln x_i^*| \leq \frac{1}{r}|x_i(t) - x_i^*|. \tag{12}$$

For $t > 0$ and $t = t_k$, $k = 1, 2, ...$, we have

$$\begin{aligned} V(t_k + 0, x(t_k + 0), x^*(t_k + 0)) &= \sum_{i=1}^{n} \left| \ln \frac{x_i(t_k + 0)}{x_i^*(t_k + 0)} \right| \\ &= \sum_{i=1}^{n} \left| \ln \frac{x_i(t_k) + P_{ik}(x_i(t_k))}{x_i^*(t_k)} \right| \\ &\leq \sum_{i=1}^{n} \left| \ln \frac{x_i(t_k)}{x_i^*(t_k)} \right| = V(t_k, x(t_k), x^*(t_k)). \end{aligned} \tag{13}$$

Consider the upper right-hand derivative $D^+_{(8)}V(t, x(t), x^*)$ of the function $V(t, x(t), x^*)$ with respect to system (8). For $t \geq 0$ and $t \neq t_k$, $k = 1, 2, ...$, we derive the estimate

$$D^+_{(8)}V(t, x(t), x^*) = \sum_{i=1}^{n} \frac{\dot{x}_i(t)}{x_i(t)} sgn\,(x_i(t) - x_i^*)\,.$$

Since x^* is the equilibrium of (8) and $b_i(t) = a_{ii}(t)x_i^* + \sum\limits_{\substack{j=1 \\ j \neq i}}^{n} a_{ij}(t)x_j^*$, then

$$D^+_{(8)}V(t, x(t), x^*) \leq \sum_{j=1}^{n} \left[-a_{jj}(t)|x_j(t) - x_j^*| + \sum_{\substack{i=1 \\ i \neq j}}^{n} a_{ij}(t)|x_j(t) - x_j^*| \right].$$

Thus in view of condition 3 of Theorem 4.2, we obtain

$$D^+_{(8)}V(t, x(t), x^*) \leq 0,$$

$t \geq 0$ and $t \neq t_k$, $k = 1, 2, ...$

Since all conditions of Theorem 3.1 hold, then the equilibrium x^* of system (8) is uniformly stable. □

THEOREM 4.3. *In addition to the assumptions of Theorem 4.2, suppose there exists a nonnegative constant μ such that*

$$a_{jj}(t) \geq \mu + \sum_{\substack{i=1 \\ i \neq j}}^{n} a_{ij}(t), \quad t \neq t_k, \quad k = 1, 2, ... \tag{14}$$

Then the equilibrium x^* of system (8) is uniformly asymptotically stable.

Proof. We consider again the Lyapunov function (11). From (13) and (14), we obtain

$$D^+_{(8)}V(t,x(t),x^*) \le -\mu\sum_{i=1}^{n}|x_i(t)-x_i^*(t)|,$$

$t \ge 0$ and $t \ne t_k$, $k = 1, 2, ...$

Since all conditions of Theorem 3.2 are satisfied, the solution x^* of system (8) is uniformly asymptotically stable. □

In order to illustrate some features of our results, in the following we will apply Theorem 4.3 to two-dimensional systems, which have been studied extensively in the literature.

EXAMPLE 4.4. *For the system*

$$\begin{cases}\dot{x}(t) = x(t)\left[8 - 14x(t) - y(t)\right],\\ \dot{y}(t) = y(t)\left[15 - 4x(t) - 13y(t)\right],\end{cases} \tag{15}$$

one can show that the point $(x^*, y^*) = (\frac{1}{2}, 1)$ *is an equilibrium which is uniformly asymptotically stable [1].*

Now, we consider the impulsive Lotka-Volterra system

$$\begin{cases}\dot{x}(t) = x(t)\left[8 - 14x(t) - y(t)\right],\ t \ne t_k,\\ \dot{y}(t) = y(t)\left[15 - 4x(t) - 13y(t)\right],\ t \ne t_k,\\ \Delta x(t_k) = -\dfrac{1}{3}\Big(x(t_k) - \dfrac{1}{2}\Big),\ k = 1, 2, ...,\\ \Delta y(t_k) = -\dfrac{3}{5}\Big(y(t_k) - 1\Big),\ k = 1, 2, ...,\end{cases} \tag{16}$$

where $0 < t_1 < t_2 < ...$ *and* $\lim\limits_{k\to\infty} t_k = \infty$.

For the system (16)*, the point* $(x^*, y^*) = (\frac{1}{2}, 1)$ *is an equilibrium and all conditions of Theorem 4.3 are satisfied. In fact, for* $\mu \le 10$, $r = \frac{1}{2}$ *and* $R = 1$, *we have*

$$\begin{aligned}\frac{1}{2} &\le \frac{3x(t_k)+1}{6} = x(t_k) + P_{1k}(x(t_k))\\ &= x(t_k) - \frac{1}{3}\Big(x(t_k) - \frac{1}{2}\Big) = \frac{2}{3}\Big(x(t_k) - \frac{1}{2}\Big) + \frac{1}{2} \le x(t_k) \le 1,\end{aligned}$$

$$\begin{aligned}\frac{1}{2} &\le \frac{2y(t_k)+3}{5} = y(t_k) + P_{2k}(y(t_k))\\ &= y(t_k) - \frac{3}{5}\Big(y(t_k) - 1\Big) = \frac{2}{5}\Big(y(t_k) - 1\Big) + 1 \le y(t_k) \le 1,\end{aligned}$$

for $\frac{1}{2} \le x(t_k) \le 1$, $\frac{1}{2} \le y(t_k) \le 1$, $k = 1, 2, ...$

Therefore, the equilibrium $(x^*, y^*) = (\frac{1}{2}, 1)$ *is uniformly asymptotically stable.*

If, in the system (16)*, we consider the impulsive perturbations of the form:*

$$\begin{cases} \Delta x(t_k) = -3\Big(x(t_k) - \dfrac{1}{2}\Big), \ k = 1, 2, ..., \\ \Delta y(t_k) = -\dfrac{3}{5}\Big(y(t_k) - 1\Big), \ k = 1, 2, ..., \end{cases}$$

then the point $(x^*, y^*) = (\frac{1}{2}, 1)$ *is again an equilibrium, but there is nothing we can say about its uniform asymptotic stability, because for* $\frac{1}{2} \le x(t_k) \le 1$*, we have* $-\dfrac{1}{2} \le x(t_k) + P_{1k}(x(t_k)) \le \dfrac{1}{2}$, $k = 1, 2, ...$

The example shows that by means of appropriate impulsive perturbations we can control the system's population dynamics. We can see that impulses are used to keep the stability properties of the system.

Example 4.5. *The system*

$$\begin{cases} \dot{x}(t) = x(t)\left[2 - 6x(t) - y(t)\right], \\ \dot{y}(t) = y(t)\left[3 - 2x(t) - 5y(t)\right]. \end{cases} \tag{17}$$

has a boundary equilibrium point $(x^*, y^*) = (\frac{1}{3}, 0)$*. We point out that efficient sufficient conditions which guarantee the stability of such solutions of predator-prey systems are given in [7, 10].*

However, for the impulsive Lotka-Volterra system

$$\begin{cases} \dot{x}(t) = x(t)\left[2 - 6x(t) - y(t)\right], \ t \ne t_k, \\ \dot{y}(t) = y(t)\left[3 - 2x(t) - 5y(t)\right], \ t \ne t_k, \\ \Delta x(t_k) = -\dfrac{1}{2}\Big(x(t_k) - \dfrac{1}{4}\Big), \ k = 1, 2, ..., \\ \Delta y(t_k) = -\dfrac{1}{3}\Big(y(t_k) - \dfrac{1}{2}\Big), \ k = 1, 2, ..., \end{cases}$$

where $0 < t_1 < t_2 < ...$ *and* $\lim\limits_{k\to\infty} t_k = \infty$*, the point* $(x^*, y^*) = (\frac{1}{4}, \frac{1}{2})$ *is an equilibrium which is uniformly asymptotically stable. In fact, all conditions of*

Theorem 4.3 are satisfied for $\mu \le 3$, $r = \frac{1}{4}$, $R = \frac{1}{2}$ *and*

$$\begin{aligned}\frac{1}{4} &\le \frac{4x(t_k)+1}{8} = x(t_k) + P_{1k}(x(t_k)) \\ &= x(t_k) - \frac{1}{2}\Big(x(t_k) - \frac{1}{4}\Big) = \frac{1}{2}\Big(x(t_k) - \frac{1}{4}\Big) + \frac{1}{4} \le x(t_k) \le \frac{1}{2},\end{aligned}$$

$$\begin{aligned}\frac{1}{4} &\le \frac{4y(t_k)+1}{6} = y(t_k) + P_{2k}(y(t_k)) \\ &= y(t_k) - \frac{1}{3}\Big(y(t_k) - \frac{1}{2}\Big) = \frac{2}{3}\Big(y(t_k) - \frac{1}{2}\Big) + \frac{1}{2} \le y(t_k) \le \frac{1}{2},\end{aligned}$$

for $\frac{1}{4} \le x(t_k) \le \frac{1}{2}$, $\frac{1}{4} \le y(t_k) \le \frac{1}{2}$, $k = 1, 2, ...$

This shows that the impulsive perturbations can prevent the population from going extinct.

References

[1] S. Ahmad and A.C. Lazer, *Average conditions for global asymptotic stability in a nonautonomous Lotka-Volterra system*, Nonlinear Anal. **40** (2000), 37–49.

[2] S. Ahmad and I.M. Stamova, *Asymptotic stability of an N-dimensional impulsive competitive system*, Nonlinear Anal. Real World Appl. **8** (2007), 654–663.

[3] S. Ahmad and I.M. Stamova, *Asymptotic stability of competitive systems with delays and impulsive perturbations*, J. Math. Anal. Appl. **334** (2007), 686–700.

[4] J. O. Alzabut, G. T. Stamov and E. Sermutlu, *On almost periodic solutions for an impulsive delay logarithmic population model*, Math. Comput. Modelling **51** (2010), 625–631.

[5] G. Ballinger and X. Liu, *Permanence of population growth models with impulsive effects*, Math. Comput. Modelling **26** (1997), 59–72.

[6] A. Battauz and F. Zanolin, *Coexistence states for periodic competitive Kolmogorov systems*, J. Math. Anal. Appl. **219** (1998), 179–199.

[7] L. Dong, L. Chen and L. Sun, *Extinction and permanence of the predator-prey system with stocking of prey and harvesting of predator impulsively*, Math. Methods Appl. Sci. **29** (2006), 415–425.

[8] T. Faria, *An asymptotic stability result for delayed population model*, Proc. Amer. Math. Soc. **132** (2003), 1163–1169.

[9] H.I. Freedman, *A perturbed Kolmogrov-type model for the growth problem*, Math. Biosci. **12** (1975), 721–732.

[10] B. S. Goh, *Global stability in two species interactions*, J. Math. Biol. **3** (1976), 313-318.

[11] V. Lakshmikantham, D.D. Bainov and P.S. Simeonov, *Theory of Impulsive Differential Equations*, World Scientific, Singapore, 1989.

[12] J. Petela, *Average conditions for Kolmogorov systems*, Appl. Math. Comput. **215** (2009), 481–494.

[13] I. M. Stamova, *Stability Analysis of Impulsive Functional Differential Equations*, Walter de Gruyter, Berlin, New York, 2009.

[14] B. Tang and Y. Kuang, *Permanence in Kolmogorov-type systems of nonautonomous functional differential equations*, J. Math. Anal. Appl. **197** (1996), 427–447.
[15] Z. Teng, L. Nie and X. Fang, *The periodic solutions for general periodic impulsive population systems of functional differential equations and its applications,* Comput. Math. Appl. **61** (2011), 2690–2703.
[16] A. Tineo, *Persistence of a class of periodic Kolmogorov systems*, J. Math. Anal. Appl. **246** (2000), 89–99.
[17] Y. Xiao, D. Chen and H. Qin, *Optimal impulsive control in periodic ecosystem*, Systems Control Lett. **55** (2006), 558–565.
[18] R.R. Vance and E.A. Coddington, *A nonautonomous model of population growth*, J. Math. Biol. **27** (1989), 491-506.
[19] F. Zanolin, *Continuation theorems for the periodic problem via the translation operator*, Rend. Sem. Mat. Univ. Politec. Torino **54** (1996), 1–23.
[20] F. Zanolin, *Permanence and positive periodic solutions for Kolmogorov competing species systems*, Results Math. **21** (1992), 224–250.
[21] L. Zhang, Z. Teng and H. Jiang, *Permanence for general nonautonomous impulsive population systems of functional differential equations and its applications*, Acta Appl. Math. **110** (2010), 1169–1197.

Authors' addresses:

Shair Ahmad
Department of Mathematics
University of Texas at San Antonio
One UTSA Circle, San Antonio TX 78249, USA
E-mail: shair.ahmad@utsa.edu

Ivanka Stamova
Department of Mathematics
University of Texas at San Antonio
One UTSA Circle, San Antonio TX 78249, USA
E-mail: ivanka.stamova@utsa.edu

Received March 3, 2012
Revised April 30, 2012

Rend. Istit. Mat. Univ. Trieste
Volume 44 (2012), 33–44

Index and persistence of stable Cantor sets[1]

Rafael Ortega, Alfonso Ruiz-Herrera

Dedicated to Professor Fabio Zanolin on the occasion of his sixtieth birthday

Abstract. *A theorem by Bell and Meyer says that a stable and transitive Cantor set in the plane can be approximated by periodic points. We prove that the periodic points can be chosen with index one. As a consequence these Cantor sets are always persistent invariant sets.*

Keywords: Lyapunov stability, Cantor set, fixed point index, translation arc
MS Classification 2010: 37E30

1. Introduction

Cantor sets often appear as invariant sets of planar homeomorphisms. Well known examples are the Bernoulli shift in Smale's horseshoe, Aubry-Mather sets in non-integrable twist maps or adding machines obtained as sections of a solenoid. Some concrete constructions can be found in [1, 3, 6]. In general we will consider a homeomorphism $h : \mathbb{R}^2 \longrightarrow \mathbb{R}^2$ and a Cantor set $\Lambda \subset \mathbb{R}^2$ with

$$h(\Lambda) = \Lambda.$$

In this paper homeomorphisms are understood as surjective maps, so that $h(\mathbb{R}^2) = \mathbb{R}^2$. Also, to avoid trivialities, it will be assumed that Λ is transitive. This means that for some $p \in \Lambda$,

$$L_\omega(p, h) = \Lambda,$$

where $L_\omega(p, h)$ is the corresponding ω-limit set. A Cantor set is a compact, perfect and totally disconnected metric space. All Cantor sets are homeomorphic but they can support many different transitive dynamics. In the examples mentioned above one can find chaos, Denjoy dynamics or almost-periodicity.

[1]Supported by the research project MTM2011-23652, Spain

An invariant set $\Lambda \subset \mathbb{R}^2$ is stable (in the sense of Lyapunov) if each neighborhood U of Λ contains another neighborhood V such that

$$h^n(V) \subset U \quad \text{for every } n \geq 1.$$

In [2], Bell and Meyer obtained a remarkable result: in the plane, stable Cantor sets are never isolated, in fact they can be approximated by periodic points lying outside Λ. The purpose of our paper is to prove that these periodic points have non-zero index. Here we refer to the fixed point index that can be expressed in terms of Brouwer's degree. As a consequence we will prove that stable Cantor sets are persistent as invariant sets. An invariant compact set Λ is persistent if, given any positive $\varepsilon > 0$, there exists $\delta > 0$ such that for any homeomorphism $\widetilde{h} : \mathbb{R}^2 \longrightarrow \mathbb{R}^2$ with

$$\|h(x) - \widetilde{h}(x)\| \leq \delta$$

for each $x \in \mathbb{R}^2$, there exists a compact set $\widetilde{\Lambda} \subset \mathbb{R}^2$ such that

$$\widetilde{h}(\widetilde{\Lambda}) = \widetilde{\Lambda} \quad \text{and} \quad D_H(\Lambda, \widetilde{\Lambda}) \leq \epsilon.$$

The symbol D_H refers to the Hausdorff distance between compact subsets of the plane. In our result, $\widetilde{\Lambda}$ will be composed by periodic points derived from the properties of degree. Summing up we can say that stable Cantor sets in the plane are simultaneously non-isolated and persistent. This is in contrast with the properties enjoyed by stable finite sets. At the end of the paper we will present an example of a fixed point that is stable and non-persistent. The structure of the paper is as follows. The main theorem on index and a corollary on persistence are stated in Section 2. The proofs of both results are presented in Section 3. Finally, in Section 4 we discuss some connections with the literature. To finish this introduction we notice that an example constructed in [2] shows that our results do not admit a direct extension to higher dimensions.

2. Main results

Given a Jordan curve $\Gamma \subset \mathbb{R}^2$, the bounded component of $\mathbb{R}^2 \backslash \Gamma$ will be indicated by $\widehat{\Gamma}$. Brouwer's degree in the plane will be denoted by $d[f, G, 0]$ where $G \subset \mathbb{R}^2$ is a bounded and open set and $f : cl(G) \longrightarrow \mathbb{R}^2$ is a continuous function defined on the closure of G. We must also assume that f does not vanish on ∂G, the boundary of G. We recall two properties of the degree that will be employed later,

i) existence of zeros: the function f has at least one zero on G if $d[f, G, 0] \neq 0$,

ii) continuity of the degree: there exists $\eta > 0$, depending on f, such that if $g : cl(G) \longrightarrow \mathbb{R}^2$ is a continuous function with

$$\|f(x) - g(x)\| \leq \eta$$

for each $x \in \partial G$, then g does not vanish on ∂G and $d[g, G, 0] = d[f, G, 0]$.

We refer to [10] for more information on degree theory. Given a continuous function $\phi : cl(G) \longrightarrow \mathbb{R}^2$, the fixed point index is defined as the degree of $f = id - \phi$. The zeros of f are precisely the fixed points of ϕ.
We will prove that the existence of a stable Cantor set has strong consequences on the fixed point index of the map $h^N = h \circ \overset{(N)}{\cdots} \circ h$. Notice that the fixed points of h^N are the periodic points of h whose minimal period is a divisor of N.

THEOREM 2.1. *Assume that $h : \mathbb{R}^2 \longrightarrow \mathbb{R}^2$ is a homeomorphism and Λ is an invariant Cantor set that is stable and has a transitive point. Then for every $\delta > 0$ and $p \in \Lambda$ there exist a Jordan curve $\Gamma = \Gamma(\delta, p)$ and an integer $N = N(\delta, p) \geq 1$ such that the following properties hold,*

$$D_H(\Gamma, \{p\}) \leq \delta, \;\; h^N(x) \neq x \;\text{ if } x \in \Gamma, \;\; d[id - h^N, \widehat{\Gamma}, 0] = 1.$$

The existence property of the degree implies that each region $\widehat{\Gamma}(\delta, p)$ contains a periodic point. This implies that Λ can be obtained as a limit of periodic points.

THEOREM 2.2. *(Bell and Meyer) In the assumptions of Theorem 2.1 and given $p \in \Lambda$, there exist a sequence of points $\{x_n\}$ in $\mathbb{R}^2$ and integers $\sigma(n) \geq 1$ such that*

$$x_n \longrightarrow p \;\text{ and }\; h^{\sigma(n)}(x_n) = x_n.$$

The persistence of Λ will be deduced from the continuity of the degree.

COROLLARY 2.3. *In the assumptions of Theorem 2.1, the set Λ is persistent.*

3. Proofs

The proof by Bell and Meyer in [2] is based on a well known fixed point theorem due to Cartwright and Littlewood. This theorem deals with orientation preserving homeomorphisms and it has been extended to the orientation reversing case by Bell. We will employ a strategy similar to that in [2] but without making use of this fixed point theorem. Instead we will use the following result which is a consequence of Brouwer's theory on translations arcs.

LEMMA 3.1. *Assume that $\Omega \subset \mathbb{R}^2$ is an open and simply connected set and let $H : \Omega \longrightarrow \Omega$ be an orientation preserving embedding. In addition, assume that H has a recurrent point that is not fixed. Then there exists a Jordan curve $\Gamma \subset \Omega$ such that $H(x) \neq x$ if $x \in \Gamma$ and*

$$d[id - H, \widehat{\Gamma}, 0] = 1.$$

Let us recall that an embedding is a continuous and one-to-one map. In contrast to homeomorphisms, embeddings are not necessarily onto, that is $H(\Omega) \subset \Omega$. For this reason, orbits are well defined for the future but not necessarily for the past. The embedding is orientation-preserving if

$$d[H, B, y] = 1,$$

where y is any point in $H(\Omega)$ and B is an open ball centered at $H^{-1}(y)$. Given any embedding H, the second power $H^2 = H \circ H$ is always orientation-preserving. This is well known and follows from the properties of the degree of a composition of maps, see for instance [10].
By a recurrent point $x_* \in \Omega$ we mean a point such that $H^{\sigma_n}(x_*) \to x_*$ for some increasing sequence of positive integers $\{\sigma_n\}$. Notice that the sequence $\{H^n(x_*)\}_{n\geq 0}$ could be unbounded.

Proof of Lemma 3.1. This is a well known result and we refer to [4, 8, 9] for the case of homeomorphisms. The proof for the case of embeddings is similar. We sketch it. Since Ω is homeomorphic to $\mathbb{R}^2$ we can restrict to the case $\Omega = \mathbb{R}^2$. For this reduction we are using the invariance of the fixed point index under topological conjugation. This is again a consequence of the properties of the degree of a composition.
Let C be a connected component of $\mathbb{R}^2 \setminus Fix(H)$ containing the recurrent point x_*. We can find a small and closed disk D centered at x_* and such that $D \subset C$ and $D \cap H(D) = \emptyset$. This is possible because x_* is not fixed. From [15, Chapter 3, Proposition 20] we know that $H(D)$ is contained in C. The recurrence of x_* allows us to obtain an integer $\sigma \geq 2$ such that $y_* = H^\sigma(x_*)$ belongs to the interior of D. The points x_* and y_* lie on D and so it is possible to apply [15, Chapter 3, Proposition 17] to deduce the existence of a translation arc α containing x_* and y_*. In consequence, y_* belongs to $\alpha \cap H^\sigma(\alpha)$ and Brouwer's Arc Translation Lemma is applicable. An adaptation to embeddings of the proof by Brown of this lemma can be found in [15]. □

We will also use the following result on minimal homeomorphisms.

LEMMA 3.2. *Assume that K is a compact metric space and $\phi : K \longrightarrow K$ is a minimal homeomorphism. Then, for each integer $N \geq 1$, the set*

$$\mathcal{R}_N = \{k \in K : k \in L_\omega(k, \phi^N)\}$$

is dense in K.

We recall that ϕ is minimal if every point is transitive; that is, $L_\omega(k,\phi) = K$ for each $k \in K$.

Proof. First of all we prove that $\mathcal{R}_N$ is non-empty. The existence of minimal sets for general homeomorphisms implies that there exists a non-empty compact set $M \subset K$ that is minimal for ϕ^N. This means that $\phi^N(M) = M$ and if N is a compact subset of M with $\phi^N(N) = N$ then either $N = \emptyset$ or $N = M$. In particular, the set $L_\omega(m, \phi^N)$ has to coincide with M for each $m \in M$. This implies that M is contained in $\mathcal{R}_N$. The second observation is that $\mathcal{R}_N$ is invariant under ϕ. This is easily checked and leads to the identity $\phi(cl(\mathcal{R}_N)) = cl(\mathcal{R}_N)$. The minimality of ϕ implies that $cl(\mathcal{R}_N) = K$. □

We need two more lemmas. The setting and the assumptions correspond to those of the main theorem.

LEMMA 3.3. *The restricted homeomorphism $h_\Lambda : \Lambda \longrightarrow \Lambda$ is minimal.*

Proof. This is a particular case of [5, Lemma 2] but we present the proof for completeness. Assume by contradiction that h is not minimal on Λ. Then there exists a point $p \in \Lambda$ such that the limit set $L_\omega(p,h)$ is a proper subset of Λ. Let us fix another point $q \in \Lambda \setminus L_\omega(p,h)$. The compact sets $L_\omega(p,h)$ and $\{q\}$ can be separated by two open sets U and V of $\mathbb{R}^2$. Since Λ is totally disconnected they can be chosen so that

- $\Lambda \subset U \cup V$,
- $cl(V) \cap cl(U) = \emptyset$,
- $L_\omega(p,h) \subset U$,
- $q \in V$.

Let V_* be the connected component of V containing q. Notice that this is also a component of the larger set $U \cup V$. The stability of Λ implies the existence of an open set $W \subset \mathbb{R}^2$ satisfying that

$$\Lambda \subset W \subset U \cup V, \;\; h^n(W) \subset U \cup V$$

for each $n \geq 2$. Let W_* be the connected component of W containing p. By assumption we know that Λ contains a transitive point. All the points in the orbit will be transitive and therefore we know that transitive points are dense in Λ. Let $r \in \Lambda$ be a transitive point close enough to p in order to guarantee that $r \in W_*$. Let (σ_n) be an increasing sequence of positive integers with $h^{\sigma_n}(r) \longrightarrow q$. This implies that $h^{\sigma_n}(r)$ belongs to V_* for large n and so $h^{\sigma_n}(W_*) \cap V_* \neq \emptyset$. Since $h^{\sigma_n}(W_*)$ is a connected subset of $U \cup V$ we conclude that it must be contained in one component. Hence $h^{\sigma_n}(W_*) \subset V_*$. Finally, we

observe that the iterates $h^{\sigma_n}(p)$ belong to $h^{\sigma_n}(W_*) \subset V_*$ and therefore $L_\omega(p,h)$ has to contain a point in $cl(V_*)$. This is a contradiction with the conditions imposed on U and V. □

The last lemma needs some preliminary remarks on the topology of $\mathbb{R}^2$. Given an open set G in $\mathbb{R}^2$, the set $\widehat{G} \subset \mathbb{R}^2$ is the smallest open and simply connected set containing G. We refer to [14] for an elementary construction of this set. In [2], this set $\widehat{G}$ is called the topological hull of G. In fact its construction is purely topological and this explains the property $h(\widehat{G}) = \widehat{h(G)}$.

LEMMA 3.4. *Given a point $p \in \Lambda$ and a disk D centered at p, there exists an integer $N \geq 1$ and an open and simply connected domain $\Omega \subset \mathbb{R}^2$ satisfying that*

$$p \in \Omega \subset D, \quad h^N(\Omega) \subset \Omega.$$

Proof. Since Λ is totally disconnected it is possible to find open sets A and B in $\mathbb{R}^2$ satisfying that

$$p \in A \subset int(D),$$
$$\Lambda \subset A \cup B,$$
$$cl(A) \cap cl(B) = \emptyset.$$

The open set $A \cup B$ is a neighborhood of Λ and the stability of this set implies the existence of another open set $V \subset \mathbb{R}^2$ with $\Lambda \subset V \subset A \cup B$ and $h^n(V) \subset A \cup B$ if $n \geq 1$. Define $W = \bigcup_{n \geq 0} h^n(V)$. This is also a neighborhood of Λ satisfying

$$\Lambda \subset W \subset A \cup B \text{ and } h^n(W) \subset W \text{ if } n \geq 1.$$

Let G be the connected component of W containing p. This component has to be contained in A, and hence in D. In consequence $\widehat{G}$ is also contained in D. We know by Lemma 3.3 that the limit set $L_\omega(p,h)$ is the whole Cantor set Λ. From here we deduce that $p \in L_\omega(p,h)$ and there exists an integer $N \geq 1$ such that $h^N(p)$ belongs to G. This implies that $G \cap h^N(G) \neq \emptyset$. But $h^N(G)$ is a connected set inside W and so it must be contained in one component of W. This component is obviously G. From $h^N(G) \subset G$ we obtain that $h^N(\widehat{G}) = \widehat{h^N(G)} \subset \widehat{G}$ and the set $\widehat{G}$ is the searched domain Ω. □

Proof of Theorem 2.1. We fix $p \in \Lambda$ and a disk D of radius $\delta > 0$. From Lemma 3.4 we obtain a simply connected domain $\Omega \subset \mathbb{R}^2$ and an integer $N \geq 1$ with

$$p \in \Omega \subset D, \quad h^N(\Omega) \subset \Omega.$$

Consider the orientation preserving embedding $H = h^{2N} : \Omega \longrightarrow \Omega$. We know from Lemmas 3.3 and 3.2 that the set

$$\mathcal{R}_{2N} = \{q \in \Lambda : q \in L_\omega(q, h^{2N})\}$$

is dense in Λ. In consequence we can find a point lying in $\Omega \cap \mathcal{R}_{2N}$. This point is recurrent for H and Lemma 3.1 applies. □

Proof of Corollary 2.3. We fix $\varepsilon > 0$. The stability of Λ as an invariant set of h guarantees the existence of $\delta_* > 0$ such that

$$dist(x, \Lambda) \leq \delta_* \Longrightarrow dist(h^i(x), \Lambda) \leq \frac{\varepsilon}{2}$$

for each $i \geq 0$. In particular, $\delta_* \leq \frac{\varepsilon}{2}$. Since Λ is compact it can be covered by a finite number of open balls $B_1, ..., B_k$ of radius δ_* and centered at points $p_1, ..., p_k$ lying in Λ. Next we apply Theorem 2.1 at each p_i to find Jordan curves Γ_1,...,Γ_k and integers $N_1, ..., N_k \geq 1$ such that $\Gamma_j \subset B_j$ and $d[id - h^{N_j}, \widehat{\Gamma}_j, 0] = 1$, $j = 1, ..., k$. Define $K = \bigcup_{j=1}^{k} (\Gamma_j \cup \widehat{\Gamma}_j)$ and $N = \max\{N_1, ..., N_k\}$.
We consider the family $\mathcal{F}_1$ composed by homeomorphisms $\widetilde{h} : \mathbb{R}^2 \longrightarrow \mathbb{R}^2$ satisfying

$$\|h - \widetilde{h}\|_\infty := \sup_{x \in \mathbb{R}^2} \|h(x) - \widetilde{h}(x)\| \leq 1.$$

We need some properties of the iterates of $\widetilde{h}$ which are common to the whole family $\mathcal{F}_1$.
Claim 1: There exists a compact set $K_* \subset \mathbb{R}^2$ such that

$$\widetilde{h}^i(K) \subseteq K_*$$

for all $i = 0, 1, ..., N$ and for each $\widetilde{h} \in \mathcal{F}_1$.
Let $C_0 > 0$ be a large number so that K is contained in the ball of radius C_0 centered at the origin. By induction, we define

$$C_{i+1} = 1 + \max_{\|x\| \leq C_i} \|h(x)\|, \quad i \geq 0.$$

We claim that

$$\|\widetilde{h}^i(x)\| \leq C_i \quad \text{if} \quad x \in K.$$

Indeed, using the induction method,

$$\begin{aligned} \|\widetilde{h}^{i+1}(x)\| &\leq \|\widetilde{h}(\widetilde{h}^i(x)) - h(\widetilde{h}^i(x))\| + \|h(\widetilde{h}^i(x))\| \\ &\leq \|\widetilde{h} - h\|_\infty + \max_{\|x\| \leq C_i} \|h(x)\|. \end{aligned}$$

Claim 2: Given $\Delta > 0$ there exists $\delta_2 > 0$ such that $\widetilde{h} \in \mathcal{F}_1$ and $\|h - \widetilde{h}\|_\infty \leq \delta_2$ implies that $\|h^i(x) - \widetilde{h}^i(x)\| \leq \Delta$ if $x \in K$, $i = 1, ..., N$.
In view of Claim 1 we can find a modulus of continuity for h on K_*. This means a function $\omega : [0, \infty[\longrightarrow \mathbb{R}$ with $\lim_{r \to 0^+} \omega(r) = 0$ and

$$\|h(x) - h(y)\| \leq \omega(\|x - y\|) \quad \text{if } x, y \in K_*.$$

Define $D_i = \max_{x\in K} \|\widetilde{h}^i(x) - h^i(x)\|$. Then, by induction, we prove that

$$D_{i+1} \leq \|\widetilde{h} - h\|_\infty + \omega(D_i), \;\; i = 1, ..., N-1$$

and the claim follows easily. Notice that

$$\|\widetilde{h}^{i+1}(x) - h^{i+1}(x)\| \leq \|\widetilde{h}(\widetilde{h}^i(x)) - h(\widetilde{h}^i(x))\| + \|h(\widetilde{h}^i(x)) - h(h^i(x))\|.$$

After these claims we are ready to prove the existence of $\widetilde{\Lambda}$. First we apply the continuity of the degree to find positive numbers $\eta_1, ..., \eta_k$ such that if

$$\|h^{N_j}(x) - \widetilde{h}^{N_j}(x)\| \leq \eta_j, \;\; x \in \Gamma_j,$$

then

$$d[id - \widetilde{h}^{N_j}, \widehat{\Gamma}_j, 0] = d[id - h^{N_j}, \widehat{\Gamma}_j, 0] = 1.$$

Next we apply Claim 2 with $\Delta = \min\{\frac{\epsilon}{2}, \eta_1, ..., \eta_k\}$ and find $\delta_2 \in]0, \delta_*[$ such that the conclusion of the claim holds if $\|h - \widetilde{h}\|_\infty \leq \delta_2$. The existence property of the degree allows us to select points $\widetilde{x}_j \in \widehat{\Gamma}_j$ such that $\widetilde{h}^{N_j}(\widetilde{x_j}) = \widetilde{x}_j$. The set

$$\widetilde{\Lambda} = \{\widetilde{h}^i(\widetilde{x}_j) : \; j = 1, ..., k, \; 0 \leq i < N_j\}$$

is finite and invariant under $\widetilde{h}$. It remains to prove that $D_H[\Lambda, \widetilde{\Lambda}] \leq \epsilon$. Assume first that p is a point in Λ. Since Λ is covered by $B_1, ..., B_k$ we find an index j such that $p \in B_j$. The ball B_j also contains the point $\widetilde{x}_j$. In consequence,

$$dist(p, \widetilde{\Lambda}) \leq \|p - \widetilde{x}_j\| \leq 2\delta_* \leq \epsilon.$$

Consider now a point in $\widetilde{\Lambda}$, say $\widetilde{h}^i(\widetilde{x}_j)$. From

$$dist(\widetilde{x}_j, \Lambda) \leq \|\widetilde{x}_j - p_j\| \leq \delta_*,$$

we deduce that

$$dist(h^i(\widetilde{x}_j), \Lambda) \leq \frac{\epsilon}{2}.$$

Hence, using Claim 2 and this estimate, if $\|h - \widetilde{h}\|_\infty \leq \delta_2$,

$$\begin{aligned} dist(\widetilde{h}^i(\widetilde{x}_j), \Lambda) &\leq \|\widetilde{h}^i(\widetilde{x}_j) - h^i(\widetilde{x}_j)\| + dist(h^i(\widetilde{x}_j), \Lambda) \\ &\leq \frac{\epsilon}{2} + \frac{\epsilon}{2}. \end{aligned}$$

□

4. Miscelaneous remarks

4.1. Invariant finite sets can be stable and non-persistent

A finite and invariant set Λ has to be composed by periodic points. We consider the simple case of a singleton $\Lambda = \{p\}$ and present an example of a stable fixed point that is not persistent as invariant set.
Consider the map

$$h : \mathbb{C} \longrightarrow \mathbb{C}$$

$$h(z) = z \exp\left(\frac{iy}{1+|z|^2}\right)$$

with $z = x + iy$. We have expressed it in complex notation but for many purposes it is more convenient the use of polar coordinates,

$$h : \begin{cases} \theta_1 = \theta + \frac{r}{1+r^2} \sin\theta\,, \\ r_1 = r\,. \end{cases}$$

It is not hard to prove that h is a real analytic diffeomorphism of the plane. We also observe that every disk of the type $|z| \leq constant$ is invariant under h and so the fixed point $z = 0$ is stable. An useful property of h is that $V(z) = \Re e\, z = x$ is a Lyapunov function. This means that

$$V(h(z)) \leq V(z)$$

for each $z \in \mathbb{C}$. Let us now consider the perturbed map $h_\varepsilon = T_\varepsilon \circ h$ where $T_\varepsilon(z) = z - \varepsilon$ is a horizontal translation with $\varepsilon > 0$. Again V is a Lyapunov function with

$$V(h_\varepsilon(z)) = V(h(z)) - \varepsilon \leq V(z) - \varepsilon.$$

More generally, if $n \geq 1$,

$$V(h_\varepsilon^n(z)) \leq V(z) - n\varepsilon$$

and so all the orbits for h_ε are unbounded. This shows that h_ε has no compact invariant sets. Since $\|h - h_\varepsilon\|_\infty = \varepsilon$, the maps h and h_ε are close and $\Lambda = \{0\}$ is not persistent.
Incidentally, we notice that the set of fixed points $Fix(h)$ is the real axis and so $z = 0$ is not an isolated fixed point. This is no surprise because stable fixed points are persistent as soon as they are isolated in $Fix(h)$. This is a consequence of the main result in [7]: if $h : \mathbb{R}^2 \longrightarrow \mathbb{R}^2$ is an orientation-preserving homeomorphism and $p = h(p)$ is a stable fixed point which is isolated in $Fix(h)$, then

$$d[id - h, \widehat{\Gamma}, 0] = 1$$

for each Jordan curve $\Gamma \subset \mathbb{R}^2$ with $\widehat{\Gamma} \cap Fix(h) = \{p\}$, $\Gamma \cap Fix(h) = \emptyset$. The case of orientation-reversing homeomorphisms was treated by Ruiz del Portal in [16].

4.2. Unstable Cantor sets can be isolated and non-persistent

With the help of a Denjoy homeomorphism on $\mathbb{S}^1$, it is possible to construct homeomorphisms $h : \mathbb{R}^2 \longrightarrow \mathbb{R}^2$ having a unique fixed point p_* and an invariant Cantor set Λ. In addition, the limit set of any point $x \in \mathbb{R}^2$ is either the fixed point, $L_\omega(x, h) = \{p_*\}$, or the Cantor set, $L_\omega(x, h) = \Lambda$. In particular, Λ is minimal. The details of the construction can be found in [11]. The map h has not periodic points and this implies that

$$d[id - h^N, \widehat{\Gamma}, 0] = 0$$

for any $N \geq 1$ and any Jordan curve $\Gamma \subset \mathbb{R}^2$ such that p_* lies in the exterior, that is, $p_* \notin \Gamma \cup \widehat{\Gamma}$. This example shows that the conclusion of Theorem 2.1 does not hold if we drop the stability assumption. In the example constructed in [11], the fixed point was placed at the origin, $p_* = 0$, and the Cantor set was inside the unit circumference, $\Lambda \subset \mathbb{S}^1$. Moreover the Euclidean norm $V(x) = \|x\|$ was a Lyapunov function satisfying

$$V(h(x)) < V(x)$$

if $x \in \mathbb{R}^2 \backslash (\Lambda \cup \{0\})$. Consider the perturbed homeomorphism $h_\varepsilon = D_\varepsilon \circ h$, with $\varepsilon > 0$ and

$$D_\varepsilon(x) = \begin{cases} (1-\varepsilon)x, & \text{if } \|x\| \leq 2\,; \\ (1 - 3\varepsilon + \varepsilon\|x\|)x, & \text{if } 2 \leq \|x\| \leq 3\,; \\ x, & \text{if } \|x\| \geq 3\,. \end{cases}$$

Then $\|h_\varepsilon - h\|_\infty = 2\varepsilon$ and

$$V(h_\varepsilon(x)) < V(x)$$

if $x \in \mathbb{R}^2 \backslash \{0\}$. La Salle's invariance principle implies that the origin is a global attractor for h_ε. This shows that Λ is not persistent.

The dynamics of h_Λ in the preceding example is of Denjoy type, a case that can be excluded if Λ is stable. The reason for this exclusion lies in a result by Buescu and Stewart [5] implying that stable Cantor sets are conjugate to adding machines. The family of adding machines is composed by certain explicit maps describing all possible almost periodic dynamics on a Cantor set. Denjoy dynamics is presented in [13] as the prototype of minimal dynamics that is not almost periodic and so it is not conjugate to an adding machine.

4.3. Adding machines cannot be isolated

In [17], Thomas obtained a result on the dynamics of solenoids in 3D flows that can be adapted to a 2D discrete setting for adding machines. Assume now that $h : \mathbb{R}^2 \longrightarrow \mathbb{R}^2$ is a C^1 diffeomorphism that is orientation-preserving and has an invariant Cantor set Λ such that h_Λ is almost periodic. Then it is possible to construct a T-periodic differential equation in the plane such that h is the Poincaré map. See [12] for an explicit construction. In this way, we obtain a $\mathcal{C}^1$ flow on the manifold $M = (\mathbb{R}/T\mathbb{Z}) \times \mathbb{R}^2$ and the results in [17] are applicable. The closure of the orbit starting at any point of Λ is a solenoid $S \subset M$ and [17, Theorem 3] implies that S is not isolated as an invariant set of the flow. The invariant sets accumulating on S must intersect the global section $M_0 = \{0\} \times \mathbb{R}^2$ and so Λ cannot be isolated as an invariant set of h. Notice that the result by Bell and Meyer does not follow from [5] and [17] because in principle one could find invariant sets without periodic points. The smoothness of h was needed in [17] to work with a smooth isolating block. At the end of that paper it is mentioned that the smoothness hypotheses can be weakened. It seems reasonable to expect that the previous discussion can be extended to homeomorphisms. We do not know if the conclusion of Bell and Meyer is also valid when the assumption of stability for Λ is replaced by almost periodicity.

References

[1] D.K. Arrowsmith, C.M. Place, *An Introduction to Dynamical Systems*, Cambridge University Press, Cambridge, 1990.

[2] H. Bell, K.R. Meyer, *Limit periodic functions, adding machines and solenoids*, J. Dynam. Differential Equations **7** (1995), 409–422.

[3] P. Boyland, T. Hall, *Isotopy stable dynamics relative to compact invariant sets*, Proc. London Math. Soc. **79** (1999), 673–693.

[4] M. Brown, *A new proof of Brouwer's lemma on translation arcs*, Houston J. Math. **10** (1984), 35–41.

[5] J. Buescu, M. Kulczycki, I. Stewart, *Liapunov stability and adding machines revisited*, Dyn. Syst. **21** (2006), 379–384.

[6] M.L. Cartwright, *Almost-periodic flows and solutions of differential equations*, Proc. London Math. Soc. **17** (1967), 355–380; Corrigenda: p. 768.

[7] E.N. Dancer, R. Ortega, *The index of Lyapunov stable fixed points in two dimensions*, J. Dynam. Differential Equations **6** (1994), 631–637.

[8] A. Fathi, *An orbit closing proof of Brouwer's lemma on translation arcs*, Enseign. Math. **33** (1987), 315–322.

[9] J. Franks, *A new proof of the Brouwer plane translation theorem*, Ergodic Theory Dynam. Systems **12** (1992), 217–226.

[10] A. Granas, J. Dugundji, *Fixed Point Theory*, Springer, Berlin, 2003.

[11] L. Hernández-Corbato, R. Ortega, F. Ruiz del Portal, *Attractors with irrational rotation number*, Math. Proc. Cambridge Philos. Soc. **153** (2012), 59-77.
[12] K.R. Meyer and G.R. Hall, *Introduction to Hamiltonian Dynamical System and the N-Body Problem*, Springer, Berlin, 1992.
[13] V.V. Nemytskii, V.V. Stepanov, *Qualitative Theory of Differential Equations*, Princeton Univ. Press, Princeton, 1960.
[14] R. Ortega, *The number of stable periodic solutions of time-dependent Hamiltonian systems with one degree of freedom*, Ergodic Theory Dynam. Systems **18** (1998), 1007–1018.
[15] R. Ortega, *Topology of the plane and periodic differential equations*, www.ugr.es/local/ecuadif/fuentenueva.htm
[16] F. Ruiz del Portal, *Planar isolated and stable fixed points have index =1*, J. Differential Equations **199** (2004), 179–188.
[17] E.S. Thomas, *One-dimensional minimal sets*, Topology **12** (1973) 233–242.

Authors' addresses:

Rafael Ortega
Departamento de Matemática Aplicada
Universidad de Granada
18071 Granada, Spain
E-mail: rortega@ugr.es

Alfonso Ruiz-Herrera
Departamento de Matemática Aplicada
Universidad de Granada
18071 Granada, Spain
E-mail: alfonsoruiz@ugr.es

Received March 13, 2012
Revised April 30, 2012

Rend. Istit. Mat. Univ. Trieste
Volume 44 (2012), 45–74

A Whiteheadian-type description of Euclidean spaces, spheres, tori and Tychonoff cubes[1]

Georgi D. Dimov

Dedicated to Prof. Fabio Zanolin on the occasion of his 60th birthday

Abstract. *In the beginning of the 20th century, A. N. Whitehead [39, 40] and T. de Laguna [9] proposed a new theory of space, known as* region-based theory of space. *They did not present their ideas in a detailed mathematical form. In 1997, P. Roeper [33] has shown that the locally compact Hausdorff spaces correspond bijectively (up to homeomorphism and isomorphism) to some algebraical objects which represent correctly Whitehead's ideas of* region *and* contact relation, *generalizing in this way a previous analogous result of de Vries [10] concerning compact Hausdorff spaces (note that even a duality for the category of compact Hausdorff spaces and continuous maps was constructed by de Vries [10]). Recently, a duality for the category of locally compact Hausdorff spaces and continuous maps, based on Roeper's results, was obtained in [11] (it extends de Vries' duality mentioned above). In this paper, using the dualities obtained in [10, 11], we construct directly (i.e. without the help of the corresponding topological spaces) the dual objects of Euclidean spaces, spheres, tori and Tychonoff cubes; these algebraical objects completely characterize the mentioned topological spaces. Thus, a mathematical realization of the original philosophical ideas of Whitehead [39, 40] and de Laguna [9] about Euclidean spaces is obtained.*

Keywords: Euclidean spaces, Tychonoff cubes, spheres, tori, (locally) compact Hausdorff spaces, duality, regular closed sets, sums of local contact algebras, sums of normal contact algebras
MS Classification 2010: 54D45, 54D30, 54B10, 06E99, 18A40, 54E05

---

[1]This paper was supported by the project no. DID 02/32/2009 "Theories of the space and time: algebraic, topological and logical approaches" of the Bulgarian Ministry of Education and Science.

1. Introduction

The region-based theory of space is a kind of point-free geometry and can be considered as an alternative to the well known Euclidean point-based theory of space. Its main idea goes back to Whitehead [40] (see also [39]) and de Laguna [9] and is based on a certain criticism of the Euclidean approach to the geometry, where the points (as well as straight lines and planes) are taken as the basic primitive notions. A. N. Whitehead and T. de Laguna noticed that points, lines and planes are quite abstract entities which have not a separate existence in reality and proposed to put the theory of space on the base of some more realistic spatial entities. In Whitehead [40], the notion of a *region* is taken as a primitive notion: it is an abstract analog of a spatial body; also some natural relations between regions are regarded. In [39], Whitehead considered some mereological relations like "part-of", "overlap" and some others, while in [40] he adopted from de Laguna [9] the relation of "*contact*" ("*connectedness*" in Whitehead's original terminology) as the only primitive relation between regions except the relation "part-of". The *regular closed* (or, equivalently, *regular open*) subsets of a topological space X are usually considered as a standard model of the regions in the point-based approach, and the *standard contact relation* ρ_X between regular closed subsets of X is defined (again in the point-based approach) as follows: $F \rho_X G \Leftrightarrow F \cap G \neq \emptyset$.

Let us note that neither Whitehead nor de Laguna presented their ideas in a detailed mathematical form. This was done by some other mathematicians and mathematically oriented philosophers who presented various versions of region-based theory of space at different levels of abstraction. Here we can mention Tarski [36], who rebuilt Euclidean geometry as an extension of mereology with the primitive notion of a *ball*. Remarkable is also Grzegorczyk's paper [27]. Models of Grzegorczyk's theory are complete Boolean algebras of regular closed sets of certain topological spaces equipped with the relation of separation which in fact is the complement of Whitehead's contact relation. On the same line of abstraction is also the point-free topology [28]. Survey papers describing various aspects and historical remarks on region-based theory of space are [5, 24, 31, 37].

Let us mention that Whitehead's ideas about region-based theory of space flourished and in a sense were reinvented and applied in some areas of computer science: Qualitative Spatial Reasoning (QSR), knowledge representation, geographical information systems, formal ontologies in information systems, image processing, natural language semantics etc. The reason is that the language of region-based theory of space allows the researches to obtain a more simple description of some qualitative spatial features and properties of space bodies. Survey papers concerning various applications are [6, 7] (see also the special issues of "Fundamenta Informaticae" [14] and "Journal of Applied Non-

classical Logics" [4]). One of the most popular among the community of QSR-researchers is the system of Region Connection Calculus (RCC) introduced by Randell, Cui and Cohn [32]. RCC attracted quite intensive research in the field of region-based theory of space, both on its applied and mathematical aspects. For instance it was unknown for some time which topological models correspond adequately to RCC; this fact stimulated the investigations of a topological representation theory of RCC and RCC-like systems (see [13, 15]). Another impact of region-based theory of space is that it stimulated the appearance of a new area in logic, namely "Spatial Logics" [2], called sometimes "Logics of Space".

The ideas of de Laguna and Whitehead lead naturally to the following general programme (or *general region-based theory of space*):

- for every topological space X belonging to some class $\mathcal{C}$ of topological spaces, define in topological terms:

  (a) a family $\mathcal{R}(X)$ of subsets of X that will serve as models of Whitehead's "regions" (and call the elements of the family $\mathcal{R}(X)$ *regions of* X);

  (b) a relation ρ_X on $\mathcal{R}(X)$ that will serve as a model of Whitehead's relation of "contact" (and call the relation ρ_X a *contact relation on* $\mathcal{R}(X)$);

- choose some (algebraic) structure which is inherent to the families $\mathcal{R}(X)$ and contact relations ρ_X, for $X \in \mathcal{C}$, fix some kind of morphisms between the obtained (algebraic) objects and build in this way a category $\boldsymbol{A}$;

- find a subcategory $\boldsymbol{T}$ of the category of topological spaces and continuous maps which is equivalent or dually equivalent to the category $\boldsymbol{A}$ trough a (contravariant) functor that assigns to each object X of $\boldsymbol{T}$ the chosen (algebraic) structure of the family of all regions of X.

If all of this is done then, in particular, the chosen (algebraic) structure of the regions of any object X of $\boldsymbol{T}$ is sufficient for recovering completely (of course, up to homeomorphism) the whole space X. Hence, in this way, a "region-based theory" of the objects and morphisms of the category $\boldsymbol{T}$ is obtained.

Of course, during the realization of this programme, one can find the category $\boldsymbol{A}$ starting with the category $\boldsymbol{T}$, if the later is the desired one.

The M. Stone [35] duality between the category of Boolean algebras and their homomorphisms and the category of compact zero-dimensional Hausdorff spaces and continuous maps can be regarded as a first realization of this programme, although M. Stone came to his results guided by ideas which are completely different from those of Whitehead and de Laguna. In M. Stone's theory, the clopen (= closed and open) subsets of a topological space serve as models of the regions; here, however, the contact relation ρ is hidden, because it can be

defined by the Boolean operations (indeed, we have that $a\rho b \iff a \wedge b \neq 0$). The *localic duality* (see, e.g., [28, Corollary II.1.7]) between the category of spatial frames and functions preserving finite meets and arbitrary joins and the category of sober spaces and continuous maps can also be regarded as a realization of the ideas of the general region-based theory of space: in it the open subsets of a topological space serve as models of the regions and, as above, the contact relation ρ between the regions is hidden because it can be recovered by the lattice operations (indeed, we have that $a\rho b \iff a \wedge b \neq 0$). The de Vries duality [10] for the category **HC** of compact Hausdorff spaces and continuous maps is the first realization of the ideas of the general region-based theory of space in their full generality and strength (and again, as it seems, de Vries was unaware of the papers [9] and [40]): the models of the regions in de Vries' theory are the regular closed sets and, in contrast to the case of the Stone duality and localic duality, the contact relation between regions, which is in the basis of de Vries' duality theorem, cannot be derived from the Boolean structure on the regions. (Note that in [10], instead of the Boolean algebra $RC(X)$ of regular closed sets, the Boolean algebra $RO(X)$ of regular open sets was regarded ($RO(X)$ and $RC(X)$ are isomorphic); also, instead of the relation ρ_X on the set $RC(X)$ which was described above (let us recall it: $F\rho_X G \iff F \cap G \neq \emptyset$), de Vries used in [10] the so-called "*compingent relation*" between regular open sets whose counterpart for $RC(X)$ is the relation $\ll_X$, defined by $F \ll_X G \iff F \subseteq \mathrm{int}(G)$, for $F, G \in RC(X)$; the relations ρ_X and $\ll_X$ are inter-definable.) It is natural to try to extend de Vries' Duality Theorem to the category **HLC** of locally compact Hausdorff spaces and continuous maps. An important step in this direction was done by P. Roeper [33]. Being guided by the ideas of de Laguna [9] and Whitehead [40], he proved that there is a bijective correspondence between all (up to homeomorphism) locally compact Hausdorff spaces and all (up to isomorphism) algebras of some sort called by him "*region-based topologies*" (we call them *complete LC-algebras*). The notion of a complete LC-algebra, introduced by Roeper [33], is an abstraction of the triples $(RC(X), \rho_X, CR(X))$, where X is a locally compact Hausdorff space and $CR(X)$ is the ideal of all compact regular closed subsets of X. P. Roeper [33] showed that every complete LC-algebra can be realized as a triple $(RC(X), \rho_X, CR(X))$, where X is a uniquely (up to homeomorphism) determined locally compact Hausdorff space. In [11], using Roeper's result, we obtained a duality between the category **HLC** and the category **DHLC** of complete LC-algebras and appropriate morphisms between them; it is an extension of de Vries' duality mentioned above; the dual object of a locally compact Hausdorff space X is the triple $(RC(X), \rho_X, CR(X))$ which will be called *the Roeper triple of the space* X. Let us note that the famous Gelfand duality [20, 21, 22, 23] also gives an algebraical description of (locally) compact Hausdorff spaces but it is not in the spirit of the ideas of Whitehead and de

Laguna.

A description of the dual object of the real line under the localic duality (i.e., a description of the frame (or locale) determined by the topology of the real line) without the help of the real line was given by Fourman and Hyland [19] (see, also, Grayson [26] and Johnstone [28, IV.1.1-IV.1.3]), assuming the set of rationals as given. As we have seen above, the ideas of the localic duality are in the spirit of the ideas of the *general* region-based theory of space but, nevertheless, they are far from the well-known and commonly accepted interpretations of the *original* philosophical ideas of Whitehead [39, 40] and de Laguna [9] given in [27] and [33] (see also [32]).

In this paper we construct directly the dual objects of Euclidean spaces, spheres, tori and Tychonoff cubes under the dualities obtained in [10, 11], i.e. we construct the complete LC-algebras isomorphic to the Roeper triples (see [33]) of these spaces without the help of the corresponding spaces, assuming the set of natural numbers as given. For doing this, we first obtain some direct descriptions of the **DHLC**-sums of complete LC-algebras and the **DHC**-sums of complete NC-algebras (where **DHC** is the de Vries category dual to the category **HC**, and the objects of the category **DHC** are the complete NC-algebras) using the dualities obtained in [10] and [11]. Let us note explicitly that, as it follows from the results of de Vries [10] and Roeper [33], the Euclidean spaces, spheres, tori and Tychonoff cubes can be completely reconstructed as topological spaces from the algebraical objects which we describe in this paper. Therefore, our results can be regarded as a mathematical realization of the original philosophical ideas of Whitehead [39, 40] and de Laguna [9] about Euclidean spaces; this realization is in accordance with the Grzegorczyk's [27] and Roeper's [33] mathematical interpretations of these ideas.

We now fix the notation.

If $\mathcal{C}$ denotes a category, we write $X \in |\mathcal{C}|$ if X is an object of $\mathcal{C}$, and $f \in \mathcal{C}(X, Y)$ if f is a morphism of $\mathcal{C}$ with domain X and codomain Y.

All lattices are with top (= unit) and bottom (= zero) elements, denoted respectively by 1 and 0. We do not require the elements 0 and 1 to be distinct.

If (X, τ) is a topological space and M is a subset of X, we denote by $\mathrm{cl}_{(X,\tau)}(M)$ (or simply by $\mathrm{cl}(M)$ or $\mathrm{cl}_X(M)$) the closure of M in (X, τ) and by $\mathrm{int}_{(X,\tau)}(M)$ (or briefly by $\mathrm{int}(M)$ or $\mathrm{int}_X(M)$) the interior of M in (X, τ). The Alexandroff compactification of a locally compact Hausdorff non-compact space X will be denoted by αX. The positive natural numbers are denoted by $\mathbb{N}^+$, the real line (with its natural topology) – by $\mathbb{R}$, the n-dimensional sphere (with its natural topology) – by $\mathbb{S}^n$ (here $n \in \mathbb{N}^+$).

2. Preliminaries

DEFINITION 2.1. *An algebraic system* $(B, 0, 1, \vee, \wedge, ^*, C)$ *is called a* contact Boolean algebra *or, briefly,* contact algebra *(abbreviated as CA or C-algebra) ([13]) if the system* $(B, 0, 1, \vee, \wedge, ^*)$ *is a Boolean algebra (where the operation "complement" is denoted by " * ") and* C *is a binary relation on* B*, satisfying the following axioms:*

(C1) If $a \neq 0$ *then* aCa*;*
(C2) If aCb *then* $a \neq 0$ *and* $b \neq 0$*;*
(C3) aCb *implies* bCa*;*
(C4) $aC(b \vee c)$ *iff* aCb *or* aCc*.*

We shall simply write (B, C) *for a contact algebra. The relation* C *is called a* contact relation*. When* B *is a complete Boolean algebra, we will say that* (B, C) *is a* complete contact Boolean algebra *or, briefly,* complete contact algebra *(abbreviated as CCA or CC-algebra). If* $a \in B$ *and* $D \subseteq B$*, we will write "*aCD*" for "*$(\forall d \in D)(aCd)$*".*

We will say that two C-algebras (B_1, C_1) *and* (B_2, C_2) *are* CA-isomorphic *iff there exists a Boolean isomorphism* $\varphi : B_1 \longrightarrow B_2$ *such that, for each* $a, b \in B_1$*,* aC_1b *iff* $\varphi(a)C_2\varphi(b)$*. Note that in this paper, by a "Boolean isomorphism" we understand an isomorphism in the category* **Bool** *of Boolean algebras and Boolean homomorphisms.*

A contact algebra (B, C) *is called a* normal contact Boolean algebra *or, briefly,* normal contact algebra *(abbreviated as NCA or NC-algebra) ([10, 18]) if it satisfies the following axioms which are very similar to the Efremovič [16] axioms of proximity spaces (we will write " −* C*" for "not* C*"):*

(C5) If $a(-C)b$ *then* $a(-C)c$ *and* $b(-C)c^*$ *for some* $c \in B$*;*
(C6) If $a \neq 1$ *then there exists* $b \neq 0$ *such that* $b(-C)a$*.*

A normal CA is called a complete normal contact Boolean algebra *or, briefly,* complete normal contact algebra *(abbreviated as CNCA or CNC-algebra) if it is a CCA. The notion of a normal contact algebra was introduced by Fedorchuk [18] under the name* Boolean δ-algebra *as an equivalent expression of the notion of a* compingent Boolean algebra *of de Vries (see its definition below). We call such algebras "normal contact algebras" because they form a subclass of the class of contact algebras and naturally arise in normal Hausdorff spaces.*

Note that if $0 \neq 1$ *then the axiom (C2) follows from the axioms (C6) and (C4).*

For any CA (B, C)*, we define a binary relation " ≪*$_C$*" on* B *(called* non-tangential inclusion*) by " *$a \ll_C b \leftrightarrow a(-C)b^*$* ". Sometimes we will write simply " ≪" instead of " ≪*$_C$*".*

The relations C and $\ll$ are inter-definable. For example, normal contact

algebras could be equivalently defined (and exactly in this way they were introduced (under the name of *compingent Boolean algebras*) by de Vries in [10]) as a pair of a Boolean algebra $B = (B, 0, 1, \vee, \wedge, ^*)$ and a binary relation $\ll$ on B subject to the following axioms:

($\ll$1) $a \ll b$ implies $a \leq b$;
($\ll$2) $0 \ll 0$;
($\ll$3) $a \leq b \ll c \leq d$ implies $a \ll d$;
($\ll$4) $a \ll c$ and $b \ll c$ implies $a \vee b \ll c$;
($\ll$5) If $a \ll c$ then $a \ll b \ll c$ for some $b \in B$;
($\ll$6) If $a \neq 0$ then there exists $b \neq 0$ such that $b \ll a$;
($\ll$7) $a \ll b$ implies $b^* \ll a^*$.

Note that if $0 \neq 1$ then the axiom ($\ll$2) follows from the axioms ($\ll$3), ($\ll$4), ($\ll$6) and ($\ll$7).

Obviously, contact algebras could be equivalently defined as a pair of a Boolean algebra B and a binary relation $\ll$ on B subject to the axioms ($\ll$1)-($\ll$4) and ($\ll$7).

It is easy to see that axiom (C5) (resp., (C6)) can be stated equivalently in the form of ($\ll$5) (resp., ($\ll$6)).

EXAMPLE 2.2. *Recall that a subset F of a topological space (X, τ) is called* regular closed *if $F = \mathrm{cl}(\mathrm{int}(F))$. Clearly, F is regular closed iff it is the closure of an open set.*

For any topological space (X, τ), the collection $RC(X, \tau)$ (we will often write simply $RC(X)$) of all regular closed subsets of (X, τ) becomes a complete Boolean algebra $(RC(X, \tau), 0, 1, \wedge, \vee, ^*)$ under the following operations:

$$1 = X, 0 = \emptyset, F^* = \mathrm{cl}(X \setminus F), F \vee G = F \cup G, F \wedge G = \mathrm{cl}(\mathrm{int}(F \cap G)).$$

The infinite operations are given by the formulae:

$$\bigvee\{F_\gamma \mid \gamma \in \Gamma\} = \mathrm{cl}\left(\bigcup\{F_\gamma \mid \gamma \in \Gamma\}\right) \left(= \mathrm{cl}\left(\bigcup\{\mathrm{int}(F_\gamma) \mid \gamma \in \Gamma\}\right)\right),$$

and

$$\bigwedge\{F_\gamma \mid \gamma \in \Gamma\} = \mathrm{cl}\left(\mathrm{int}\left(\bigcap\{F_\gamma \mid \gamma \in \Gamma\}\right)\right).$$

It is easy to see that setting $F \rho_{(X,\tau)} G$ iff $F \cap G \neq \emptyset$, we define a contact relation $\rho_{(X,\tau)}$ on $RC(X, \tau)$; it is called a standard contact relation. *So, $(RC(X, \tau), \rho_{(X,\tau)})$ is a CCA (it is called a* standard contact algebra*). We will often write simply ρ_X instead of $\rho_{(X,\tau)}$. Note that, for $F, G \in RC(X)$, $F \ll_{\rho_X} G$ iff $F \subseteq \mathrm{int}_X(G)$.*

Clearly, if (X, τ) is a normal Hausdorff space then the standard contact algebra $(RC(X, \tau), \rho_{(X,\tau)})$ is a complete NCA.

A subset U of (X,τ) such that $U = \mathrm{int}(\mathrm{cl}(U))$ is said to be regular open. *The set of all regular open subsets of (X,τ) will be denoted by $RO(X,\tau)$ (or briefly, by $RO(X)$).*

The following notion is a lattice-theoretical counterpart of Leader's notion of a *local proximity* ([30]):

DEFINITION 2.3 ([33]). *An algebraic system $\underline{B}_l = (B, 0, 1, \vee, \wedge, ^*, \rho, \mathbb{B})$ is called a* local contact Boolean algebra *or, briefly,* local contact algebra *(abbreviated as LCA or LC-algebra) if $(B, 0, 1, \vee, \wedge, ^*)$ is a Boolean algebra, ρ is a binary relation on B such that (B,ρ) is a CA, and $\mathbb{B}$ is an ideal (possibly non proper) of B, satisfying the following axioms:*

(BC1) If $a \in \mathbb{B}$, $c \in B$ and $a \ll_\rho c$ then $a \ll_\rho b \ll_\rho c$ for some $b \in \mathbb{B}$ (see Definition 2.1 for "$\ll_\rho$");
(BC2) If $a\rho b$ then there exists an element c of $\mathbb{B}$ such that $a\rho(c \wedge b)$;
(BC3) If $a \neq 0$ then there exists $b \in \mathbb{B} \setminus \{0\}$ such that $b \ll_\rho a$.

We shall simply write $(B,\rho,\mathbb{B})$ for a local contact algebra. We will say that the elements of $\mathbb{B}$ are bounded *and the elements of $B \setminus \mathbb{B}$ are* unbounded. *When B is a complete Boolean algebra, the LCA $(B,\rho,\mathbb{B})$ is called a* complete local contact Boolean algebra *or, briefly,* complete local contact algebra *(abbreviated as CLCA or CLC-algebra).*

We will say that two local contact algebras $(B,\rho,\mathbb{B})$ and $(B_1,\rho_1,\mathbb{B}_1)$ are LCA-isomorphic *if there exists a Boolean isomorphism $\varphi : B \longrightarrow B_1$ such that, for $a,b \in B$, $a\rho b$ iff $\varphi(a)\rho_1\varphi(b)$, and $\varphi(a) \in \mathbb{B}_1$ iff $a \in \mathbb{B}$. A map $\varphi : (B,\rho,\mathbb{B}) \longrightarrow (B_1,\rho_1,\mathbb{B}_1)$ is called an* LCA-embedding *if $\varphi : B \longrightarrow B_1$ is an injective Boolean homomorphism (i.e. Boolean monomorphism) and, moreover, for any $a,b \in B$, $a\rho b$ iff $\varphi(a)\rho_1\varphi(b)$, and $\varphi(a) \in \mathbb{B}_1$ iff $a \in \mathbb{B}$.*

REMARK 2.4. *Note that if $(B,\rho,\mathbb{B})$ is a local contact algebra and $1 \in \mathbb{B}$ then (B,ρ) is a normal contact algebra. Conversely, any normal contact algebra (B,C) can be regarded as a local contact algebra of the form (B,C,B).*

DEFINITION 2.5 ([38]). *Let $(B,\rho,\mathbb{B})$ be a local contact algebra. Define a binary relation "$C_{\rho,\mathbb{B}}$" on B by*

$$aC_{\rho,\mathbb{B}}b \quad \textit{iff} \quad a\rho b \quad \textit{or} \quad a,b \notin \mathbb{B}. \tag{1}$$

It is called the Alexandroff extension of ρ relatively to the LCA $(B,\rho,\mathbb{B})$ *(or, when there is no ambiguity, simply, the* Alexandroff extension of ρ*).*

The following lemma is a lattice-theoretical counterpart of a theorem from Leader's paper [30].

LEMMA 2.6 ([38]). *Let $(B,\rho,\mathbb{B})$ be a local contact algebra. Then $(B, C_{\rho,\mathbb{B}})$, where $C_{\rho,\mathbb{B}}$ is the Alexandroff extension of ρ, is a normal contact algebra.*

Notation. Let (X,τ) be a topological space. We denote by $CR(X,\tau)$ the family of all compact regular closed subsets of (X,τ). We will often write $CR(X)$ instead of $CR(X,\tau)$.

PROPOSITION 2.7 ([33]). *Let (X,τ) be a locally compact Hausdorff space. Then the triple $(RC(X,\tau),\rho_{(X,\tau)},CR(X,\tau))$ (see Example 2.2 for $\rho_{(X,\tau)}$) is a complete local contact algebra; it is called a* standard local contact algebra.

The next theorem was proved by Roeper [33] (but its particular case concerning compact Hausdorff spaces and NC-algebras was proved by de Vries [10]).

THEOREM 2.8 (P. Roeper [33] for locally compact spaces and de Vries [10] for compact spaces). *There exists a bijective correspondence Ψ^t between the class of all (up to homeomorphism) locally compact Hausdorff spaces and the class of all (up to isomorphism) CLC-algebras; its restriction to the class of all (up to homeomorphism) compact Hausdorff spaces gives a bijective correspondence between the later class and the class of all (up to isomorphism) CNC-algebras.*

Let us recall the definition of the correspondence Ψ^t mentioned in the above theorem: if (X,τ) is a locally compact Hausdorff space then

$$\Psi^t(X,\tau)=(RC(X,\tau),\rho_{(X,\tau)},CR(X,\tau)) \tag{2}$$

(see Proposition 2.7 for the notation).

DEFINITION 2.9 (De Vries [10]). *Let* **HC** *be the category of all compact Hausdorff spaces and all continuous maps between them.*

Let **DHC** *be the category whose objects are all complete NC-algebras and whose morphisms are all functions $\varphi:(A,C)\longrightarrow(B,C')$ between the objects of* **DHC** *satisfying the conditions:*

(DVAL1) $\varphi(0)=0$;
(DVAL2) $\varphi(a\wedge b)=\varphi(a)\wedge\varphi(b)$, for all $a,b\in A$;
(DVAL3) If $a,b\in A$ and $a\ll_C b$, then $(\varphi(a^*))^*\ll_{C'}\varphi(b)$;
(DVAL4) $\varphi(a)=\bigvee\{\varphi(b)\mid b\ll_C a\}$, for every $a\in A$,

and let the composition "$\diamond$" of two morphisms $\varphi_1:(A_1,C_1)\longrightarrow(A_2,C_2)$ and $\varphi_2:(A_2,C_2)\longrightarrow(A_3,C_3)$ of **DHC** *be defined by the formula*

$$\varphi_2\diamond\varphi_1=(\varphi_2\circ\varphi_1)^{\check{}}, \tag{3}$$

where, for every function $\psi:(A,C)\longrightarrow(B,C')$ between two objects of **DHC**, *$\psi^{\check{}}:(A,C)\longrightarrow(B,C')$ is defined as follows:*

$$\psi^{\check{}}(a)=\bigvee\{\psi(b)\mid b\ll_C a\}, \tag{4}$$

for every $a\in A$.

De Vries [10] proved the following duality theorem:

THEOREM 2.10 ([10]). *The categories* **HC** *and* **DHC** *are dually equivalent.*

In [11], an extension of de Vries' Duality Theorem to the category of locally compact Hausdorff spaces and continuous maps was obtained. Let us recall its formulation.

DEFINITION 2.11 ([11]). *Let* **HLC** *be the category of all locally compact Hausdorff spaces and all continuous maps between them.*

Let **DHLC** *be the category whose objects are all complete LC-algebras and whose morphisms are all functions* $\varphi : (A, \rho, \mathbb{B}) \longrightarrow (B, \eta, \mathbb{B}')$ *between the objects of* **DHLC** *satisfying conditions*

(DLC1) $\varphi(0) = 0$*;*
(DLC2) $\varphi(a \wedge b) = \varphi(a) \wedge \varphi(b)$*, for all* $a, b \in A$*;*
(DLC3) If $a \in \mathbb{B}, b \in A$ *and* $a \ll_\rho b$*, then* $(\varphi(a^*))^* \ll_\eta \varphi(b)$*;*
(DLC4) For every $b \in \mathbb{B}'$ *there exists* $a \in \mathbb{B}$ *such that* $b \leq \varphi(a)$*;*
(DLC5) $\varphi(a) = \bigvee\{\varphi(b) \mid b \in \mathbb{B}, b \ll_\rho a\}$*, for every* $a \in A$*;*

let the composition "$\diamond$*" of two morphisms* $\varphi_1 : (A_1, \rho_1, \mathbb{B}_1) \longrightarrow (A_2, \rho_2, \mathbb{B}_2)$ *and* $\varphi_2 : (A_2, \rho_2, \mathbb{B}_2) \longrightarrow (A_3, \rho_3, \mathbb{B}_3)$ *of* **DHLC** *be defined by the formula*

$$\varphi_2 \diamond \varphi_1 = (\varphi_2 \circ \varphi_1)^{\check{}}, \tag{5}$$

where, for every function $\psi : (A, \rho, \mathbb{B}) \longrightarrow (B, \eta, \mathbb{B}')$ *between two objects of* **DHLC**, $\psi^{\check{}} : (A, \rho, \mathbb{B}) \longrightarrow (B, \eta, \mathbb{B}')$ *is defined as follows:*

$$\psi^{\check{}}(a) = \bigvee\{\psi(b) \mid b \in \mathbb{B}, b \ll_\rho a\}, \tag{6}$$

for every $a \in A$*.*

(We used here the same notation as in Definition 2.9 for the composition between the morphisms of the category **DHLC** *and for the functions of the type* $\psi^{\check{}}$ *because the NC-algebras can be regarded as those LC-algebras* $(A, \rho, \mathbb{B})$ *for which* $A = \mathbb{B}$*, and hence the right sides of the formulae (6) and (4) coincide in the case of NC-algebras.)*

It can be shown that condition (DLC3) in Definition 2.11 can be replaced by any of the following four constrains:

(DLC3′) If $a, b \in \mathbb{B}$ and $a \ll_\rho b$, then $(\varphi(a^*))^* \ll_\eta \varphi(b)$.
(DLC3S) If $a, b \in A$ and $a \ll_\rho b$, then $(\varphi(a^*))^* \ll_\eta \varphi(b)$.
(LC3) If, for $i = 1, 2$, $a_i \in \mathbb{B}$, $b_i \in A$ and $a_i \ll_\rho b_i$, then $\varphi(a_1 \vee a_2) \ll_\eta \varphi(b_1) \vee \varphi(b_2)$.
(LC3S) If, for $i = 1, 2$, $a_i, b_i \in A$ and $a_i \ll_\rho b_i$, then $\varphi(a_1 \vee a_2) \ll_\eta \varphi(b_1) \vee \varphi(b_2)$.

THEOREM 2.12 ([11]). *The categories* **HLC** *and* **DHLC** *are dually equivalent.*

The duality, constructed in Theorem 2.12 and denoted by $\Psi^t : \mathbf{HLC} \longrightarrow \mathbf{DHLC}$, is an extension of the Roeper's correspondence Ψ^t defined by (2) (i.e. the definition of the contravariant functor Ψ^t *on the objects of the category* **HLC** coincides with the definition of the Roeper's correspondence).

We will also need a lemma from [8]:

LEMMA 2.13. *Let X be a dense subspace of a topological space Y. Then the functions $r : RC(Y) \longrightarrow RC(X)$, $F \mapsto F \cap X$, and $e : RC(X) \longrightarrow RC(Y)$, $G \mapsto \mathrm{cl}_Y(G)$, are Boolean isomorphisms between Boolean algebras $RC(X)$ and $RC(Y)$, and $e \circ r = id_{RC(Y)}$, $r \circ e = id_{RC(X)}$.*

For the notions and notation not defined here see [1, 17, 28, 34].

3. Sums in the categories DHLC and DHC

In [12], we described the **DHLC**-products of complete local contact algebras. Here we will describe the **DHLC**-sums of finite families of complete local contact algebras and the **DHC**-sums of arbitrarily many complete contact algebras using the notion of a *sum of a family of Boolean algebras* (see [25]) which is known also as a *free product* (see [29]). (We will denote the sum of a family $\{A_\gamma \mid \gamma \in \Gamma\}$ of Boolean algebras by $\bigoplus_{\gamma\in\Gamma} A_\gamma$ (as in [29]).) Note that the sums (resp., finite sums) in the category **DHC** (resp., **DHLC**) surely exist because the dual category **HC** (resp., **HLC**) of all compact (resp., locally compact) Hausdorff spaces and continuous maps has products (resp., finite products).

Let us recall the definition of the notion of a sum of a family $(A_i)_{i\in I}$ of Boolean algebras (see, e.g. [29]): a pair $(A, (e_i)_{i\in I})$ is a *sum of* $(A_i)_{i\in I}$ if A is a Boolean algebra, each e_i is a homomorphism from A_i into A and, for every family $(f_i)_{i\in I}$ of homomorphisms from A_i into any Boolean algebra B, there is a unique homomorphism $f : A \longrightarrow B$ such that $f \circ e_i = f_i$ for $i \in I$. It is well known that every family of Boolean algebras has, up to isomorphism, a unique sum. Recall, as well, that a family $(B_i)_{i\in I}$ of subalgebras of a Boolean algebra A is *independent* if, for arbitrary $n \in \mathbb{N}^+$, pairwise distinct $i(1), \ldots, i(n) \in I$ and non-zero elements $b_{i(k)}$ of $B_{i(k)}$, for $k = 1, \ldots, n$, $b_{i(1)} \wedge \ldots \wedge b_{i(n)} > 0$ in A. The following characterization of the sums holds (see, e.g., [29]):

PROPOSITION 3.1. *Let A be a Boolean algebra and, for $i \in I$, $e_i : A_i \longrightarrow A$ a homomorphism; assume that no A_i is trivial. The pair $(A, (e_i)_{i\in I})$ is a sum of $(A_i)_{i\in I}$ iff each of (a) through (c) holds:*

(a) each $e_i : A_i \longrightarrow A$ is an injection,

(b) $(e_i(A_i))_{i\in I}$ is an independent family of subalgebras of A,

(c) A is generated by $\bigcup_{i\in I} e_i(A_i)$.

Moreover, if $(A, (e_i)_{i\in I})$ is a sum of $(A_i)_{i\in I}$ then

(d) $e_i(A_i) \cap e_j(A_j) = \{0, 1\}$, for $i \neq j$.

We start with a proposition which should be known, although I was not able to find it in the literature. Recall that a topological space X is called *semiregular* if $RO(X)$ is a base of X. By a *completion* of a Boolean algebra A, we will understand the *MacNeille completion* of A.

PROPOSITION 3.2. *Let $\{X_\gamma \mid \gamma \in \Gamma\}$ be a family of semiregular topological spaces and $X = \prod\{X_\gamma \mid \gamma \in \Gamma\}$. Then the Boolean algebra $RC(X)$ is isomorphic to the completion of $\bigoplus_{\gamma\in\Gamma} RC(X_\gamma)$.*

Proof. Let, for every $\gamma \in \Gamma$, $\pi_\gamma : X \longrightarrow X_\gamma$ be the projection. Using the fact that π_γ is an open map (and, thus, the formulae $\mathrm{cl}(\pi_\gamma^{-1}(M)) = \pi_\gamma^{-1}(\mathrm{cl}(M))$ and $\mathrm{int}(\pi_\gamma^{-1}(M)) = \pi_\gamma^{-1}(\mathrm{int}(M))$ hold for every $M \subseteq X_\gamma$) (see, e.g., [17]), it is easy to show, that the map $\varphi_\gamma : RC(X_\gamma) \longrightarrow RC(X)$, $F \mapsto \pi_\gamma^{-1}(F)$, is a complete monomorphism for every $\gamma \in \Gamma$. Set $A_\gamma = \varphi_\gamma(RC(X_\gamma))$, for every $\gamma \in \Gamma$, and let A be the subalgebra of $RC(X)$ generated by $\bigcup\{A_\gamma \mid \gamma \in \Gamma\}$. It is easy to check that, for every finite non-empty subset Γ_0 of Γ, we have that if $a_\gamma \in A_\gamma \setminus \{0\}$ for every $\gamma \in \Gamma_0$, then $\bigwedge\{a_\gamma \mid \gamma \in \Gamma_0\} \neq 0$ (i.e. the family $\{A_\gamma \mid \gamma \in \Gamma\}$ is an *independent family* (see, e.g., [29])). Thus, by [29, Proposition 11.4], we get that $A = \bigoplus_{\gamma\in\Gamma} RC(X_\gamma)$. Since $RO(X_\gamma)$ is a base of X_γ, for every $\gamma \in \Gamma$, we obtain that A is a dense subalgebra of $RC(X)$. Thus, $RC(X)$ is the completion of A. □

The proof of this proposition shows that the following is even true:

COROLLARY 3.3. *Let $\{X_\gamma \mid \gamma \in \Gamma\}$ be a family of semiregular topological spaces and $X = \prod\{X_\gamma \mid \gamma \in \Gamma\}$. Let, for every $\gamma \in \Gamma$, B_γ be a subalgebra of $RC(X_\gamma)$ such that $\{\mathrm{int}(F) \mid F \in B_\gamma\}$ is a base of X_γ. Then the Boolean algebra $RC(X)$ is isomorphic to the completion of $\bigoplus_{\gamma\in\Gamma} B_\gamma$.*

DEFINITION 3.4. *Let $n \in \mathbb{N}^+$ and let, for every $i = 1, \ldots, n$, $(A_i, \rho_i, \mathbb{B}_i)$ be a CLCA. Let*

$$(A, (\varphi_i)_{i=1}^n) = \bigoplus_{i=1}^n A_i,$$

where, for every $i \in \{1, \ldots, n\}$,

$$\varphi_i : A_i \longrightarrow A$$

is the canonical complete monomorphism, and let $\tilde{A}$ be the completion of A. We can suppose, without loss of generality, that $A \subseteq \tilde{A}$. Set

$$E = \left\{ \bigwedge_{i=1}^n \varphi_i(a_i) \mid a_i \in \mathbb{B}_i \right\}$$

and let $\widetilde{\mathbb{B}}$ be the ideal of $\tilde{A}$ generated by E (thus,

$$\widetilde{\mathbb{B}} = \{x \in \tilde{A} \mid x \leq e_1 \vee \ldots \vee e_n \text{ for some } n \in \mathbb{N}^+ \text{ and } e_1, \ldots, e_n \in E\}).$$

For every two elements $a = \bigwedge_{i=1}^{n} \varphi_i(a_i)$ *and* $b = \bigwedge_{i=1}^{n} \varphi_i(b_i)$ *of* E, *set*

$$a\tilde{\rho}b \Leftrightarrow (a_i \rho_i b_i, \forall i \in \{1, \dots, n\}).$$

Further, for every two elements c *and* d *of* $\widetilde{\mathbb{IB}}$, *set*

$$c(-\tilde{\rho})d \Leftrightarrow \Big(\exists k, l \in \mathbb{N}^+ \text{ and } \exists c_1, \dots, c_k, d_1, \dots, d_l \in E \text{ such that}$$
$$c \leq \bigvee_{i=1}^{k} c_i,\ d \leq \bigvee_{j=1}^{l} d_j \text{ and } c_i(-\tilde{\rho})d_j,\ \forall i = 1, \dots, k \text{ and } \forall j = 1, \dots, l\Big).$$

Finally, for every two elements a *and* b *of* $\tilde{A}$, *set*

$$a\tilde{\rho}b \Leftrightarrow (\exists c, d \in \widetilde{\mathbb{IB}} \text{ such that } c \leq a,\ d \leq b \text{ and } c\tilde{\rho}d).$$

Then the triple $(\tilde{A}, \tilde{\rho}, \widetilde{\mathbb{IB}})$ *will be denoted by* $\bigoplus_{i=1}^{n}(A_i, \rho_i, \mathbb{IB}_i)$.

THEOREM 3.5. *Let* $n \in \mathbb{N}^+$ *and* $\mathcal{A} = \{(A_i, \rho_i, \mathbb{IB}_i) \mid i = 1, \dots, n\}$ *be a family of CLCAs. Then* $\bigoplus_{i=1}^{n}(A_i, \rho_i, \mathbb{IB}_i)$ *is a* **DHLC***-sum of the family* $\mathcal{A}$.

Proof. As the Duality Theorem 2.12 shows, for every $i \in \{1, \dots, n\}$ there exists a $X_i \in |\mathbf{HLC}|$ such that the CLCAs $(RC(X_i), \rho_{X_i}, CR(X_i))$ and $(A_i, \rho_i, \mathbb{IB}_i)$ are LCA-isomorphic. Let $X = \prod_{i=1}^{n} X_i$. Then we have, in the notation of Definition 3.4, that the Boolean algebras $RC(X)$ and $\tilde{A}$ are isomorphic (see Proposition 3.2). Also, again in the notation of Definition 3.4, $(A, (\varphi_i)_{i=1}^{n})$ is isomorphic to $(\bigoplus_{i=1}^{n} RC(X_i), (\psi_i)_{i=1}^{n})$, where $\psi_i : RC(X_i) \longrightarrow RC(X),\ F \mapsto \pi_i^{-1}(F)$, and $\pi_i : X \longrightarrow X_i$ is the projection, for every $i \in \{1, \dots, n\}$ (this follows from Proposition 3.1). Thus, the set E from Definition 3.4 corresponds to the following set:

$$E' = \left\{ \bigwedge_{i=1}^{n} \psi_i(F_i) \mid F_i \in CR(X_i) \right\}.$$

Let $F \in E'$. Then there exist $F_i \in CR(X_i)$, for $i = 1, \dots, n$, such that $F = \bigwedge_{i=1}^{n} \psi_i(F_i)$. Set $U_i = \operatorname{int}_{X_i}(F_i)$, for $i = 1, \dots, n$. Then $F = \bigwedge_{i=1}^{n} \pi_i^{-1}(F_i) = \operatorname{cl}_X(\bigcap_{i=1}^{n} \operatorname{int}_X(\pi_i^{-1}(F_i))) = \operatorname{cl}_X(\bigcap_{i=1}^{n} \pi_i^{-1}(U_i)) = \operatorname{cl}(\prod_{i=1}^{n} U_i) = \prod_{i=1}^{n} F_i$ (note that we used [17, 1.4.C,2.3.3] here). Hence, for every $F, G \in E'$, where $F = \prod_{i=1}^{n} F_i$ and $G = \prod_{i=1}^{n} G_i$, we have that

$$F \rho_X G \Leftrightarrow F \cap G \neq \emptyset \Leftrightarrow (F_i \cap G_i \neq \emptyset, \forall i = 1, \dots, n) \Leftrightarrow (F_i \rho_{X_i} G_i, \forall i = 1, \dots, n).$$

Further, since $\{\prod_{i=1}^{n} U_i \mid U_i \in RO(X_i), \forall i = 1, \dots, n\}$ is a base of X and X is regular, we obtain that $CR(X)$ coincides with the ideal of $RC(X)$ generated by E'. The fact that every two disjoint compact subsets of X can be separated

by open sets implies that if $F, G \in CR(X)$ then $F(-\rho_X)G$ (i.e. $F \cap G = \emptyset$) iff there exists finitely many elements $F_1, \ldots, F_k, G_1, \ldots, G_l \in E'$ such that $F \subseteq \bigcup_{i=1}^{k} F_i$, $G \subseteq \bigcup_{i=1}^{l} G_i$ and $F_i \cap G_j = \emptyset$ (i.e. $F_i(-\rho_X)G_j$) for all $i = 1, \ldots, k$ and all $j = 1, \ldots, l$. Finally, since $(RC(X), \rho_X, CR(X))$ is an LCA (see 2.7), we have (by (BC2)) that for any $F', G' \in RC(X)$, $F'\rho_X G' \Leftrightarrow \exists F, G \in CR(X)$ such that $F \subseteq F'$, $G \subseteq G'$ and $F\rho_X G$. All this shows that the triple $(\tilde{A}, \tilde{\rho}, \widetilde{\mathbb{B}})$ from 3.4 is an LCA which is LCA-isomorphic to $(RC(X), \rho_X, CR(X))$. Now, using Theorem 2.12 and the facts that $\Psi^t(X) = (RC(X), \rho_X, CR(X))$, $\Psi^t(X_i) = (RC(X_i), \rho_{X_i}, CR(X_i))$ for all $i = 1, \ldots, n$, and X is a **HLC**-product of the family $\{X_i \mid i = 1, \ldots, n\}$, we get that $(RC(X), \rho_X, CR(X))$ is a **DHLC**-sum of the family $\{(RC(X_i), \rho_{X_i}, CR(X_i)) \mid i = 1, \ldots, n\}$. Thus $(\tilde{A}, \tilde{\rho}, \widetilde{\mathbb{B}})$ is a **DHLC**-sum of the family $\{(A_i, \rho_i, \mathbb{B}_i) \mid i = 1, \ldots, n\}$. □

Definition 3.6. *Let J be a set and let, for every $j \in J$, (A_j, ρ_j) be a CNCA. Let*

$$(A, (\varphi_j)_{j \in J}) = \bigoplus_{j \in J} A_j,$$

where, for every $j \in J$,

$$\varphi_j : A_j \longrightarrow A$$

is the canonical complete monomorphism, and let $\tilde{A}$ be the completion of A. We can suppose, without loss of generality, that $A \subseteq \tilde{A}$. Set

$$E = \left\{ \bigwedge_{i \in I} \varphi_i(a_i) \mid I \subseteq J, |I| < \aleph_0, a_i \in A_i, \forall i \in I \right\}.$$

For every two elements $a = \bigwedge_{i \in I_1} \varphi_i(a_i)$ and $b = \bigwedge_{i \in I_2} \varphi_i(b_i)$ of E, set

$$a\tilde{\rho}b \Leftrightarrow (a_i \rho_i b_i, \forall i \in I_1 \cap I_2).$$

Further, for every two elements c and d of $\tilde{A}$, set

$$c(-\tilde{\rho})d \Leftrightarrow \Big(\exists k, l \in \mathbb{N}^+ \text{ and } \exists c_1, \ldots, c_k, d_1, \ldots, d_l \in E \text{ such that}$$
$$c \leq \bigvee_{i=1}^{k} c_i,\ d \leq \bigvee_{j=1}^{l} d_j \text{ and } c_i(-\tilde{\rho})d_j,\ \forall i = 1, \ldots, k \text{ and } \forall j = 1, \ldots, l\Big).$$

Then the pair $(\tilde{A}, \tilde{\rho})$ will be denoted by $\bigoplus_{j \in J}(A_j, \rho_j)$.

Theorem 3.7. *Let $\mathcal{A} = \{(A_j, \rho_j) \mid j \in J\}$ be a family of complete normal contact algebras. Then $\bigoplus_{j \in J}(A_j, \rho_j)$ is a* **DHC***-sum of the family $\mathcal{A}$.*

Proof. The proof is similar to that one of Theorem 3.5. In it de Vries' Duality Theorem 2.10 instead of Theorem 2.12 can be used. □

4. A Whiteheadian-type description of Euclidean spaces

Notation. We will denote by $\mathbb{Z}$ the set of all integers with the natural order, by $\mathbb{I}$ the unit interval $[0,1]$ with its natural topology and by $\mathbb{I}'$ – the open interval $(0,1)$ with its natural topology, by $\mathbb{N}$ the set of natural numbers, by $\mathbb{J}$ the subspace of the real line consisting of all irrational numbers, and by $\mathbb{D}$ the set of all dyadic numbers in the interval $(0,1)$. We set $\mathbb{Z}_0 = \mathbb{Z}\setminus\{0\}$, $\mathbb{Z}^- = \mathbb{Z}\setminus\mathbb{N}$ and $\mathbb{J}_2 = \mathbb{I}'\setminus\mathbb{D}$. If $(X,<)$ is a linearly ordered set and $x \in X$, then we set

$$succ(x) = \{y \in X \mid x < y\},\ pred(x) = \{y \in X \mid y < x\};$$

also, we denote by x^+ the *successor of* x (when it exists) and by x^- – the *predecessor of* x (when it exists). If M is a set, then we will denote by $P(M)$ the power set Boolean algebra of M; the cardinality of M will be denoted by $|M|$. If X is a topological space, then we will denote by $CO(X)$ the set of all clopen (= closed and open) subsets of X.

Now we will construct a CLCA $(\tilde{A}, \tilde{\sigma}, \widetilde{\mathbb{IB}})$ and we will show that it is LCA-isomorphic to $\Psi^t(\mathbb{R})$.

The construction of $(\tilde{A}, \tilde{\sigma}, \widetilde{\mathbb{IB}})$. Let $A_i = P(\mathbb{Z}_0)$, for every $i \in \mathbb{N}^+$. Thus, if $i \in \mathbb{N}^+$ and $a_i \in A_i$, then a_i is a subset of $\mathbb{Z}_0$ and its cardinality will be denoted by $|a_i|$. Let $(A, (\varphi_i)_{i\in\mathbb{N}^+})$ be the sum of Boolean algebras $\{A_i \mid i \in \mathbb{N}^+\}$; then, by Proposition 3.1, for every $i \in \mathbb{N}^+$, $\varphi_i : A_i \longrightarrow A$ is a monomorphism, the family $\{\varphi_i(A_i) \mid i \in \mathbb{N}^+\}$ is an independent family and the set $\bigcup_{i\in\mathbb{N}^+} \varphi_i(A_i)$ generates A. Let $\tilde{A}$ be the completion of A. We can suppose, without loss of generality, that $A \subseteq \tilde{A}$.

The following subset of A will be important for us:

$$B_0 = \{\varphi_1(a_1) \wedge \ldots \wedge \varphi_k(a_k) \mid k \in \mathbb{N}^+, \\ (\forall i = 1,\ldots,k)(a_i \in A_i \text{ and } |a_i| = 1)\}. \tag{7}$$

If $b \in B_0$ and $b = \varphi_1(a_1) \wedge \ldots \wedge \varphi_k(a_k)$, where $a_k = \{p\}$, then we set

$$b_- = \varphi_1(a_1) \wedge \varphi_2(a_2) \wedge \ldots \wedge \varphi_{k-1}(a_{k-1}) \wedge \varphi_k(\{p^-\}). \tag{8}$$

For every $b \in B_0$, where $b = \varphi_1(a_1) \wedge \ldots \wedge \varphi_k(a_k)$, and every $n \in \mathbb{N}^+$, we set

$$q_{bn} = (b_- \wedge \varphi_{k+1}(succ(n))) \vee (b \wedge \varphi_{k+1}(pred(-n))). \tag{9}$$

Now we set

$$B_1 = \{q_{bn} \mid b \in B_0, n \in \mathbb{N}^+\}. \tag{10}$$

Let $\widetilde{\mathbb{IB}}$ be the ideal of $\tilde{A}$ generated by the set $B_0 \cup B_1$. Now, we will define a relation $\tilde{\sigma}$ on $\tilde{A}$. It will be, by definition, a symmetric relation.

Let $r, r' \in \mathbb{N}^+$, $b, b' \in B_0$, $b = \varphi_1(a_1) \wedge \ldots \wedge \varphi_k(a_k)$, $b' = \varphi_1(a'_1) \wedge \ldots \wedge \varphi_l(a'_l)$ and $a_k = \{n\}$, $a'_k = \{m\}$. We can suppose, without loss of generality, that $k \leq l$. If $k < l$, then let $a'_{k+1} = \{p\}$. Now we set

$$b\tilde{\sigma}b' \Leftrightarrow \quad \left[\Big(a_i = a'_i,\ \forall i \in \{1, \ldots, k-1\}\Big) \right. \\ \left. \& \left(\begin{cases} m \in \{n^-, n, n^+\}, & \text{if } k = l \\ m = n, & \text{if } k < l \end{cases}\right)\right], \tag{11}$$

and

$$q_{br}\tilde{\sigma}q_{b'r'} \Leftrightarrow \left[\Big(a_i = a'_i,\ \forall i \in \{1, \ldots, k-1\}\Big)\right. \\ \left. \& \left(\begin{cases} m{=}n, & \text{if } l{=}k \\ (m{=}n \text{ and } p{\leq}{-}r) \text{ or } (m{=}n^- \text{ and } p{>}r), & \text{if } l{=}k{+}1 \\ (m{=}n \text{ and } p{<}{-}r) \text{ or } (m{=}n^- \text{ and } p{>}r), & \text{if } l{>}k{+}1 \end{cases}\right)\right]. \tag{12}$$

Let $r \in \mathbb{N}^+$, $b, b' \in B_0$, $b = \varphi_1(a_1) \wedge \ldots \wedge \varphi_k(a_k)$, $b' = \varphi_1(a'_1) \wedge \ldots \wedge \varphi_l(a'_l)$ and $a_k = \{n\}$, $a'_k = \{m\}$. If $k < l$, then let $a'_{k+1} = \{p\}$. Now, if $k > l$, we set

$$q_{br}\tilde{\sigma}b' \Leftrightarrow (a_i = a'_i,\ \forall i \in \{1, \ldots, l\}); \tag{13}$$

if $k \leq l$, we set

$$q_{br}\tilde{\sigma}b' \Leftrightarrow \left[\Big(a_i = a'_i,\ \forall i \in \{1, \ldots, k-1\}\Big)\right. \\ \left. \& \left(\begin{cases} m \in \{n^-, n\}, & \text{if } l{=}k \\ (p{\geq}r \text{ and } m{=}n^-) \text{ or } (p{\leq}{-}r \text{ and } m{=}n), & \text{if } l{=}k{+}1 \\ (p{>}r \text{ and } m{=}n^-) \text{ or } (p{<}{-}r \text{ and } m{=}n), & \text{if } l{>}k{+}1 \end{cases}\right)\right]. \tag{14}$$

Further, for every two elements c and d of $\widetilde{\mathbb{B}}$, set

$$c(-\tilde{\sigma})d \Leftrightarrow \Big(\exists k, l \in \mathbb{N}^+ \text{ and } \exists c_1, \ldots, c_k, d_1, \ldots, d_l \in B_0 \cup B_1 \text{ such that} \\ c \leq \bigvee_{i=1}^{k} c_i,\ d \leq \bigvee_{j=1}^{l} d_j \text{ and } c_i(-\tilde{\sigma})d_j,\ \forall i{=}1, \ldots, k \text{ and } \forall j{=}1, \ldots, l\Big). \tag{15}$$

Finally, for every two elements a and b of $\tilde{A}$, set

$$a\tilde{\sigma}b \Leftrightarrow (\exists c, d \in \widetilde{\mathbb{B}} \text{ such that } c \leq a,\ d \leq b \text{ and } c\tilde{\sigma}d). \tag{16}$$

THEOREM 4.1. *The triple $(\tilde{A}, \tilde{\sigma}, \widetilde{\mathbb{B}})$ (constructed above) is a CLCA; it is LCA-isomorphic to the CLCA $(RC(\mathbb{R}), \rho_{\mathbb{R}}, CR(\mathbb{R}))$. Thus, the triple $(\tilde{A}, \tilde{\sigma}, \widetilde{\mathbb{B}})$ completely determines the real line $\mathbb{R}$ with its natural topology.*

Proof. In this proof, we will use the notation introduced in the construction of $(\tilde{A}, \tilde{\sigma}, \widetilde{\mathbb{B}})$.

Let $\mathbb{Z}_0$ be endowed with the discrete topology. Then $RC(\mathbb{Z}_0) = P(\mathbb{Z}_0)$ and Proposition 3.2 shows that the algebra $\tilde{A}$ is isomorphic to $RC(\mathbb{Z}_0^{\mathbb{N}^+})$. Since the space $\mathbb{Z}_0^{\mathbb{N}^+}$ is homeomorphic to $\mathbb{J}$ (see, e.g., [17]), we get, by Lemma 2.13, that $\tilde{A}$ is isomorphic to $RC(\mathbb{R})$. Clearly, $RC(\mathbb{J})$ can be endowed with an LCA-structure LCA-isomorphic to the LCA $(RC(\mathbb{R}), \rho_{\mathbb{R}}, CR(\mathbb{R}))$. Then, using the homeomorphism between $\mathbb{J}$ and $\mathbb{Z}_0^{\mathbb{N}^+}$, we can transfer this structure to $RC(\mathbb{Z}_0^{\mathbb{N}^+})$ and, hence, to $\tilde{A}$. For technical reasons, this plan will be slightly modified. We will use the homeomorphism between $\mathbb{Z}_0^{\mathbb{N}^+}$ and $\mathbb{J}_2$ described in [3]. Since $\mathbb{J}_2$ is dense in the open interval $\mathbb{I}'$, and $\mathbb{I}'$ is homeomorphic to $\mathbb{R}$, we can use $\mathbb{J}_2$ instead of $\mathbb{J}$ for realizing the desired transfer. So, we start with the description (given by P. S. Alexandroff [3]) of the homeomorphism $f : \mathbb{Z}_0^{\mathbb{N}^+} \longrightarrow \mathbb{J}_2$. Let, for every $j \in \mathbb{N}^+$, $\Delta_j = [1 - \frac{1}{2^j}, 1 - \frac{1}{2^{j+1}}]$ and let, for every $j \in \mathbb{Z}^-$, $\Delta_j = [2^{j-1}, 2^j]$. Set $\delta_1 = \{\Delta_j \mid j \in \mathbb{Z}_0\}$. Further, for every $\Delta_j \in \delta_1$, where $\Delta_j = [a_j, b_j]$, set $d_j = b_j - a_j$ and $\Delta_{jk} = [b_j - \frac{d_j}{2^k}, b_j - \frac{d_j}{2^{k+1}}]$ when $k \in \mathbb{N}^+$, $\Delta_{jk} = [a_j + d_j.2^{k-1}, a_j + d_j.2^k]$ when $k \in \mathbb{Z}^-$. Let $\delta_2 = \{\Delta_{jk} \mid j, k \in \mathbb{Z}_0\}$. In the next step we construct analogously the family δ_3, and so on. Set $\delta = \bigcup\{\delta_i \mid i \in \mathbb{N}^+\}$. It is easy to see that the set of all end-points of the elements of the family δ coincides with the set $\mathbb{D}$. Now we define the function $f : \mathbb{Z}_0^{\mathbb{N}^+} \longrightarrow \mathbb{J}_2$ by the formula

$$f(n_1, n_2, \ldots, n_k, \ldots) = \Delta_{n_1} \cap \Delta_{n_1 n_2} \cap \ldots \cap \Delta_{n_1 n_2 \ldots n_k} \cap \ldots.$$

One can prove that the definition of f is correct and that f is a homeomorphism. Set $X_i = \mathbb{Z}_0$, for every $i \in \mathbb{N}^+$. Let $X = \prod\{X_i \mid i \in \mathbb{N}^+\}$ and let

$$\pi_i : X \longrightarrow X_i,$$

where $i \in \mathbb{N}^+$, be the projection. Then, for every $k \in \mathbb{N}^+$ and every $n_i \in X_i$, where $i = 1, \ldots, k$, we have that (writing, for short, "$\pi_i^{-1}(n_i)$" instead of "$\pi_i^{-1}(\{n_i\})$")

$$f\left(\bigcap_{i=1}^{k} \pi_i^{-1}(n_i)\right) = \Delta_{n_1 n_2 \ldots n_k} \cap \mathbb{J}_2. \tag{17}$$

Let $\psi_i : RC(X_i) \longrightarrow RC(X)$, $F \mapsto \pi_i^{-1}(F)$, where $i \in \mathbb{N}^+$; then, as we have seen in the proof of Proposition 3.2, ψ_i is a complete monomorphism. Set $A'_i = \psi_i(RC(X_i))$. Since X_i is a discrete space, we have that $A_i = RC(X_i)$ and $A'_i \subseteq CO(X)$, for all $i \in \mathbb{N}^+$. Thus, for the elements of the subset $\bigcup_{i\in\mathbb{N}^+} A'_i$ of $RC(X)$, the Boolean operation "meet in $RC(X)$" coincides with the set-theoretic operation "intersection" between the subsets of X, and the same for the Boolean complement in $RC(X)$ and the set-theoretic complement in X. We also have that the Boolean algebras A_i and A'_i are isomorphic. Let A' be the subalgebra of $P(X)$ generated by $\bigcup_{i\in\mathbb{N}^+} A'_i$. Then A' is isomorphic to A. Note that A' is a subalgebra of $CO(X)$. Also, A' is a dense subalgebra of $RC(X)$; therefore, $RC(X)$ is the completion of A'. Thus, $\tilde{A}$ is isomorphic to $RC(X)$. So, without loss of generality, we can think that $\tilde{A}$ is $RC(X)$, A is A', $\varphi_i = \psi_i$ and hence $\varphi_i(A_i)$ is A'_i, for $i \in \mathbb{N}^+$. We will now construct an LCA $(RC(X), \sigma, \mathbb{B})$ LCA-isomorphic to $(RC(\mathbb{R}), \rho_{\mathbb{R}}, CR(\mathbb{R}))$. Then, identifying $RC(X)$ with $\tilde{A}$, we will show that $\sigma = \tilde{\sigma}$ and $\mathbb{B} = \widetilde{\mathbb{B}}$.

Let $\mathbb{B}_2 = \{M \in RC(\mathbb{J}_2) \mid \mathrm{cl}_{\mathbb{I}'}(M) \text{ is compact}\}$. For every two elements M and N of $RC(\mathbb{J}_2)$, set $M\rho_2 N \Leftrightarrow \mathrm{cl}_{\mathbb{I}'}(M) \cap \mathrm{cl}_{\mathbb{I}'}(N) \neq \emptyset$. Then, using Lemma 2.13, we get that the triple $(RC(\mathbb{J}_2), \rho_2, \mathbb{B}_2)$ is LCA-isomorphic to the LCA $(RC(\mathbb{I}'), \rho_{\mathbb{I}'}, CR(\mathbb{I}'))$ (which, in turn, is LCA-isomorphic to the local contact algebra $(RC(\mathbb{R}), \rho_{\mathbb{R}}, CR(\mathbb{R}))$). Now, for every two elements $F, G \in RC(X)$, we set

$$F\sigma G \Leftrightarrow f(F)\rho_2 f(G). \tag{18}$$

Also, we put

$$\mathbb{B} = \{f^{-1}(M) \mid M \in \mathbb{B}_2\}. \tag{19}$$

Obviously, $(RC(X), \sigma, \mathbb{B})$ is LCA-isomorphic to $(RC(\mathbb{R}), \rho_{\mathbb{R}}, CR(\mathbb{R}))$. In the rest of this proof, we will show that the definitions of $\mathbb{B}$ and σ given above agree with the corresponding definitions of $\widetilde{\mathbb{B}}$ and $\tilde{\sigma}$ given in the construction of $(\tilde{A}, \tilde{\sigma}, \widetilde{\mathbb{B}})$.

Note first that the subset B'_0 of A', which corresponds to the subset B_0 of A described in the construction of $(\tilde{A}, \tilde{\sigma}, \widetilde{\mathbb{B}})$, is the following:

$$B'_0 = \left\{ \bigcap_{i=1}^{k} \pi_i^{-1}(n_i) \mid k \in \mathbb{N}^+, (\forall i = 1, \ldots, k)(n_i \in X_i) \right\}. \tag{20}$$

Let $F, G \in B'_0$ and $F = \bigcap_{i=1}^{k} \pi_i^{-1}(n_i)$, $G = \bigcap_{i=1}^{l} \pi_i^{-1}(m_i)$. We can suppose, without loss of generality, that $k \leq l$. Then, by (17) and Lemma 2.13, $\mathrm{cl}_{\mathbb{I}'}(f(F)) = \Delta_{n_1 n_2 \ldots n_k}$ and $\mathrm{cl}_{\mathbb{I}'}(f(G)) = \Delta_{m_1 m_2 \ldots m_l}$. If $k = l$, then, clearly, $\Delta_{n_1 n_2 \ldots n_k} \cap \Delta_{m_1 m_2 \ldots m_k} \neq \emptyset$ iff $(n_i = m_i$, for all $i = 1, \ldots, k-1$, and $m_k \in \{n_k^-, n_k, n_k^+\})$. If $k < l$, then, obviously, $\Delta_{n_1 n_2 \ldots n_k} \cap \Delta_{m_1 m_2 \ldots m_l} \neq \emptyset$ iff $(n_i = m_i$, for all $i = 1, \ldots, k)$. Then, using (18) and the formula (11), we get that σ and $\tilde{\sigma}$ agree on B'_0 (or, equivalently, on B_0).

Let $F \in B'_0$, $F = \bigcap_{i=1}^{k} \pi_i^{-1}(n_i)$ and $n \in \mathbb{N}^+$. Then the element Q_{Fn} of A' corresponding to the element q_{bn} of A, where $b \in B_0$ corresponds to F, is the following:

$$Q_{Fn} = \left[\left(\bigcap_{i=1}^{k-1} \pi_i^{-1}(n_i)\right) \cap \pi_k^{-1}(n_k^-) \cap \pi_{k+1}^{-1}(succ(n))\right] \cup \left[F \cap \pi_{k+1}^{-1}(pred(-n))\right].$$

Clearly,

$$\begin{aligned} Q_{Fn} = \quad & \left[\bigcup_{s \in succ(n)} \left(\bigcap_{i=1}^{k-1} \pi_i^{-1}(n_i) \cap \pi_k^{-1}(n_k^-) \cap \pi_{k+1}^{-1}(s)\right)\right] \\ & \cup \left[\bigcup_{s \in pred(-n)} \left(\bigcap_{i=1}^{k} \pi_i^{-1}(n_i) \cap \pi_{k+1}^{-1}(s)\right)\right]. \end{aligned} \tag{21}$$

(It is easy to see, as well, that in the formula (21) the sign of the union can be replaced everywhere with the sign of the join in $RC(X)$.) Thus,

$$f(Q_{Fn}) = \left[\left(\bigcup_{s \in succ(n)} \Delta_{n_1 n_2 \ldots n_{k-1} n_k^- s}\right) \cup \left(\bigcup_{s \in pred(-n)} \Delta_{n_1 n_2 \ldots n_k s}\right)\right] \cap \mathbb{J}_2. \tag{22}$$

Let d be the left end-point of the closed interval $\Delta_{n_1 n_2 \ldots n_k}$. Then it is easy to see that

$$\mathrm{cl}_{\mathbb{I}'}(f(Q_{Fn})) = [d - \varepsilon_n, d + \varepsilon'_n], \tag{23}$$

where ε_n and ε'_n depend from n and also from $n_1, \ldots, n_k$ (for simplicity, we don't reflect this dependence on the notation), but for fixed $n_1, \ldots, n_k$, we have that $\varepsilon_n > \varepsilon_{n+1} > 0$, $\varepsilon'_n > \varepsilon'_{n+1} > 0$, for all $n \in \mathbb{N}^+$, and $\lim_{n\to\infty} \varepsilon_n = 0$, $\lim_{n\to\infty} \varepsilon'_n = 0$; also, the closed interval $[d - \varepsilon_n, d + \varepsilon'_n]$ lies in the open interval having as end-points the middles of the closed intervals $\Delta_{n_1 n_2 \ldots n_{k-1} n_k^-}$ and $\Delta_{n_1 n_2 \ldots n_k}$. Since the family $\{D \cap \mathbb{J}_2 \mid D \in \delta\}$ is a base of $\mathbb{J}_2$ and every element of $\mathbb{D}$ appears as a left end-point of some element of the family δ, we get that the family

$$\mathcal{B} = \{\mathrm{int}_{\mathbb{I}'}(\mathrm{cl}_{\mathbb{I}'}((f(F))), \mathrm{int}_{\mathbb{I}'}(\mathrm{cl}_{\mathbb{I}'}((f(Q_{Fn}))) \mid n \in \mathbb{N}^+, F \in B'_0\}$$

is a base of $\mathbb{I}'$. Also, if

$$\boldsymbol{B} = \{\mathrm{cl}_{\mathbb{I}'}((f(F)), \mathrm{cl}_{\mathbb{I}'}((f(Q_{Fn})) \mid n \in \mathbb{N}^+, F \in B'_0\},$$

then $\boldsymbol{B} = \{\mathrm{cl}_{\mathbb{I}'}(U) \mid U \in \mathcal{B}\}$ and $\boldsymbol{B} \subseteq CR(\mathbb{I}')$. Hence, $\boldsymbol{B}$ generates the ideal $CR(\mathbb{I}')$ of $RC(\mathbb{I}')$. Clearly, the family

$$B'_1 = \{Q_{Fn} \mid F \in B'_0, n \in \mathbb{N}^+\} \tag{24}$$

corresponds to the subset B_1 of A constructed above (before the formulation of our theorem). Since $\boldsymbol{B} = \{\mathrm{cl}_{\mathbb{I}'}(G) \mid G \in f(B'_0 \cup B'_1)\}$, we get that the subset $f(B'_0 \cup B'_1)$ of $RC(\mathbb{J}_2)$ generates the ideal $\mathbb{B}_2$ of $RC(\mathbb{J}_2)$. Thus, the subset $B'_0 \cup B'_1$ of $RC(X)$ generates the ideal $\mathbb{B}$ of $RC(X)$. Therefore, $\mathbb{B}$ corresponds to $\widetilde{\mathbb{B}}$; we can even write that $\mathbb{B} = \widetilde{\mathbb{B}}$.

Let now $r, r' \in \mathbb{N}^+$, $F, F' \in B'_0$, $F = \pi_1^{-1}(n_1) \cap \ldots \cap \pi_k^{-1}(n_k)$ and $F' = \pi_1^{-1}(n'_1) \cap \ldots \cap \pi_l^{-1}(n'_l)$. We can suppose, without loss of generality, that $k \leq l$. Let d and d' be the left end-points of the closed intervals $\Delta_{n_1 n_2 \ldots n_k}$ and $\Delta_{n'_1 n'_2 \ldots n'_l}$, respectively. Then, using (23), we get that $\mathrm{cl}_{\mathbb{I}'}(f(Q_{Fr})) = [d - \varepsilon_r, d + \varepsilon'_r]$ and $\mathrm{cl}_{\mathbb{I}'}(f(Q_{F'r'})) = [d' - \varepsilon_{r'}, d' + \varepsilon'_{r'}]$. If $k = l$, then it is easy to see that $\mathrm{cl}_{\mathbb{I}'}(f(Q_{Fr})) \cap \mathrm{cl}_{\mathbb{I}'}(f(Q_{F'r'})) \neq \emptyset$ iff $(n_i = n'_i$, for all $i = 1, \ldots, k)$. If $l = k + 1$, then one readily checks that $\mathrm{cl}_{\mathbb{I}'}(f(Q_{Fr})) \cap \mathrm{cl}_{\mathbb{I}'}(f(Q_{F'r'})) \neq \emptyset$ iff $[(n_i = n'_i$, for all $i = 1, \ldots, k-1)$ and $((n_k = n'_k$ and $n'_{k+1} \leq -r)$ or $(n'_k = (n_k)^-$ and $n'_{k+1} > r))]$. Finally, if $l > k + 1$, then $\mathrm{cl}_{\mathbb{I}'}(f(Q_{Fr})) \cap \mathrm{cl}_{\mathbb{I}'}(f(Q_{F'r'})) \neq \emptyset$ iff $[(n_i = n'_i$, for all $i = 1, \ldots, k-1)$ and $((n_k = n'_k$ and $n'_{k+1} < -r)$ or $(n'_k = (n_k)^-$ and $n'_{k+1} > r))]$. All this shows that the relations σ and $\tilde{\sigma}$ agree on B'_1 (or, equivalently, on B_1).

Let $r \in \mathbb{N}^+$, $F, F' \in B'_0$, $F = \pi_1^{-1}(n_1) \cap \ldots \cap \pi_k^{-1}(n_k)$ and $F' = \pi_1^{-1}(n'_1) \cap \ldots \cap \pi_l^{-1}(n'_l)$. If $l < k$, then we get that $\mathrm{cl}_{\mathbb{I}'}(f(Q_{Fr})) \cap \mathrm{cl}_{\mathbb{I}'}(f(F')) \neq \emptyset$ iff $(n_i = n'_i$, for all $i = 1, \ldots, l)$. If $l = k$, then $\mathrm{cl}_{\mathbb{I}'}(f(Q_{Fr})) \cap \mathrm{cl}_{\mathbb{I}'}(f(F')) \neq \emptyset$ iff $(n_i = n'_i$, for all $i = 1, \ldots, k-1$, and $n'_k \in \{n_k^-, n_k\})$. If $l = k + 1$, then $\mathrm{cl}_{\mathbb{I}'}(f(Q_{Fr})) \cap \mathrm{cl}_{\mathbb{I}'}(f(F')) \neq \emptyset$ iff $[(n_i = n'_i$, for all $i = 1, \ldots, k-1)$, and $((n'_k = n_k^-$ and $n'_{k+1} \geq r)$ or $(n'_k = n_k$ and $n'_{k+1} \leq -r))]$. Finally, if $l > k + 1$, then $\mathrm{cl}_{\mathbb{I}'}(f(Q_{Fr})) \cap \mathrm{cl}_{\mathbb{I}'}(f(F')) \neq \emptyset$ iff $[(n_i = n'_i$, for all $i = 1, \ldots, k-1)$, and $((n'_k = n_k^-$ and $n'_{k+1} > r)$ or $(n'_k = n_k$ and $n'_{k+1} < -r))]$. We get that the relations σ and $\tilde{\sigma}$ agree on $B'_0 \cup B'_1$ (or, equivalently, on $B_0 \cup B_1$).

Now, using the facts that $\mathcal{B}$ is a base of $\mathbb{I}'$, $\mathbb{I}'$ is a regular space, and $\mathrm{cl}_{\mathbb{I}'}(f(F))$ is a compact set for all $F \in \mathbb{B}$, we get that for all $F, G \in \mathbb{B}$, $\mathrm{cl}_{\mathbb{I}'}(f(F)) \cap \mathrm{cl}_{\mathbb{I}'}(f(G)) = \emptyset$ iff (there exist $F_1, \ldots, F_k, G_1, \ldots, G_l \in B'_0 \cup B'_1$ such that $F \subseteq \bigcup_{i=1}^k F_i$, $G \subseteq \bigcup_{j=1}^l G_j$ and $\mathrm{cl}_{\mathbb{I}'}(f(F_i)) \cap \mathrm{cl}_{\mathbb{I}'}(f(G_j)) = \emptyset$ for all $i = 1, \ldots, k$ and all $j = 1, \ldots, l)$. This shows that the relations σ and $\tilde{\sigma}$ agree on $\mathbb{B}$ (or, equivalently, on $\widetilde{\mathbb{B}}$).

Finally, as in every LCA, for every $F, G \in RC(X)$, we have that $F \sigma G$ iff (there exist $F', G' \in \mathbb{B}$ such that $F' \subseteq F$, $G' \subseteq G$ and $F' \sigma G'$). Therefore, the relations σ and $\tilde{\sigma}$ agree on $RC(X)$ (or, equivalently, on $\tilde{A}$). □

THEOREM 4.2. *For every $n \in \mathbb{N}^+$, the CLCA $(RC(\mathbb{R}^n), \rho_{\mathbb{R}^n}, CR(\mathbb{R}^n))$ $(= \Psi^t(\mathbb{R}^n))$ is LCA-isomorphic to the* **DHLC***-sum $(\tilde{A}_n, \tilde{\sigma}_n, \widetilde{\mathbb{B}}_n)$ of n copies of the CLCA $(\tilde{A}, \tilde{\sigma}, \widetilde{\mathbb{B}})$ (see Theorem 4.1 for it); thus, the CLCA $(\tilde{A}_n, \tilde{\sigma}_n, \widetilde{\mathbb{B}}_n)$ completely determines the Euclidean space $\mathbb{R}^n$ with its natural topology. For every $n \in \mathbb{N}^+$, the Boolean algebras $\tilde{A}_n$ and $\tilde{A}$ are isomorphic.*

Proof. Since $\mathbb{J}^n$ is homeomorphic to $\mathbb{J}$ and is dense in $\mathbb{R}^n$, we get that $RC(\mathbb{R}^n)$

is isomorphic to $RC(\mathbb{J})$, and thus, to $\tilde{A}$ (see "**The construction of** $(\tilde{A}, \tilde{\sigma}, \widetilde{\mathrm{IB}})$" and the proof of Theorem 4.1). Now all follows from Theorems 4.1 and 3.5. □

We will now present the description of the CLCA $(RC(\mathbb{R}), \rho_{\mathbb{R}}, CR(\mathbb{R}))$ in two new forms; the notation used in them permits to obtain a more compact form of the definitions of the corresponding relations. As we have already mentioned, $RC(\mathbb{R})$ is isomorphic to $RC(\mathbb{J})$, i.e. to $RC(\mathbb{Z}_0^{\mathbb{N}^+})$ or, equivalently, to $RC(\omega^\omega)$. The last algebra, which is one of the collapsing algebras $RC(k^\omega)$ (where k is an infinite cardinal equipped with the discrete topology), has many abstract descriptions. The one, which is the most appropriate for our purposes, is the following: a complete Boolean algebra C is isomorphic to the Boolean algebra $RC(k^\omega)$ iff it has a dense subset isomorphic to T^*, for the normal tree $T = \bigcup\{k^n \mid n \in \mathbb{N}^+\}$ (here T^* is the tree T with the opposite partial order and $k^n \cap k^m = \emptyset$ for $n \neq m$) (see, e.g., [29, 14.16(a),(b)]). (Recall that a partially ordered set $(T, \leq_T)$ is called a *tree* if for every $t \in T$, the set $pred(t)$ is well-ordered by $\leq_T$.) This shows that $RC(k^\omega)$ is isomorphic to the Boolean algebra $RC(T^*)$, where the ordered set T^* is endowed with the *left topology*, i.e. that one generated by the base $\{L_{T^*}(t) \mid t \in T\}$ (here $L_{T^*}(t) = \{t' \in T \mid t' \leq_{T^*} t\} = \{t' \in T \mid t \leq_T t'\}$, for every $t \in T$) (see, e.g., [29, 4.11-4.16] and [17, 1.7.2]).

Let us add some details and introduce some notation.

Notation. For any $n \in \mathbb{N}^+$, we set

$$\underline{n} = \{1, \ldots, n\}.$$

We set

$$T_0 = \bigcup\{\mathbb{Z}_0^n \mid n \in \mathbb{N}^+\},$$

where $\mathbb{Z}_0^n \cap \mathbb{Z}_0^m = \emptyset$ for $n \neq m$. Any element $t \in \mathbb{Z}_0^n$ is interpreted, as usual, as a function $t : \underline{n} \longrightarrow \mathbb{Z}_0$. Further, we let $\perp \subseteq t$ and $\perp \neq t$, for any $t \in T_0$; if $n, n' \in \mathbb{N}^+$, $t \in \mathbb{Z}_0^n$ and $t' \in \mathbb{Z}_0^{n'}$, then we set $t \subseteq t'$ iff t' is an extension of t, i.e. iff $n \leq n'$ and $t(i) = t'(i)$ for any $i \in \underline{n}$. Then the ordered set $(T_0 \cup \{\perp\}, \subseteq)$ is a normal tree of height ω with $\mathbb{Z}_0^n$ as its nth level (it will be denoted by L_n). We also put, for any $t, t' \in T_0 \cup \{\perp\}$,

$$t \leq t' \Leftrightarrow t' \subseteq t.$$

We set

$$T_0^* = (T_0 \cup \{\perp\}, \leq).$$

Let T_0^* be endowed with its left topology (i.e. let $(T_0 \cup \{\perp\}, \subseteq)$ be equipped with its right topology (which is defined analogously to the left topology (see [17, 1.7.2])))). Further, for any $t \in T_0 \cup \{\perp\}$, put

$$c_t = \{t' \in T_0 \mid t \text{ and } t' \text{ are } T_0^*\text{-compatible}\}.$$

(Recall that two elements x and y of a partially ordered set $(M, \preceq)$ are *compatible* if there is some $z \in M$ such that $z \preceq x$ and $z \preceq y$.) Then, as it is well known (see, e.g., [29, 4.13, 4.16, the formula for $\mathrm{cl}(u_p)$ in the proof of 4.16]), the embedding e of the partially ordered set T_0^* into the Boolean algebra $RC(T_0^*)$ is given by the formula

$$e(t) = c_t, \ \forall t \in T_0 \cup \{\bot\}.$$

(Note that the map e is an embedding because T_0^* is a *separative partial order* (see, e.g., [29, 4.15,4.16,p.226]).) Also, let us recall that the left topology on $T_0 \cup \{\bot\}$ induced by the ordered set T_0^* is an *Alexandroff topology*, i.e. the union of arbitrarily many closed sets is a closed set (see, e.g., [17, 1.7.2]). Thus, the (finite or infinite) joins $\bigvee\{F_j \mid j \in J\}$ in $RC(T_0^*)$ are just the unions $\bigcup\{F_j \mid j \in J\}$.

Finally, for every $n \in \mathbb{N}^+ \setminus \{1\}$ and every $t \in L_n$ (i.e. $t : \underline{n} \longrightarrow \mathbb{Z}_0$), define

$$t_\lambda : \underline{n} \longrightarrow \mathbb{Z}_0 \text{ by the formulas } (t_\lambda)_{|\ \underline{n-1}} = t_{|\ \underline{n-1}} \text{ and } t_\lambda(n) = (t(n))^-; \quad (25)$$

let, for $t \in L_1$, $t_\lambda : \underline{1} \longrightarrow \mathbb{Z}_0$ be defined by $t_\lambda(1) = (t(1))^-$.

Remark 4.3. *As we have already mentioned, the Boolean algebra $RC(\mathbb{Z}_0^{\mathbb{N}^+})$ is isomorphic to the Boolean algebra $RC(T_0^*)$ (see, e.g., [29, 14.16(a),(b),4.11-4.16]). We will recall the proof of this fact since we will use it later. For every $t \in T_0$, set*

$$a_t = \{x \in \mathbb{Z}_0^{\mathbb{N}^+} \mid t \subseteq x\}. \quad (26)$$

Note that if $t : \underline{n} \longrightarrow \mathbb{Z}_0$, where $n \in \mathbb{N}^+$, then

$$a_t = \bigcap_{i=1}^{n} \pi_i^{-1}(t(i)) \quad (27)$$

and thus a_t is a clopen subset of $\mathbb{Z}_0^{\mathbb{N}^+}$. Set

$$S = \{a_t \mid t \in T_0\} \cup \mathbb{Z}_0^{\mathbb{N}^+}. \quad (28)$$

Then $S \subseteq CO(\mathbb{Z}_0^{\mathbb{N}^+}) \subseteq RC(\mathbb{Z}_0^{\mathbb{N}^+})$. Now it is easy to see that the set S is dense in $RC(\mathbb{Z}_0^{\mathbb{N}^+})$ and isomorphic to T_0^* (indeed, the map

$$s : T_0^* \longrightarrow S, \text{ where } s(\bot) = \mathbb{Z}_0^{\mathbb{N}^+} \text{ and } s(t) = a_t, \forall t \in T_0 \quad (29)$$

is an isomorphism). Therefore, $RC(\mathbb{Z}_0^{\mathbb{N}^+})$ is isomorphic to the Boolean algebra $RC(T_0^*)$.

We will now equip the Boolean algebra $RC(T_0^*)$ defined above with an LCA-structure $(RC(T_0^*), \theta, \mathbb{B}_T)$ and will prove that the obtained CLCA is LCA-isomorphic to the CLCA $(RC(\mathbb{R}), \rho_{\mathbb{R}}, CR(\mathbb{R}))$. Recall that two elements x and y of a partially ordered set $(M, \preccurlyeq)$ are *comparable* if $x \preccurlyeq y$ or $y \preccurlyeq x$.

The construction of $(RC(T_0^*), \theta, \mathbb{B}_T)$. For every $k, n \in \mathbb{N}^+$ and for every $t \in L_k$ (recall that $L_k = \mathbb{Z}_0^k$), set

$$d_{tn} = \bigcup \Big\{ c_{t'} \mid (t' \in L_{k+1}) \\ \& \Big((t_\lambda \subseteq t' \ \& \ t'(k+1) > n) \text{ or } (t \subseteq t' \ \& \ t'(k+1) < -n) \Big) \Big\}.$$

Note that the fact that the left topology on T_0^* is an Alexandroff topology implies that

$$d_{tn} = \bigvee \Big\{ c_{t'} \mid (t' \in L_{k+1}) \\ \& \Big((t_\lambda \subseteq t' \text{ and } t'(k+1) > n) \text{ or } (t \subseteq t' \text{ and } t'(k+1) < -n) \Big) \Big\}. \tag{30}$$

Let

$$C_0 = \{c_t \mid t \in T_0\} \text{ and } C_1 = \{d_{tn} \mid t \in T_0,\ n \in \mathbb{N}^+\}. \tag{31}$$

Denote by $\mathbb{B}_{T_0}$ the ideal of $RC(T_0^*)$ generated by $C_0 \cup C_1$.
For every $k, k', n, n' \in \mathbb{N}^+$ and every $t \in L_k$, $t' \in L_{k'}$, set

$$c_t \theta c_{t'} \Leftrightarrow \begin{cases} t = t' \text{ or } t = t'_\lambda \text{ or } t' = t_\lambda, & \text{if } k = k' \\ t \text{ and } t' \text{ are comparable}, & \text{if } k \neq k', \end{cases} \tag{32}$$

and

$$d_{tn} \theta d_{t'n'} \Leftrightarrow \tag{33}$$

$$\begin{cases} (t' \subseteq t \text{ and } t(k'+1) < -n') \text{ or } (t'_\lambda \subseteq t \text{ and } t(k'+1) > n'), & \text{if } k > k'+1 \\ (t' \subseteq t \text{ and } t(k) \leq -n') \text{ or } (t'_\lambda \subseteq t \text{ and } t(k) > n'), & \text{if } k = k'+1 \\ t = t', & \text{if } k = k' \\ (t \subseteq t' \text{ and } t'(k') \leq -n) \text{ or } (t_\lambda \subseteq t' \text{ and } t'(k') > n), & \text{if } k = k'-1 \\ (t \subseteq t' \text{ and } t'(k+1) < -n) \text{ or } (t_\lambda \subseteq t' \text{ and } t'(k+1) > n), & \text{if } k < k'-1; \end{cases}$$

and also

$$d_{tn} \theta c_{t'} \Leftrightarrow c_{t'} \theta d_{tn} \Leftrightarrow \tag{34}$$

$$\begin{cases} t' \subseteq t, & \text{if } k' < k \\ t' = t \text{ or } t' = t_\lambda, & \text{if } k' = k \\ (t_\lambda \subseteq t' \text{ and } t'(k') \geq n) \text{ or } (t \subseteq t' \text{ and } t'(k') \leq -n), & \text{if } k' = k+1 \\ (t_\lambda \subseteq t' \ \& \ t'(k+1) > n) \text{ or } (t \subseteq t' \ \& \ t'(k+1) < -n), & \text{if } k' > k+1. \end{cases}$$

Further, for every two elements c and d of $\mathbb{B}_{T_0}$, set

$$c(-\theta)d \Leftrightarrow \Big(\exists k,l \in \mathbb{N}^+ \text{ and } \exists c_1,\dots,c_k,d_1,\dots,d_l \in C_0 \cup C_1 \text{ such that } c \subseteq \bigcup_{i=1}^{k} c_i,\ d \subseteq \bigcup_{j=1}^{l} d_j \text{ and } c_i(-\theta)d_j,\ \forall i=1,\dots,k \text{ and } \forall j=1,\dots,l\Big). \tag{35}$$

Finally, for every two elements a and b of $RC(T_0^*)$, set

$$a\theta b \Leftrightarrow (\exists c,d \in \mathbb{B}_{T_0} \text{ such that } c \subseteq a,\ d \subseteq b \text{ and } c\theta d). \tag{36}$$

THEOREM 4.4. *The triple $(RC(T_0^*),\theta,\mathbb{B}_{T_0})$ (constructed above) is a CLCA; it is LCA-isomorphic to the complete local contact algebra $(RC(\mathbb{R}),\rho_{\mathbb{R}},CR(\mathbb{R}))$. Thus, the triple $(RC(T_0^*),\theta,\mathbb{B}_{T_0})$ completely determines the real line $\mathbb{R}$ with its natural topology.*

Proof. In this proof, we will use the notation introduced in the following places of this paper: in Remark 4.3 and in the "**Notation**" before it, in "**The construction of** $(\tilde{A},\tilde{\sigma},\widetilde{\mathbb{B}})$" and in "**The construction of** $(RC(T_0^*),\theta,\mathbb{B}_T)$". As it follows from Remark 4.3 and [29, the proof of 4.14], there is an isomorphism $h : RC(T_0^*) \longrightarrow RC(\mathbb{Z}_0^{\mathbb{N}^+})$ defined by the formula $h(c) = \bigvee_{RC(\mathbb{Z}_0^{\mathbb{N}^+})}\{a_t \mid t \in T_0^*, c_t \subseteq c\}$, for every $c \in RC(T_0^*)$. Thus, $h(c_t) = a_t = \bigcap_{i=1}^{k} \pi_i^{-1}(t(i))$ and c_t corresponds to $\bigwedge_{i=1}^{k} \varphi_i(t(i))$ (see "**The construction of** $(\tilde{A},\tilde{\sigma},\widetilde{\mathbb{B}})$"), where $t \in L_k \subseteq T_0^*$ (i.e., $t : \underline{k} \longrightarrow \mathbb{Z}_0$). This implies that $h(C_0) = B_0' = \{a_t \mid t \in T_0\}$ and C_0 corresponds to $B_0 = \{\bigwedge_{i=1}^{k} \varphi_i(t(i)) \mid k \in \mathbb{N}^+, t \in L_k\}$ (see (31), (20), (7)). Note that t_λ corresponds to b_- (see (25) and (8)). Since h is a complete homomorphism, we get that $h(d_{tn}) = Q_{a_t n}$ and thus d_{tn} corresponds to $q_{a_t n}$, for every $k,n \in \mathbb{N}^+$ and every $t \in L_k$ (see (30), (21), (9)). Then $h(C_1) = B_1'$ and hence C_1 corresponds to B_1 (see (31), (24), (10)). Hence, $h(\mathbb{B}_{T_0}) = \mathbb{B}$ and therefore $\mathbb{B}_{T_0}$ corresponds to $\widetilde{\mathbb{B}}$ (see the line after (31), (19) and the paragraph after (24), the line after (10)). Having all these facts in mind, we obtain easily that the formula (32) follows from the formula (11), (33) from (12), (34) from (14), (35) from (15) and (36) from (16). This completes the proof of our theorem. □

THEOREM 4.5. *A CLCA $(M,\mu,\mathbb{M})$ is LCA-isomorphic to the complete local contact algebra $(RC(\mathbb{R}),\rho_{\mathbb{R}},CR(\mathbb{R}))$ iff there exists an embedding (between partially ordered sets) $\zeta : T_0^* \longrightarrow M$ such that the following two conditions are satisfied:*
(a) $\zeta(T_0)$ is dense in M, and
(b) let $\zeta(t) = z_t$, for every $t \in T_0$, and let the elements $\widetilde{d_{tn}}$ be defined by the formula (30) in which d_{tn} is replaced by $\widetilde{d_{tn}}$, and c_t is replaced by z_t; then the ideal $\mathbb{M}$ is generated by the set $Z = \zeta(T_0) \cup \{\widetilde{d_{tn}} \mid t \in T_0, n \in \mathbb{N}^+\}$ and the

formulas (32), (33), (34), (15), (16) hold with θ and $\tilde{\sigma}$ replaced by μ, c_t by z_t, d_{tn} by $\widetilde{d_{tn}}$, $\widetilde{\mathbb{B}}$ by $\mathbb{M}$, $B_0 \cup B_1$ by Z, and $\tilde{A}$ by M.

Proof. It follows from Theorem 4.4 and [29, 4.14,14.16]. □

5. A Whiteheadian-type description of Tychonoff cubes, spheres and tori

THEOREM 5.1. *For every $n \in \mathbb{N}^+$, the CNCA $(RC(\mathbb{S}^n), \rho_{\mathbb{S}^n})$ $(= \Psi^t(\mathbb{S}^n))$ is CA-isomorphic to the CNCA $(\tilde{A}_n, C_{\tilde{\sigma}_n, \widetilde{\mathbb{B}}_n})$ (see 4.2 for the LCA $(\tilde{A}_n, \tilde{\sigma}_n, \widetilde{\mathbb{B}}_n)$, and 2.5 for $C_{\tilde{\sigma}_n, \widetilde{\mathbb{B}}_n}$); thus, the CNCA $(\tilde{A}_n, C_{\tilde{\sigma}_n, \widetilde{\mathbb{B}}_n})$ completely determines the n-dimensional sphere $\mathbb{S}^n$ with its natural topology. Note that $\tilde{A}_n$ is isomorphic to $\tilde{A}$, for every $n \in \mathbb{N}^+$.*

Proof. As it follows from the proof of [38, Theorem 4.8], if X is a locally compact Hausdorff space then the complete normal contact algebra $(RC(\alpha X), \rho_{\alpha X})$ is CA-isomorphic to the complete normal contact algebra $(RC(X), C_{\rho_X, CR(X)})$. Now, since $\alpha\mathbb{R}^n$ is homeomorphic to $\mathbb{S}^n$, our result follows from Theorem 4.2. □

For every cardinal number τ, denote by $\mathbb{T}^\tau$ the space $(\mathbb{S}^1)^\tau$ (for finite τ, this is just the τ-dimensional torus).

THEOREM 5.2. *For every cardinal number τ, the complete normal contact algebra $(RC(\mathbb{T}^\tau), \rho_{\mathbb{T}^\tau})$ $(= \Psi^t(\mathbb{T}^\tau))$ is CA-isomorphic to the* **DHC***-sum of τ copies of the CNCA $(\tilde{A}, C_{\tilde{\sigma}, \widetilde{\mathbb{B}}})$ (see Theorem 5.1 for it); therefore, this* **DHC***-sum completely determines the space $\mathbb{T}^\tau$.*

Proof. Since the CNCA $(RC(\mathbb{S}^1), \rho_{\mathbb{S}^1})$ is CA-isomorphic to the complete normal contact algebra $(\tilde{A}, C_{\tilde{\sigma}, \widetilde{\mathbb{B}}})$ (see Theorem 5.1), our result follows from Theorem 3.7. □

Recall that if A is a Boolean algebra and $a \in A$ then the set $\downarrow(a) = \{b \in A \mid b \leq a\}$ endowed with the same meets and joins as in A and with complement b' defined by the formula $b' = b^* \wedge a$, for every $b \leq a$, is a Boolean algebra; it is denoted by $A|a$. If $J = \downarrow(a^*)$ then $A|a$ is isomorphic to the factor algebra A/J; the isomorphism $h : A|a \longrightarrow A/J$ is the following: $h(b) = [b]$, for every $b \leq a$ (see, e.g., [29]).

In [12], we proved the following theorem:

THEOREM 5.3 ([12, Theorem 6.8]). *Let X be a locally compact Hausdorff space and $F \in RC(X)$. Set $B = RC(X)|F$, $\mathbb{B}' = \{G \wedge F \mid G \in CR(X)\}$ and let, for every $a, b \in B$, $a\eta b$ iff $a\rho_X b$ (i.e. $a \cap b \neq \emptyset$). Then $(B, \eta, \mathbb{B}')$ is LCA-isomorphic to $\Psi^t(F)$.*

Using this assertion, we obtain the following result:

THEOREM 5.4. *Let $(M, \mu, \mathbb{M})$ be a CLCA which is LCA-isomorphic to the CLCA $(RC(\mathbb{R}), \rho_{\mathbb{R}}, CR(\mathbb{R}))$ and $\zeta : T_0^* \longrightarrow M$ be the embedding described in Theorem 4.5. Then, for each $t \in T_0$, the CNCA $(M|\zeta(t), \mu')$, where μ' is the restriction of the relation μ to $M|\zeta(t)$, is NCA-isomorphic to the CNCA $(RC(\mathbb{I}), \rho_{\mathbb{I}})$.*

Proof. By (17), (27) and the beginning of the proof of Theorem 4.1, if $t \in T_0$, i.e. $t : \underline{n} \longrightarrow \mathbb{Z}_0$ for some $n \in \mathbb{N}^+$, then the element $\zeta(t)$ coresponds to the element $\Delta_{t(1)\ldots t(n)}$ of $RC(\mathbb{I}')$ (see also the proofs of theorems 4.4 and 4.5). Since $\Delta_{t(1)\ldots t(n)}$ is homeomorphic to $\mathbb{I}$, our assertion follows from Theorem 5.3. □

The last theorem shows, in particular, that the following assertion holds (the notation from "**The construction of** $(\tilde{A}, \tilde{\sigma}, \widetilde{\mathbb{B}})$" will be used in it):

THEOREM 5.5. *Let $m \in \mathbb{N}^+$, $n_1, \ldots, n_m \in \mathbb{Z}_0$, $a_j = \{n_j\}$ for $j = 1, \ldots, m$, $u = \bigwedge_{j=1}^{m} \varphi_j(a_j)$ and $B = \tilde{A}|u$. Then the CNCA $(B, \tilde{\sigma}')$, where $\tilde{\sigma}'$ is the restriction of the relation $\tilde{\sigma}$ to B, is NCA-isomorphic to the CNCA $(RC(\mathbb{I}), \rho_{\mathbb{I}})$. In particular, the CNCA $(RC(\mathbb{I}), \rho_{\mathbb{I}})$ is NCA-isomorphic to the CNCA $(\tilde{A}|\varphi_1(\{1\}), \tilde{\sigma}')$.*

A direct description of the CNCA $(RC(\mathbb{I}), \rho_{\mathbb{I}})$ is given below.

The construction of $(\tilde{A}, \tilde{\sigma}')$. We will use the notation from "**The construction of** $(\tilde{A}, \tilde{\sigma}, \widetilde{\mathbb{B}})$".

We will define a relation $\tilde{\sigma}'$ on the Boolean algebra $\tilde{A}$.

For every $n \in \mathbb{N}^+$, set

$$u_n^{\uparrow} = \varphi_1(succ(n)) \text{ and } u_n^{\downarrow} = \varphi_1(pred(-n))$$

and let

$$B_2 = \{u_n^{\uparrow}, u_n^{\downarrow} \mid n \in \mathbb{N}^+\}.$$

For every $a, b \in B_0 \cup B_1 \cup B_2$, set

$$a\tilde{\sigma}'b \Leftrightarrow a\tilde{\sigma}b$$

(see the construction of $(\tilde{A}, \tilde{\sigma}, \widetilde{\mathbb{B}})$ for the definition of the relation $\tilde{\sigma}$). For convenience of the reader, we will write down the corresponding formulae. For every $n, m \in \mathbb{N}^+$,

$$u_n^{\uparrow}\tilde{\sigma}'u_m^{\uparrow}, \quad u_n^{\downarrow}\tilde{\sigma}'u_m^{\downarrow} \text{ and } u_n^{\downarrow}(-\tilde{\sigma}')u_m^{\uparrow}.$$

Further, for every $n, r \in \mathbb{N}^+$ and every $b = \varphi_1(a_1) \wedge \ldots \wedge \varphi_k(a_k) \in B_0$, where $a_1 = \{m\}$,

$$b\tilde{\sigma}'u_n^{\uparrow} \Leftrightarrow \begin{cases} m \geq n, & \text{if } k = 1 \\ m > n, & \text{if } k > 1 \end{cases}, \quad b\tilde{\sigma}'u_n^{\downarrow} \Leftrightarrow \begin{cases} m \leq -n, & \text{if } k = 1 \\ m < -n, & \text{if } k > 1 \end{cases} \tag{37}$$

and

$$q_{br}\tilde{\sigma}'u_n^{\uparrow} \Leftrightarrow m > n, \quad q_{br}\tilde{\sigma}'u_n^{\downarrow} \Leftrightarrow \begin{cases} m \leq -n, & \text{if } k = 1 \\ m < -n, & \text{if } k > 1. \end{cases} \tag{38}$$

Now, for every $c, d \in \tilde{A}$, set

$$c(-\tilde{\sigma}')d \Leftrightarrow \Big(\exists k, l \in \mathbb{N}^+ \text{ and } \exists c_1, \ldots, c_k, d_1, \ldots, d_l \in B_0 \cup B_1 \cup B_2 \text{ such that}$$
$$c \leq \bigvee_{i=1}^{k} c_i,\ d \leq \bigvee_{j=1}^{l} d_j \text{ and } c_i(-\tilde{\sigma}')d_j,\ \forall i = 1, \ldots, k \text{ and } \forall j = 1, \ldots, l\Big). \tag{39}$$

THEOREM 5.6. *The pair $(\tilde{A}, \tilde{\sigma}')$ (constructed above) is a complete normal contact algebra; it is CA-isomorphic to the CNCA $(RC(\mathbb{I}), \rho_{\mathbb{I}})$. Thus, the pair $(\tilde{A}, \tilde{\sigma}')$ completely determines the closed interval $\mathbb{I}$ with its natural topology.*

Proof. The proof of this assertion is analogous to the proof of Theorem 4.1. We will use in it the notation introduced in Theorem 4.1, in "**The construction of** $(\tilde{A}, \tilde{\sigma}, \widetilde{\mathbb{IB}})$" and in the above construction.

Clearly, $RC(\mathbb{R})$ is isomorphic to $RC(\mathbb{I})$ (by Lemma 2.13). Thus, $RC(\mathbb{I})$ is isomorphic to $RC(X)$, where $X = \mathbb{Z}_0^{\mathbb{N}^+}$ (see the proof of Theorem 4.1). We will now construct an NCA $(RC(X), \sigma')$ CA-isomorphic to $(RC(\mathbb{I}), \rho_{\mathbb{I}})$. Then, identifying $RC(X)$ with $\tilde{A}$, we will show that $\sigma' = \tilde{\sigma}'$.

For every two elements M and N of $RC(\mathbb{J}_2)$, set $M\rho_1 N \Leftrightarrow \mathrm{cl}_{\mathbb{I}}(M) \cap \mathrm{cl}_{\mathbb{I}}(N) \neq \emptyset$. Then, using Lemma 2.13, we get that the pair $(RC(\mathbb{J}_2), \rho_1)$ is CA-isomorphic to the NCA $(RC(\mathbb{I}), \rho_{\mathbb{I}})$. Now, for every two elements $F, G \in RC(X)$, we set

$$F\sigma'G \Leftrightarrow f(F)\rho_1 f(G), \tag{40}$$

where $f : X \longrightarrow \mathbb{J}_2$ is the homeomorphism constructed in the proof of Theorem 4.1. Obviously, $(RC(X), \sigma')$ is CA-isomorphic to $(RC(\mathbb{I}), \rho_{\mathbb{I}})$. In the rest of this proof, we will show that the definition of σ' given above agrees with the definition of $\tilde{\sigma}'$ given in the construction of $(\tilde{A}, \tilde{\sigma}')$.

Using the proof of Proposition 3.2, it is easy to see that the set

$$B_2' = \Big\{\pi_1^{-1}(succ(n)), \pi_1^{-1}(pred(-n)) \mid n \in \mathbb{N}^+\Big\}$$

corresponds to the set B_2 introduced in the construction of $(\tilde{A}, \tilde{\sigma}')$. Now, the formula (17) implies that, for every $n \in \mathbb{N}^+$,

$$cl_{\mathbb{I}}(f(\pi_1^{-1}(succ(n)))) = \left[1 - \frac{1}{2^{n+1}}\,,\, 1\right] \tag{41}$$

and

$$\mathrm{cl}_{\mathbb{I}}(f(\pi_1^{-1}(pred(-n)))) = \left[0\,,\ \frac{1}{2^{n+1}}\right]. \tag{42}$$

Thus, for every $m, n \in \mathbb{N}^+$, $\mathrm{cl}_{\mathbb{I}}(f(\pi_1^{-1}(succ(n)))) \cap \mathrm{cl}_{\mathbb{I}}(f(\pi_1^{-1}(pred(-m)))) = \emptyset$. Also, for every $m, n \in \mathbb{N}^+$, we have that $f(\pi_1^{-1}(succ(n))) \cap f(\pi_1^{-1}(succ(m))) \neq \emptyset$ and $f(\pi_1^{-1}(pred(-n))) \cap f(\pi_1^{-1}(pred(-m))) \neq \emptyset$. Having in mind these formulae and the fact that $\mathrm{cl}_{\mathbb{I}}(f(F)) = \mathrm{cl}_{\mathbb{I}'}(f(F))$, for every $F \in B_0' \cup B_1'$ (see the proof of Theorem 4.1 for the notation), we get that $G\sigma H \Leftrightarrow G\sigma' H$, for every $G, H \in B_0' \cup B_1' \cup B_2'$. This shows that $a\tilde{\sigma}'b \Leftrightarrow a\tilde{\sigma}b$, for every $a, b \in B_0 \cup B_1 \cup B_2$. Hence, the definitions of σ' and $\tilde{\sigma}'$ agree on $B_0' \cup B_1' \cup B_2'$ (or, equivalently, on $B_0 \cup B_1 \cup B_2$).

Further, using (41) and (42), we get that the family

$$\mathcal{B}_1 = \mathcal{B} \cup \{\mathrm{int}_{\mathbb{I}}(\mathrm{cl}_{\mathbb{I}}(f(F))) \mid F \in B_2'\}$$

(see the proof of Theorem 4.1 for the notation and for the fact that $\mathcal{B}$ is a base of $\mathbb{I}'$) is a base of $\mathbb{I}$. Thus, by the regularity of $\mathbb{I}$, every two disjoint elements of $RC(\mathbb{I})$ can be separated by the finite unions of the elements of the family $\{\mathrm{cl}_{\mathbb{I}}(f(F)) \mid F \in B_0' \cup B_1' \cup B_2'\}$. This implies that the definitions of σ' and $\tilde{\sigma}'$ agree on $RC(X)$ (or, equivalently, on $\tilde{A}$). □

THEOREM 5.7. *For every cardinal number τ, the complete normal contact algebra $(RC(\mathbb{I}^\tau), \rho_{\mathbb{I}^\tau})$ $(= \Psi^t(\mathbb{I}^\tau))$ is CA-isomorphic to the* **DHC***-sum of τ copies of the CNCA $(\tilde{A}, \tilde{\sigma}')$ (see Theorem 5.6 for it); therefore, this* **DHC***-sum completely determines the space $\mathbb{I}^\tau$.*

Proof. It follows from Theorems 5.6 and 3.7. □

Acknowledgements. The author is very grateful to the referee for the helpful suggestions.

REFERENCES

[1] J. ADÁMEK, H. HERRLICH, AND G. E. STRECKER, *Abstract and Concrete Categories*, Wiley Interscience, New York, 1990.

[2] M. AIELLO, I. PRATT-HARTMANN, AND J. VAN BENTHEM, *(Eds.), Handbook of Spatial Logics*, Springer, Berlin Heidelberg, 2007.

[3] P. S. ALEXANDROFF, *Outline of Set Theory and General Topology*, Nauka, Moscow, 1977, (In Russian).

[4] PH. BALBIANI, *(Ed.), Special Issue on Spatial Reasoning*, J. Appl. Non-Classical Logics, vol. 12, 2002.

[5] B. BENNETT AND I. DÜNTSCH, *Axioms, algebras and topology*, In: M. Aiello, I. Pratt-Hartmann and J. van Benthem (Eds.), Handbook of Spatial Logics (Berlin Heidelberg), Springer, 2007, pp. 99–160.

[6] A. Cohn and S. Hazarika, *Qualitative spatial representation and reasoning: an overview*, Fund. Inform. **46** (2001), 1–29.
[7] A. Cohn and J. Renz, *Qualitative spatial representation and reasoning*, In: F. van Hermelen, V. Lifschitz and B. Porter (Eds.), Handbook of Knowledge Representation, Elsevier, 2008, pp. 551–596.
[8] W. Comfort and S. Negrepontis, *Chain Conditions in Topology*, Cambridge Univ. Press, Cambridge, 1982.
[9] T. de Laguna, *Point, line and surface as sets of solids*, J. Philos. **19** (1922), 449–461.
[10] H. de Vries, *Compact Spaces and Compactifications, an Algebraic Approach*, Van Gorcum and Comp. N.V., Assen, 1962.
[11] G. Dimov, *A de Vries-type duality theorem for the category of locally compact spaces and continuous maps – I*, Acta Math. Hungar. **129** (2010), 314–349.
[12] G. Dimov, *A de Vries-type duality theorem for the category of locally compact spaces and continuous maps – II*, Acta Math. Hungar. **130** (2011), 50–77.
[13] G. Dimov and D. Vakarelov, *Contact algebras and region-based theory of space: a proximity approach – I*, Fund. Inform. **74** (2006), 209–249.
[14] I. Düntsch, *(Ed.), Special issue on Qualitative Spatial Reasoning*, Fund. Inform., vol. 46, 2001.
[15] I. Düntsch and M. Winter, *A representation theorem for Boolean contact algebras*, Theoret. Comput. Sci. **347** (2005), 498–512.
[16] V. Efremovič, *Infinitesimal spaces*, Doklady Akad. Nauk SSSR **76** (1951), 341–343.
[17] R. Engelking, *General Topology*, PWN, Warszawa, 1977.
[18] V. V. Fedorchuk, *Boolean δ-algebras and quasi-open mappings*, Sibirsk. Mat. Ž. **14** (1973), 1088–1099.
[19] M. P. Fourman and J. M. E. Hyland, *Sheaf models for analysis*, In: Applications of Sheaves, Springer LNM, vol. 753, 1979, pp. 280–301.
[20] I. M. Gelfand, *On normed rings*, Doklady Akad. Nauk USSR **23** (1939), 430–432.
[21] I. M. Gelfand, *Normierte Ringe*, Mat. Sbornik **9** (1941), 3–24.
[22] I. M. Gelfand and M. A. Naimark, *On the embedding of normed rings into the ring of operators in Hilbert space*, Mat. Sbornik **12** (1943), 197–213.
[23] I. M. Gelfand and G. E. Shilov, *Über verschiedene Methoden der Einführung der Topologie in die Menge der maximalen Ideale eines normierten Ringes*, Mat. Sbornik **9** (1941), 25–39.
[24] G. Gerla, *Pointless geometries*, In: Handbook of Incidence Geometry, F. Buekenhout (Ed.), Elsevier Science B.V., 1995, pp. 1015–1031.
[25] St. Givant and P. Halmos, *Introduction to Boolean Algebras*, Springer, 2009.
[26] R. J. Grayson, *Intuitionistic Set Theory*, Ph.D. thesis, Oxford University, 1978.
[27] A. Grzegorczyk, *Axiomatization of geometry without points*, Synthese **12** (1960), 228–235.
[28] P. T. Johnstone, *Stone Spaces*, Cambridge Univ. Press, Cambridge, 1982.
[29] S. Koppelberg, *Handbook on Boolean algebras, vol. 1: General Theory of Boolean Algebras*, North Holland, 1989.
[30] S. Leader, *Local proximity spaces*, Math. Ann. **169** (1967), 275–281.

[31] I. Pratt-Hartmann, *First-Order Mereotopology*, In: M. Aiello, I. Pratt-Hartmann and J. van Benthem (Eds.), Handbook of Spatial Logics (Berlin Heidelberg), Springer-Verlag, 2007, pp. 13–97.
[32] D. A. Randell, Z. Cui, and A. G. Cohn, *A spatial logic based on regions and connection*, In: B. Nebel, W. Swartout and C. Rich (Eds.), Proceedings of the 3rd International Conference Knowledge Representation and Reasoning (Los Allos, CA), Morgan Kaufmann, 1992, pp. 165–176.
[33] P. Roeper, *Region-based topology*, J. of Philos. Logic **26** (1997), 251–309.
[34] R. Sikorski, *Boolean Algebras*, Springer-Verlag, Berlin, 1964.
[35] M. H. Stone, *Applications of the theory of Boolean rings to general topology*, Trans. Amer. Math. Soc. **41** (1937), 375–481.
[36] A. Tarski, *Les fondements de la géométrie des corps*, First Polish Mathematical Congress (Lwów), 1927.
[37] D. Vakarelov, *Region-based theory of space: Algebras of regions, representation theory and logics*, In: Dov Gabbay et al. (Eds.), Mathematical Problems from Applied Logics. New Logics for the XXIst Century. II (Berlin Heidelberg), Springer-Verlag, 2007, pp. 267–348.
[38] D. Vakarelov, G. Dimov, I. Düntsch, and B. Bennett, *A proximity approach to some region-based theories of space*, J. Appl. Non-Classical Logics **12** (2002), 527–559.
[39] A. N. Whitehead, *An Enquiry Concerning the Principles of Natural Knowledge*, University Press, Cambridge, 1919.
[40] A. N. Whitehead, *Process and Reality*, MacMillan, New York, 1929.

Author's address:

Georgi D. Dimov
Department of Mathematics and Informatics
Sofia University
5 J. Bourchier Blvd., 1164 Sofia, Bulgaria
E-mail: gdimov@fmi.uni-sofia.bg

Received March 18, 2012
Revised May 5, 2012

Rend. Istit. Mat. Univ. Trieste
Volume 44 (2012), 75–87

Periodic solutions for quasilinear complex-valued differential systems involving singular ϕ-Laplacians

Jean Mawhin

Cordially dedicated to Fabio Zanolin, for his sixtieth birthday anniversary, and the twentieth anniversary of our first joint paper

Abstract. *Topological degree is used to obtain sufficient conditions for the existence of periodic solutions of systems of second order complex-valued ordinary differential equations involving a singular ϕ-Laplacian. Corresponding results for first order equations are also obtained.*

Keywords: periodic solutions, complex-valued systems, topological degree, singular ϕ-Laplacian
MS Classification 2010: 34C25, 55M25

1. Introduction

In [8], Manásevich, Zanolin and the author have used topological degree arguments to study the existence of periodic solutions for some complex-valued differential equations of the form

$$z'' = f(t, z, z'). \tag{1}$$

or for systems of such equations, where the nonlineary $f : [0, T] \times \mathbb{C}^2 \to \mathbb{C}$ has some special structure inspired by the equations of Liénard or Rayleigh. The existence conditions, as well as the technicalities to obtain the requested a priori bounds, are rather involved.

On the other hand, Bereanu and the author [1, 2, 3] have considered the existence of solutions of quasilinear differential equations or systems of the form

$$(\phi(u'))' = f(t, u, u'), \tag{2}$$

where $f : [0, T] \times \mathbb{R}^{2n} \to \mathbb{R}^n$ satisfies Carathéodory conditions and $\phi : B(a) \to \mathbb{R}^n$ belongs to a suitable class of so-called *singular* homeomorphisms between the open ball $B(a) \subset \mathbb{R}^n$ of center 0 and radius $a > 0$ and $\mathbb{R}^n$. A *solution* of (2)

on $[0,T]$ is a function $u \in C^1([0,T],\mathbb{R}^n)$ such that $u'(t) \in B(a)$ for all $t \in [0,T]$, $\phi \circ u'$ is absolutely continuous and equation (2) holds almost everywhere. A motivating example of singular homeomorphism comes from the relativistic acceleration, associated to the homeomorphism

$$\phi : B(1) \to \mathbb{R}^n,\ v \mapsto \frac{v}{\sqrt{1-|v|^2}}\,.$$

Despite of the apparent greater complexity of equation (2) with respect to (1), existence conditions for periodic solutions of (2) are in general weaker than those for (1).

Hence it may be of interest to study the problem of the existence of periodic solutions for quasilinear complex-valued differential systems of the form

$$(\phi(z'))' = f(t,z,z'). \tag{3}$$

where $\phi : B(a) \subset \mathbb{C}^m \to \mathbb{C}^m$ is a singular homeomorphism and $f : [0,T] \times \mathbb{C}^{2m} \to \mathbb{C}^m$ is a Carathéodory function. This is done in Section 3, where we state and prove fairly general results for nonlinearities containing the Liénard or Rayleigh types. A very special case is the existence of a solution for the problem

$$\left(\frac{z'}{\sqrt{1-|z|^2}}\right)' = \alpha z^n + h(t), \quad z(0) = z(T),\ z'(0) = z'(T) \tag{4}$$

for every integer $n \geq 1$, $\alpha \in \mathbb{C}\backslash\{0\}$, and $h \in L^1([0,T],\mathbb{C})$. Such a result is sharp because, when $\alpha = 0$, problem (4) has no solution when $T^{-1}\int_0^T h(t)\,dt \neq 0$.

On the other hand, motivated by some work of Szrednicki [10, 11], Manásevich, Zanolin and the author have proved in [7] existence conditions for periodic solutions of some first order complex-valued differential equations. In the special case of the complex Riccati equation

$$z' = z^2 + h(t), \quad z(0) = z(T),$$

interesting existence and non-existence results have been subsequently obtained by Campos and Ortega [4, 5]. Hence it may be of interest to consider first order periodic problems of the type

$$(\phi(z))' = f(t,z), \quad z(0) = z(T),$$

where $\phi : B(a) \subset \mathbb{C} \to \mathbb{C}$ is a suitable singular homeomorphism. This is done in Section 4, where a very special case of the obtained results is the existence of a solution for the problem

$$\left(\frac{z}{\sqrt{1-|z|^2}}\right)' = \alpha z^n + h(t), \quad z(0) = z(T), \tag{5}$$

for every $n \geq 1$, $\alpha \in \mathbb{C} \setminus \{0\}$ and $h \in L^1([0,T],\mathbb{C})$ such that

$$\left| T^{-1} \int_0^T h(t)\,dt \right| < |\alpha|.$$

Again, this condition is sharp because, when $\alpha = 0$, problem (5) has no solution when $T^{-1}\int_0^T h(t)\,dt \neq 0$.

We end this introduction with some notations. We denote some norm in $\mathbb{R}^n$ by $|\cdot|$, and the usual norm in $L^p := L^p(0,T;\mathbb{R}^n)$ $(1 \leq p \leq \infty)$ by $|\cdot|_p$. For $k \geq 0$, we set $C^k := C^k([0,T],\mathbb{R}^n)$ and $W^{1,1} := W^{1,1}([0,T],\mathbb{R}^n)$. The usual norm $|\cdot|_\infty$ is considered on C, and the space C^1 is endowed with the norm

$$|v|_{1,\infty} = |v|_\infty + |v'|_\infty.$$

Each $v \in C$ can be written $v(t) = v_0 + \widehat{v}(t)$, with $v_0 = v(0)$ and $\widehat{v}(0) = 0$. For $u \in W^{1,1}$ such that $u(0) = u(T)$, we have

$$\widehat{u}(t) = \int_0^t u'(s)\,ds = -\int_t^T u'(s)\,ds,$$

and $\max_{[0,T]} |\widehat{u}|$ being reached either in $[0,T/2]$ or in $[T/2,T]$, this gives

$$|\widehat{u}|_\infty \leq \frac{T}{2}|u'|_\infty \tag{6}$$

It is easily shown that the constant $T/2$ is optimal. We define the mean value $\overline{u}$ of $u \in L^1$ by

$$\overline{u} := T^{-1}\int_0^T u(t)\,dt,$$

2. A continuation theorem for periodic solutions of quasilinear systems involving singular ϕ-Laplacians

Let us consider now the periodic problem

$$(\phi(u'))' = f(t,u,u'), \quad u(0) = u(T),\ u'(0) = u'(T), \tag{7}$$

where $f : [0,T] \times \mathbb{R}^{2n} \to \mathbb{R}^n$ is a Carathéodory function and $\phi : B(a) \to \mathbb{R}^n$ $(a < +\infty)$ satisfies the following assumption introduced in [3].

(H_Φ) *ϕ is a homeomorphism from $B(a) \subset \mathbb{R}^n$ onto $\mathbb{R}^n$ such that $\phi(0) = 0$, $\phi = \nabla\Phi$, with $\Phi : \overline{B(a)} \to \mathbb{R}$ of class C^1 on $B(a)$, continuous, strictly convex on $\overline{B(a)}$, and such that $\Phi(0) = 0$.*

The motivating example is given by the C^∞-mapping $\Phi : \overline{B(1)} \subset \mathbb{R}^n \to \mathbb{R}$ defined by

$$\Phi(u) = 1 - \sqrt{1 - |u|^2} \quad (u \in \overline{B(1)}),$$

so that

$$\phi(u) = \nabla\Phi(u) = \frac{u}{\sqrt{1 - |u|^2}} \quad (u \in B(1)).$$

Hence $(\phi(u'))'$ describes the relativistic acceleration.

Notice that the scalar problem

$$(\phi(u'))' = 1, \quad u(0) = u(T), \quad u'(0) = u'(T)$$

has no solution, because the existence of a solution would imply, by integration over $[0, T]$ of both members of the differential equation and use of the boundary conditions, that $0 = T$. Hence we cannot expect an existence result for *any* right-hand side of the differential system in (7).

The following continuation result essentially comes from [1], and its present form is given in [9]. We denote by d_B the Brouwer degree for continuous mappings in $\mathbb{R}^n$ (see e.g. [6]).

LEMMA 1. *Assume that there exists an open bounded set $\Omega \subset C$ such that the following conditions hold :*

1. *For each $\lambda \in (0, 1]$, there is no solution of the problem*

$$(\phi(u'))' = \lambda f(t, u, u'), \quad u(0) = u(T),\ u'(0) = u'(T) \tag{8}$$

such that $u \in \partial\Omega$.

2. *There is no solution $u_0 \in \partial\Omega \cap \mathbb{R}^n$ of the system in $\mathbb{R}^n$*

$$\overline{f}(u_0) := T^{-1} \int_0^T f(t, u_0, 0)\, dt = 0, \tag{9}$$

where, in $\partial\Omega \cap \mathbb{R}^n$, $\mathbb{R}^n$ is identified with the subspace of constant functions in C.

3. *$d_B[\overline{f}, \Omega \cap \mathbb{R}^n, 0] \neq 0$.*

Then problem (7) has at least one solution such that $u \in \Omega$.

3. Periodic solutions of complex-valued quasilinear systems involving singular ϕ-Laplacians

In this section, let us provide $\mathbb{R}^2$ with the multiplication structure of the complex plane $\mathbb{C}$, and consider the complex-valued periodic system in $\mathbb{C}^m \simeq \mathbb{R}^{2m}$ with $m \geq 1$ an integer,

$$\begin{aligned}(\phi_k(z'))' = \alpha_k(t) z_k^{n_k} + [F_k(t,z)]' + h_k(t,z,z') \qquad (k = 1,2,\dots,m)\\ z(0) = z(T),\ z'(0) = z'(T), \end{aligned} \tag{10}$$

where $z' = (z_1', \dots, z_m')$, $z = (z_1, \dots, z_m)$, $\phi = (\phi_1, \dots, \phi_m) : B(a) \subset \mathbb{C}^m \to \mathbb{C}^m$ satisfies Assumption (H_ϕ), $n_k \geq 1$ is an integer, $\alpha_k \in L^1$, $F_k : [0,T] \times \mathbb{C}^m \to \mathbb{C}^m$ is of class C^1, and $h_k : [0,T] \times \mathbb{C}^{2m} \to \mathbb{C}^m$ is a Carathéodory function $(k = 1,2,\dots,m)$. For $z = (z_1, \dots, z_m)$, we take

$$|z| = \max\{|z_1|, \dots, |z_m|\},$$

and for $z \in C$,

$$|z|_\infty = \max_{t \in [0,T]} |z(t)|.$$

We set

$$n = \min\{n_1, \dots, n_m\}, \quad N = \max\{n_1, \dots, n_m\}.$$

THEOREM 1. *Assume that, for each $k = 1,2,\dots,m$, $\overline{\alpha}_k \neq 0$, and there exist $1 \leq \sigma_k < n$ and $\beta_k, \gamma_k \in L^1$ such that*

$$|h_k(t,z,v)| \leq \beta_k(t)|z|^{\sigma_k} + \gamma_k(t) \tag{11}$$

for a.e. $t \in [0,T]$, all $z \in \mathbb{C}^m$ and all $v \in \mathbb{C}^m$ such that $|v| < a$. Then problem (10) has at least one solution.

Proof. Following Lemma 1, we introduce the homotopy

$$\begin{aligned}(\phi_k(z'))' = \lambda[\alpha_k(t) z^{n_k} + [F_k(t,z)]' + h_k(t,z,z')] \quad (k = 1,2,\dots,m)\\ z(0) = z(T),\ z'(0) = z'(T) \quad (\lambda \in (0,1]). \end{aligned} \tag{12}$$

If $z(t) = z_0 + \widehat{z}(t)$ with $z_0 = z(0)$ is a possible solution of (12), then z' satisfies the inequality,

$$|z'|_\infty < a. \tag{13}$$

and hence by (6) the inequality

$$|\widehat{z}|_\infty < \frac{aT}{2}. \tag{14}$$

On the other hand, integrating both members of (12) over one period and using the periodicity gives

$$0 = \int_0^T \alpha_k(t)[z_{0,k} + \widehat{z}_k(t)]^{n_k}\,dt + \int_0^T h_k[t, z_0 + \widehat{z}(t), z'(t)]\,dt \qquad (k = 1, 2, \dots, m),$$

and hence, letting $C_n^j = \frac{n!}{j!(n-j)!}$,

$$\begin{aligned}\overline{\alpha}_k z_{0,k}^{n_k} &= -T^{-1}\int_0^T \left[\sum_{j=0}^{n_k-1} C_{n_k}^j z_{0,k}^j \widehat{z}_k(t)^{n_k-j}\right] dt \\ &\quad -T^{-1}\int_0^T h_k(t, z_0 + \widehat{z}(t), z'(t))\,dt \quad (k = 1, \dots, m).\end{aligned}$$

Consequently, using (11), (13) and (14),

$$|\overline{\alpha}_k||z_{0,k}|^{n_k} \le \sum_{j=0}^{n_k-1} C_{n_k}^j (aT/2)^{n_k-j}|z_{0,k}|^j + \overline{\beta}_k 2^{\sigma_k}[|z_0|^{\sigma_k} + (aT/2)^{\sigma}] + \overline{\gamma}_k \qquad (k = 1, \dots, m). \quad (15)$$

Let $k_0 \in \{1, \dots, m\}$ be such that $|z_{0,k_0}| = |z_0|$. Then, either $|z_0| < 1$ or, using (15) with $k = k_0$, $|z_0| \ge 1$ and

$$\alpha|z_0|^n \le \sum_{j=0}^{N-1} C_N^j \eta(a,T)^{N-j}|z_0|^j + 2^{\sigma}\beta[|z_0|^{\sigma} + \eta(a,T)^{\sigma}] + \gamma,$$

where

$$\alpha = \min\{|\overline{\alpha}_1|, \dots, |\overline{\alpha}_m|\},\ \beta = \max\{\beta_1, \dots, \beta_m\},\ \gamma = \max\{\gamma_1, \dots, \gamma_m\},$$

$$\sigma = \max\{\sigma_1, \dots, \sigma_m\},\ \eta(a,T) = \max\{1, aT/2\}.$$

Hence there exists $\rho > 0$ depending only upon a, T, α, β and γ such that

$$|z_0| < \rho$$

which, together with (14) gives

$$|z|_\infty < \max\{1, \rho\} + \frac{aT}{2} := R. \quad (16)$$

Thus Assumption (1) of Lemma 1 holds with $\Omega = B(R) \subset C$. System (9) can be written

$$\overline{f}_k(z_0) := \overline{\alpha}_k z_{0,k}^{n_k} + T^{-1}\int_0^T h_k(t, z_0, 0)\,dt = 0 \quad (k = 1, \dots, m),$$

and any of its possible solution is such that either $|z_0| < 1$ or $|z_0| \geq 1$ and

$$\alpha|z_0|^n \leq \beta|z_0|^\sigma + \gamma. \tag{17}$$

Consequently, $|z_0| < \max\{1, \rho\} < R$ and Assumption (2) of Lemma 1 is satisfied. Finally, introducing the homotopy $\mathcal{F} : \mathbb{C} \times [0,1] \to \mathbb{C}$ defined by

$$\mathcal{F}_k(z_0, \mu) = \overline{\alpha}_k z_{0,k}^{n_k} + \frac{\mu}{T}\int_0^T h_k(t, z_0, 0)\, dt \quad (k = 1, \ldots, m;\ \mu \in [0,1])$$

we see that any possible solution z_0 of $\mathcal{F}(z_0, \mu) = 0$ again is such that (17) holds, so that $|z_0| < R$ and, by the homotopy invariance of Brouwer degree, with

$$p(z) = (z_1^{n_1}, z_2^{n_2}, \ldots, z_m^{n_m})$$

and A is the diagonal matrix

$$A = \operatorname{diag}(\overline{\alpha}_1, \ldots, \overline{\alpha}_m),$$

we obtain

$$\begin{aligned} d_B[\overline{f}, B(R), 0] &= d_B[\mathcal{F}(\cdot, 1), B(R), 0] = d_B[\mathcal{F}(\cdot, 0), B(R), 0] \\ &= d_B[Ap, B(R), 0] = d_B[p, B(R), 0] = n_1\, n_2 \ldots\, n_m, \end{aligned}$$

and Assumption (3) of Lemma 1 holds. □

The special case of Theorem 1 with $m = 1$ states as follows. Consider the complex-valued periodic equation

$$(\phi(z'))' = \alpha(t)z^n + [F(t,z)]' + h(t, z, z'), \quad z(0) = z(T),\ z'(0) = z'(T), \tag{18}$$

where $\phi : B(a) \subset \mathbb{C} \to \mathbb{C}$ satisfies Assumption (H_ϕ), $n \geq 1$ is an integer, $\alpha \in L^1$, $F : [0,T] \times \mathbb{C} \to \mathbb{C}$ is of class C^1 and $h : [0,T] \times \mathbb{C}^2 \to \mathbb{C}$ is a Carathéodory function.

COROLLARY 1. *Assume that $\overline{\alpha} \neq 0$, and that there exist $1 \leq \sigma < n$ and $\beta, \gamma \in L^1$ such that*

$$|h(t, z, v)| \leq \beta(t)|z|^\sigma + \gamma(t)$$

for a.e. $t \in [0,T]$, all $z \in \mathbb{C}$ and all $v \in \mathbb{C}$ such that $|v| < a$. Then problem (18) has at least one solution.

REMARK 1. Such a result does not hold in classical case. The problem

$$z'' = -z + \sin t, \quad z(0) = z(2\pi),\ z'(0) = z'(2\pi),$$

has no solution, as shown by multiplying each member by $\sin t$ and integrating the result over $[0, 2\pi]$.

REMARK 2. Such a result does not hold in the real case. The problem

$$(\phi(u'))' = u^2 + 1, \quad u(0) = u(T),\ u'(0) = u'(T)$$

has no solution, as shown by integrating each member of the differential equation over $[0, 2\pi]$ and using the boundary conditions.

REMARK 3. The periodic problem (18) is of course equivalent to a periodic problem for a system of two *real-valued* differential equation. Getting the requested a priori bounds for the solutions from the real form is less apparent, showing the help of the complex structure in their obtention.

It follows from Corollary 1 that, for any integer $n \geq 1$, any C^1 function $F : \mathbb{C} \to \mathbb{C}$ and any $h \in L^1$ the periodic problem for the Liénard-type equation

$$(\phi(z'))' = \alpha(t)z^n + [F(z)]' + h(t), \quad z(0) = z(T),\ z'(0) = z'(T),$$

has a solution when $\overline{\alpha} \neq 0$. This is in particular the case for the complex-valued relativistic van der Pol equation

$$\left(\frac{z'}{1 - |z'|^2}\right)' + (\beta + \gamma z^2)z' + \alpha z = h(t), \quad z(0) = z(T),\ z'(0) = z'(T) \qquad (19)$$

when $\alpha \neq 0$, $\beta, \gamma \in \mathbb{R}$ and $h \in L^1$. When $\alpha = 0$, problem (19) has no solution when $\overline{h} \neq 0$.

Another consequence of Corollary 1 is that the problem

$$(\phi(z'))' = \alpha_n(t)z^n + \sum_{k=0}^{n-1} \alpha_k(t, z')z^k, \quad z(0) = z(T),\ z'(0) = z'(T),$$

where $n \geq 1$, $\alpha_n \in L^1$ and the $\alpha_k : [0, T] \times \mathbb{C} \to \mathbb{C}$ are Carathéodory functions $(k = 1, \ldots, n-1)$, has at least one solution if $\overline{\alpha}_n \neq 0$.

In particular, for any integer $n \geq 1$ and any $h \in L^1$, the periodic problem

$$(\phi(z'))' = \alpha(t)z^n + h(t), \quad z(0) = z(T),\ z'(0) = z'(T)$$

has a solution for any $\alpha \in L^1$ such that $\overline{\alpha} \neq 0$, and the periodic problem for the complex-valued relativistic Rayleigh equation

$$\left(\frac{z'}{1 - |z'|^2}\right)' + \beta z' + \gamma z'^3 + \alpha z = h(t), \quad z(0) = z(T),\ z'(0 = z'(T),$$

has a solution when $\alpha \neq 0$, $\beta, \gamma \in \mathbb{R}$ and $h \in L^1$.

4. The case of first order equations

Let us consider the periodic problem for first order quasilinear systems of the form

$$(\phi(u))' = f(t,u), \quad u(0) = u(T) \tag{20}$$

where $\phi : B(a) \subset \mathbb{R}^n \to \mathbb{R}^n$ satisfies Assumption (H_ϕ) and $f : [0,T] \times \mathbb{R}^n \to \mathbb{R}^n$ is a Carathéodory function. By *solution* of (20) we mean a continuous function $u : [0,T] \to B(a)$ such that $\phi \circ u \in W^{1,1}$ and equation (20) holds almost everywhere. We keep the notations of the previous sections, and define the mapping $N_f : C \to W^{1,1}$ by

$$N_f(u)(t) := \int_0^t f(s,u(s))\,ds \quad (t \in [0,T]).$$

The following result is the analog of Lemma 1 for problem (20).

LEMMA 2. *Assume that the following conditions hold.*

(i) There is no solution $u_0 \in \partial B(a) \subset \mathbb{R}^n$ of equation

$$\overline{f}(u_0) := T^{-1} \int_0^T f(t,u_0)\,dt = 0.$$

(ii) $d_B[\overline{f}, B(a) \cap \mathbb{R}^n, 0] \neq 0$.

Then problem (20) has at least one solution in $B(a)$.

Proof. Let us consider the family of problems

$$(\phi(u))' = \lambda f(t,u), \quad u(0) = u(T) \quad (\lambda \in [0,1]). \tag{21}$$

We first show that, for $\lambda \in (0,1]$, problem (21) is equivalent to the fixed point problem in C

$$u(t) = \phi^{-1} \circ [\phi(u(0)) - N_f(u)(T) + \lambda N_f(u)(t)] \quad (t \in [0,T]). \tag{22}$$

Indeed, if u is a solution of (21), then by integrating the differential equation from 0 to t, and from 0 to T and using boundary conditions, we get

$$\phi(u(t)) - \phi(u(0)) - \lambda N_f(u)(t) = 0, \quad N_f(u)(T) = 0,$$

hence, both equations taking values in supplementary subspaces,

$$\phi(u(t)) = \phi(u(0)) - N_f(u)(T) + \lambda N_f(u)(t),$$

which is equivalent to (22). Conversely, if u satisfies (22), then $u \in B(a)$ (as $\phi^{-1} : \mathbb{R}^n \to B(a)$), and

$$\phi(u(t)) = \phi(u(0)) - N_f(u)(T) + \lambda N_f(u)(t) \quad (t \in [0,T]). \tag{23}$$

Differentiating, we get the differential equation in (21), taking $t = 0$ we obtain

$$N_f(u)(T) = 0, \tag{24}$$

and taking $t = T$ and using (24) we get

$$\phi(u(T)) = \phi(u(0)),$$

which is equivalent to the boundary condition in (21).

For $\lambda = 0$, equation (22) reduces to

$$u(t) = \phi^{-1} \circ [\phi(u(0)) - N_f(u)(T)] \quad (t \in [0,T])$$

which means that any solution $u = u(0)$ is constant with $u(0) \in B(a) \subset \mathbb{R}^n$ and $u(0)$ solution of (24). Conversely, the solutions of (24) in $B(a)$ are the solutions of (22) with $\lambda = 0$.

Now, the operator $\mathcal{M} : C \times [0,1] \to B(a) \subset C$ defined by

$$\mathcal{M}(u)(t) := \phi^{-1} \circ [\phi(u(0)) - N_f(u)(T) + \lambda N_f(u)(t)] \quad (t \in [0,T])$$

is easily seen to be completely continuous on C, using Arzela-Ascoli's theorem. Hence, if Assumption (i) holds, we have

$$u \neq \mathcal{M}(u,\lambda) \quad \forall\, (u,\lambda) \in \partial B(a) \times [0,1],$$

and the homotopy invariance and reduction property of Leray-Schauder degree d_{LS}, together with Brouwer degree results for homeomorphisms (see e.g. [6]), imply, with $P : C \to C \cap \mathbb{R}^n, u \mapsto u(0)$, that

$$\begin{aligned}
d_{LS}[I - \mathcal{M}(\cdot,1), B(a), 0] &= d_{LS}[I - \mathcal{M}(\cdot,0), B(a), 0] \\
&= d_{LS}[I - \phi^{-1} \circ \{\phi \circ P - N_f(\cdot)(T)\}, B(a), 0] \\
&= d_B[(I - \phi^{-1} \circ \{\phi - N_f(\cdot)(T)\})|_{\mathbb{R}^n}, B(a) \cap \mathbb{R}^n, 0] \\
&= \pm d_B[\phi \circ \{I - \phi^{-1} \circ [\phi - N_f(\cdot)(T)], B(a), 0] \\
&= \pm d_B[N_f(\cdot)(T), B(a), 0] = \pm d_B[\overline{f}, B(a), 0] \neq 0,
\end{aligned}$$

using Assumption (ii). The result follows from the existence property of Leray-Schauder's degree. □

Let us apply Lemma 2 to the periodic problem for the complex-valued differential equation

$$(\phi(z))' = \alpha(t) z^n + h(t,z), \quad z(0) = z(T) \tag{25}$$

where $\phi : B(a) \subset \mathbb{C} \to \mathbb{C}$ satisfies condition (H_ϕ), $\alpha \in L^1$, $n \geq 1$ is an integer, and $h : [0,T] \times \mathbb{C} \to \mathbb{C}$ is a Carathéodory function.

THEOREM 2. *Assume that* $\overline{\alpha} \neq 0$ *and that there exists* $0 \leq \sigma < n$ *and* $\beta \geq 0$, $\gamma \geq 0$ *such that*

(a) $\left| T^{-1} \int_0^T h(t,z)\,dt \right| \leq \beta |z|^\sigma + \gamma$ *for all* $z \in B(a) \subset \mathbb{C}$.

(b) the unique positive root u_0 *of equation*

$$|\overline{\alpha}| u^n = \beta u^\sigma + \gamma$$

is such that $u_0 < a$.

Then problem (25) has at least one solution z.

Proof. With the notations of Lemma 2, we have

$$\overline{f}(z_0) = \overline{\alpha} z_0^n + T^{-1} \int_0^T h(t, z_0)\,dt,$$

so that any possible zero z_0 of $\overline{f}$ is such that

$$|\overline{\alpha}||z_0|^n \leq \beta |z_0|^\sigma + \gamma, \tag{26}$$

and hence, by Assumption (b), $|z_0| < a$. Now, let us consider the homotopy

$$\mathcal{F} : \mathbb{C} \times [0,1] \to \mathbb{C},\ (z_0, \mu) \mapsto \overline{\alpha} z_0^n + \mu T^{-1} \int_0^T h(t, z_0)\,dt \quad (\mu \in [0,1]).$$

If $\mathcal{F}(z_0, \mu) = 0$, then z_0 satisfies inequality (26) and hence $|z_0| < a$. By the homotopy invariance of Brouwer degree, we get , with $p(z) := z^n$,

$$\begin{aligned} d_B[\overline{f}, B(a), 0] &= d_B[\mathcal{F}(\cdot, 1), B(a), 0] = d_B[\mathcal{F}(\cdot, 0), B(a), 0] \\ &= d_B[\overline{\alpha} p, B(a), 0] = d_B[p, B(a), 0] = n. \end{aligned}$$

The result follows from Lemma 2. □

COROLLARY 2. *Let* $\phi : B(a) \to \mathbb{C}$ *satisfy condition* (H_ϕ), $n \geq 1$ *be an integer and* $\alpha \in L^1$. *Then the periodic problem*

$$(\phi(z))' = \alpha(t) z^n + h(t), \quad z(0) = z(T) \tag{27}$$

has at least one solution when $\overline{\alpha} \neq 0$ *and* $|\overline{h}| < |\overline{\alpha}| a^n$.

In particular, the problem

$$\left(\frac{z}{\sqrt{1 - |z|^2}} \right)' = \alpha z^n + h(t), \quad z(0) = z(T) \tag{28}$$

has at least one solution when $\alpha \in \mathbb{C} \setminus \{0\}$ and $|\overline{h}| < |\alpha|$. This result is sharp because if (28) has a solution z, then letting

$$y = \frac{z}{\sqrt{1-|z|^2}} \quad \text{so that} \quad z = \frac{y}{\sqrt{1+|y|^2}}$$

we have

$$y' = \alpha \left(\frac{y}{\sqrt{1+|y|^2}} \right)^n + h(t), \quad y(0) = y(T).$$

Hence, taking the mean value of the differential equation and using the boundary conditions,

$$0 = \alpha T^{-1} \int_0^T \left(\frac{y(t)}{\sqrt{1+|y(t)|^2}} \right)^n dt + \overline{h},$$

which gives

$$|\overline{h}| \leq |\alpha| T^{-1} \int_0^T \left(\frac{|y(t)|}{\sqrt{1+|y(t)|^2}} \right)^n dt < |\alpha|.$$

Remark 4. A result like Corollary 2 does not hold in the classical case

$$z' = \alpha(t) z^n + h(t), \quad z(0) = z(T),$$

as shown by

$$z' = iz + e^{it}, \quad z(0) = z(2\pi)$$

which has no solution, because if it were the case, we would have

$$(e^{-it} z)' = e^{-it} z' - i e^{-it} z = 1, \quad z(0) = z(2\pi)$$

leading to a contradiction by integration over $[0, 2\pi]$.

Remark 5. By analogy with the results of Section 3, the reader will easily state and proof the extension of Theorem 2 to complex-valued systems of the form

$$(\phi_k(z))' = \alpha_k(t) z_k^{n_k} + h_k(t, z), \quad z(0) = z(T) \quad (k = 1, \ldots, m).$$

References

[1] C. Bereanu and J. Mawhin, *Existence and multiplicity results for some nonlinear problems with singular ϕ-laplacian*, J. Differential Equations **243** (2007), 536–557.

[2] C. Bereanu and J. Mawhin, *Boundary value problems for some nonlinear systems with singular ϕ-laplacian*, J. Fixed Point Theory Appl. **4** (2008), 57–75.

[3] C. Bereanu and J. Mawhin, *Periodic solutions of nonlinear perturbations of ϕ-laplacian with possibly bounded ϕ*, Nonlinear Anal. **68** (2008), 1668–1681.
[4] J. Campos, *Möbius transformation and periodic solutions of complex riccati equations*, Bull. London Math. Soc. **9** (1997), 205–213.
[5] J. Campos and R. Ortega, *Nonexistence of periodic solutions of a complex riccati equation*, Differential Integral Equations **9** (1996), 247–250.
[6] K. Deimling, *Nonlinear functional analysis*, Springer, Berlin, 1985.
[7] R. Manásevich, Mawhin J., and F. Zanolin, *Periodic solutions of complex-valued differential equations and systems with periodic coefficients*, J. Differential Equations **126** (1996), 355–373.
[8] R. Manásevich, J. Mawhin, and Zanolin F., *Periodic solutions of some complex-valued liénard and rayleigh equations*, Nonlinear Anal. **36** (1999), 997–1014.
[9] J. Mawhin, *Resonance problems for some non-autonomous ordinary differential equations*, Non-autonomous differential equations, Cetraro 2011 (Berlin), CIME Lecture Notes in Math., vol. 2065, Springer, 2012, pp. 103–184.
[10] R. Srzednicki, *On periodic solutions of planar polynomial differential equations with periodic coefficients*, J. Differential Equations **114** (1994), 77–100.
[11] R. Srzednicki, *Periodic and bounded solutions in blocks for time-periodic non-autonomous ordinary differential equations*, Nonlinear Anal. **22** (1994), 707–737.

Author's address:

Jean Mawhin
Institut de recherche en mathématique et physique
Université Catholique de Louvain
B-1348 Louvain-la-Neuve, Belgium
E-mail: jean.mawhin@uclouvain.be

Received April 16, 2012
Revised May 7, 2012

Rend. Istit. Mat. Univ. Trieste
Volume 44 (2012), 89–116

Remarks concerning the Lyapunov exponents of linear cocycles

Russell Johnson and Luca Zampogni

Dedicated to Professor Fabio Zanolin on the occasion of his 60th birthday

Abstract. *We impose a condition of pointwise convergence on the Lyapunov exponents of a d-dimensional cocycle over a compact metric minimal flow. This condition turns out to have significant consequences for the dynamics of the cocycle. We make use of such classical ODE techniques as the Lyapunov-Perron triangularization method, and the ergodic-theoretical techniques of Krylov and Bogoliubov.*

Keywords: Lyapunov exponent, Sacker-Sell spectrum, discrete spectrum.
MS Classification 2010: 37B55, 34D08, 34D09

1. Introduction

The question of the continuity properties of the Lyapunov exponents of a linear differential system under perturbation of the coefficient matrix is of intrinsic interest and is of importance in various applications. Many important results concerning this theme are due to the "Moscow school" centered around the Nemytskii seminar; we mention some representative papers ([3, 4, 26]) and refer especially to the book [5] by Bylov-Vinograd-Grobman-Nemytskii. In the works of the Moscow school, attention is not restricted to the Lyapunov exponents; other quantities such the upper and lower characteristic indexes and the Bohl exponent are also studied in a systematic way, both from the point of view of continuity and from that of intrinsic properties.

More recent work of Bochi-Viana [2] and of Bessa [1] permits one to make statements concerning the discontinuity of the Lyapunov exponents of certain topological/ergodic families of linear systems. The paper [1] adapts to the continuous setting certain important results of [2] for discrete cocycles. The basic object of study in [1, 2] is the set of Lyapunov exponents determined by the Oseledets theorem relative to a discrete or continuous cocycle and an ergodic measure defined on a compact metric flow. Generally speaking, it is shown that, if the cocycle does not admit a dominated splitting (a.k.a. an exponential separation), and if the Lyapunov exponents are not all equal, then

those exponents do not vary continuously under C^0-perturbation of the cocycle. See also ([28, 30]) for results in this vein.

In a somewhat different vein, Furman [14] studied the case of a discrete cocycle over a strictly ergodic flow. He considered the time averages which define the maximal Lyapunov exponent of the cocycle; that exponent is well-defined by the subadditive ergodic theorem. He shows that, if the cocycle has dimension $d = 2$, and if the time averages converge uniformly with respect to the phase point of the flow, then the maximal Lyapunov exponent varies continuously if the cocycle is perturbed. If in addition the flow is equicontinuous, then the converse statement holds as well.

In the present paper, our point of departure is similar to that of [14], though we work with the usual Lyapunov exponents and not with the maximal exponent. We assume that, for each phase point in the flow, each Lyapunov exponent is defined by a true limit (and not by a non-convergent limit superior). Let $d \geq 2$ be the dimension of the cocycle. We show that, if the flow is minimal, and if the Oseledets spectrum of the cocycle is simple (i.e., consists of d distinct numbers), then the cocycle has the discrete spectrum property of Sacker and Sell. If $d = 2$, we do not need to assume that the Oseledets spectrum is simple (but need slightly more information concerning the limits defining the Lyapunov exponents). We are able to strengthen the continuity result of [14] in the sense that the compact metric flow is minimal but need not be strictly ergodic.

We wish to emphasize that our results will be proved by using quite classical techniques in the theory of linear differential and discrete systems. These include the method of Krylov and Bogoliubov for constructing invariant measures, and the Lyapunov-Perron triangularization procedure. We will also adapt a small part of that proof of the Oseledets theorem which is based on those methods. Beyond that, we will apply some specific results, including an ergodic oscillation result of [16], and two statements of [10] which concern smoothing of real cocycles and the untwisting of invariant vector bundles.

The paper is organized as follows. In Section 2 we prepare the ground by recalling the statement of the Oseledets theorem, and some elements of the spectral theory of Sacker and Sell for linear cocycles. In Section 3 we work out some consequences, regarding the continuity of Lyapunov exponents, of the hypothesis that a cocycle Φ have discrete spectrum. These results are (mostly) known, but perhaps not *well*-known.We also discuss a specific situation in which the results of [1, 2] imply the discontinuity of the Lyapunov exponents under a C^0-perturbation of the cycle.

In Section 4 we present our main result. We show that, if Φ is a cocycle over a compact minimal flow of dimension $d = 2$, and if the time averages which define its Lyapunov exponents all converge, then Φ has discrete spectrum. If the dimension d of Φ is greater than two, we encounter technical problems

when attempting to prove the above result. We are, however, able to prove a theorem which has the following corollary. Suppose that $(\Omega, \{\tau_t\})$ is strictly ergodic with unique ergodic measure μ. Suppose that the cocycle Φ has simple Oseledets spectrum with respect to μ. Finally, suppose that the time averages which define the Lyapunov exponents of Φ all converge. Then Φ has discrete spectrum, and in fact the Sacker-Sell spectrum of Φ is simple. In classical language, this means that Φ has the Lillo property [23].

We finish this Introduction by listing some notational conventions which will be in force throughout the paper. First, the brackets $\langle\ ,\ \rangle$ will indicate the Euclidean inner product on $\mathbb{R}^d$. Second, the symbol $|\cdot|$ will denote a norm whose significance will be clear from the context if it is not explicitly defined. Third, we let $GL(\mathbb{R}^d)$ denote the set of invertible $d \times d$ matrices. Fourth, we let $\mathbb{L}(\mathbb{R}^d)$ denote the set of all $d \times d$ real matrices with the operator norm: if $A \in \mathbb{L}(\mathbb{R}^d)$, then $|A| = \sup\{|Ax| \mid x \in \mathbb{R}^d,\ |x| = 1\}$.

2. Preliminaries

In this section, we introduce basic concepts and results, and express in a precise way the issue to be discussed in this paper.

Let Ω be a compact metric space, and let T be either the reals ($T = \mathbb{R}$) or the integers ($T = \mathbb{Z}$). For each $t \in T$, let $\tau_t : \Omega \to \Omega$ be a continuous map. We say that the family $\{\tau_t \mid t \in T\}$ defines a *topological flow* on Ω if the following conditions are satisfied:

(i) $\tau_0(\omega) = \omega$ for all $\omega \in \Omega$;

(ii) $\tau_t \circ \tau_s = \tau_{t+s}$ for all $t, s \in T$;

(iii) the map $\tau : \Omega \times T \to \Omega : (t, \omega) \mapsto \tau_t(\omega)$ is continuous.

It is clear that, if these conditions are satisfied, then for each $t \in T$, the map $\tau_t : \Omega \to \Omega$ is a homeomorphism and $(\tau_t)^{-1} = \tau_{-t}$ $(t \in T)$. If $T = \mathbb{Z}$, then the topological flow $\{\tau_t \mid t \in \mathbb{Z}\}$ is generated by τ_1, in the sense that $\tau_n = (\tau_1)^n$ if $n > 0$ and $\tau_n = (\tau_{-1})^{-n}$ if $n < 0$. We will refer to a pair $(\Omega, \{\tau_t \mid t \in T\})$ consisting of a compact metric space Ω and a flow $\{\tau_t \mid t \in T\}$ on Ω as a compact metric flow.

Important examples of flows are obtained via the following construction. Let $\mathbb{T}^g = \mathbb{R}^g/\mathbb{Z}^g$ be the g-dimensional torus, and let $\gamma_1, \ldots, \gamma_g$ be rationally independent numbers. Let $\theta_1, \ldots, \theta_g$ be 1-periodic coordinates on $\mathbb{T}^g$. If $T = \mathbb{R}$ or $\mathbb{Z}$, set $\tau_t(\theta_1, \ldots, \theta_g) = (\theta_1 + t\gamma_1, \ldots, \theta_g + t\gamma_g)$ $(t \in T)$. Then $\{\tau_t \mid t \in T\}$ is a flow on $\mathbb{T}^g$, called a *Kronecker flow*.

A compact metric flow $(\Omega, \{\tau_t\})$ is called *minimal* or *Birkhoff recurrent* if Ω is nonempty and for each $\omega \in \Omega$, the *orbit* $\{\tau_t(\omega) \mid t \in T\}$ is dense in Ω. A Kronecker flow as defined above on $\Omega = \mathbb{T}^g$ is minimal. Actually a Kronecker

flow satisfies a stronger property, namely that of *Bohr almost periodicity*: thus, in addition to minimality, there is a metric d on Ω, which is compatible with its topology, such that $d(\tau_t(\omega_1), \tau_t(\omega_2)) = d(\omega_1, \omega_2)$ for all points $\omega_1, \omega_2 \in \Omega$ and all $t \in T$. Clearly the Euclidean metric d on $\Omega = \mathbb{T}^g$ satisfies this last condition.

Let $(\Omega, \{\tau_t\})$ be a compact metric flow, and let μ be a regular Borel probability measure on Ω (thus in particular $\mu(\Omega) = 1$). The measure μ is called $\{\tau_t\}$-*invariant* if $\mu(\tau_t(B)) = \mu(B)$ for each Borel set $B \subset \Omega$ and $t \in T$. An invariant measure μ is called *ergodic* if it satisfies the following indecomposibility condition: whenever $B \subset \Omega$ is a Borel set and $\mu(\tau_t(B)\Delta B) = 0$ for all $t \in T$, there holds $\mu(B) = 0$ or $\mu(B) = 1$ (Δ = symmetric difference of sets).

A classical construction of Krylov and Bogoliubov ([20, 29]) shows that a compact metric flow $(\Omega, \{\tau_t\})$ always admits an ergodic measure μ. If $(\Omega, \{\tau_t\})$ is minimal and admits exactly one ergodic measure, then it is called *strictly ergodic*. A Kronecker flow $\{\tau_t\}$ on $\Omega = \mathbb{T}^g$ is strictly ergodic: the unique ergodic measure is the normalized Haar measure on $\mathbb{T}^g$.

Next we discuss cocycles. A T-*cocycle* over a compact metric flow $(\Omega, \{\tau_t\})$ with values in the general linear group $GL(\mathbb{R}^d)$ is a continuous map $\Phi : \Omega \times T \to GL(\mathbb{R}^d)$ such that:

(i) $\Phi(\omega, 0) = I$= identity for all $\omega \in \Omega$;

(ii) $\Phi(\omega, t + s) = \Phi(\tau_t(\omega), s)\Phi(\omega, t)$ for all $\omega \in \Omega$ and $t, s \in T$.

One obtains an important class of *real* cocycles ($T = \mathbb{R}$) from appropriate families of linear nonautonomous differential systems. Let $(\Omega, \{\tau_t\})$ be a compact metric real flow, and let $A : \Omega \to \mathbb{L}(\mathbb{R}^d)$ be a continuous function. Let $\Phi(\omega, t)$ be the fundamental matrix solution of the ODE

$$\frac{dx}{dt} = A(\tau_t(\omega))x \qquad (x \in \mathbb{R}^d); \tag{1_ω}$$

thus $\Phi(\omega, 0) = I$ and $\dfrac{d}{dt}\Phi(\omega, t) = A(\tau_t(\omega))\Phi(\omega, t)$ for all $\omega \in \Omega$ and $t \in T = \mathbb{R}$. It can be checked that Φ is a real cocycle.

Actually the general real cocycle can be obtained in this way, up to "cohomology". We explain this. Let $(\Omega, \{\tau_t\})$ be a compact metric real flow, and let $\Psi : \Omega \times \mathbb{R} \to GL(\mathbb{R}^d)$ be a real cocycle. Then there exist continuous functions $A : \Omega \to \mathbb{L}(\mathbb{R}^d)$ and $F : \Omega \to GL(\mathbb{R}^d)$ such that, if $\Phi(\omega, t)$ is the cocycle generated by the family (1_ω) corresponding to $A(\cdot)$, then

$$\Psi(\omega, t) = F(\tau_t(\omega))\Phi(\omega, t)F(\omega)^{-1} \qquad (\omega \in \Omega,\ t \in \mathbb{R}).$$

See [10] for a proof; in fact one defines $F(\omega) = \dfrac{1}{\varepsilon}\displaystyle\int_0^\varepsilon \Phi(\omega, s)ds$ for sufficiently small ε. The function F is called a cohomology between Ψ and Φ. It turns out

that the properties of a real cocycle which are of interest to us are preserved under a cohomology. So we will always be able to assume that the real cocycles which we study are derived from a family (1_ω) of linear differential systems in the manner described above.

An *integer* cocycle ($T = \mathbb{Z}$) is obtained from a nonautonomous difference equation, as follows. Set $A(\omega) = \Phi(\omega, 1)$, $\tau(\omega) = \tau_1(\omega)$, and consider

$$x_{n+1} = A(\tau^n(\omega))x_n \qquad (n \in \mathbb{Z},\ x \in \mathbb{R}^d). \tag{2_ω}$$

Then the family of difference equations (2_ω) generates the cocycle Φ in the sense that

$$\Phi(\omega, n) = A(\tau^{n-1}(\omega)) \dots A(\omega) \qquad n > 0,$$
$$\Phi(\omega, 0) = I,$$
$$\Phi(\omega, n) = A^{-1}(\tau^{n-1}(\omega)) \dots A^{-1}(\tau^{-1}(\omega)) \qquad n < 0$$

for all $\omega \in \Omega$. Note that an integer cocycle $\Phi(\omega, n)$ is determined once $\Phi(\omega, 1)$ is known.

Next let $T = \mathbb{R}$ or $\mathbb{Z}$, and let $(\Omega, \{\tau_t\})$ be a compact metric flow. Let $\Phi : \Omega \times T \to GL(\mathbb{R}^d)$ be a cocycle. We recall the definition and some basic properties of the Lyapunov exponents of Φ. Fix $\omega \in \Omega$. For each $0 \neq x \in \mathbb{R}^d$, let

$$\beta(\omega, x) = \limsup_{t \to \infty} \frac{1}{t} \ln |\Phi(\omega, t)x|.$$

The number $\beta(\omega, x)$ is called a Lyapunov exponent of Φ at ω. It is well-known that, as x varies over $\mathbb{R}^d \setminus \{0\}$, $\beta(\omega, x)$ takes on only finitely many values, say $\beta_1(\omega) \leq \beta_2(\omega) \leq \dots \leq \beta_s(\omega)$ where $1 \leq s \leq d$. Moreover, for each $1 \leq r \leq s$, one has that $W_r(\omega) = \{0\} \cup \{0 \neq x \in \mathbb{R}^d \mid \beta(\omega, x) \leq \beta_r(\omega)\}$ is a vector subspace of $\mathbb{R}^d$. One says that $\{0\} = W_0(\omega) \subset W_1(\omega) \subset \dots \subset W_s(\omega) = \mathbb{R}^d$ is the filtration associated to Φ at ω. Set $d_1 = \dim W_1(\omega), \dots, d_r = \dim W_r(\omega) - \dim W_{r-1}(\omega)$ $(2 \leq r \leq s)$; the integer d_r is called the multiplicity of $\beta_r(\omega)$ $(1 \leq r \leq s)$.

Continuing the discussion, we now define the upper Lyapunov exponent of Φ at ω to be

$$\beta_*(\omega) = \limsup_{t \to \infty} \frac{1}{t} \ln |\Phi(\omega, t)|.$$

It is clear that $\beta_s(\omega) \leq \beta_*(\omega)$ for each $\omega \in \Omega$. According to the regularity theory of Lyapunov [24], one has the following. Let d_r be the multiplicity of $\beta_r(\omega)$ $(1 \leq r \leq s)$, and suppose that $d_1\beta_1(\omega) + \dots + d_s\beta_s(\omega) = \liminf_{t \to \infty} \frac{1}{t} \ln \det \Phi(\omega, t)$. Then $\beta_s(\omega) = \beta_*(\omega)$, and the limit $\lim_{t \to \infty} \frac{1}{t} \ln |\Phi(\omega, t)x|$ exists for each $0 \neq x \in \mathbb{R}^d$. One says that Φ is regular at ω. The regularity concept is important in the study of the stability of $x = 0$ relative to nonlinear perturbations of Φ.

There is a considerable body of Russian literature concerning the theory of the Lyapunov exponents, as well as other exponents related to a T-cocycle, namely the central exponents and the Bohl exponents. We will not discuss these important concepts, but refer the reader to [5].

It is useful to consider the Lyapunov exponents associated with the exterior products of the cocycle Φ. For this, let $\Lambda_1\mathbb{R}^d \cong \mathbb{R}^d$, $\Lambda_2\mathbb{R}^d, \ldots, \Lambda_d\mathbb{R}^d \cong \mathbb{R}$ be the exterior product spaces of $\mathbb{R}^d$. These spaces have natural inner products and norms induced by the Euclidean inner product and Euclidean norm in $\mathbb{R}^d$; (see [13, Chapter 1]). The cocycle Φ induces a cocycle with values in $GL(\mathbb{R}^d)$ for each $1 \leq k \leq d$, via the formula $\Lambda_k\Phi(\omega,t)(x_1 \wedge \cdots \wedge x_k) = \Phi(\omega,t)x_1 \wedge \cdots \wedge \Phi(\omega,t)x_k$. Each of these cocycles admits Lyapunov exponents which are analogues of these introduced above for Φ. In this paper, we will only make use of the upper Lyapunov exponents of these cocycles, which are determined as follows

$$\lambda_k(\omega) = \limsup_{t\to\infty} \frac{1}{t} \ln |\Lambda_k\Phi(\omega,t)| \qquad (\omega \in \Omega,\ 1 \leq k \leq d).$$

Of course, $\lambda_1(\omega) = \beta_*(\omega)$ and $\lambda_k(\omega) = \limsup_{t\to\infty} \frac{1}{t} \ln \det \Phi(\omega,t)$.

Let us state a corollary of a result of Ruelle ([36, Proposition 1.3]).

PROPOSITION 2.1. *Let $T = \mathbb{R}$ or $\mathbb{Z}$, let $(\Omega, \{\tau_t\})$ be a compact metric flow, and let $\Phi : \Omega \times T \to GL(n,\mathbb{R})$ be a T-cocycle. Let $\omega \in \Omega$. Suppose that, for each $k = 1, 2, \ldots, d$, the following limit exists:*

$$\lim_{t\to\infty} \frac{1}{t} \ln |\Lambda_k(\omega,t)| = \lambda_k(\omega).$$

Let $\beta_1(\omega) < \ldots < \beta_s(\omega)$ be the Lyapunov exponents of Φ at ω, and let $\{0\} = W_0(\omega) \subset W_1(\omega) \subset \cdots \subset W_s(\omega) = \mathbb{R}^d$ be the corresponding filtration. Then if $1 \leq r \leq s$ and if $0 \neq x \in W_r(\omega) \setminus W_{r-1}(\omega)$, one has

$$\lim_{t\to\infty} \frac{1}{t} \ln |\Phi(\omega,t)x| = \beta_r(\omega) \qquad (1 \leq r \leq s).$$

Thus the limit $\lim_{t\to\infty} \frac{1}{t} \ln |\Phi(\omega,t)x|$ exists *for each* $0 \neq x \in \mathbb{R}^d$.

We now recall certain results concerning T-cocycles, namely the Oseledets theorem [31] and the spectral theorem of Sacker and Sell [38].

THEOREM 2.2 (**Oseledets**). *Let $T = \mathbb{R}$ or $\mathbb{Z}$, let $(\Omega, \{\tau_t\})$ be a compact metric flow, and let μ be a $\{\tau_t\}$-ergodic measure on Ω. Let $\Phi : \Omega \times T \to GL(\mathbb{R}^d)$ be a T-cocycle over $(\Omega, \{\tau_t\})$. If $\omega \in \Omega$, let $\beta_1(\omega), \ldots, \beta_s(\omega)$ be the Lyapunov exponents of Φ at ω.*

There is a $\{\tau_t\}$-invariant μ-measurable subset $\Omega_1 \subset \Omega$ with $\mu(\Omega_1) = 1$, such that, if $\omega \in \Omega_1$, then $\mathbb{R}^d$ admits a direct sum decomposition

$$\mathbb{R}^d = V_1^{(m)}(\omega) \oplus V_2^{(m)}(\omega) \oplus \cdots \oplus V_s^{(m)}(\omega),$$

such that the following statements are valid. First, if $0 \neq x \in V_r^{(m)}(\omega)$, then

$$\lim_{t \to \pm\infty} \frac{1}{t} \ln |\Phi(\omega, t)x| = \beta_r(\omega);$$

note the two-sidedness of the limit. The dimension of $V_r^{(m)}(\omega)$ equals the multiplicity d_r of $\beta_r(\omega)$. Second, the number s and the multiplicities $d_1, \ldots, d_s$ do not depend on $\omega \in \Omega_1$, and moreover $\beta_r(\omega)$ is constant on Ω_1 $(1 \leq r \leq s)$. Third, the correspondence $\omega \mapsto V_r^{(m)}(\omega)$ is μ-measurable in the Grassmann sense $(1 \leq r \leq s)$. Fourth, the "measurable bundle"

$$V_r^{(m)} = \bigcup_{\omega \in \Omega_1} \Big\{ (\omega, x) \mid x \in V_r^{(m)}(\omega) \Big\}$$

is Φ invariant in the sense that, if $\omega \in \Omega_1$, $t \in T$ and $x \in V_r^{(m)}(\omega)$, then $(\tau_t(\omega), \Phi(\omega, t)x) \in V_r^{(m)}$.

This is not the most general form of the Oseledets theorem but it will be sufficient for our purposes. We note that the "μ-measurability" of $\omega \mapsto V_r^{(m)}(\omega)$ has the following meaning. – For each $\omega \in \Omega_1$, $V_r^{(m)}(\omega)$ defines an element of the Grassmannian manifold $Gr(d, d_r)$ of d_r-dimensional subspaces of $\mathbb{R}^d$; the mapping $\Omega_1 \mapsto Gr(d, d_r) : \omega \mapsto V_r^{(m)}(\omega)$ is μ-measurable. – The numbers $\beta_1 < \ldots < \beta_s$, which do not depend on $\omega \in \Omega_1$, are collectively referred to as the Oseledets spectrum or μ-spectrum of Φ.

The Oseledets theorem is a basic result in the theory of real or discrete cocycles. It has been proved using two distinct approaches. One method of proof uses the triangularization technique of Lyapunov-Perron; see [18, 31]. The other approach makes use of the subadditive ergodic theorem of Kingman [15, 36]. Both methods offer advantages and important information.

Next we review some aspects of the Sacker-Sell spectral theory, which taken together can be thought of as a continuous analogue of the Oseledets theory. First recall that a T-cocycle Φ over a compact metric flow $(\Omega, \{\tau_t\})$ is said to have an *exponential dichotomy* if there are positive constants $k > 0$, $\gamma > 0$ and a continuous, projection-valued function $\omega \mapsto P_\omega = P_\omega^2 : \Omega \to \mathbb{L}(\mathbb{R}^d)$ such that the following estimates hold:

$$|\Phi(\omega, t)P_\omega\Phi(\omega, s)^{-1}| \leq ke^{-\gamma(t-s)} \qquad t \geq s$$

$$|\Phi(\omega, t)(I - P_\omega)\Phi(\omega, s)^{-1}| \leq ke^{\gamma(t-s)} \qquad t \leq s$$

for all $\omega \in \Omega$ and $t, s \in T$.

The following basic theorem was proved by Sacker-Sell [37] and Selgrade [39]. Recall that a compact metric flow $(\Omega, \{\tau_t\})$ is said to be chain recurrent [7] if for each $\omega \in \Omega$, $\varepsilon > 0$ and $T > 0$, there are points $\omega = \omega_0, \omega_1, \ldots, \omega_N = \omega$ and times $t_1 > T, \ldots, t_N > T$ such that $d(\tau_{t_i}(\omega_{i-1}), \omega_i) \leq \varepsilon$ $(1 \leq i \leq N)$. A minimal flow $(\Omega, \{\tau_t\})$ is chain recurrent.

THEOREM 2.3. *Suppose that the compact metric flow $(\Omega, \{\tau_t\})$ is chain recurrent, where $t \in T = \mathbb{R}$ or $\mathbb{Z}$. Let $\Phi : \Omega \times T \to GL(\mathbb{R}^d)$ be a T-cocycle. Suppose that, for each $\omega \in \Omega$, the condition $\sup_{t \in T} |\Phi(\omega, t)x| < \infty$ implies that $x = 0$; i.e., the cocycle Φ admits no nontrivial "bounded orbits". Then Φ admits an exponential dichotomy over Ω.*

Let us define the *dynamical* (or *Sacker-Sell*) spectrum σ_Φ of the T-cocycle Φ over the compact metric flow $(\Omega, \{\tau_t\})$ to be $\{\lambda \in \mathbb{R} \mid$ the translated cocycle $e^{\lambda t}\Phi(\omega, t)$ does not admit an exponential dichotomy over $\Omega\}$. Let us also recall that a compact metric flow $(\Omega, \{\tau_t\})$ is said to be invariantly connected [21] if Ω cannot be expressed as the union of two nonempty disjoint compact invariant subsets. We state the spectral theorem of Sacker-Sell.

THEOREM 2.4 ([38]). *Let $(\Omega, \{\tau_t\})$ be a compact metric invariantly connected flow, where $T = \mathbb{R}$ or $\mathbb{Z}$. Let $\Phi : \Omega \times T \to GL(\mathbb{R}^d)$ be a T-cocycle. Then the dynamical spectrum σ_Φ of Φ is a disjoint union of finitely many compact intervals:*

$$\sigma_\Phi = [a_1, b_1] \cup [a_2, b_2] \cup \cdots \cup [a_q, b_q]$$

where $1 \leq q \leq d$ and $-\infty < a_1 \leq b_1 < a_2 \leq \ldots < a_q \leq b_q < \infty$. To each interval $[a_p, b_p]$ there corresponds a Φ-invariant topological vector subbundle $V_p^{(c)} \subset \Omega \times \mathbb{R}^d$ with the property that

$$(\omega, x) \in V_p^{(c)} \quad \text{and} \quad x \neq 0$$

$$\Updownarrow$$

$$a_p \leq \liminf_{t \to \infty} \frac{1}{t} \ln |\Phi(\omega, t)x| \leq \limsup_{t \to \infty} \frac{1}{t} \ln |\Phi(\omega, t)x| \leq b_p$$

and

$$a_p \leq \liminf_{t \to -\infty} \frac{1}{t} \ln |\Phi(\omega, t)x| \leq \limsup_{t \to -\infty} \frac{1}{t} \ln |\Phi(\omega, t)x| \leq b_p$$

$(1 \leq p \leq q)$.

One has further

$$\Omega \times \mathbb{R}^d = V_1^{(c)} \oplus V_2^{(c)} \oplus \cdots \oplus V_q^{(c)} \qquad \text{(Whitney sum)}.$$

We will emphasize the following concept:

DEFINITION 2.5. *Let $T = \mathbb{R}$ or $\mathbb{Z}$, let $(\Omega, \{\tau_t\})$ be a compact metric flow, and let Φ be a T-cocycle over $(\Omega, \{\tau_t\})$. Suppose that $(\Omega, \{\tau_t\})$ is invariantly connected. Then Φ is said to have* discrete spectrum *if each spectral interval $[a_p, b_p]$ reduces to a point: $a_p = b_p$ for each $1 \leq p \leq q$.*

The discrete spectrum concept is related to but weaker than that of the "Lillo property" [23]. See [19] in this regard.

In Section 3, we will state and prove some results to the effect that, if a T-cocycle Φ has discrete spectrum, then its Lyapunov exponents vary continuously under perturbation of Φ. We claim no particular originality for these results as many statements of this type appear in the literature; e.g., [4, 26]. We do wish to emphasize our use of the Krylov-Bogoliubov method in our proofs, and the fact that one result (Proposition 3.4) appears to be more general than most. We also note that quite recent papers [1, 2, 14] have taken up the theme of the continuity of Lyapunov exponents, so it may not be inappropriate if we do so as well.

In Section 4, we give conditions which are sufficient in order that a cocycle Φ have discrete spectrum. One of our results (Theorem 4.4) generalizes a result of Furman [14] when $d = 2$.

To our knowledge, the connection between the expressibility of $\beta(\omega, x)$ as a limit for all $\omega \in \Omega$, $0 \neq x \in \mathbb{R}^d$, and the discrete spectrum property has not received much attention in the literature. However that may be, the said connection has turned out to be important in the spectral theory of quasicrystals. In this context $d = 2$. For example, in the paper [8] by Damanik-Lenz, the authors use the so-called avalance principle and detailed properties of certain strictly ergodic shift flows to verify that $\lim_{t\to\infty} \frac{1}{t} \ln|\Phi(\omega, t)|$ exists for all $\omega \in \Omega$. One can then use Proposition 2.1 to show that $\beta(\omega, x)$ is expressible as a limit for all $\omega \in \Omega$, $0 \neq x \in \mathbb{R}^2$. In [8], the authors use the Furman result mentioned above to show that Φ has discrete spectrum; that result is subsumed in ours. They go on to show that, for certain quasicrystals, the spectrum of the associated Schrödinger operator has zero Lebesgue measure and is purely singular and continuous.

Perhaps our results will be useful in the study of higher-dimensional spectral problems of Atkinson type. We plan to investigate this issue in future work.

3. Discrete spectrum and Lyapunov exponents

In this section, we derive some continuity results for the Lyapunov exponents of a T-cocycle Φ ($T = \mathbb{R}$ or $\mathbb{Z}$) when Φ has discrete spectrum. As stated above, we make no claims concerning the originality of these results, as there is a

very substantial literature on the subject. On the other hand, we think it is appropriate to present them here since they generalize some theorems in the recent literature. Also our proofs differ from some others in our systematic use of the classical Krylov-Bogoliubov method.

We begin the discussion with a simple consequence of Theorem 2.4.

PROPOSITION 3.1. *Let $T = \mathbb{R}$ or $\mathbb{Z}$, let $(\Omega, \{\tau_t\})$ be a compact metric flow which is invariantly connected, and let $\Phi : \Omega \times T \to GL(\mathbb{R}^d)$ be a T-cocycle. Suppose that the dynamical spectrum σ_Φ of Φ is discrete:*

$$\sigma_\Phi = \{a_0 < a_2 < \ldots < a_q\} \qquad (1 \leq q \leq d).$$

Then for each $\omega \in \Omega$ and $0 \neq x \in \mathbb{R}^d$ the limits $\lim_{t\to\pm\infty} \frac{1}{t} \ln|\Phi(\omega,t)x|$ exist. In fact, if $(\omega, x) \in V_p^{(c)}$ then $\lim_{t\to\pm\infty} \frac{1}{t} \ln|\Phi(\omega,t)x| = a_p$ $(1 \leq p \leq q)$, while if $x \notin V_p^{(c)}(\omega)$ for all $p = 1, 2, \ldots, q$, then $\lim_{t\to\infty} \frac{1}{t} \ln|\Phi(\omega,t)x| = a_m$ and $\lim_{t\to-\infty} \frac{1}{t} \ln|\Phi(\omega,t)x| = a_l$ where $l \leq m$ and $x \in V_l^{(c)}(\omega) \oplus \cdots \oplus V_m^{(c)}(\omega)$. .

Actually, if one restricts attention to the dynamics of Φ on a subbundle $V_p^{(c)}$, then the limits defining the Lyapunov exponents converge uniformly, in a sense which we now make precise. We first consider real cocycles, and carry out a preliminary discussion concerning them.

Let $\mathbb{L}$ be the usual projective space of lines through the origin in $\mathbb{R}^d$, so that $\mathbb{L}$ is a compact $(d-1)$-dimensional manifold. Let $\mathbb{B} = \Omega \times \mathbb{L}$. We assume that the $\mathbb{R}$-cocycle $\Phi = \Phi(\omega, t)$ is defined by the family of linear ordinary differential equations

$$x' = A(\tau_t(\omega))x \qquad \omega \in \Omega,\ x \in \mathbb{R}^d \tag{3_ω}$$

where $A : \Omega \to \mathbb{L}(\mathbb{R}^d)$ is a continuous function. Define a flow $\{\hat{\tau}_t \mid t \in \mathbb{R}\}$ on $\mathbb{B}$ by setting $\hat{\tau}_t(\omega, l) = (\tau_t(\omega), \Phi(\omega, t)l)$ for $\omega \in \Omega$, $l \in \mathbb{L}$. Then define $f : \mathbb{B} \to \mathbb{R} : f(\omega, l) = \langle A(\omega)x, x\rangle / \langle x, x\rangle$ where $0 \neq x \in l$. It is easy to check that, if $x \in \mathbb{R}^d$ has norm 1, and if $l \in \mathbb{L}$ is the line containing x, then

$$\ln|\Phi(\omega,t)x| = \int_0^t f(\hat{\tau}_t(\omega, l))ds. \tag{4}$$

This formula allows one to use ergodic theory (in particular the method of Krylov-Bogoliubov) to study the limiting expressions which define Lyapunov exponents.

PROPOSITION 3.2. *Let $(\Omega, \{\tau_t,\ t \in \mathbb{R}\})$ be a compact metric invariantly connected flow, and let $\Phi : \Omega \times \mathbb{R} \to GL(\mathbb{R}^d)$ be a real cocycle. Let $[a_p, b_p]$ be the p-th interval in the dynamical spectrum σ_Φ of Φ, and let the corresponding*

spectral subbundle be $V_p^{(c)}$ *(*$1 \le p \le q$*). Suppose that* $[a_p, b_p]$ *degenerates to a point for some* $p \in \{1, 2, \dots, q\}$*: thus* $a_p = b_p$*. Then*

$$\lim_{t \to \pm\infty} \frac{1}{t} \ln |\Phi(\omega, t)x| = a_p$$

where the limit is uniform with respect to pairs $(\omega, x) \in V_p^{(c)}$ *with* $|x| = 1$*.*

Proof. It follows from (4) that it is sufficient to prove that $\frac{1}{t}\int_0^t f(\hat{\tau}_s(b))ds$ converges uniformly to a_p with respect to $b = (\omega, l) \in \mathbb{B}_p = \{(\omega, l) \mid l \subset V_p^{(c)}(\omega)\}$. We do this by using arguments of the classical Krylov-Bogoliubov type (see, e.g., [29]).

Suppose for contradiction that, for some $\varepsilon > 0$, there exist a sequence $\{t_n\} \subset \mathbb{R}$ with $|t_n| \to \infty$ and a sequence $\{b_n = (\omega_n, l_n)\} \subset \mathbb{B}_p$ such that

$$\left| \frac{1}{t_n} \int_0^{t_n} f(\hat{\tau}_s(b_n))ds - a_p \right| \ge \varepsilon \qquad (n = 1, 2, \dots).$$

Let $C(\mathbb{B}_p)$ be the space of continuous, real-valued functions on $\mathbb{B}_p$ with the uniform norm. Let $\mathcal{F} \subset C(\mathbb{B}_p)$ be a countable dense set: $\mathcal{F} = \{f_1, f_2, \dots, f_k, \dots\}$ with $f_1 = f$. Using a Cantor diagonal argument, we can determine a subsequence $\{t_m\}$ of $\{t_n\}$ such that

$$\lim_{m \to \infty} \frac{1}{t_m} \int_0^{t_m} f_k(\hat{\tau}_s(b_m))ds$$

exists for $k = 1, 2, \dots$. Call the limit $\nu_*(f_k)$ $(1 \le k < \infty)$. One shows easily that ν_* extends to a bounded nonnegative linear functional on $C(\mathbb{B}_p)$, which we also denote by ν_*. It is clear that $\nu_*(c) = c$ for each constant function c on $\mathbb{B}_p$. This functional is $\{\hat{\tau}_t\}$-invariant in the sense that $\nu_*(g \circ \hat{\tau}_t) = \nu_*(g)$ for each $g \in C(\mathbb{B}_p)$ and each $t \in \mathbb{R}$. Using the Riesz representation theorem, one can find a $\{\hat{\tau}_t\}$-invariant measure ν on $\mathbb{B}_p$ such that

$$\left| \int_{\mathbb{B}_p} f d\nu - a_p \right| \ge \varepsilon.$$

We claim that there exists a $\{\hat{\tau}_t\}$-ergodic measure e on $\mathbb{B}_p$ such that

$$\left| \int_{\mathbb{B}_p} f de - a_p \right| \ge \varepsilon.$$

To see this, use the Krein-Mil'man theorem to represent the weak-$*$ compact convex set I of $\{\hat{\tau}_t\}$-invariant linear functionals on $\mathbb{B}_p$ as the closed convex hull

of its set E of extreme points. It is easy to see that $e_* \in E$ if and only if its associated measure e is ergodic. By the Choquet representation theorem [35]:

$$\int_{\mathbb{B}_p} f d\nu = \int_E \left(\int_{\mathbb{B}_p} f de_* \right) dm(e_*)$$

where m is the representing measure of ν_* on E. It is now clear that e can be found.

Changing notation, let ν be a $\{\hat{\tau}_t\}$-ergodic measure on $\mathbb{B}_p$ such that

$$\left| \int_{\mathbb{B}_p} f d\nu - a_p \right| \geq \varepsilon.$$

By the Birkhoff ergodic theorem there is a set $\mathbb{B}_* \subset \mathbb{B}_p$ of full ν-measure such that, if $b_* \in \mathbb{B}_*$, then

$$\frac{1}{t} \int_0^t f(\hat{\tau}_s(b_*)) ds \to \int_{\mathbb{B}_p} f d\nu \neq a_p$$

as $t \to \infty$. This contradicts Proposition 3.1 and completes the proof of Proposition 3.2. □

REMARK 3.3. *(a) We can prove the $T = \mathbb{Z}$-analogue of Proposition 3.2 in the following way. Set $A(\omega) = \Phi(\omega, 1)$, then define $f_* : \mathbb{B}_p \to \mathbb{R} : f_*(\omega, l) = \frac{1}{2} \ln \langle A(\omega)x, A(\omega)x \rangle$ for each $(\omega, l) \in \mathbb{B}_p$ and $x \in l$, $|x| = 1$. One can check that Proposition 3.2 and its proof remain valid if one considers an integer cocycle Φ and if f is substituted with the above function f_*.*

(b) Let $T = \mathbb{R}$ or $\mathbb{Z}$, let $(\Omega, \{\tau_t\})$ be an invariantly connected compact metric flow, and let $\Phi : \Omega \times T \to GL(\mathbb{R}^d)$ be a T-cocycle for which the hypotheses of Proposition 3.2 are valid. Let $\Phi_*(\omega, t)$ be the restriction of $\Phi(\omega, t)$ to $V_p^{(c)}$, so that for each $\omega \in \Omega$ and $t \in T$ one has the linear transformation $\Phi_*(\omega, t) : V_p^{(c)}(\omega) \to V_p^{(c)}(\tau_t(\omega))$. Define the norm $|\Phi_*(\omega, t)|$ in the usual way. Then

$$\lim_{t \to \pm\infty} \frac{1}{t} \ln |\Phi_*(\omega, t)| = a_p,$$

where the limit is uniform in $\omega \in \Omega$. This statement is a consequence of Proposition 3.2, because for each $\omega \in \Omega$ and $t \in \mathbb{R}$, there exists a unit vector $x \in V_p^{(c)}(\omega)$ such that $|\Phi_*(\omega, t)| = |\Phi_*(\omega, t)x|$.

(c) By combining Propositions 3.1 and 3.2, one obtains a continuity result for the Lyapunov exponents of Φ with respect to variation of $\omega \in \Omega$. In fact, let $\{\beta_1(\omega), \ldots, \beta_s(\omega)\}$ be the Lyapunov exponents of Φ at ω, with

multiplicities $d_1, \ldots, d_s$. If the hypotheses of Propositions 3.1 and 3.2 are valid, then the multiplicities and the exponents $\beta_r(\omega)$ themselves do not depend on $\omega \in \Omega$.

Now we consider another type of continuity result for the Lyapunov exponents of a T-cocycle Φ. We will see that it is possible to vary the matrix function $A(\omega)$ in a non-uniform way, and still retain continuous variation of the exponents. We formulate a result along these lines which illustrate the power of a perturbation theorem due to Sacker and Sell ([38]; see also Palmer [34]).

For this, let Ω be the g-torus $\mathbb{T}^g = \mathbb{R}^g/\mathbb{Z}^g$. Let $\gamma_1, \ldots, \gamma_g$ be rationally independent numbers. Consider the Kronecker flow $\{\tau_t\}$ on $\mathbb{T}^g$ defined by $\gamma = (\gamma_1, \ldots, \gamma_g)$. Thus if $\omega \in \mathbb{R}^g/\mathbb{Z}^g$, then $\tau_t(\omega) = \omega + \gamma t$ $(t \in \mathbb{R})$.

Next, let $A : \mathbb{T}^g \to \mathbb{L}(\mathbb{R}^d)$ be a continuous function. Let $\Phi(\omega, t)$ be the cocycle defined by the family of differential systems (1_ω):

$$x' = A(\tau_t(\omega))x.$$

Suppose that Φ has discrete spectrum; $\sigma_\Phi = \{a_1 < a_2 < \ldots < a_q\}$.

Let $\gamma^{(n)}$ be a sequence in $\mathbb{R}^g$ such that $\gamma^{(n)} \to \gamma$. Each $\gamma^{(n)}$ defines a flow $\{\tau_t^{(n)}\}$ on $\mathbb{T}^g$ via the formula $\tau_t^{(n)}(\omega) = \omega + \gamma^{(n)}t$. However these flows need not be minimal because we do not assume that the components $\gamma_1^{(n)}, \ldots, \gamma_g^{(n)}$ of $\gamma^{(n)}$ are rationally independent. Let $\Phi^{(n)}(\omega, t)$ be the cocycle generated by the family of linear systems

$$x' = A(\tau_t^{(n)}(\omega))x.$$

Note that, if $\gamma^{(n)} \neq \gamma$ for $n = 1, 2, \ldots$, then $A(\tau_t^{(n)}(\omega))$ certainly does not converge uniformly in $t \in \mathbb{R}$ to $A(\tau_t(\omega))$ $(\omega \in \Omega)$. Nevertheless we have the following result.

PROPOSITION 3.4. *For each $\omega \in \Omega$ and $n \geq 1$, let $\{\beta_r^{(n)}(\omega) \mid 1 \leq r \leq s = s(n)\}$ be the Lyapunov exponents of $\Phi^{(n)}$. Also let $\beta_*^{(n)}(\omega)$ be the upper Lyapunov exponent of $\Phi^{(n)}$ at ω $(\omega \in \Omega, n \geq 1)$.*

Given $\varepsilon > 0$, there exists $n_0 \geq 1$ such that, if $n \geq n_0$, then each Lyapunov exponent $\beta_r^{(n)}(\omega)$ is in the ε-neighborhood of σ_Φ $(\omega \in \Omega)$ and $\beta_*^{(n)}(\omega)$ is in the ε-neighborhood of a_q.

We sketch the proof of Proposition 3.4. Let $\mathcal{C} = \{c : \mathbb{R} \to \mathbb{L}(\mathbb{R}^d) \mid c$ is continuous and bounded$\}$ with the the topology of uniform convergence on compact sets. Introduce the Bebutov (translation) flow $\{\hat{\tau}_t\}$ on $\mathcal{C}$: thus $\hat{\tau}_t c(\cdot) = c(\cdot + t)$ for each $t \in \mathbb{R}$ and $c \in \mathcal{C}$.

Next let $U \subset \mathbb{R}^g$ be a compact neighborhood of γ. For each $\hat{\gamma} \in U$ and each $\omega \in \Omega$, set $c(t, \omega, \hat{\gamma}) = A(\omega + \hat{\gamma}t)$ $(t \in \mathbb{R})$. Set $C_{\hat{\gamma}} = \{c(\cdot, \omega, \hat{\gamma}) \mid \omega \in \Omega\} \subset \mathcal{C}$, and

further set $C = \bigcup\{C_{\hat{\gamma}} \mid \hat{\gamma} \in U\} \subset \mathcal{C}$. It can be checked that C is a compact, $\{\hat{\tau}_t\}$-invariant subset of $\mathcal{C}$ which is invariantly connected.

Define a cocycle $\hat{\Phi}$ on C in the following way: $\hat{\Phi}(c,t)$ is the fundamental matrix solution of the linear differential equation $x' = c(t)x$ ($c \in C,\ t \in \mathbb{R},\ x \in \mathbb{R}^d$). Let $C_\gamma = \{t \mapsto A(\omega + \gamma t) \mid \omega \in \Omega\} \subset C$; it can be checked that the dynamical spectrum of the restriction $\hat{\Phi}_\gamma = \hat{\Phi}|_{C_\gamma \times \mathbb{R}}$ equals σ_Φ. Similarly, let $C_{\gamma_n} = \{t \mapsto A(\omega + \gamma_n t) \mid \omega \in \Omega\}$. Then the dynamical spectrum of the restriction $\hat{\Phi}_n = \hat{\phi}|_{C_{\gamma_n} \times \mathbb{R}}$ of $\hat{\Phi}$ to $C_{\gamma_n} \times \mathbb{R}$ equals $\sigma_{\Phi^{(n)}}$.

We are now in a position to apply the perturbation Theorem 6 of [38]. According to this theorem, there is a neighborhood $W \subset C$ of C_γ with the property that, if C_* is a $\{\hat{\tau}_t\}$-invariant subset of W, then the dynamical spectrum of $\hat{\Phi}_{C_*}$ is contained in the ε-neighborhood of $\sigma_{\hat{\Phi}_\gamma} = \sigma_\Phi = \{a_1 < a_2 < \ldots < a_q\}$. Now if n is sufficiently large, then $C_\gamma \subset W$. So the remarks of the preceding paragraph and Proposition 3.1 imply that the thesis of Proposition 3.4 is true.

REMARK 3.5. *Let $T = \mathbb{Z}$, let $A : \Omega = \mathbb{T}^g \to GL(\mathbb{R}^d)$ be a continuous map, let $\gamma \in \mathbb{R}^g$ have rationally independent components, and let $\Phi(\omega, t)$ be the cocycle generated by the family of difference equations*

$$x_{t+1} = A(\omega + \gamma t)x_t \qquad (\omega \in \Omega,\ t \in \mathbb{Z}).$$

Similarly, let $\Phi^{(n)}(\omega, t)$ be the cocycle generated by the family

$$x_{t+1} = A(\omega + \gamma^{(n)} t)x_t \qquad (\omega \in \Omega,\ t \in \mathbb{Z})$$

where $\gamma^{(n)} \in \mathbb{R}^g$ ($n = 1, 2, \ldots$). Then Proposition 3.4 is true as stated for Φ and $\Phi^{(n)}$. The proof is practically identical to that given above for real cocycles (one must introduce a discrete Bebutov flow, and one must note that [38, Theorem 6] holds also for integer cocycles).

We have shown that the discrete spectrum condition has significant consequences for the convergence of the limits which define the Lyapunov exponents, and for the continuity of those Lyapunov exponents. Our results can be viewed as generalizations of [14, Theorem 3].

If the discrete spectrum condition does not hold, then one cannot expect the Lyapunov exponents of Φ to vary continuously when Φ is subjected to a C^0-perturbation. We indicate a concrete result along these lines, the proof of which uses important theorems of Bochi-Viana [2] and Bessa [1]. These papers were motivated by a well-known conjecture of Mañe [25].

Let $T = \mathbb{R}$ or $\mathbb{Z}$. Let $(\Omega, \{\tau_t\})$ be a compact metric flow which is strictly ergodic with unique ergodic measure μ. Thus for example it can be a Kronecker flow as defined in Section 2.

Let $\Phi : \Omega \times T \to GL(\mathbb{R}^d)$ be a cocycle over $(\Omega, \{\tau_t\})$. Suppose that the dynamical spectrum σ_Φ of Φ is a single interval: $\sigma_\Phi = [a, b]$. Suppose that

$a < b$. Let $\{\beta_1 < \ldots < \beta_s\}$ be the Oseledets spectrum of Φ with respect to μ, and let $V_1^{(m)}, \ldots, V_s^{(m)}$ be the corresponding Oseledets bundles.

According to the results of [1] and [2], there is a C^0-residual set $\{\Psi\}$ of $GL(\mathbb{R}^d)$-valued cocycles over $(\Omega, \{\tau_t\})$ for which one of the following alternatives holds.

(i) The Oseledets spectrum of Ψ reduces to a single point;

(ii) The Oseledets bundles give rise to a dominated splitting (or exponential separation) of Ψ over $(\Omega, \{\tau_t\})$.

Moreover, it is shown that, if Ψ does not admit a dominated splitting, then an arbitrarily small C^0-perturbation of Ψ has property (i). See also [28, 30] for related results. We will not define the concept of dominated splitting/exponential separation here. For this we refer to [1, 2] or to the older literature on exponential separation (e.g., [3, 4, 5, 32, 33]).

Now, one can use a Krylov-Bogoliubov argument to show that, if the Oseledets bundles of Ψ give rise to a dominated splitting, then the dynamical spectrum σ_Ψ of Ψ consists of at least two disjoint intervals. We omit the proof, but note that it uses the hypothesis that $(\Omega, \{\tau_t\})$ admits just one ergodic measure.

Returning to the cocycle Φ, one can use another Krylov-Bogoliubov argument to show that the endpoints a and b of $\sigma_\Phi = [a, b]$ are in the Oseledets spectrum; see [18]. But an arbitrarily small C^0-perturbation of Φ has the property that its Oseledets spectrum reduces to a single point. This implies that the Lyapunov exponents of Φ cannot vary continuously if Φ is varied in the C^0-sense.

4. Consequences of convergence

In this section, we consider a problem which is inverse to that taken up in Section 3. Namely, suppose that Φ is a cocycle over a compact metric flow $(\Omega, \{\tau_t\})$, and suppose that $\lim_{t\to\infty} \frac{1}{t} \ln|\Phi(\omega, t)|$ exists for all $\omega \in \Omega$ and all $0 \neq x \in \mathbb{R}^d$. We ask if the cocycle Φ has discrete spectrum. In general this is not true, as the following example shows.

EXAMPLE 4.1. *Let Ω be the annulus $0 < \alpha \leq r \leq \beta$, $0 \leq \theta \leq 2\pi$ in the plane $\mathbb{R}^2$ with polar coordinates (r, θ). Let $a : \Omega \to \mathbb{R}$ be a continuous function such that the correspondence $r \mapsto \int_0^{2\pi} a(r, \bar{\theta}) d\bar{\theta}$ takes on more than one value. Consider the family of one-dimensional ODEs*

$$x' = a(r, \theta + t)x \qquad x \in \mathbb{R} \tag{5_ω}$$

where $\omega = (r, \theta) \in \Omega$. *The family* (5_ω) *has the form of the family* (1_ω) *if we put* $\tau_t(r, \theta) = (r, \theta + t)$ *for* $t \in \mathbb{R}$ *and* $(r, \theta) \in \Omega$. *It it clear that the cocycle* Φ *which is determined by equations* (5_ω) *has the form*

$$\Phi(\omega, t) = \exp\left(\int_0^t a(r, \theta + s)ds\right) \qquad (\omega = (r, \theta) \in \Omega,\ t \in \mathbb{R}).$$

We see that, if $\omega = (r, \theta) \in \Omega$ *and* $0 \neq x \in \mathbb{R}$, *then* $\lim_{t\to\infty} \frac{1}{t} \ln |\Phi(\omega, t)x|$ *exists and equals* $\frac{1}{2\pi}\int_0^{2\pi} a(r, \bar{\theta})d\bar{\theta}$. *This integral traces out a nondegenerate interval* I *as* r *varies from* α *to* β. *It turns out that* I *is the dynamical spectrum of the family* (5_ω).

This example is in fact "too simple" and only indicates that we must specify our inverse problem in a more detailed way. So let us suppose that $(\Omega, \{\tau_t\})$ is minimal, and that, for each $\omega \in \Omega$ and each $0 \neq x \in \mathbb{R}^d$, the limit $\lim_{t\to\infty} \frac{1}{t} \ln |\Phi(\omega, t)x|$ exists. We ask: does Φ have discrete spectrum?

This question has an affirmative answer if $d = 1$. It may well be that the answer is still affirmative if $d \geq 2$. We have not been able to prove this, however. Here is what we can and will do.

(1) If $d \geq 2$, we suppose (in addition to the conditions already listed) that, for each ergodic measure μ on Ω, the corresponding Oseledets spectrum $\{\beta_1(\mu) < \beta_2(\mu) < \ldots < \beta_s(\mu)\}$ is simple. That is, $s = d$, or equivalently all the multiplicities d_r are equal to 1 ($1 \leq r \leq s = d$: see Theorem 2.2). Under these conditions, we will show that Φ has discrete spectrum. In fact, it will turn out that the numbers $\beta_1(\mu) = \beta_1, \ldots, \beta_d(\mu) = \beta_d$ do not depend on the choice of the ergodic measure μ, and that $\sigma_\Phi = \{\beta_1 < \beta_2 < \ldots < \beta_d\}$. Thus in particular Φ satisfies the classical Lillo property [23].

(2) If $d = 2$, we make no a priori hypothesis regarding the Oseledets spectrum: we suppose that $(\Omega, \{\tau_t\})$ is minimal, and that, for each $\omega \in \Omega$ and each $0 \neq x \in \mathbb{R}^2$, the limits $\lim_{t\to\infty} \frac{1}{t} \ln |\Phi(\omega, t)x|$ and $\lim_{t\to-\infty} \frac{1}{t} \ln |\Phi(\omega, t)x|$ exist (they need not be equal). We will prove that, subject to these hypotheses, Φ has discrete spectrum. As noted in the Introduction, we generalize a result of Furman [14], who assumes that $(\Omega, \{\tau_t\})$ is strictly ergodic. He uses certain properties of the projective flow defined by Φ when $d = 2$. See also [17] in this regard.

To our knowledge, our inverse problem has not been frequently discussed in the literature. We point out that the hypothesis concerning the existence of

the limits $\lim_{t\to\infty}\frac{1}{t}\ln|\Phi(\omega,t)x|$ (and $\lim_{t\to-\infty}\frac{1}{t}\ln|\Phi(\omega,t)x|$ in point (2)) is rather delicate since no uniformity is assumed. We also remark that there are various results in the literature to the effect that the set of cocycles over a given compact metric flow which have simple Oseledets spectrum is dense in various topologies. See, e.g., [12, 27].

After these preliminary remarks, we express point (1) in a formal statement:

THEOREM 4.2. *Let $T=\mathbb{R}$ or $\mathbb{Z}$. Let $(\Omega,\{\tau_t\})$ be a compact metric minimal flow, and let $\Phi:\Omega\times T\to GL(\mathbb{R}^d)$ be a T-cocycle over $(\Omega,\{\tau_t\})$. Suppose that, for every $\{\tau_t\}$-ergodic measure μ on Ω, the Oseledets spectrum is simple. This means that it consists of d distinct points $\beta_1<\ldots<\beta_d$ (which may depend on μ). Suppose that, for each $\omega\in\Omega$ and $0\neq x\in\mathbb{R}^d$, the limit $\lim_{t\to\infty}\frac{1}{t}\ln|\Phi(\omega,t)x|$ exists. Then the dynamical spectrum σ_Φ of Φ consists of d distinct points (and in particular is discrete).*

Note that, if $(\Omega,\{\tau_t\})$ is minimal, then it is invariantly connected and chain recurrent. So the results stated in Section 2 will be available to us in the proof of Theorem 4.2, to which we now turn.

Before beginning the proof of Theorem 4.2, we describe several convenient constructions. Let Φ be a T-cocycle over $(\Omega,\{\tau_t\})$, let $\sigma_\Phi=[a_1,b_1]\cup\cdots\cup[a_q,b_q]$, and let $V_1^{(c)},\ldots,V_q^{(c)}$ be the corresponding spectral subbundles of Theorem 2.4. These are topological vector subbundles of $\Omega\times\mathbb{R}^d$, of fiber dimension $1\leq d_1,\ldots,d_q$ where $d_1+\cdots+d_q=d$. They need not be topologically trivial; i.e., they need not be equivalent to product bundles $\Omega\times\mathbb{R}^{d_p}$, $1\leq p\leq q$.

However, it is explained in [10] how these bundles can be trivialized via an appropriate cohomology. We explain the relevant constructions of [10].

Let us recall that a minimal flow $(\hat\Omega,\{\hat\tau_t\})$ is said to be an extension of the minimal flow $(\Omega,\{\tau_t\})$ if there is a continuous map $\pi:\hat\Omega\to\Omega$ such that $\pi\circ\hat\tau_t=\tau_t\circ\pi$ for all $t\in T$ (one says that π is a flow homomorphism). Using the minimality of $(\Omega,\{\tau_t\})$ one sees that π must be surjective.

The cocycle Φ can be lifted to a cocycle $\hat\Phi$ on $\hat\Omega$ via the formula $\hat\Phi(\hat\omega,t)=\Phi(\pi(\hat\omega),t)$ $(\hat\omega\in\hat\Omega,\ t\in T)$. Moreover the bundles $V_1^{(c)},\ldots,V_q^{(c)}$ lift to $\hat\Omega$ via the usual pullback construction. Call the lifted bundles $\hat V_1^{(c)},\ldots,\hat V_q^{(c)}$; they are $\hat\Phi$-invariant and it is easy to see that they are the spectral subbundles of $\hat\Phi$. Let us write $\hat V_p^{(c)}(\hat\omega)=\hat V_p^{(c)}\cap(\{\hat\omega\}\times\mathbb{R}^d)$ for the fiber of $\hat V_p(c)$ at $\hat\omega\in\hat\Omega$.

Next let $\mathcal{O}(d)$ be the group of orthogonal $d\times d$ matrices. According to [10, Theorem 4.5], one can find a minimal extension $(\hat\Omega,\{\hat\tau_t\})$ of $(\Omega,\{\tau_t\})$ together with a continuous map $F:\hat\Omega\to\mathcal{O}(d)$ such that, if $\tilde V_p^{(c)}(\hat\omega)=F(\hat\omega)\hat V_p^{(c)}(\hat\omega)$, then the bundle $\tilde V_p^{(c)}=\bigcup_{\hat\omega\in\hat\Omega}\tilde V_p^{(c)}(\hat\omega)$ is a product bundle. In fact, let $e_1,\ldots,e_d$ be the standard basis of $\mathbb{R}^d$. For each $p\in\{2,3,\ldots,q\}$, let us identify $\mathbb{R}^{d_p}$ with the

span of the set of unit vectors $\{e_{d_1+\dots d_{p-1}+1}, \dots, e_{d_1+\dots+d_p}\}$; if $p = 1$ we identify $\mathbb{R}^{d_1}$ with $Span\{e_1, \dots, e_{d_1}\}$. Then F can be chosen so that $\tilde{V}_p^{(c)} = \Omega \times \mathbb{R}^{d_p}$ $(1 \le p \le q)$.

Define the cocycle $\tilde{\Phi}$ by

$$\tilde{\Phi}(\hat{\omega}, t) = F(\tilde{\tau}_t(\hat{\omega}))\hat{\Phi}(\hat{\omega}, t)F(\hat{\omega})^{-1} \qquad (\hat{\omega} \in \hat{\Omega},\ t \in T);$$

thus $\tilde{\Phi}$ is cohomologous to the cocycle Φ via the cohomology F. We see that $\tilde{\Phi}$ admits the spectral decomposition $\tilde{V}_1^{(c)} = \Omega \times \mathbb{R}^{d_1}, \dots, \tilde{V}_q^{(c)} = \Omega \times \mathbb{R}^{d_q}$.

We conclude that, to prove Theorem 4.2, there is no loss of generality in assuming that the spectral subbundles of Φ are product bundles: $V_p^{(c)} = \Omega \times \mathbb{R}^{d_p}$ $(1 \le p \le q)$. This is equivalent to saying that there is no loss of generality in assuming that: (i) Φ has block-diagonal form:

$$\Phi = \begin{pmatrix} \Phi_1 & & 0 \\ & \ddots & \\ 0 & & \Phi_q \end{pmatrix} \tag{6}$$

where Φ_p is a T-cocycle over $(\Omega, \{\tau_t\})$ with values in $GL(\mathbb{R}^d)$, and (ii) the dynamical spectrum of Φ_p is the single interval $[a_p, b_p]$ $(1 \le p \le q)$. (The reader is warned that, if $(\Omega, \{\tau_t\})$ is strictly ergodic, then the extension $(\hat{\Omega}, \{\hat{\tau}_t\})$ of the above construction need not be strictly ergodic.)

We pass to a second construction. Say that Φ is *upper triangular* if $\Phi = (\Phi_{ij})$ where $\Phi_{ij} = 0$ if $i > j$ and $\Phi_{ii} > 0$ $(1 \le i \le d)$. Our construction will give rise to a cohomology between a suitable lifted version of Φ, and an upper triangular cocycle.

Let $\mathcal{O}(d)$ be the group of orthogonal $d \times d$ matrices. If $u_0 \in \mathcal{O}(d)$, then $\Phi(\omega, t)u_0$ can be uniquely decomposed in the form

$$\Phi(\omega, t)u_0 = U(\omega, u_0, t)\Delta(\omega, u_0, t) \qquad (\omega \in \Omega,\ t \in T)$$

where $U \in \mathcal{O}(d)$ and Δ is upper triangular with positive diagonal elements. This follows from the Gram-Schmidt decomposition of $\Phi(\omega, t)u_0$. It turns out that, if one sets $\hat{\tau}_t(\omega, u_0) = (\tau_t(\omega), U(\omega, u_0, t))$ then $\{\hat{\tau}_t \mid t \in T\}$ is a flow on $\Omega \times \mathcal{O}(d)$, and Δ is a $\{\hat{\tau}_t\}$-cocycle.

Note that, if Φ has a block diagonal structure as in (6), then U and Δ have corresponding block-diagonal structures.

Next let $\hat{\Omega} \subset \Omega \times \mathcal{O}(d)$ be a minimal $\{\hat{\tau}_t\}$-subflow (such a subflow exists by Zorn's Lemma). Then the projection $\pi : \hat{\Omega} \to \Omega : (\omega, u_0) \mapsto \omega$ is continuous, and $\pi \circ \hat{\tau}_t = \tau_t \circ \pi$. We introduce the lifted cocycle $\hat{\Phi} : \hat{\Omega} \times T \to GL(\mathbb{R}^d) : \hat{\Phi}(\omega, t) = \Phi(\pi(\hat{\omega}), t)$ where $\hat{\omega} = (\omega, t) \in \hat{\Omega}$. Note that the map $F : \hat{\Omega} \to \mathcal{O}(d) : F(\omega, u_0) = u_0$ defines a cohomology between $\hat{\Phi}$ and Δ. In fact, $F(\hat{\tau}_t(\hat{\omega}))\Delta(\hat{\omega})F(\hat{\omega})^{-1} = \hat{\Phi}(\hat{\omega}, t)$ for $\hat{\omega} = (\omega, t) \in \hat{\Omega}$ and $t \in T$.

Our third and final construction was already discussed in Section 2. Namely, assume that $T = \mathbb{R}$. Then there exists a continuous function $A : \Omega \to \mathbb{L}(\mathbb{R}^d)$ such that Φ is cohomologous to the cocycle generated by the family of linear ODEs (1_ω):

$$x' = A(\tau_t(\omega))x.$$

We observe that, if a given cocycle Φ has a block-triangular form as in the first construction, then the coefficient matrix $A(\cdot)$ in (1_ω) may be chosen to have the corresponding block-diagonal form. Moreover, if Φ has an upper triangular from as in the second construction, then $A(\cdot)$ may be chosen to have the corresponding upper triangular form.

We assume until further notice that $T = \mathbb{R}$. Using the above constructions, we see that by introducing a suitable minimal extension of $(\Omega, \{\tau_t\})$, and by introducing a suitable cohomology, it can be arranged that Φ satisfies the following conditions.

HYPOTHESES 4.3. (*a*) *The cocycle* Φ *is generated by a family of linear ODEs*

$$x' = A(\tau_t(\omega))x \qquad \omega \in \Omega,\ x \in \mathbb{R}^d \tag{7_ω}$$

where the matrix function $A(\cdot)$ *has block-diagonal form:* $A = \begin{pmatrix} A_1 & & 0 \\ & \ddots & \\ 0 & & A_q \end{pmatrix}$.

(*b*) *If* Φ_p *is the cocycle over* $(\Omega, \{\tau_t\})$ *which is generated by the family* $x' = A_p(\tau_t(\omega))x$, *then the dynamical spectrum* σ_p *of* Φ_p *is the single interval* $[a_p, b_p]$ *(*$1 \le p \le q$*).*

(*c*) *Each matrix function* A_p *is upper triangular (*$1 \le p \le q$*).*

It can be shown that, if Φ and Ψ are cohomologous cocycles, and if Φ satisfies the hypotheses of Theorem 4.2, then so does Ψ. It can also be shown that, if Φ and Ψ are cohomologous, and if Φ satisfies the thesis of Theorem 4.2, then so does Ψ.

We pass to the proof of Theorem 4.2 in the case when $T = \mathbb{R}$. According to the above constructions and remarks, we can assume that Φ satisfies any or all of Hypotheses 4.3 (a)–(c), when it is appropriate to do so.

We proceed by induction on the dimension d of the cocycle Φ. Suppose that $d = 1$. There is no loss of generality in assuming that Φ is generated by a family of one dimensional systems of the form (1_ω). The family (1_ω) has the form $x' = A(\tau_t(\omega))x$ where $A : \Omega \to \mathbb{R}$ is a continuous scalar function. Using the hypothesis concerning the existence of the limits which define the Lyapunov exponents of Φ, we see that $\lim_{t\to\infty} \frac{1}{t} \int_o^t A(\tau_s(\omega))ds$ exists for all $\omega \in \Omega$.

Now the flow $(\Omega, \{\tau_t\})$ is by assumption minimal, so one can use an oscillation result of Johnson [16] to show that the quantity $\bar{a} = \lim_{t\to\infty} \frac{1}{t} \int_0^t A(\tau_s(\omega))ds$

does not depend on $\omega \in \Omega$, and the limit is uniform in ω. Moreover $\bar{a} = \lim_{t\to-\infty} \frac{1}{t}\int_0^t A(\tau_s(\omega))ds$, where again the limit is uniform in $\omega \in \Omega$. One can now check directly that the dynamical spectrum of Φ satisfies $\sigma_\Phi = \{\bar{a}\}$; i.e., it is discrete.

Next, suppose that Theorem 4.2 is valid for all continuous $\mathbb{R}$-cocycles of dimension $\leq d-1$, over all minimal flows $(\Omega, \{\tau_t\})$. We suppose without loss of generality that our given cocycle Φ satisfies Hypotheses 4.3 (a) and (b). Suppose first that the number of diagonal blocks of the (d-dimensional) matrix function $A(\cdot)$ is at least 2. Each block $A_1, \ldots, A_q$ then has dimension $\leq d-1$. So by the induction hypothesis, the family

$$x' = A_p(\tau_t(\omega))x \qquad (\omega \in \Omega,\ x \in \mathbb{R}^d)$$

has discrete spectrum ($1 \leq p \leq q$). By Hypotheses 4.3 (2), this spectrum is the singleton $\{a_p\}$, and it follows that the cocycle Φ has discrete spectrum: $\sigma_\Phi = \{a_1, \ldots, a_q\}$. So Theorem 4.2 is proved in this case.

We now assume that $q = 1$, which means that the spectrum σ_Φ of Φ consists of a single interval $[a, b]$. We must show that $a = b$. We assume w.l.o.g. that Hypotheses $(a), (b)$ and (c) are valid. The matrix function $A(\cdot)$ has values in $\mathbb{L}(\mathbb{R}^d)$ and is upper triangular.

Let us write

$$A(\omega) = \begin{pmatrix} A_*(\omega) & a_{1d}(\omega) \\ \\ 0 & a_{dd}(\omega) \end{pmatrix}$$

where A_* takes values in $\mathbb{L}(\mathbb{R}^{d-1})$ and is upper triangular. Consider the family of subsystems

$$y' = A_*(\tau_t(\omega))y \qquad \omega \in \Omega,\ y \in \mathbb{R}^{d-1}. \tag{8_ω}$$

Note that a solution $y(t)$ of (8_ω) determines a solution $x(t) = \begin{pmatrix} y(t) \\ x_n(t) \end{pmatrix}$ of (7_ω) by setting $x_n(t) = 0$; that is, $x(t) = \begin{pmatrix} y(t) \\ 0 \end{pmatrix}$ is a solution of (7_ω) if and only if $y(t)$ is a solution of (8_ω).

We see that the family (8_ω) has the property that $\lim_{t\to\infty} \frac{1}{t}\ln|y(t)|$ exists whenever $y(t)$ is a nonzero solution of equation (8_ω) ($\omega \in \Omega$). By the induction hypothesis, the dynamical spectrum σ_* of the family (8_ω) is discrete, say

$$\sigma_* = \{\alpha_1 < \alpha_2 < \ldots < \alpha_j\}$$

where $1 \leq j \leq d-1$. By Proposition 2.2, the set of Lyapunov exponents of (8_ω) is exactly $\{\alpha_1, \ldots, \alpha_j\}$ for each $\omega \in \Omega$. Moreover, if d_i is the multiplicity of α_i for $1 \leq i \leq j$, then $d_1 + \cdots + d_j = d - 1$.

Now, for each ergodic measure μ on Ω, the Oseledets spectrum of the family (8_ω) is contained in the dynamical spectrum σ_* of that family. Moreover, the Oseledets spectrum equals the set of averages

$$\left\{\int_\Omega a_{ii}(\omega)d\mu(\omega) \mid 1 \le i \le d-1\right\}$$

of the diagonal elements of A_*; see [18, 31]. By hypothesis, the μ-Oseledets spectrum of Φ is simple, and therefore the μ-Oseledets spectrum of the cocycle Φ_* generated by equations (8_ω) is also simple. Using the fact that $\sigma_* = \{\alpha_1 < \alpha_2 < \ldots < \alpha_j\}$, we see that each multiplicity $d_i = 1$, and that $\sigma_* = \{\alpha_1 < \alpha_2 < \ldots < \alpha_{d-1}\}$ consists of $d-1$ distinct real numbers. It is clear that these numbers are just a reordered version of the numbers $\left\{\int_\Omega a_{ii}(\omega)d\mu(\omega) \mid 1 \le i \le d-1\right\}$. One can show (by applying Proposition 3.2, or by carrying out a "secondary" induction on j, $1 \le j \le d-1$) that $\int_\Omega a_{ii}(\omega)d\mu(\omega)$ does not depend on the choice of the $\{\tau_t\}$-ergodic measure μ, if $1 \le j \le d-1$.

We must now study the significance of the numbers $\int_\Omega a_{dd}(\omega)d\mu(\omega)$ as μ ranges over the set of $\{\tau_t\}$-ergodic measures on Ω. To do this, it is convenient to introduce a projective flow. The construction is quite similar to that carried out in the proof of Theorem 3.2 above. Let $\mathbb{L}$ be the $(d-1)$-dimensional manifold of lines through the origin in $\mathbb{R}^d$. Let $\mathbb{B} = \Omega \times \mathbb{L}$, and define a flow $\{\hat\tau_t\}$ on $\mathbb{B}$ by setting $\hat\tau_t(\omega, l) = (\tau_t(\omega), \Phi(\omega, t)l)$ $(\omega \in \Omega,\ l \in \mathbb{L})$. Define $f : \mathbb{B} \to \mathbb{R} : f(\omega, l) = \langle A(\omega)x, x\rangle / \langle x, x\rangle$ if $0 \ne x \in l$. Then if $x(t)$ is a solution of (7_ω), and if $l \in \mathbb{L}$ is the line containing $x(0) \ne 0$, then

$$\int_0^t f(\hat\tau_s(\omega, l))ds = \ln \frac{|x(t)|}{|x(0)|}. \tag{9}$$

By the hypothesis concerning the existence of the limits defining the Lyapunov exponents, and by (9), one has that the limit $\lim_{t\to\infty} \frac{1}{t}\int_0^t f(\hat\tau_s(b))ds$ exists for each $b \in \mathbb{B}$. Let us denote the limit by $f_*(b)$. Since f_* is the pointwise limit of a sequence of continuous functions, it admits a residual set of continuity points [6]. Let b_* be a point of continuity of f_*. For each $\varepsilon > 0$, there is an open neighborhood $U = U(\varepsilon) \subset \Omega \times \mathbb{L}$ of b_* such that, if $b \in U$, then $|f_*(b) - f_*(b_*)| < \varepsilon$. There is no loss of generality in assuming that $U = U_1 \times U_2$ where $U_1 \subset \Omega$ and $U_2 \subset \mathbb{L}$ are open sets. There is also no loss of generality in assuming that U does not intersect the $\{\hat\tau_t\}$-invariant set $\mathbb{B}_1 = \{(\omega, l) \in \mathbb{B} \mid l \subset \mathbb{R}^{d-1} \subset \mathbb{R}^d\}$.

For each $\omega \in \Omega$, there is a real number $\beta_*(\omega)$ such that the set of Lyapunov exponents of equation (7_ω) equals $\{\alpha_1, \alpha_2, \ldots, \alpha_{d-1}, \beta_*(\omega)\}$. Let $\beta_{\max}(\omega) = \max\{\alpha_1, \ldots, \alpha_{d-1}, \beta_*(\omega)\}$ be the largest Lyapunov exponent of (7_ω). Write

the continuity point b_* of f_* in the form $b_* = (\omega_*, l_*)$. It follows from the continuity of f_* at b_* that $f_*(b_*)$ equals $\beta_{\max}(\omega_*)$. In fact, this is a consequence of the observation that, if $\bar{\beta}(\omega_*)$ is the maximum of the Lypunov exponents of (7_{ω_*}) which are distinct from $\beta_{\max}(\omega_*)$, then $\left\{x \in \mathbb{R}^d \mid \ x = 0 \text{ or } \lim_{t\to\infty} \frac{1}{t} \ln|\Phi(\omega_*, t)x| \leq \bar{\beta}(\omega_*)\right\}$ is a proper vector subspace of $\mathbb{R}^d$, so its complement in $\mathbb{R}^d$ is open and dense. This means that there is an open dense subset $W \subset \mathbb{L}$ such that, if $l \in W$, then $\lim_{t\to\infty} \int_0^t f(\hat{\tau}_s(\omega_*, l)ds = \beta_{\max}(\omega_*)$.

Recall that we are working under the hypothesis that σ_Φ is a single interval $[a, b]$. Using Theorem 2.4, we see that the numbers $\alpha_1, \ldots, \alpha_{d-1}$ all lie in $[a, b]$. Suppose for the time being that b is greater than α_{d-1}.

According to a result of [18], there is a $\{\tau_t\}$-ergodic measure μ on Ω for which b is a Lyapunov exponent of Φ, for μ-a.a. $\omega \in \Omega$. By Theorem 2.4, we have that $\beta_{\max}(\omega) = b$ for μ-a.a. $\omega \in \Omega$. Fix a point $\bar{\omega} \in \Omega$ such that $\beta_{\max}(\bar{\omega}) = b$. If $x \in \mathbb{R}^d$, we write $x = \begin{pmatrix} y \\ x_d \end{pmatrix}$ where $y \in \mathbb{R}^{d-1}$ and $x_d \in \mathbb{R}$. Let x be a vector such that $\lim_{t\to\infty} \frac{1}{t} \ln|\Phi(\bar{\omega}, t)x| = b$. Writing $\Phi(\bar{\omega}, t)x = \Phi(\bar{\omega}, t)\begin{pmatrix} y \\ x_d \end{pmatrix} = \begin{pmatrix} y(t) \\ x_d(t) \end{pmatrix}$, and using the fact that $b > \alpha_{d-1} = \max\{\alpha_i \mid 1 \leq i \leq d-1\}$, we see that $x_d \neq 0$, and that $\lim_{t\to\infty} \frac{1}{t} \ln|x_d(t)| = b$. (For later use, we note that $b = \lim_{t\to\infty} \frac{1}{t} \ln|x_d(t)| = \lim_{t\to\infty} \frac{1}{t} \int_0^t a_{dd}(\tau_s(\bar{\omega})ds.)$

One checks that, if $\begin{pmatrix} y \\ x_d \end{pmatrix}$ is *any* vector with $x_d \neq 0$, then $\lim_{t\to\infty} \frac{1}{t} \ln|\Phi(\bar{\omega}, t)x| = b$.

Return to the continuity point $(\omega_*, l_*) \in \mathbb{B}$ of f_* which was introduced previously. Let $\varepsilon > 0$, and choose $U(\varepsilon) = U = U_1 \times U_2$ as before. Let $\bar{\omega}$ be the point of the preceding two paragraphs. Since $(\Omega, \{\tau_t\})$ is minimal, the positive semiorbit $\{\tau_t(\bar{\omega}) \mid t \geq 0\}$ is dense in Ω, hence it enters U_1. Using the fact that U does not intersect $\mathbb{B}$ together with the result of the previous paragraph, we can find a vector $\begin{pmatrix} y \\ x_d \end{pmatrix} \in \mathbb{R}^d$, whose projective image l lies in U_2, such that $\lim_{t\to\infty} \frac{1}{t} \ln|\Phi(\omega_*, t)x| = b$. This means that $|f_*(\omega_*, l_*) - b| \leq \varepsilon$. Since $\varepsilon > 0$ is arbitrary, we have that $f_*(\omega_*, l_*) = b$.

Next, let $\varepsilon > 0$, and let $\omega \in \Omega$ be any point of Ω. Again the positive semiorbit $\{\tau_t(\omega) \mid t \geq 0\}$ enters U_1. So there exists a vector $x = \begin{pmatrix} y \\ x_d \end{pmatrix} \in \mathbb{R}^d$ with $x_d \neq 0$ such that $\left|\lim_{t\to\infty} \frac{1}{t} \ln|\Phi(\omega, t)x| - f_*(\omega_*, l_*)\right| \leq \varepsilon$. Hence $\left|\lim_{t\to\infty} \frac{1}{t} \ln|\Phi(\omega, t)x| - b\right| \leq \varepsilon$. At this point choose $0 < \varepsilon < \frac{1}{2}(b - \alpha_{d-1})$, and

write $\Phi(\omega,t)x = \begin{pmatrix} y(t) \\ x_d(t) \end{pmatrix}$. It can be checked that the limit $\lim_{t\to\infty} \frac{1}{t} \ln |x_d(t)|$ exists and is $\geq b - \varepsilon > \alpha_{d-1} + \varepsilon$.

Now, $\lim_{t\to\infty} \frac{1}{t} \ln |x_d(t)| = \lim_{t\to\infty} \frac{1}{t} \int_0^t a_{dd}(\tau_s(\omega))ds$. We are able to conclude that the limit $\lim_{t\to\infty} \frac{1}{t} \int_0^t a_{dd}(\tau_s(\omega))ds$ exists for all $\omega \in \Omega$. The limit equals b if $\omega = \bar{\omega}$. By the oscillation result of [16], $\int_\Omega a_{dd} d\mu = b$ for all ergodic measures μ on Ω. By using a Krylov-Bogoliubov argument, one proves that $\lim_{t\to\infty} \frac{1}{t} \int_0^t a_{dd}(\tau_s(\omega))ds = b$, and the limit is uniform in $\omega \in \Omega$.

Let $\alpha_{d-1} < \lambda < b$. Let us show that $\lambda \notin \sigma_\Phi$. Using Theorem 2.3, we see that it is sufficient to show that, if $\omega \in \Omega$ and $0 \neq x \in \mathbb{R}^d$, then $e^{\lambda t}\Phi(\omega,t)x$ is not bounded in $-\infty < t < \infty$. To do this, note first that, if $x = \begin{pmatrix} y \\ 0 \end{pmatrix} \in \mathbb{R}^d$, then $|e^{-\lambda t}\Phi(\omega,t)x| \to \infty$ as $t \to -\infty$, because $\sigma_* = \{\alpha_1, \ldots, \alpha_{d-1}\}$. On the other hand, if $x = \begin{pmatrix} y \\ x_d \end{pmatrix}$ with $x_d \neq 0$, and if $x_d(t)$ is defined by $\Phi(\omega,t)x = \begin{pmatrix} y(t) \\ x_d(t) \end{pmatrix}$, then $|x_d(t)| \to \infty$ as $t \to \infty$. So in fact $\lambda \notin \sigma_\Phi$.

However, σ_Φ is by hypothesis the interval $[a,b]$, and we know that $\alpha_{d-1} \in \sigma_\Phi$. So we have arrived at a contradiction, and must conclude that $b \leq \alpha_{d-1}$.

There remains to study the situation when $b \leq \alpha_{d-1}$. For this, let us first recall that, if $1 \leq i \leq d-1$, then $\int_\Omega a_{ii} d\mu$ does not depend on the choice of the ergodic measure μ on Ω. Second, we recall that, if μ is an ergodic measure on Ω, then the corresponding Oseledets spectrum equals $\left\{ \int_\Omega a_{11} d\mu, \ldots, \int_\Omega a_{dd} d\mu \right\}$. By hypothesis, the Oseledets spectrum is simple for each $\{\tau_t\}$-ergodic measure μ on Ω. So $\int_\Omega a_{dd} d\mu < \alpha_{d-1}$ for each such μ. Let us define $\bar{\alpha} = \sup \left\{ \int_\Omega a_{dd} d\mu \;\middle|\; \mu \right.$ is a $\{\tau_t\}$-ergodic measure on $\left. \Omega \right\}$. We claim that $\bar{\alpha} < \alpha_{d-1}$. Here is a sketch of the proof. Since the set $\{\nu\}$ of $\{\tau_t\}$-invariant measures on Ω is compact and convex in the weak-$*$ topology, and since μ is an extreme point of $\{\nu\}$ if and only if μ is ergodic, we can use the Choquet theorem [35] to show that $\int_\Omega a_{dd} d\nu \leq \alpha_{d-1}$ for each $\{\tau_t\}$-invariant measure ν on Ω. If $\bar{\alpha}$ equals α_{d-1}, then the weak-$*$ compactness of $\{\nu\}$ allows us to find an invariant measure ν on Ω such that $\int_\Omega a_{dd} d\mu = \alpha_{d-1}$. Using the Choquet theorem again, we determine an ergodic measure μ on Ω such that $\int_\Omega a_{dd} d\mu = \alpha_{d-1}$. This is not possible,

so indeed $\bar{\alpha} < \alpha_{d-1}$.

We can now use a Krylov-Bogoliubov argument to prove the following statement: Let $\varepsilon > 0$; then there exists $T > 0$ such that, if $t \leq -T$ and $\omega \in \Omega$, then $\frac{1}{t}\int_0^t a_{dd}(\tau_s(\omega))ds \leq \bar{\alpha} + \varepsilon$.

Next choose $\lambda \in (\bar{\alpha}, \alpha_{d-1})$ such that $\lambda > \alpha_{d-2}$. We claim that λ is not in the spectrum σ_Φ of Φ. As before, it is sufficient to show that, if $\omega \in \Omega$ and $0 \neq x \in \mathbb{R}^d$, then $e^{-\lambda t}\Phi(\omega, t)x$ is unbounded on $-\infty < t < \infty$. So let $x = \begin{pmatrix} y \\ 0 \end{pmatrix}$ where $y \in \mathbb{R}^{d-1}$. Then $e^{-\lambda t}\Phi(\omega, t)x$ is unbounded because $\lambda \notin \sigma_* = \{\alpha_1, \ldots, \alpha_{d-1}\}$. On the other hand, if $x = \begin{pmatrix} y \\ x_d \end{pmatrix}$ with $x_d \neq 0$, then $e^{-\lambda t}\Phi(\omega, t)x$ is unbounded as $t \to -\infty$.

We conclude as before that σ_Φ cannot be an interval, which contradicts the assumption that $\sigma_\Phi = [a, b]$. This completes the proof of Theorem 4.2 in the case $T = \mathbb{R}$.

There remains to prove Theorem 4.2 in the case when $T = \mathbb{Z}$. One can do this by following the steps of the above proof for $T = \mathbb{R}$. The proof when $T = \mathbb{Z}$ is actually somewhat simpler, since one need not effect a cohomology which transforms the cocycle Φ into the cocycle defined by a family of differential systems (1_ω). We omit the details. □

We finish the paper with a discussion of the case $d = 2$. We are able to strengthen Theorem 4.2 in the sense that we do not need the hypothesis of simple Oseledets spectrum. On the other hand, we need the convergence of the time averages which define the Lyapunov exponents at $t = -\infty$.

THEOREM 4.4. *Let $T = \mathbb{R}$ or $\mathbb{Z}$, and let $(\Omega, \{\tau_t\})$ be a minimal flow. Let Φ be a T-cocycle over $(\Omega, \{\tau_t\})$ with values in $GL(\mathbb{R}^2)$. Suppose that, for each $\omega \in \Omega$ and $0 \neq x \in \mathbb{R}^2$, the limits*

$$\lim_{t\to\infty} \frac{1}{t} \ln |\Phi(\omega, t)x|, \qquad \lim_{t\to-\infty} \frac{1}{t} \ln |\Phi(\omega, t)x|$$

both exist (they may or may not be equal). Then Φ has discrete spectrum.

Proof. We consider the case $T = \mathbb{R}$. There is no loss of generality in assuming that Hypotheses 4.3 $(a), (b)$ and (c) are satisfied. In particular the spectrum σ_Φ consists of a single interval; $\sigma_\Phi = [a, b]$ with $a \leq b$.

Let us write equations (7_ω) in the form

$$x' = \begin{pmatrix} a_{11}(\tau_s(\omega)) & a_{12}(\tau_s(\omega)) \\ 0 & a_{22}(\tau_s(\omega)) \end{pmatrix} x,$$

where $x \in \mathbb{R}^2$. It follows from the hypothesis concerning the existence of the limits that, for each $\omega \in \Omega$, the limit $\lim_{t\to\infty} \frac{1}{t}\int_0^t a_{11}(\tau_s(\omega))ds$ exists. By [16], there is a real number $\bar{a}_1$ such that $\lim_{t\to\pm\infty} \frac{1}{t}\int_0^t a_{11}(\tau_s(\omega))ds = \bar{a}_1$, where the limits are uniform in $\omega \in \Omega$.

It follows that $\bar{a}_1 = \int_\Omega a_{11}d\mu$ for each $\{\tau_t\}$-ergodic measure on Ω. Now, by [18] there is an ergodic measure μ_a on Ω such that a is an element of the μ_a-Oseledets spectrum. Similarly, there is an ergodic measure μ_b on Ω such that b is an element of the μ_b-Oseledets spectrum. Therefore $\bar{a}_1 \in \{a, b\}$.

Suppose first that $\bar{a}_1 = a$, and assume for contradiction that $b > a$. Then we can argue as in the proof of Theorem 4.2 to show that $\int_\Omega a_{22}d\mu = b$ for every ergodic measure μ on Ω and so $\frac{1}{t}\int_0^t a_{22}(\tau_s(\omega))ds = b$ uniformly in $\omega \in \Omega$. Again, arguing as in the proof of Theorem 4.2, one shows that, if $\lambda \in (a, b)$, then λ is not in σ_Φ. This is a contradiction, so $b = a$ and in fact Φ has discrete spectrum.

If $\bar{a}_1 = b$, then we use the hypothesis that $\lim_{t\to-\infty} \frac{1}{t}\ln|\Phi(\omega, t)x|$ exists for all $\omega \in \Omega$ and all $0 \neq x \in \mathbb{R}^2$. One assumes for contradiction that $a < b$, then repeats the steps of the proof of Theorem 4.2, using the negative-time Lyapunov exponents $\lim_{t\to-\infty} \frac{1}{t}\ln|\Phi(\omega, t)x|$ in place of the positive-time exponents $\lim_{t\to\infty} \frac{1}{t}\ln|\Phi(\omega, t)x|$. The end result is that, if $a < \lambda < b$. then $\lambda \notin \sigma_\Phi$. So one again concludes that σ_Φ is discrete.

□

References

[1] M. Bessa, *Dynamics of generic multidimensional linear differential systems*, Adv. Nonlinear Stud. **8** (2008), 191–211.

[2] J. Bochi and M. Viana, *The Lyapunov exponents of generic volume-preserving and symplectic maps*, Ann. of Math. **161** (2005), 1–63.

[3] B. Bylov, *A reduction to block-triangular form, and necessary and sufficient conditions for the stability of the characteristic exponents of a linear system of differential equations*, Differencial'nye Uravnenija **6** (1970), 243–252.

[4] B. Bylov and N. Izobov, *Necessary and sufficient stability conditions for the characteristic indices of a linear system*, Differencial'nye Uravnenija **5** (1969), 1794–1803.

[5] B. Bylov, R. Vinograd, D. Grobman and V. Nemytskii, *Theory of Lyapunov Exponents and its Applications to Stability*, Nauka, Moscow, 1966.

[6] G. Choquet, *Lectures on Analysis*, Vol. 1, Benjamin, New York, 1969.
[7] C. Conley, *Isolated invariant sets and the Morse index*, CBMS Regional Conference Series in Mathematics, Vol. 38, SIAM, Philadelphia, 1978.
[8] D. Damanik and D. Lenz, *A condition of Boshernitzan and uniform convergence in the multiplicative ergodic theorem*, Duke Math. J. **133** (2006), 95–123.
[9] R. Ellis, *Lectures on Topological Dynamics*, Benjamin, New York, 1969.
[10] R. Ellis and R. Johnson, *Topological dynamics and linear differential systems*, J. Differential Equations **44** (1982), 21–39.
[11] R. Fabbri, R. Johnson, S. Novo and C. Núnez, *Some remarks concerning weakly disconjugate linear Hamiltonian systems*, J. Math. Anal. Appl. **380** (2011), 853–864.
[12] R. Fabbri, R. Johnson and L. Zampogni, *On the Lyapunov exponents of certain $SL(2,\mathbb{R})$-cocycles II*, Differ. Equ. Dyn. Syst. **18** (2010), 135–161.
[13] H. Federer, *Geometric Measure Theory*, Springer-Verlag, New York, 1969.
[14] A. Furman, *On the multiplicative ergodic theorem for uniquely ergodic systems*, Ann. Inst. H. Poincaré Probab. Statist. **33** (1997), 797–815.
[15] H. Furstenberg and H. Kesten, *Products of random matrices*, Ann. Math. Statist. **31** (1960), 457–489.
[16] R. Johnson, *Minimal functions with unbounded integral*, Israel J. Math. **31** (1978), 133–141.
[17] R. Johnson, *Ergodic theory and linear differential equations*, J. Differential Equations **28** (1978), 23–34.
[18] R. Johnson, K. Palmer and G. Sell, *Ergodic properties of linear dynamical systems*, SIAM J. Math. Anal. **18** (1987), 1–33.
[19] R. Johnson and G. Sell, *Smoothness of spectral subbundles and reducibility of quasi-periodic linear differential systems*, J. Differential Equations **41** (1981), 262–288.
[20] N. Krylov and N. Bogoliubov, *La théorie générale de la mesure dans son application à l'étude des systémes dynamiques de la méchanique non lineaire*, Ann. Math. **38** (1937), 65–113.
[21] J. La Salle, *The stability of dynamical systems*, SIAM Regional Conference Series on Applied Mathematics, SIAM, Philadelphia, 1977.
[22] D. Lenz, *Singular spectrum of Lebesgue measure zero for one-dimensional quasicrystals*, Comm. Math. Phys. **227** (2002). 119–130.
[23] J. Lillo, *Approximate similarity and almost periodic matrices*, Proc. Amer. Math. Soc. **12** (1961), 400–407.
[24] A. Lyapunov, *Problème Générale de la Stabilité du Mouvement*, Éditions Jacques Gabay, Sceaux, 1988.
[25] R. Mané, *Oseledets theorem from the generic viewpoint*, Proc. Int. Cong. Math., Warsaw 1983, pp. 1269–1276.
[26] V. Millionshchikov, *A stability criterion for the probable spectrum of linear systems of differential equations with recurrent coefficients and a criterion for the almost reducibility of systems with almost periodic coefficients*, Math. USSR-Sb. **78** (1969), 171–193.
[27] V. Millionshchikov, *A certain dense set in a space of linear systems*, Diff. Urav. **11** (1975), 755–757.

[28] V. Millionshchikov, *Baire classes of functions and Lyapunov exponents IX*, Differential Equations **18** (1982), 1507–1548.
[29] V. Nemytskii and V. Stepanov, *Qualitative Theory of Differential Equations*, Princeton University Press, Princeton USA, 1960.
[30] V. Novikov, *On almost reducible systems with almost periodic coefficients*, Math. Notes **16** (1975), 1065–1071.
[31] V. Oseledets, *A multiplicative ergodic theorem. Lyapunov characteristic numbers for dynamical systems*, Trans. Moscow Math. Soc. **19** (1968), 197–231.
[32] K. Palmer, *On the reducibility of almost periodic systems of linear differential equations*, J. Differential Equations **36** (1980), 374–390.
[33] K. Palmer, *Exponential separation, exponential dichotomy and spectral theory for linear systems of ordinary differential equations*, J. Differential Equations **46** (1982), 324–345.
[34] K. Palmer, *A perturbation theorem for exponential dichotomy*, Proc. Royal Soc. Edinburgh Sect. A **106** (1987), 25–37.
[35] R. Phelps, *Lectures on Choquet Theory*, American Book Co., New York, 1966.
[36] D. Ruelle, *Ergodic theory of differentiable dynamical systems*, Inst. Hautes Études Sci. Publ. Mat. **50** (1979), 27–58.
[37] R. Sacker and G. Sell, *Existence of dichotomies and invariant splittings for linear differential systems II*, J. Differential Equations **22** (1976), 478–496.
[38] R. Sacker and G. Sell, *A spectral theory for linear differential systems*, J. Differential Equations **27** (1978), 320–358.
[39] J. Selgrade, *Isolated invariant sets for flows on vector bundles*, Trans. Amer. Math. Soc. **203** (1975), 359–390.
[40] R. Vinograd, *On the central characteristic exponent of a system of differential equations*, Math. USSR-Sb. **42** (1957), 207–222.

Authors' addresses:

Russell Johnson
Dipartimento di Sistemi e Informatica
Università di Firenze, Italy
E-mail: johnson@dsi.unifi.it

Luca Zampogni
Dipartimento di Matematica e Informatica
Università degli Studi di Perugia, Itay
E-mail: zampoglu@dmi.unipg.it

Received April 12, 2012
Revised June 9, 2012

Rend. Istit. Mat. Univ. Trieste
Volume 44 (2012), 117–132

A boundary value problem on the half-line for superlinear differential equations with changing sign weight[1]

Mauro Marini and Serena Matucci

Dedicated to the 60th birthday of Professor Fabio Zanolin

Abstract. *The existence of positive solutions x for a superlinear differential equation with p-Laplacian is here studied, satisfying the boundary conditions $x(0) = x(\infty) = 0$. Under the assumption that the weight changes its sign from nonpositive to nonnegative, necessary and sufficient conditions for the existence are derived by combining Kneser-type properties for solutions of an associated boundary value problem on a compact set, a-priori bounds for solutions of suitable boundary value problems on noncompact intervals, and continuity arguments.*

Keywords: differential equation with p-Laplacian, positive solutions, decaying solutions
MS Classification 2010: 34B15, 34B18

1. Introduction

In this paper we study the existence of solutions for the second order nonlinear differential equation with p-Laplacian

$$\big(r(t)\Phi(x')\big)' = q(t)f(x), \tag{1}$$

satisfying the boundary conditions

$$x(0) = 0, \quad \lim_{t\to\infty} x(t) = 0, \quad x(t) > 0 \quad \text{for } t > 0. \tag{2}$$

We will assume the following conditions:

H1. $\Phi(u) = |u|^p \operatorname{sgn} u$, for $u \in \mathbb{R}$ and $p > 0$;

[1]Both authors are supported by the Research Project *PRIN09-Area 01 "Equazioni differenziali ordinarie e applicazioni"* of the Italian Ministry of Education.

H2. f is a continuous function on $\mathbb{R}$ such that $uf(u) > 0$ for $u \neq 0$, and

$$\text{(a)} \quad \lim_{u\to 0^+} \frac{f(u)}{\Phi(u)} = 0, \qquad \text{(b)} \quad \lim_{u\to\infty} \frac{f(u)}{\Phi(u)} = \infty; \tag{3}$$

H3. r, q are continuous functions for $t \geq 0$, $r(t) > 0$ for $t \geq 0$, and q satisfies the sign condition

$$q(t) \leq 0,\ q(t) \not\equiv 0, \text{ for } t \in [0,1],$$
$$q(t) \geq 0 \text{ for } t > 1,\ q(t) \not\equiv 0 \text{ for large } t.$$

Boundary values problems (BVPs) associated to (1) on infinite intervals have been considered in many papers. For instance, in [14, 18, 20] some asymptotic problems for second-order equations with the Sturm-Liouville operator, possibly singular, are studied and BVPs, concerning equations with p-Laplacian, are considered, e.g., in [9, 11, 17]. For other contributions we refer to the monograph [1] and references therein.

As usual, by a solution of (1), we mean a continuously differentiable function x such that $r(t)\Phi(x')$ has a continuous derivative satisfying (1). For any solution x of (1), denote its quasiderivative as

$$x^{[1]}(t) = r(t)\Phi(x').$$

Let

$$R(t) = \int_0^t r^{-\frac{1}{p}}(s)\, ds.$$

The limit $\lim_{t\to\infty} R(t)$ will be denoted by $R(\infty)$; both the cases $R(\infty) < \infty$ and $R(\infty) = \infty$ will be considered. If $R(\infty) < \infty$, we put

$$\rho(t) = \int_t^\infty r^{-\frac{1}{p}}(s)\, ds.$$

The sign condition on q is motivated by the following. When q has constant sign on the whole half-line, and $q \not\equiv 0$, we can distinguish three cases: i_1) $q(t) \geq 0$ for $t \geq 0$, i_2) $q(t) \leq 0$ for $t \geq 0$ and $R(\infty) = \infty$, i_3) $q(t) \leq 0$ for $t \geq 0$ and $R(\infty) < \infty$. In cases i_1) or i_2), the problem (1)-(2) is not solvable. To see this, if i_1) holds, consider the function $G(t) = r(t)\Phi(x')x$, where x is a solution of (1)-(2). Since $G'(t) = q(t)f(x)x + r(t)|x'|^{p+1}$, then G is nondecreasing, and, as $G(0) = 0$, we obtain $G(t) \geq 0$ for $t > 0$. Thus, the positivity of x yields the existence of a point $t_0 > 0$ such that $G(t_0) > 0$. Since G is nondecreasing, x' is eventually positive, which contradicts the asymptotic condition in (2). In case i_2), for any solution x of (1)-(2) the quasiderivative $x^{[1]}$ is nonincreasing. If $\lim_{t\to\infty} x^{[1]}(t) = k \geq 0$, we immediately get a contradiction with the

boundary conditions (2), since x should be eventually nondecreasing. Therefore $\lim_{t\to\infty} x^{[1]}(t) = -k < 0$, which implies $x^{[1]}(t) < -k/2$ for large t. Integrating the inequality $x'(t) < -r(t)^{-1/p}(k/2)^{1/p}$ on $[T,t]$, with T sufficiently large, we get

$$x(t) - x(T) < -\left(\frac{k}{2}\right)^{\frac{1}{p}} \int_T^t r^{-\frac{1}{p}}(s)ds,$$

which contradicts as $t \to \infty$ the positivity of x.

Finally, if the case i_3) holds, the change of variable

$$\tau(t) = R(t)$$

transforms (1) into

$$\frac{d}{d\tau}\big(\Phi(\dot{x})\big) = q(t(\tau))f(x(t(\tau))),$$

where $\dot{} = d/d\tau$, and $t(\tau)$ is the inverse function of $\tau(t)$. Since τ is an increasing bounded function, the problem (1)-(2) is transformed into a boundary value problem, possibly singular, on a bounded interval, and a very wide literature is devoted to this kinds of problems.

Therefore, the most interesting case for the solvability of (1)-(2) is that the function q changes its sign at least once.

Let

$$J =: \lim_{T\to\infty} \int_1^T \left(r^{-1}(t) \int_t^T q(s)\,ds\right)^{1/p} dt.$$

The main result of this paper is the following.

THEOREM 1.1. *Assume either $R(\infty) = \infty$ and $J = \infty$, or $R(\infty) < \infty$. Then the BVP (1)-(2) has a solution. Further, in the remaining case $J < \infty$ and $R(\infty) = \infty$, the BVP (1)-(2) has no solution.*

The tools used for proving Theorem 1.1 are a combination of a shooting method in a compact interval, following some ideas by Gaudenzi, Habets and Zanolin [12], a study of some topological properties of positive solutions of (1) in the half-line $[1,\infty)$, and some arguments in the phase space.

More in detail, we will consider two auxiliary BVPs, the first one on the compact interval $[0,1]$, where q is nonpositive, and the second one on the half-line $[1,\infty)$, where q is nonnegative. The existence of solutions for (1), emanating from zero, positive in the interval $(0,1)$, and satisfying additional assumptions at $t = 1$, is considered in the first problem, namely

$$\begin{cases} \big(r(t)\Phi(x')\big)' = q(t)f(x), \quad t \in [0,1], \\ x(0) = 0, \quad x(t) > 0 \text{ for } t \in (0,1), \\ \gamma x(1) + \delta x'(1) = 0, \end{cases} \tag{4}$$

where $\gamma + \delta > 0$, $\delta\gamma = 0$. The boundary conditions in (4) are a particular case of the well known Sturm-Liouville conditions. A wide literature has been devoted to the existence and the multiplicity of solutions of second order linear and nonlinear equations with Sturm-Liouville boundary conditions, see for instance [2, 15, 16] and the references therein. On the half-line $[1, \infty)$, we analyze the existence of positive decreasing solutions for (1), starting from a given positive value, and approaching zero as $t \to \infty$, namely the BVP

$$\begin{cases} \big(r(t)\Phi(x')\big)' = q(t)f(x), \quad t \in [1, \infty) \\ x(1) = x_0, \lim_{t\to\infty} x(t) = 0, \ x(t) > 0, \ x'(t) < 0. \end{cases} \tag{5}$$

The existence of a solution of (1)-(2) is obtained, roughly speaking, as the intersection of two connected sets in the space $\mathbb{R}^2$, the first set representing the final values of the solutions (x, x') of (4), and the other set representing the initial values of solutions for (5).

Our method is based on a Kneser type property, concerning solutions emanating from a continuum set of initial data; moreover, principal solutions of suitable associated half-linear equations play a crucial role for obtaining suitable upper and lower bounds.

The paper is organized as follows. In Section 2 we recall the notion of principal solutions in the half-linear case and some properties which will be used in the following. In Section 3 the BVPs (4) and (5) are solved and some additional properties of solutions are proved. The proof of Theorem 1.1 is given in Section 4. Finally, some comments and suggestions for future researches complete the paper.

2. Preliminary results

As claimed, a key role will be played by the so-called *principal solutions* of some half-linear equations associated to (1).

The notion of principal solution, introduced by Leighton and Morse for second-order linear nonoscillatory differential equations, see, e.g., [13, Ch. 11], has been extended to the half-linear equation

$$\big(r(t)\Phi(x')\big)' = q(t)\Phi(x) \quad (t \geq 1) \tag{6}$$

in [10] (see also [19, Ch. 4.15]) by using the Riccati equation approach, and reads as follows.

Definition 2.1. *A nontrivial solution z of (6) is said to be* principal solution of *(6) if for every nontrivial solution x of (6), such that $x \neq \lambda z$, $\lambda \in \mathbb{R}$, it holds*

$$\frac{z'(t)}{z(t)} < \frac{x'(t)}{x(t)} \quad \text{as } t \to \infty. \tag{7}$$

Observe that, in view of the sign assumptions on q, the equation (6) is nonoscillatory. The set of principal solutions of (6) is nonempty ([10, 19]) and for any $\mu \neq 0$ there exists a unique principal solution z such that $z(1) = \mu$, i.e. principal solutions are determined up to a constant factor.

The characteristic properties of principal solutions for (6), when q is positive for $t \geq 1$, are investigated in [4]. In particular, it is shown that, roughly speaking, principal solutions of (6) are the smallest solutions in a neighborhood of infinity. Here we summarize further properties which will be useful in the sequel. Observe that these properties continue to hold also when $q(t) \geq 0$ for $t > 1$, $q(t) \not\equiv 0$ for large t.

PROPOSITION 2.2 ([4, Theorem 3.1, Corollary 1]). *Assume either $R(\infty) = \infty$ and $J = \infty$ or $R(\infty) < \infty$. Then any principal solution z of (6) satisfies $z(t)z'(t) < 0$ on $[1, \infty)$ and $\lim_{t\to\infty} z(t) = 0$.*

A comparison between principal solutions of a suitable half-linear equation, and the solutions of (5) is needed for proving our main result, and is given in the following. The argument is similar to the one given in [3, Theorem 5].

LEMMA 2.3. *Let $c > 0$ be a fixed constant, and assume that $M > 0$ (depending on c) exists, such that*

$$f(u) \leq Mu^p \text{ on } [0, c]. \tag{8}$$

Further, assume either $R(\infty) = \infty$ and $J = \infty$, or $R(\infty) < \infty$. Let z_γ be the principal solution of the half-linear equation

$$\big(r(t)\Phi(z')\big)' = Mq(t)\Phi(z)$$

with $z_\gamma(1) = \gamma$, $0 < \gamma \leq c$. Then for any solution x of (5) with $x_0 = c$ we have

$$x(t) \geq z_\gamma(t), \qquad t \geq 1, \tag{9}$$

$$x'(1) \geq \frac{c}{\gamma}\, z'_\gamma(1). \tag{10}$$

Moreover, if $R(\infty) < \infty$, then

$$x(t) \leq \frac{c}{\rho(1)}\rho(t). \tag{11}$$

Proof. Set $g(t) = x(t) - z_\gamma(t)$. Since $g(1) \geq 0$, and, in view of Proposition 2.3, it holds $\lim_{t\to\infty} g(t) = 0$, for proving (9) it is sufficient to show that g does not have negative minima. By contradiction, let $T > 1$ be a point of negative minimum for g. Hence $g(T) < 0$, $g'(T) = 0$. Moreover, there exists $t_0 > T$

such that $g'(t_0) > 0$ and $g(t) < 0$ on $[T, t_0]$. Thus

$$r(t_0)\left(\Phi(x'(t_0)) - \Phi(z'_\gamma(t_0))\right) = \int_T^{t_0} q(s)\left(f(x(s)) - M\Phi(z_\gamma(s))\right) ds$$
$$\leq M \int_T^{t_0} q(s)\left(\Phi(x(s)) - \Phi(z_\gamma(s))\right) ds.$$

Since $g(t) < 0$ on $[T, t_0]$, we obtain $\Phi(x'(t_0)) - \Phi(z'_\gamma(t_0)) \leq 0$, which contradicts $g'(t_0) > 0$.

Now let us show that (10) holds. Consider $g_c(t) = x(t) - z_c(t)$. Using the same argument as above, since $g_c(1) = 0$, we obtain $x'(1) \geq z'_c(1)$. Since principal solutions of a half-linear equation are uniquely determined up to a constant factor, and being z_c and z_γ two principal solutions of the same half-linear equation, we have for any $t \geq 1$

$$z_c(t) = \frac{c}{\gamma}\, z_\gamma(t),$$

from which (10) follows.

Finally, considering the function

$$h(t) = x(t) - \frac{c}{\rho(1)}\rho(t),$$

the inequality (11) follows by observing that $h(1) = 0 = \lim_{t\to\infty} h(t)$ and observing that the function $c\rho(t)/\rho(1)$ is the principal solution of $(r(t)\Phi(z'))' = 0$, $z(1) = c$. □

We close this section with a result which describes a general asymptotic property of solutions for (1), depending on the behavior of the nonlinear term f in a neighborhood of zero.

LEMMA 2.4. *Assume that f satisfies*

$$\limsup_{u\to 0^+} \frac{f(u)}{\Phi(u)} < \infty. \tag{12}$$

Then any nontrivial solution x of (1), defined on $[1, \infty)$, satisfies

$$\sup_{t\in[\tau,\infty)} |x(t)| > 0 \quad \text{for any } \tau \geq 1,$$

that is, x is not eventually zero.

Proof. The assertion follows, from instance, from [19, Theorem 1.2 and Remark 1.1] with minor changes. For sake of completeness, we give here another simple alternative proof. By contradiction, let $x(t) = 0$ for $t \geq T > 1$. Since

the function $G(t) = r(t)\Phi(x'(t))x(t)$ is not decreasing and $G(T) = 0$, we have $x(t)x'(t) \leq 0$ on $[1, T]$. Without loss of generality, suppose $x(1) = x_0 > 0$. In view of (12), there exists $M > 0$ such that

$$f(u) \leq Mu^p \text{ on } [0, x_0]. \tag{13}$$

By integration of (1), taking into account (13) and that x is positive nonincreasing on $[1, T)$, we get

$$\begin{aligned} x(t) &= \int_t^T \left(\frac{1}{r(s)} \int_s^T q(\sigma) f(x(\sigma))\, d\sigma \right)^{\frac{1}{p}} ds \\ &\leq M^{\frac{1}{p}} x(t) \int_t^T \left(\frac{1}{r(s)} \int_s^T q(\sigma)\, d\sigma \right)^{\frac{1}{p}} ds, \end{aligned}$$

that is

$$1 - M^{\frac{1}{p}} \int_t^T \left(\frac{1}{r(s)} \int_s^T q(\sigma)\, d\sigma \right)^{\frac{1}{p}} ds \leq 0$$

for all $t \in [1, T)$, which is a contradiction as $t \to T$. □

REMARK 2.5. *The assumption (12) plays a crucial role in Lemma 2.4. Indeed, if the estimation (13) does not hold, then (1) can have solutions x such that $x(t) \equiv 0$ for large t, the so-called singular solutions, see, e.g., [6].*

3. Some Auxiliary Boundary Value Problems

In this section we study the existence of positive solutions for the problems (4) and (5).

The existence of solutions for (4) follows from a classical result by Wang [22], which makes use of the Krasnoselskii fixed point theorem on cone compressions or expansions. Here, by means of a change of variable, we show how it is possible to apply that result, overcoming the problems due to the lack of concavity of the positive solutions of (1), due to the presence of the coefficient r.

THEOREM 3.1. *If f satisfies (3), then the BVP (4) has at least one positive solution.*

Proof. Let

$$\tau(t) = \frac{R(t)}{R(1)}.$$

Since r is a positive continuous function on $[0, 1]$, it follows that τ is a positive C^1-function, with $\tau' > 0$ on the whole interval, and $\tau(0) = 0$, $\tau(1) = 1$. It

therefore defines a change of the independent variable $\tau = \tau(t)$. Consider the function $y(\tau) = x(t(\tau))$, where $t = t(\tau)$ is the inverse function of τ. Simple calculations show that x is a solution of (4) if and only if y is a solution of the problem

$$\begin{cases} \dfrac{d}{d\tau}\big(\Phi(\dot{y})\big) = \hat{q}(\tau) f(y), \quad \tau \in [0,1], \\ y(0) = 0, \quad y(\tau) > 0 \text{ for } \tau \in (0,1), \\ \gamma y(1) + \hat{\delta} y'(1) = 0, \end{cases} \tag{14}$$

where $\dot{} = d/d\tau$, $\hat{q}(\tau) = (R(1))^{p+1} \left(r(t(\tau))\right)^{1/p} q(t(\tau))$, and $\hat{\delta} = \delta \left(r(1)\right)^{-1/p} (R(1))^{-1}$. Clearly, $\hat{q}(\tau) \le 0$, $\hat{q}(\tau) \not\equiv 0$ in $[0,1]$, and $\gamma + \hat{\delta} > 0$, $\gamma\hat{\delta} = 0$.

Problem (14) is a particular case of the BVPs studied in [22]. The assumption

$$0 < \int_0^{1/2} \left(\int_s^{1/2} q(t)\, dt \right)^{\frac{1}{p}} ds + \int_{1/2}^1 \left(\int_{1/2}^s q(t)\, dt \right)^{\frac{1}{p}} ds < \infty,$$

which plays a key role in [22], is satisfied in our setting, since here $\hat{q}$ is continuous in $[0,1]$, and at least an interval $(\tau_1, \tau_2) \subseteq (0,1)$ exists, such that $\hat{q}(\tau) < 0$ for $\tau \in (\tau_1, \tau_2)$. Therefore Theorem 3 in [22] can be applied to (14), leading to the existence of at least a solution $\bar{y}$. Then $\bar{x}(t) = \bar{y}(\tau(t))$ is a solution of (4). □

Now, we study the properties of the solutions of the BVP on the half-line (5). The solvability of (5) is proved in the subsequent theorem, which easily follows from a well-known result of Chanturia.

THEOREM 3.2. *Assume (3)-(a). Then (5) is solvable for any $x_0 > 0$ if either $R(\infty) = \infty$ and $J = \infty$, or $R(\infty) < \infty$.*

Proof. Using [7, Theorem 1], we obtain the existence of a solution x of (1) on $[1, \infty)$ such that

$$x(1) = x_0, \ x(t) \ge 0, \ x'(t) \le 0, \tag{15}$$

for any $x_0 > 0$. The positivity of x follows from Lemma 2.4. Let us show that $\lim_{t\to\infty} x(t) = 0$. We consider separately the case $R(\infty) = \infty$ and $R(\infty) < \infty$.

Case I). Assume $R(\infty) = \infty, J = \infty$. Since $x^{[1]}$ is nondecreasing and $x^{[1]}(t) \le 0$, the limit $\lim_{t\to\infty} x^{[1]}(t)$ is finite. If $\lim_{t\to\infty} x^{[1]}(t) = x^{[1]}(\infty) < 0$, from $x^{[1]}(t) \le x^{[1]}(\infty)$ we obtain

$$x(t) \le x(1) + \Phi^* \left(x^{[1]}(\infty) \right) \int_1^t r^{-1/p}(s) ds,$$

where Φ^* is the inverse function of Φ. Letting $t \to \infty$, we get a contradiction with the positivity of x. Thus $\lim_{t\to\infty} x^{[1]}(t) = 0$. Now suppose $\lim_{t\to\infty} x(t) =$

$x(\infty) > 0$ and set $k = \min_{x(\infty)\le u\le x_0} f(u)$. Hence $k > 0$. Integrating (1) we have

$$x(t) \le x(1) - k^{1/p} \int_1^t \left(r^{-1}(s) \int_s^\infty q(\sigma) d\sigma \right)^{1/p} ds,$$

which gives again a contradiction as $t \to \infty$.

Case II). Assume $R(\infty) < \infty$. The assertion follows reasoning as in the proof of [9, Theorem 1.1], with minor changes.

Finally, let us prove that $x'(t) < 0$ on $[1, \infty)$. Assume, by contradiction, that $\bar t \ge 1$ exists, such that $x'(\bar t) = 0$. Let $G(t) = r(t)\Phi(x')x$. Since $G'(t) = q(t)f(x) + r(t)|x'|^{p+1} \ge 0$, then G is nondecreasing, with $G(\bar t) = 0$. Assuming that $G(t) = 0$ for every $t \ge \bar t$, we immediately get a contradiction, since the positivity of r yields $x' \equiv 0$ on $[\bar t, \infty)$, i.e. x is eventually constant and positive. Then $t_1 > \bar t$ exists, such that $G(t) > 0$ for every $t > t_1$. Thus, $x'(t) > 0$ for every $t > t_1$, which is again a contradiction. □

REMARK 3.3. *When $R(\infty) = \infty$, condition $J = \infty$ is necessary for the existence of solutions of the BVP (5). Indeed, if $J < \infty$, then any bounded solution x of (1) satisfies $\lim_{t\to\infty} |x(t)| = |x(\infty)| > 0$, see, e.g., [3, Th. 6] with minor changes. When $R(\infty) < \infty$ and $J < \infty$, this fact does not occur, because in this case (1) can have positive (bounded) solutions both approaching zero and a non-zero limit when t tends to infinity, as the Emden-Fowler equation*

$$(r(t)\Phi(x'))' = q(t)|x|^\beta sgn\, x, \qquad p < \beta,$$

illustrates, see, e.g. [5, Theorem 3].

REMARK 3.4. *If (3)-(a) holds and f is increasing for $u > 0$, then (5) is uniquely solvable for any $x_0 > 0$. This property is a consequence of the fact that, in this case, two positive solutions of (1) defined for $t \ge 1$, can cross at most in one point, including $t = \infty$. We refer the reader to a classical result by Mambriani (see, e.g., [21, Cap. XII, Section 5]), in which the same property is proved for a generalized Thomas-Fermi equation.*

Finally, the following "continuity" result holds for solutions of (5).

THEOREM 3.5. *Assume (3)-(a) and either $R(\infty) = \infty$ and $J = \infty$, or $R(\infty) < \infty$. Then the set*

$$S = \left\{ (x(1), x^{[1]}(1)) \right\},$$

where x is a solution of (5) for some $x_0 > 0$, contains a connected subset S_1 such that $P(S_1) = (0, \infty)$, where P is the projection $P(u, v) = u$. Moreover, if $(c_n, d_n) \in S_1$ and $\lim_n c_n = 0$, then $\lim_n d_n = 0$, and S_1 is contained in the set $\pi = \{(u, v) : u > 0, v < 0\}$.

Proof. Let $c > 0$ be fixed. In virtue of Theorem 3.2, the boundary value problem

$$\begin{cases} \big(r(t)\Phi(x')\big)' = q(t)f(x), \quad t \in [1,\infty) \\ x(1) = c - n^{-1}, \ \lim\limits_{t\to\infty} x(t) = 0, \\ x(t) > 0, \ x'(t) < 0, \end{cases} \tag{16}$$

is solvable for any positive integer n. Let $\{x_n\}$ be a solution of (16). Fixed $\gamma < c$, choose n large so that $\gamma \le c - n^{-1}$. In view of (3)-(a), the inequality (8) holds, and so, from Lemma 2.3, taking into account that x_n is nonincreasing, we obtain for $t \ge 1$

$$z_\gamma(t) \le x_n(t) \le c,$$

i.e. $\{x_n\}$ is equibounded on $C[1,\infty)$. Moreover, in view of Proposition 2.2, $z'_\gamma(1) < 0$, and again from Lemma 2.3 we have

$$x'_n(1) \ge \frac{c - n^{-1}}{\gamma}\, z'_\gamma(1) \ge \frac{c}{\gamma}\, z'_\gamma(1),$$

and so $0 \ge x_n^{[1]}(1) \ge c z_\gamma^{[1]}(1)/\gamma$, i.e. $\left\{x_n^{[1]}(1)\right\}$ is bounded on $\mathbb{R}$. Integrating (1), we get

$$x_n^{[1]}(t) = x_n^{[1]}(1) + \int_1^t q(s)f(x_n(s))ds. \tag{17}$$

Thus, since $\{x_n\}$ is equibounded and $\left\{x_n^{[1]}(1)\right\}$ is bounded in $\mathbb{R}$, also $\left\{x_n^{[1]}\right\}$ is equibounded on $C[1,\infty)$, i.e. $\{x_n\}$ is compact on $C[1,T]$ for every $T > 1$. Fixed $T > 1$, without loss of generality, suppose $\lim_n x_n(t) = x(t)$ for $t \in [0,T]$ and $\lim_n x_n^{[1]}(1) = d$. Thus, from (17) the sequence $\left\{x_n^{[1]}\right\}$ uniformly converges on $[1,T]$ and

$$\lim_n x_n^{[1]}(t) = x^{[1]}(t).$$

Hence from

$$\begin{aligned} x_n(t) &= \left(c - \frac{1}{n}\right) + \int_1^t \left(\frac{1}{a(s)}\left(x_n^{[1]}(1) + \int_1^s q(\sigma)f(x_n(\sigma))d\sigma\right)\right)^{1/p} ds = \\ &= \left(c - \frac{1}{n}\right) + \int_1^t \left(\frac{x_n^{[1]}(s)}{a(s)}\right)^{1/p} ds, \end{aligned}$$

we obtain for $t \in [1,T]$

$$x(t) = c + \int_1^t \left(\frac{x^{[1]}(s)}{a(s)}\right)^{1/p} ds,$$

that is x is solution of (1).

Now, let us prove that $\lim_{t\to\infty} x(t) = 0$. If $R(\infty) = \infty$, $J = \infty$, since x is bounded, this property can be proved using the same argument to that given in the proof of Theorem 3.2, case I). If $R(\infty) < \infty$, being x_n a solution of (16), from Lemma 2.3 we get

$$x_n(t) \le \frac{c - n^{-1}}{\rho(1)}\rho(t) \le \frac{c}{\rho(1)}\rho(t).$$

Since the sequence $\{x_n\}$ uniformly converges to x on every compact interval in $[1, \infty)$ and it is dominated by a zero-convergent function, again we have $\lim_{t\to\infty} x(t) = 0$. Clearly $x'(t) \le 0$. The argument for proving that $x'(t) < 0$ is analogous to the one in the final part of the proof of Theorem 3.2. Thus, there exists at most a solution x of (5) such that

$$\lim_n x_n^{[1]}(1) = x^{[1]}(1).$$

This means that S contains a connected subset S_1, contained in π, and, in view of the arbitrariness of c, $P(S_1) = (0, \infty)$.

Finally, let $(c_n, d_n) \in S_1$, with $c_n \to 0$, and let x_n be the solution of (5) with initial data (c_n, d_n). Then, from Lemma 2.3, we obtain $0 > x_n'(1) = d_n \ge z_{c_n}'(1) = c_n z_1'(1)$, and letting $n \to \infty$ we get the assertion. □

REMARK 3.6. *Theorem 3.5 can be view also as a "selection" theorem and extends to (5) a property of principal solutions of linear equations stated by Hartman and Wintner, see [13, Corollary 6.6]. Indeed, from the proof of Theorem 3.5, if $\{c_n\}$ is a real positive sequence converging to $c > 0$, the sequence $\{x_n\}$ of solutions of (5) starting at $x_0 = c_n$ admits a subsequence which uniformly converges, on every closed interval of $[1, \infty)$, to a solution of (5) starting at $x_0 = c$. Observe that the selection is unnecessary if (5) has a unique solution, see Remark 3.4.*

4. Proof of Theorem 1.

The following generalization of the well known Kneser's theorem, see for instance [8, Section 1.3], plays a key role in the proof of Theorem 1.1.

PROPOSITION 4.1 ([8]). *Consider the system*

$$z' = F(t, z), \quad (t, z) \in [a, b] \times \mathbb{R}^n$$

where F is continuous, and let K_0 be a continuum (i.e., compact and connected) subset of $\{(t, z) : t = a\}$ and $\mathcal{Z}(K_0)$ the family of all the solutions emanating from K_0. If any solution $z \in \mathcal{Z}(K_0)$ is defined on the interval $[a, b]$, then the cross-section $\mathcal{Z}(b; K_0) = \{z(b) : z \in \mathcal{Z}(K_0)\}$ is a continuum in $\mathbb{R}^n$.

Proof of Theorem 1.1. Consider the Cauchy problem

$$\begin{cases} \big(r(t)\Phi(x')\big)' = q(t)f(x_+), \quad t \in [0,1] \\ x(0) = 0,\ x'(0) = A > 0 \end{cases}, \tag{18}$$

where $x_+ = \max\{x, 0\}$. Clearly, every nonnegative solution of (18) is also solution of (1) in $[0,1]$. Vice versa, if x is a solution of (1), with $x(0) = 0$, and $x > 0$ in $(0,1)$, then x is also solution of (18). Indeed, since $r(t)\Phi(x')$ is nonincreasing, assuming by contradiction $x'(0) = 0$, it follows that $x'(t) \le 0$ for $t \in [0,1]$, which, together with the condition $x(0) = 0$, contradicts the positivity of x in $(0,1)$.

Now, we show that all solutions of (18) are persistent, i.e., are defined for all $t \in [0,1]$. To see this, first of all notice that all the solutions of (18) have an upper bound, since from $x^{[1]}(t) \le x^{[1]}(0)$ we get

$$x(t) \le A\ r^{\frac{1}{p}}(0)\ R(t).$$

Moreover, if x is a solution of (18) such that $x(t) > 0$ in $(0, t_1)$ and $x(t_1) = 0$, $0 < t_1 \le 1$, then $x'(t_1) < 0$. Indeed, integrating the equation in (18) over $[0, t_1]$ we obtain

$$0 = x(t_1) - x(0) = \int_0^{t_1} \left(\frac{1}{r(s)}\right)^{\frac{1}{p}} \Phi^* \left(x^{[1]}(0) + \int_0^s q(r)f(x(r))\,dr\right) ds.$$

Since $x^{[1]}$ is nonincreasing, $x^{[1]}(0) > 0$, and $q(t) \le 0$ in $[0,1]$, the quasiderivative

$$x^{[1]}(t) = x^{[1]}(0) + \int_0^t q(r)f(x(r))\,dr$$

has to assume a negative value for $s = t_1$, and so $x'(t_1) < 0$. Hence, if $t_1 < 1$, x is negative in a right neighborhood (t_1, t_2) of t_1, and satisfies $(x^{[1]}(t))' = 0$ in (t_1, t_2), i.e., $x^{[1]}(t) = x^{[1]}(t_1) < 0$, which yields $x(t) < 0$ on $(t_1, 1]$. By integration we obtain for $t > t_1$:

$$x(t) = x^{[1]}(t_1) \int_{t_1}^t \left(\frac{1}{r(s)}\right)^{\frac{1}{p}} ds,$$

that is, x is also bounded from below.

Notice that, by the above argument, we get the following property, that will be used several times in the remaining part of the proof.

(P) *If x is a solution of (18), with $x(t_0) \le 0$, $0 < t_0 \le 1$, then $x'(t_0) < 0$.*

By Theorem 3.1, equation (1) have solutions y and w, which are positive in $(0,1)$ and satisfy $y(0) = 0$, $y'(1) = 0$ and $w(0) = w(1) = 0$, respectively. Let

$A_1 = y'(0)$, $A_2 = w'(0)$. Then, from the first part of the proof, $A_1, A_2 > 0$ and y, w are also solutions of (18) for $A = A_1$ and $A = A_2$, respectively. Assume, without restriction, $A_2 < A_1$ and let

$$T = \{(x(1), x'(1)) : x \text{ sol. of (18) s.t. } x'(0) = A \in [A_2, A_1]\}$$

Since all the solutions of (18) are defined on $[0, 1]$, Proposition 4.1 assures that T is a continuum in $\mathbb{R}^2$, containing the points $(y(1), 0)$ and $(0, w'(1))$. Notice that, from property (P), it results $y(1) > 0$ and $w'(1) < 0$. Further, T does not contain any point $(0, c)$ with $c \geq 0$. It follows that a continuum $T_1 \subseteq T$ exists, such that T_1 is contained in $\overline{\pi} = \{(u, v) : u \geq 0, v \leq 0\}$, $(0, 0) \notin T_1$, and there exist $R, M > 0$ such that $(R, 0) \in T_1$, $(0, -M) \in T_1$, see Figure 1.

Now consider equation (1) for $t \geq 1$. By Theorem 3.2, for every $x_0 > 0$, there exists a positive solution x of (1) which is defined on $[1, \infty)$, satisfies $x(1) = x_0$, is decreasing and tends to zero as $t \to \infty$. Further, from Theorem 3.5, the set S of the initial values of solutions of (5), contains a connected set $S_1 \subseteq \pi = \{(u, v) : u > 0, v < 0\}$, whose projection on the first component is the half-line $(0, \infty)$. Therefore it holds

$$T_1 \cap S_1 \neq \emptyset.$$

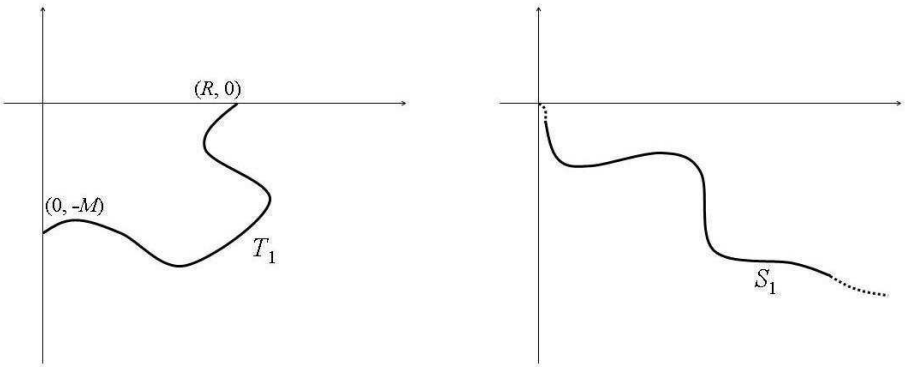


Figure 1: The connected sets T_1 and S_1.

Let us show that to any point $(c_0, c_1) \in T_1 \cap S_1$ corresponds a solution of the BVP (1)-(2). Let $(c_0, c_1) \in T_1 \cap S_1$. Then $c_0 > 0, c_1 < 0$. Moreover, there exists a solution u of (18), for a suitable $A > 0$, such that $u(1) = c_0 > 0$ and

$u'(1) = c_1 < 0$. The condition $u(1) > 0$ implies that u is positive on $(0, 1]$, because every solution of (18), which is negative at some point $T \in (0, 1)$, is negative also for $t \in [T, 1]$, see property (P). Therefore u is solution of (1) in $[0, 1]$, with $u(0) = 0$, $u(t) > 0$ for $t \in (0, 1]$. Further, as $(c_0, c_1) \in S_1$, a solution v of (5) exists, such that $v(1) = c_0$, $v'(1) = c_1$. Then v is a positive solution of (1) on $[1, \infty)$, and satisfies $\lim_{t\to\infty} v(t) = 0$. Hence the function

$$x(t) = \begin{cases} u(t), & t \in [0, 1], \\ v(t), & t > 1. \end{cases}$$

is clearly a solution of the BVP (1)-(2).

Finally, if $J < \infty$ and $R(\infty) = \infty$, the BVP (1)-(2) has no solution, since, in this case, any bounded solution of (1) has a nonzero limit at infinity, see Remark 3.3. □

5. Concluding remarks

1). If the function f satisfies

$$\lim_{u\to 0^+} \frac{f(u)}{\Phi(u)} = l > 0, \quad \lim_{u\to\infty} \frac{f(u)}{\Phi(u)} = L > 0,$$

i.e. (1) is, roughly speaking, close to an half-linear equation near zero and infinity, then all our results concerning the solvability of the second BVP (5) continue to hold. Nevertheless, the solvability of (4) is a more "delicate" problem, and the existence of positive solutions with suitable boundary conditions has been studied by different approaches. A wide literature has been devoted to this topic and we refer to [2, 15, 16] for more details.

If f is sublinear, that is

$$\lim_{u\to 0^+} \frac{f(u)}{\Phi(u)} = \infty, \quad \lim_{u\to\infty} \frac{f(u)}{\Phi(u)} = 0,$$

then the opposite situation occurs. The BVP (4) on $[0, 1]$ is now solvable, see [22], but the BVP on the half-line (5) can be not solvable, because in this case the solutions x of (1), obtained via the Chanturia result [7, Theorem 1] and satisfying on $[1, \infty)$ the boundary conditions (15), can be zero for any large t, see [6]. Moreover, under additional assumptions on r and q, the BVP (5) is solvable ([5, Theorem 2]), but not for any small $|x_0|$ and this fact makes inapplicable the crossing method used in the proof of Theorem 1.1.

2). Using an approach similar to that in the proof of Theorem 1.1, we can treat also the existence of solutions x of (1) satisfying any of the following boundary conditions

$$x(0) = 0,\ \lim_{t\to\infty} x(t) = \ell_x, 0 < \ell_x < \infty,\ x(t) > 0 \text{ for } t > 0,$$
$$x(0) = 0,\ \lim_{t\to\infty} x^{[1]}(t) = 0,\ x(t) > 0 \text{ for } t > 0,$$
$$x(0) = 0,\ \lim_{t\to\infty} x^{[1]}(t) = -d_x, 0 < d_x < \infty,\ x(t) > 0 \text{ for } t > 0.$$

In these cases, their solvability on the half-line $[1, \infty)$ requires a different approach, because for obtaining suitable upper and lower bounds, some nontrivial asymptotic properties of nonprincipal solutions of suitable associated half-linear equations are needed. This will be done in a forthcoming paper.

References

[1] R.P. Agarwal and D. O'Regan, *Infinite Interval Problems for Differential, Difference and Integral Equations*, Kluwer Academic Publishers, Dordrecht, 2001.

[2] R.P. Agarwal, D. O'Regan and P.J.Y. Wong, *Positive Solutions of Differential, Difference and Integral Equations*, Kluwer Academic Publishers, Dordrecht, 1999.

[3] M. Cecchi, Z. Došlá and M. Marini, *On nonoscillatory solutions of differential equations with p-Laplacian*, Adv. Math. Sci. Appl. **11** (2001), 419–436.

[4] M. Cecchi, Z. Došlá and M. Marini, *Principal solutions and minimal sets of quasilinear differential equations*, Dynam. Systems Appl. **13** (2004), 223-234.

[5] M. Cecchi, Z. Došlá, M. Marini and I. Vrkoč, *Integral conditions for nonoscillation of second order nonlinear differential equations*, Nonlinear Anal. **64** (2006), 278–1289.

[6] T.A. Chanturia, *On singular solutions of nonlinear systems of ordinary differential equations,* Colloq. Math. Soc. János Bolyai **15** (1976), 107–119.

[7] T.A Chanturia, *On monotonic solutions of systems of nonlinear differential equations,* Ann. Polon. Math. **37** (1980), 59–70 (Russian).

[8] W.A. Coppel, *Stability and Asymptotic Behavior of Differential Equations*, D.C. Heath and Co., Boston, 1965.

[9] Z. Došlá, M. Marini and S. Matucci, *On some boundary value problems for second order nonlinear differential equations*, Math. Bohem. **137** (2012), 113-122.

[10] A. Elbert and T. Kusano, *Principal solutions of non-oscillatory half-linear differential equations*, Adv. Math. Sci. Appl. **8** (1998), 745–759.

[11] M. García-Huidobro, R. Manásevich and C. Yarur, *On the structure of positive radial solutions to an equation containing a p-Laplacian with weight*, J. Differential Equations **223** (2006), 51–95.
[12] M. Gaudenzi, P. Habets and F. Zanolin, *An example of a superlinear problem with multiple positive solutions*, Atti Sem. Mat. Fis. Univ. Modena **51** (2003), 259–272.
[13] P. Hartman, *Ordinary Differential Equations*, 2 Ed., Birkäuser, Boston-Basel-Stuttgart, 1982.
[14] N. Kosmatov N, *Second order boundary value problems on an unbounded domain*, Nonlinear Anal. **68** (2008), 875–882.
[15] M.K. Kwong and J.S.W. Wong, *The shooting method and nonhomogeneous multipoint BVPs of second-order ODE*, Bound. Value Probl. **2007** (2007), Art. ID 64012, 16 pp.
[16] K. Lan and J.R.L. Webb, *Positive solutions of semilinear differential equations with singularities*, J. Differential Equations **148**, (1998), 407–421.
[17] H. Lian, H. Pang and W. Ge, *Triple positive solutions for boundary value problems on infinite interval*, Nonlinear Anal. **67** (2007), 2199–2207.
[18] B. Liu, L. Liu and Y. Wu, *Unbounded solutions for three-point boundary value problems with nonlinear boundary conditions on* $[0, \infty)$, Nonlinear Anal. **73** (2010), 2923–2932.
[19] J.D. Mirzov, *Asymptotic Properties of Solutions of the Systems of Nonlinear Nonautonomous Ordinary Differential Equations*, (Russian), Maikop, Adygeja Publ. 1993. English translation: Folia Fac. Sci. Nat. Univ. Masaryk Brun. Mathematica **14**, Masaryk University Brno, 2004.
[20] I. Rachůnková and J. Tomeček, *Superlinear singular problems on the half line*, Bound. Value Probl. **2010**, Art. ID 429813, 18 pp.
[21] G. Sansone, *Equazioni Differenziali nel Campo Reale*, Zanichelli, Bologna, 1941.
[22] J. Wang, *The existence of positive solutions for the one-dimensional p-Laplacian*, Proc. Amer. Math. Soc. **125** (1997), 2275–2283.

Authors' addresses:

Mauro Marini
Department of Mathematics and Informatics "U. Dini"
University of Florence
E-mail: mauro.marini@unifi.it

Serena Matucci
Department of Mathematics and Informatics "U. Dini"
University of Florence
E-mail: serena.matucci@unifi.it

Received March 22, 2012
Revised June 21, 2012

Rend. Istit. Mat. Univ. Trieste
Volume 44 (2012), 133–151

On the asymptotic behaviour of the characteristics in the codiffusion of radioactive isotopes with general initial data

Elena Comparini and Maura Ughi

Dedicated to Fabio Zanolin on the occasion of his 60th birthday

Abstract. *The large-time behaviour of the solution of a hyperbolic-parabolic problem in an isolated domain, which models the diffusion of n species of radiative isotopes of the same element, is studied, assuming general hypotheses on the initial data.*
Depending on the radiative law and on the distribution of the initial concentration, either a uniform distribution for the concentration of each isotope or the presence of oscillations may be possible when $t \to \infty$.

Keywords: isotopes, diffusion, hyperbolic equations
MS Classification 2010: 35L50, 35K57, 35B05

1. Introduction

Let us consider the following problem in $\Omega = (-L, L)$:

$$\begin{cases} c_{it} = \left(\dfrac{c_i}{c} c_x\right)_x + \displaystyle\sum_{j=1}^{n} \Lambda_{ij} c_j, & x \in \Omega,\ t > 0, \\ c_i(x,0) = c_{i0}(x) \geq 0, & x \in \Omega, \\ c_i \dfrac{c_x}{c}(-L,t) = c_i \dfrac{c_x}{c}(L,t) = 0, & t > 0, \\ i = 1, ..., n,\ \ c = \sum_{k=1}^{n} c_k. \end{cases} \tag{1}$$

The problem comes from a model for the diffusion of n species of isotopes of the same element in a medium, in the assumption that the flux of the $i-th$ species, whose concentration is c_i, is

$$J_i = -\frac{c_i}{c} c_x,\ i = 1, ..., n,\ x \in \Omega,$$

where $c = \sum_{i=1}^{n} c_i$ is the total concentration.
This assumption means that any component varies with the total gradient of the element in a relative percentage $\frac{c_i}{c}$ (see [7, 20]).

Actually the above law for the flux is an approximation of a more complete model where the flux is $J_i = -(\tilde{D}_i c_{ix} + D_i \frac{c_i}{c} c_x)$. If one assumes $D_i = 0$ then the problem becomes a classical parabolic problem whose solution does not quite agree with the experimental data (see [20]). On the other hand there are physical situations, such as self-diffusion, in which it is sensible to try the model with $\tilde{D}_i = 0$, thus obtaining solutions more in agreement with experimental data, at least qualitatively.

Moreover, it would be reasonable, for solutes, that the coefficients D_i are practically the same for all isotopic molecules of the element, as they have the same partial molar volume and the same electronic configuration, especially for the heavier chemical elements. Although it would be interesting from a mathematical point of view to study the model in the general hypothesis that the diffusion coefficients are different (see [7]), numerical simulations evidentiate no significant difference in the qualitative behaviour of the solution in dependence on the diffusion coefficients D_i, here assumed to be all equal to 1 after rescaling (see [6]). For more details on the physical motivations of the model see [5].

The coefficients Λ_{ij} are the elements of a constant $n \times n$ matrix Λ which expresses the "radiative decay law" in the case of radiative isotopes. In the physically relevant hypothesis that $\mathbf{C} = (c_1, ..., c_n)$ is regular and satisfies

$$c_{i0}(x) \geq 0, \qquad c_0 = \sum_{i=1}^{n} c_{i0}(x) > 0, \tag{2}$$

there exists a unique classical non negative solution (see Section 2 for the precise assumptions, [7] for the complete model and [5] in the present case). We remark that it has been proved that the total concentration c satisfies a parabolic equation with data $c_x(\pm L) = 0$ and it is regular and strictly positive for any $t \geq 0$. Once c is given, the concentrations c_i for the single isotopes are solutions of linear hyperbolic first order equations and they can be derived by means of the method of the characteristics, defined by the total concentration. In this case, denoted by $X(t; x_0)$ the characteristic starting in x_0 at time 0, we have:

$$\frac{dX(t;x_0)}{dt} = -\frac{c_x}{c}\bigg|_{x=X(t;x_0)}, \qquad X(0;x_0) = x_0. \tag{3}$$

Let us remark that if the initial total concentration $c_0(x)$ has zeroes, there can be effects of "loss of regularity". Actually it can happen that, also if the data are regular, c_i has discontinuities for positive time. Although from a physical point of view it is more sensible to consider c_0 small rather than

$c_0 \equiv 0$, a mathematical approach to the hyperbolic problem was performed in [5], defining, also if the data are regular, a weak solution as in [2]. Let us stress that, since the total concentration satisfies a uniform parabolic equation, it will be strictly positive for any positive time also if it is initially zero on subintervals. The problem is that this initial "holes" may possibly cause the c_i to be discontinuous for positive time (for details see [5]). Since we wish to understand first the asymptotic behaviour for physically relevant initial data, possibly strongly oscillating but smooth, we need to assume $c_0 > 0$. In this assumption, one can use the results of [4] and show that the solution constructed along the characteristics is the "viscosity solution" obtained as the limit of the complete physical model, with $\tilde{D}_i = \tilde{D} \neq 0,\ D_i = D = 1$ as $\tilde{D} \to 0$. Numerical simulations confirm this result, also for the complete physical model, in very general situations, and they have been performed using a program for solving parabolic equations, with initial data possibly zero ([6]); however the proof of existence and uniqueness of the solution of the complete parabolic problem and its convergence to the hyperbolic problem in the possible presence of zeroes in the initial total concentration is still an open problem.

We remark that the asymptotic behaviour of the solutions for $t \to \infty$ strongly depends on the decay law, that is on Λ, and on the first significative term of the asymptotic expansion for $t \to \infty$ of the solutions of the ODE

$$\begin{cases} \dot{\mathbf{C}} = \Lambda \mathbf{C}, \qquad \mathbf{C} = (c_1, ..., c_n), \\ \mathbf{C}(0) = \mathbf{C}_0. \end{cases} \tag{4}$$

These results are evidentiated in [8], under strong assumptions on the positivity of the initial data in the whole Ω. However there are physically relevant initial data that do not satisfy such assumptions in the whole Ω but still the corresponding solution should have a similar asymptotic behaviour. In the present paper we will study the problem assuming the most general hypotheses.

2. Statement of the problem

Existence and uniqueness of a classical non negative solution of Problem (1) have been obtained in [5] under the following assumptions:

H1) $c_{i0} \in H^{2+l}(\overline{\Omega}),\ l > 0, \quad i = 1, ..., n, \quad 0 \leq c_{i0} \leq K,\ c_0 = \sum_{i=1}^{n} c_{i0} > 0,$

H2) positivity property for the ODE (4):

if $c_{i0} \geq 0$, then $c_i(t) \geq 0,\ \ i = 1, ..., n,$

Since we want to consider a set of isotopes which either decay or are stable, it is natural to assume that all the eigenvalues of the matrix Λ are real non positive, actually we can assume:

H3) all the eigenvalues of Λ are real.

Due to the structure of Problem (1) it is convenient to consider instead of $\mathbf{C}$, $\tilde{\mathbf{C}} = (c_1, ...c_{n-1}, c)$, $c = \sum_{i=1}^n c_i$, then (4) is transformed in the following:

$$\begin{cases} \dot{\tilde{\mathbf{C}}} = \tilde{\mathbf{\Lambda}}\tilde{\mathbf{C}}, \\ \tilde{\mathbf{C}}(0) = \tilde{\mathbf{C}}_0, \quad \tilde{\mathbf{C}}_0 = (c_{10}, ..., c_{(n-1)0}, c_0). \end{cases} \tag{5}$$

where $\tilde{\Lambda}$, for which **H3** holds too, is given by

$$\tilde{\Lambda} = \begin{pmatrix} \Lambda_{11} - \Lambda_{1n} & \dots & \Lambda_{1n} \\ \Lambda_{21} - \Lambda_{2n} & \dots & \Lambda_{2n} \\ \vdots & \ddots & \vdots \\ \sum_{m=1}^n (\Lambda_{m1} - \Lambda_{mn}) & \dots & \sum_{m=1}^n \Lambda_{mn} \end{pmatrix}.$$

Assuming that $\tilde{\Lambda}$ has $s \leq n$ distinct eigenvalues $\lambda_s < ... < \lambda_1$, for $i = 1, ..., s$, let us denote by (see [1, 12])

$\mu(\lambda_i)$ = algebraic multiplicity of λ_i,

$\nu(\lambda_i)$ = geometric multiplicity of λ_i,

$E(\lambda_i)$ = generalized autospace of λ_i,

$h(\lambda_i)$ = the least integer k s.t. $\mathrm{Ker}\,(\tilde{\Lambda} - \lambda_i I)^{k+1} = \mathrm{Ker}\,(\tilde{\Lambda} - \lambda_i I)^k$,

so that $E(\lambda_i) = \mathrm{Ker}\,(\tilde{\Lambda} - \lambda_i I)^{h(\lambda_i)}$, with $I = Id$ matrix $n \times n$.

Any solution is a linear combination of the product of exponential functions time polynomials. Quite precisely:

$$\tilde{\mathbf{C}}(t) = \sum_{i=1}^{s} \left[\sum_{k=0}^{h(\lambda_i)-1} (\tilde{\Lambda} - \lambda_i I)^k \frac{t^k}{k!} \right] e^{\lambda_i t} \tilde{\mathbf{C}}_{0,i}, \tag{6}$$

with $\tilde{\mathbf{C}}_0 = \sum_{i=1}^s \tilde{\mathbf{C}}_{0,i}$, $\tilde{\mathbf{C}}_{0,i} \in E(\lambda_i)$.

Therefore, since λ_1 is the highest eigenvalue, we have:

$$\begin{aligned} &\lim_{t \to +\infty} t^{-(h(\lambda_1)-1)} e^{-\lambda_1 t} \tilde{\mathbf{C}}(t; \tilde{\mathbf{C}}_0) \\ &\quad = \frac{1}{(h(\lambda_1)-1)!} (\tilde{\Lambda} - \lambda_1 I)^{h(\lambda_1)-1} \tilde{\mathbf{C}}_{0,1} = \hat{B}\tilde{\mathbf{C}}_0. \end{aligned} \tag{7}$$

Here $\hat{B}$ is a constant $n \times n$ matrix, determined by the $E(\lambda_i)$ (see [8]).

Given $\tilde{\mathbf{C}}_0(x),\ x \in \overline{\Omega}$, let us define:

$$\mathbf{F}(x) = \hat{B}\tilde{\mathbf{C}}_0(x), \quad F(x) = (\hat{B}\tilde{\mathbf{C}}_0(x))_n. \tag{8}$$

Let us remark that the positivity hypothesis **H2** together with **H1** guarantees $F(x) \geq 0$, moreover, if for some x_0 $F(x_0) = 0$, then $\mathbf{F}(x_0) = \mathbf{0}$.

We proved in [8, Theorem 3.1], that, assuming **H1**, **H2**, **H3**, for any initial datum $\tilde{\mathbf{C}}_0$ such that

H4) $F(x) \geq \delta > 0$ in $\overline{\Omega}$,

we have

$$\lim_{t\to+\infty} t^{-(h(\lambda_1)-1)} e^{-\lambda_1 t} m(x,t) = \frac{x+L}{2L} M_\infty, \tag{9}$$

uniformly in Ω, where

$$m(x,t) = \int_{-L}^{x} c(\xi,t)\, d\xi, \quad M_\infty = \int_{-L}^{L} F(\xi)\, d\xi. \tag{10}$$

Then the first asymptotic term for the total concentration c is given by $t^{h(\lambda_1)-1} e^{\lambda_1 t} \dfrac{M_\infty}{2L}$, that is a uniform distribution of the total concentration, and this is in agreement with the physics of the problem.

Moreover, once the characteristics have been defined as in (3), it is possible to get their asymptotic behaviour, and precisely (see [8, Corollary 3.1]):

$$\lim_{t\to+\infty} X(t;x_0) = X_\infty(x_0) = \frac{2L}{M_\infty} \int_{-L}^{x_0} F(\xi)\, d\xi - L. \tag{11}$$

The hypothesis **H4** ensures that the function $X_\infty(x_0)$ is monotone increasing, and consequently it is possible to obtain the information on the ratioes $r_i = \dfrac{c_i}{c},\ i = 1, \ldots n-1,\ \dfrac{c_n}{c} = 1 - \displaystyle\sum_{i=1}^{n-1} r_i$, precisely:

$$\lim_{t\to+\infty} r_i(x,t) = \frac{F_i(X_\infty^{-1}(x))}{F(X_\infty^{-1}(x))}, \quad i = 1, \ldots, n-1 \tag{12}$$

uniformly in Ω (see [8, Corollary 3.2]).

Of course, if $M_\infty = 0$, that is $F \equiv 0$, the first significative term of the asymptotic expansion of m and c changes, but it is natural to investigate what happens if $F \not\equiv 0$ but e.g. it is null in a subset of Ω.

In order to better understand the question, let us consider the couple of isotopes (U^{238}, U^{234}) whose decay law is:

$$\begin{cases} \dot{c}_1 = -\gamma_1 c_1 \\ \dot{c}_2 = \gamma_1 c_1 - \gamma_2 c_2, \qquad 0 < \gamma_1 < \gamma_2, \end{cases} \tag{13}$$

that is the isotope 1, U^{238}, decays into the isotope 2, U^{234}, and the second one decays out of the element. In this example one can see that $F(x) = \dfrac{\gamma_2 - \gamma_1}{\gamma_2} c_{10}(x)$. If the isotope 1 is not present initially (i.e. $c_{10} \equiv 0$), then the solution is $c_1 \equiv 0$ and $c_2 \equiv c = e^{-\gamma_2 t} w(x,t)$, with $w(x,t)$ solution of

$$\begin{cases} w_t = w_{xx}(x), \quad x \in \Omega, \ t > 0, \\ w(x,0) = c_0(x), \quad x \in \Omega, \\ w_x(\pm L, t) = 0, \quad t > 0, \end{cases}$$

that is, for large time,

$$m(x,t) \simeq e^{-\gamma_2 t} \frac{x+L}{2L} \int_{-L}^{L} c_0(\xi)\, d\xi, \quad \text{and } r \equiv 0.$$

If on the contrary assumption **H4** holds, that is the isotope 1 is initially present everywhere in Ω, then from (9)-(12):

$$\begin{cases} m(x,t) \simeq e^{-\gamma_1 t} \dfrac{x+L}{2L} \left(1 - \dfrac{\gamma_1}{\gamma_2}\right) \displaystyle\int_{-L}^{L} c_{10}(\xi)\, d\xi, \\ r(x,t) \simeq r_E = 1 - \dfrac{\gamma_1}{\gamma_2}, \end{cases}$$

uniformly in Ω, with $0 < r_E < 1$. We have in this case the so called "secular equilibrium" of the two isotopes, that are both present in Ω for $t > 0$ and tend, for $t \to \infty$, respectively to r_E, $1 - r_E$. The question is what happens if the isotope 1 is absent only in a subset of Ω but $M_\infty > 0$. We will prove in the sequel that the asymptotic behaviour of m is still given by (9).

Other significant examples will be analyzed in Section 4.

3. Main result

Aim of this Section is to prove that the same result (9) holds if instead of **H4** we assume the following hypothesis:

H5) $\quad F(x) = (\hat{B}\tilde{\mathbf{C}}_0(x))_n \geq 0, \qquad F(x) \not\equiv 0 \text{ in } \overline{\Omega}.$

We have the following:

THEOREM 3.1. *In the assumptions* **H1**, **H2**, **H3**, **H5**, *then*

$$\lim_{t \to +\infty} t^{-(h(\lambda_1)-1)} e^{-\lambda_1 t} m(x,t) = \frac{x+L}{2L} M_\infty, \tag{14}$$

uniformly in $\overline{\Omega}$, *with* m *and* M_∞ *defined in (10).*

Proof. Taking as an unknown $\tilde{\mathbf{C}} = (c_1, ..., c_{n-1}, c)$, $c = \sum_{i=1}^{n} c_i$, the original problem (1) becomes:

$$\begin{cases} c_{it} = \left(\dfrac{c_i}{c} c_x\right)_x + (\tilde{\Lambda}\tilde{\mathbf{C}})_i,\ i = 1, ..., n-1, \quad x \in \Omega,\ t > 0, \\ c_t = c_{xx} + (\tilde{\Lambda}\tilde{\mathbf{C}})_n, \quad x \in \Omega,\ t > 0, \\ c_x(-L, t) = c_x(L, t) = 0, \quad t > 0, \\ \tilde{\mathbf{C}}(x, 0) = \tilde{\mathbf{C}}_0(x) = (c_{10}(x), ..., c_{(n-1)0}(x), c_0(x)), \\ c_0(x) = \sum_{i=1}^{n} c_{i0}(x), \quad x \in \Omega. \end{cases} \tag{15}$$

As in other problems of this kind, see [2, 5, 13, 14, 18], it is more convenient to consider, instead of (15), the problem for
$r_i = \dfrac{c_i}{c},\ i = 1, ..., n-1$:

$$\begin{cases} r_{it} = \dfrac{c_x}{c} r_{ix} + P_i(\mathbf{r}),\ i = 1, ..., n-1, \quad x \in \Omega,\ t > 0, \\ c_t = c_{xx} + b(r_1, ..., r_{n-1})c, \quad x \in \Omega,\ t > 0, \\ c_x(-L, t) = c_x(L, t) = 0, \qquad t > 0, \\ c(x, 0) = c_0(x), \qquad x \in \Omega, \\ r_i(x, 0) = \dfrac{c_{i0}(x)}{c_0(x)}, \quad i = 1, ..., n-1, \qquad x \in \Omega, \end{cases} \tag{16}$$

where P_i are polynomial expressions of degree ≤ 2 in $\mathbf{r} = (r_1, ..., r_{n-1})$, the coefficients depending on Λ, and b is defined by

$$b = (\tilde{\Lambda}\tilde{\mathbf{r}})_n, \quad \tilde{\mathbf{r}} = (r_1, ..., r_{n-1}, 1). \tag{17}$$

Let us remark that under hypotheses **H1** and **H2** we have proved in [5] the existence of a unique classical solution of problem (16).

Moreover, $c(x, t)$ is always positive, satisfying a linear parabolic equation with zero flux on the boundary and positive initial datum.
Once c is known, the characteristics depend only on c, see (3), but the r_i evolve along each characteristic, independently of c, like the solutions of the spatially omogeneous problem. Then, fixed x_0 and $\tilde{\mathbf{C}}_0(x_0)$, the r_i are given explicitely on the characteristic $X(t; x_0)$ by the ratioes c_i/c, with c_i, c given in (6) with initial datum $\tilde{\mathbf{C}}_0(x_0)$.

Moreover, we proved in [5] that the "masses" between two characteristics $X(t; x_1)$, $X(t; x_2)$ starting respectively in x_1, x_2, with $-L \leq x_1 < x_2 \leq L$, defined by

$$\tilde{\mathbf{M}}(t) = \int_{X(t;x_1)}^{X(t;x_2)} \tilde{\mathbf{C}}(\xi, t)\, d\xi, \tag{18}$$

are solutions of the ODE system:

$$\dot{\tilde{\mathbf{M}}} = \tilde{\mathbf{\Lambda}}\tilde{\mathbf{M}}, \qquad \tilde{\mathbf{M}}(\mathbf{0}) = \int_{\mathbf{x_1}}^{\mathbf{x_2}} \tilde{\mathbf{C}}_{\mathbf{0}}(\xi)\, \mathbf{d}\xi = \tilde{\mathbf{M}}_{\mathbf{0}}, \tag{19}$$

and hence are given explicitly by (6) with initial datum $\tilde{\mathbf{M}}_0$ instead of $\tilde{\mathbf{C}}_0$.

This means that, since $x = -L$ is the characteristic starting in $x_0 = -L$, we know the evolution in time of $m(x,t)$ on any characteristic $x = X(t;x_0)$ and in particular for $x = X(t;L) \equiv L$.

Then $v(x,t)$, defined by

$$v(x,t) = (1+t)^{-(h(\lambda_1)-1)} e^{-\lambda_1 t} m(x,t), \tag{20}$$

is solution of:

$$\begin{cases} v_t = v_{xx} + f(x,t), & x \in \Omega,\ t > 0, \\ v(x,0) = \int_{-L}^{x} c_0(\xi)\, d\xi, & x \in \Omega, \\ v(-L,t) = 0, & t > 0, \\ v(L,t) = H(t), & t > 0. \end{cases} \tag{21}$$

with

$$\begin{aligned} &f(x,t) = \textstyle\int_{-L}^{x} \tilde{b}\, u\, d\xi, \\ &u = (1+t)^{-(h(\lambda_1)-1)} e^{-\lambda_1 t} c, \\ &\tilde{b} = b - \lambda_1 - \frac{h(\lambda_1)-1}{1+t}, \qquad b = (\tilde{\Lambda}\tilde{\mathbf{r}})_n, \\ &H(t) = (1+t)^{-(h(\lambda_1)-1)} e^{-\lambda_1 t} \times \\ &\quad \times \textstyle\sum_{i=1}^{s} \left\{ \left[\sum_{k=0}^{h(\lambda_i)-1} (\tilde{\Lambda} - \lambda_i I)^k \frac{t^k}{k!} \right] e^{\lambda_i t} \int_{-L}^{L} \tilde{\mathbf{C}}_{0,i}(\xi)\, d\xi \right\}_n . \end{aligned} \tag{22}$$

The expression of $H(t)$ comes from (19), recalling that $x \equiv \pm L$ are the characteristics starting at $x_0 = \pm L$, since there $c_x = 0$. Then, see (21), v is solution of a Dirichlet problem for the heat equation with source $f(x,t)$ and known boundary data.

Under the hypothesis **H5**, from (7) and the definition (8) of F, we have

$$\lim_{t \to +\infty} H(t) = \int_{-L}^{L} F(\xi)\, d\xi = M_\infty. \tag{23}$$

Using a classical result ([11, Theorem 1, Chapter V]) the proof of Theorem 3.1 follows, provided that

$$\lim_{t \to +\infty} f(x,t) = 0 \tag{24}$$

uniformly in Ω.
In order to prove (24), fixed an arbitrary $\sigma > 0$, let us divide the interval $\Omega = (-L, L)$ into the two subsets:

$$\begin{aligned} \Omega_- &= \{x \in \Omega : F(x) < \sigma\}, \\ \Omega_+ &= \{x \in \Omega : F(x) \geq \sigma\}. \end{aligned} \tag{25}$$

Let us remark that, for any σ sufficiently small, Ω_+ is not empty and, if $F(x_0) = 0$, there exists a neighborhood of x_0 where $F < \sigma$ and Ω_- is not empty.
For any fixed $t > 0$, let us divide Ω into

$$\begin{aligned} \Omega_-(t) &= \{x \in \Omega : x = X(t; x_0),\ x_0 \in \Omega_-\}, \\ \Omega_+(t) &= \{x \in \Omega : x = X(t; x_0),\ x_0 \in \Omega_+\}, \end{aligned} \tag{26}$$

that is $\Omega_-(t)$, $\Omega_+(t)$ are the set of the characteristics at time t starting from Ω_-, Ω_+ respectively.

Then

$$\begin{aligned} f(x,t) &= \textstyle\int_{[-L,x]\cap\Omega_-(t)} \tilde{b}u\, d\xi + \int_{[-L,x]\cap\Omega_+(t)} \tilde{b}u\, d\xi = \\ &= f_-(x,t) + f_+(x,t). \end{aligned} \tag{27}$$

Let us consider first f_+. In [8, Lemma 3.1], we proved that if for some x_0 $F(x_0) \geq \sigma > 0$, on the characteristic $X(t; x_0)$ starting in x_0, the following estimate on $\tilde{b}$ depending on σ holds:

$$|\tilde{b}| \leq \frac{k_1}{\sigma}\left[\frac{h(\lambda_1) - 1}{t^2} + (s-1)e^{\frac{\lambda_2 - \lambda_1}{2}t}\right] = \frac{k_1}{\sigma}g(t), \tag{28}$$

for $x = X(t; x_0)$ and $t \geq 1$, where k_1 is a constant depending on Λ and on $\max_\Omega ||\tilde{\mathbf{C}}_0(x)||$.

Then, being $u > 0$, recalling (21)-(23), we have:

$$\begin{aligned} |f_+(x,t)| &\leq \frac{k_1}{\sigma}g(t)\int_{[-L,x]\cap\Omega_+(t)} u(\xi,t)\, d\xi \\ &\leq \frac{k_1}{\sigma}g(t)\int_{-L}^{L} u(\xi,t)\, d\xi \\ &\leq \frac{k_1}{\sigma}g(t)H(t) \leq 2\frac{k_1 M_\infty}{\sigma}g(t), \quad t \geq T_1. \end{aligned} \tag{29}$$

Let us consider now f_-. Notice that for any $x \in \overline{\Omega}$, $t > 0$, $\tilde{b}$ is uniformly bounded because the r_i are bounded between 0 and 1 (see (22)), that is $|\tilde{b}| \leq k_2$. Since $u > 0$ we have:

$$\begin{aligned} |f_-(x,t)| &\leq k_2 \textstyle\int_{[-L,x]\cap\Omega_-(t)} u(\xi,t)\, d\xi \\ &\leq k_2 \textstyle\int_{\Omega_-(t)} u(\xi,t)\, d\xi. \end{aligned} \tag{30}$$

From (18), (19), (6), (8), the last term in (30) can be written in the form

$$\begin{aligned}\textstyle\int_{\Omega_-(t)} u(\xi,t)\,d\xi &= (1+t)^{-(h(\lambda_1)-1)}e^{-\lambda_1 t}\times\\ \times\textstyle\sum_{i=1}^{s}&\left\{\left[\sum_{k=0}^{h(\lambda_i)-1}(\tilde\Lambda-\lambda_i I)^k\frac{t^k}{k!}\right]e^{\lambda_i t}\int_{\Omega_-}\tilde{\mathbf{C}}_{0,i}(\xi)\,d\xi\right\}_n\\ &=\left(\hat B\textstyle\int_{\Omega_-}\tilde{\mathbf{C}}_{0,1}(\xi)\,d\xi\right)_n+\tilde z=\textstyle\int_{\Omega_-}F(\xi)\,d\xi+\tilde z,\end{aligned}\tag{31}$$

with $\tilde z$ bounded for any $x\in\overline{\Omega},\ t\geq 1$ by:

$$|\tilde z|\leq k_3\left(\frac{(h(\lambda_1)-1)}{t}+(s-1)e^{\frac{\lambda_2-\lambda_1}{2}t}\right)=k_3 g_1(t),\tag{32}$$

with k_3 depending on Λ and on $\max_\Omega||\tilde{\mathbf{C}}_0||$.

Recalling that $F<\sigma$ in Ω_-, from (31), (32) it follows

$$|f_-|\leq k_4(\sigma+g_1(t)),\ x\in\overline{\Omega},\ t\geq 1.\tag{33}$$

From the estimates (29), (33) on f_+, f_- we have, for any $x\in\overline{\Omega},\ t\geq\max(1,T_1)$:

$$|f|\leq k_5\left(\sigma+\frac{g(t)}{\sigma}+g_1(t)\right).\tag{34}$$

Then, fixed an arbitrary $\epsilon>0$, e.g. $\sigma=\dfrac{\epsilon}{3}$, recalling that $g(t)$ and $g_1(t)$ tend to zero as $t\to\infty$, from (34) we have that there exists a time $T(\epsilon)$ such that

$$|f|\leq\epsilon,\ \forall x\in\overline{\Omega},\ t>T(\epsilon),$$

that gives the proof of the theorem. □

From Theorem 3.1, as in [8], it is possible to obtain the asymptotic behaviour of the characteristics, precisely we have:

COROLLARY 3.2. *In the hypotheses of Theorem 3.1 we have that*

$$\lim_{t\to+\infty}X(t;x_0)=X_\infty(x_0)=\frac{2L}{M_\infty}\int_{-L}^{x_0}F(\xi)\,d\xi-L,\tag{35}$$

uniformly in $\overline{\Omega}$.

Proof. The proof is the same as the one of [8, Corollary 3.1], let us mention here that the idea of the proof is that we know the evolution in time of $m(X(t;x_0),t)$,

since m is solution of the ODE (19). Therefore we have that, by the definition of $X_\infty(x_0)$ in (35) and by (6)-(8):

$$t^{-(h(\lambda_1)-1)}e^{-\lambda_1 t}m(X(t;x_0),t) = \int_{-L}^{x_0} F(\xi)\,d\xi + \hat{z}(x_0,t) = \frac{X_\infty(x_0)+L}{2L}M_\infty + \hat{z},$$

where

$$|\hat{z}(x_0,t)| \le k_6 g_1(t),$$

with k_6 constant depending on Λ and on $\max_\Omega ||\tilde{\mathbf{C}}_0||$, and $g_1(t)$ defined in (32).

On the other hand, Theorem 3.1 implies that, for t sufficiently large and for any x_0 in $\overline{\Omega}$, $t^{-(h(\lambda_1)-1)}e^{-\lambda_1 t}m$ on the characteristic $X(t;x_0)$ is close to $\frac{X(t;x_0)+L}{2L}M_\infty$. □

Concerning the asymptotic behaviour of the $r_i = \dfrac{c_i}{c}$, $i = 1, ..., n-1$, as in [8] we have:

COROLLARY 3.3. *In the hypotheses of Theorem 3.1, and assuming that $F(x) \ge \delta > 0$ in $[x_1, x_2], \subset \overline{\Omega}$, we have:*

$$\lim_{t\to+\infty} r_i(x,t) = \frac{F_i(X_\infty^{-1}(x))}{F(X_\infty^{-1}(x))}, \tag{36}$$

uniformly in $[X_\infty(x_1), X_\infty(x_2)]$, and

$$\left| r_i\left(X(t; X_\infty^{-1}(x)), t\right) - \frac{F_i(X_\infty^{-1}(x))}{F(X_\infty^{-1}(x))} \right| \le k(\delta) g_1(t), \tag{37}$$

for $t > T(\delta) = g_1^{-1}\left(\dfrac{\delta}{2}\right)$, g_1 defined in (32).

Proof. From the hypothesis $F(x) \ge \delta > 0$, $x \in [x_1, x_2]$, it follows that the function $X_\infty(x)$ is monotone increasing in $[x_1, x_2]$, consequently the inverse function is monotone increasing in $[X_\infty(x_1), X_\infty(x_2)]$.

Moreover the characteristics are ordered so that $\forall \bar{t} > 0$ and $\forall \bar{x} \in [X(t;x_1), X(t;x_2)]$ there exists a unique $\hat{x} \in [x_1, x_2]$ such that $\bar{x} = X(t;\hat{x})$ and $F(\hat{x}) \ge \delta > 0$. Then we can repeat the arguments of [8, Corollary 3.2]. The estimate (37) on r_i comes from the explicit expression of $\tilde{\mathbf{C}}(t)$ in (6). □

From the explicit expression of $X_\infty(x)$, see (35), we have the following

REMARK 3.4. **i)** *If $F(x) \equiv 0$ for $x \in [x_1, x_2] \subset \Omega$, then $X_\infty(x_1) = X_\infty(x_2)$. That is, if F is identically zero in a subinterval of Ω, all the subinterval asymptotically reduces to the point*

$$X^* = X_\infty(x_1) = \frac{2L}{M_\infty} \int_{-L}^{x_1} F(\xi) - L.$$

ii) *If $0 \le F(x) \le \beta,\ \beta > 0$ for $x \in [x_1, x_2] \subset \Omega$, then*

$$X_\infty(x_2) - X_\infty(x_1) = \frac{2L}{M_\infty} \int_{x_1}^{x_2} F(\xi)\, d\xi \le \frac{2L}{M_\infty}(x_2 - x_1)\beta.$$

That is the asymptotic measure of the subinterval is of the order β.

In the next Section we will consider some examples in order to make clearer the above observations concerning the asymptotic behaviour of $\mathbf{r} = (r_1, ..., r_{n-1})$.

4. Examples and comments

Let us consider the example described in Section 2, for the couple (U^{238}, U^{234}), where the matrix Λ is given by (13). If we assume in this example that $F(x) = \dfrac{\gamma_2 - \gamma_1}{\gamma_2} c_{10}(x)$ is null in a subinterval $[x_1, x_2] \subset \overline{\Omega}$ and positive out of this interval, then (see Remark 3.4), the whole interval $[x_1, x_2]$ reduces, for $t \to \infty$ to the unique point

$$X^* = X_\infty(x_1) = \frac{2L}{M_\infty} \int_{-L}^{x_1} F(\xi) - L.$$

In this case there does not exist the $\lim_{x \to X^*,\, t \to \infty} r(x,t)$, because in any neighborhood of X^* there are characteristics on which $r \equiv 0$ (precisely $X(t; x_0)$, $\forall x_0 \in [x_1, x_2]$) and characteristics on which

$$r \to r_E = \frac{\gamma_2 - \gamma_1}{\gamma_2},\ 0 < r_E < 1,$$

precisely the ones starting at a point out of $[x_1, x_2]$.

However, fixed a neighborhood of X^*, out of it r tends uniformly to r_E for $t \to \infty$, because of Corollary 3.3. From a physical point of view in this case $(0 < \gamma_1 < \gamma_2)$ there is not a uniform asymptotic distribution for c_1, c_2 and, in particular, oscillations may be present near X^* also asymptotically. However varying order of the parameters γ_1, γ_2 one can observe that:

i) if $\gamma_1 > \gamma_2 > 0$ then $F(x) = c_0(x) + \dfrac{\gamma_2}{\gamma_1 - \gamma_2} c_{10}(x) \ge c_0(x) > 0$.

Then **H1** implies that assumption **H4** is satisfied and $r \to 0$ uniformly for $t \to \infty$, that is only the isotope 2 is present asymptotically.

ii) if $\gamma_1 = \gamma_2 = \gamma > 0$ then $F(x) = \gamma c_{10}(x)$.

Then in assumption **H5** we have that $M_\infty = \int_{-L}^{L} F > 0$ depends only on the isotope 1 and the asymptotic expansion of $m(x,t)$ is

$$te^{-\gamma t}\frac{x+L}{2L}M_\infty.$$

However for any initial data satisfying **H1** we have $r \leq \dfrac{1}{\gamma t}$

for $t > 1$ and $x \in \Omega$, so that $r \to 0$ uniformly for $t \to \infty$, that is there exists a uniform asymptotic distribution of r in Ω, independently of the possible vanishing of F in a subset of Ω.

Let us remark that if assumption **H5** does not hold, that is if $F \equiv 0$ in Ω, the isotope 1 is initially absent in the explicit solution and the first asymptotic term of m is

$$e^{-\gamma t}\frac{x+L}{2L}\int_{-L}^{L} c_0(\xi)\,d\xi,\ c_0 \equiv c_{20}.$$

This example shows that depending on the form of the matrix Λ there can be three different asymptotic behaviours:

case I for any initial data satisfying **H1**, $F(x)$ is always strictly positive, and hence hypothesis **H4** holds. Then $\mathbf{r} = (r_1, ..., r_{n-1})$ has an asymptotic distribution in the whole Ω (see [8] and (12));

case II assuming hypothesis **H5**, there exists an asymptotic distribution of $\mathbf{r}$ in the whole Ω;

case III assuming hypothesis **H5**, there does not exists in general an asymptotic distribution of $\mathbf{r}$ in the whole Ω.

These three possible behaviours are present in the general case of n species with different evolutive laws. We will present some of them, interesting from a physical point of view.

case I

example Ia) The matrix Λ is a multiple of the identical matrix, defined by:

$$\dot{c}_i = -\gamma c_i, \quad i = 1, ..., n, \quad \gamma \geq 0. \tag{38}$$

This example describes both sets of stable isotopes, i.e. with $\gamma = 0$, e.g. of the couple (Cl^{37}, Cl^{35}), and of radiative isotopes that decade out of

the element with the same coefficients of decay ($\gamma > 0$), e.g. the couple (U^{235}, U^{238}).

In this case we have that $F(x) = c_0(x) > 0$ because of hypothesis **H1**.

Let us remark that in this case the asymptotic distribution of $\mathbf{r}$ strongly depends on the initial conditions, since it is given explicitely by:

$$\lim_{t\to\infty} r_i(x,t) = \frac{c_{i0}(X_\infty^{-1}(x))}{c_0(X_\infty^{-1}(x))}, \quad i = 1, ..., n. \tag{39}$$

example Ib) The matrix Λ is defined by

$$\begin{cases} \dot{c}_1 = -\gamma_1 c_1, \\ \dot{c}_i = \gamma_{i-1} c_{i-1} - \gamma_i c_i, \quad i = 2, ..., n-1, \\ \dot{c}_n = \gamma_{n-1} c_{n-1}, \end{cases} \tag{40}$$

with $\gamma_i > 0,\ i = 1, ..., n-1$.

This case describes the evolution of a chain of n isotopes such that the i^{th} one decades into the $(i+1)^{th}$ one, for $i = 1, ..., n-1$, while the n^{th} one is stable.

It is shown in [8] that also in this example $F(x) = c_0(x)$, however in this case

$$\lim_{t\to\infty} r_i(x,t) = 0 \quad i = 1, ..., n-1, \tag{41}$$

uniformly in Ω, then the unique isotope asymptotically present is the n^{th} one, that is the unique stable isotope.

example Ic) The matrix Λ is defined by

$$\begin{cases} \dot{c}_1 = -\gamma_1 c_1, \\ \dot{c}_i = \gamma_{i-1} c_{i-1} - \gamma_i c_i, \quad i = 2, ..., n. \end{cases} \tag{42}$$

with $\gamma_i > 0,\ i = 1, ..., n$ and $\gamma_n = \min \gamma_i, \mu(-\gamma_n) = 1$.

This is a generalization of the couple (U^{238}, U^{234}): we have a chain of n isotopes of which the i^{th} one decades into the $(i+1)^{th}$ one, for $i = 1, ..., n-1$, and the n^{th} one decays out of the element. In [8, Example 2, Section 4] we have shown that

$$\mathbf{F} = F(x)\mathbf{v}^n,\ \mathbf{v}^n = (0, ..., 0, 1), \qquad F(x) \geq c_0(x).$$

Then, again, for any datum satisfying **H1**, $F(x)$ is strictly positive and

$$\lim_{t\to\infty} r_i(x,t) = 0 \quad i = 1, ..., n-1, \tag{43}$$

uniformly in Ω, and the unique isotope asymptotically present is the n^{th} one.

Let us remark that the estimate on $F(x)$ can be derived directly, without a detailed analysis of the eigenvalues-eigenvectors of Λ.

In fact in this case the ODE system $\dot{\tilde{\mathbf{C}}} = \tilde{\mathbf{\Lambda}}\tilde{\mathbf{C}}$ is given by

$$\begin{cases} \dot{c}_1 = -\gamma_1 c_1, \\ \dot{c}_i = \gamma_{i-1} c_{i-1} - \gamma_i c_i, \quad i = 2, ..., n \\ \dot{c} = -\gamma_n c + \gamma_n \sum_{i=1}^{n-1} c_i(t). \end{cases} \tag{44}$$

Then the $c_i(t),\ i = 1, ..., n-1$, can be obtained from the first $n-1$ equations and depend only on $c_{i0}(t),\ i = 1, ..., n-1$, and the total concentration consequently is given by

$$ce^{\gamma_n t} = c_0 + \gamma_n \int_0^t e^{\gamma_n \tau} \sum_{i=1}^{n-1} c_i(\tau)\, d\tau. \tag{45}$$

The hypotheses $\gamma_n = \min \gamma_i,\ \mu(-\gamma_n) = 1$ ensure that the integral in (45) is bounded for $t \to \infty$, since $c_i,\ i = 1, ..., n-1$, behave at most like $e^{-\gamma_i t} Q(t)$, with $Q(t)$ polynomial in t of degree less or equal to $n-1$ (equal if the γ_i, $i = 1, ..., n-1$, are all identical).

Since $c_i \geq 0$, we have $\lim_{t\to\infty} ce^{\gamma_n t} = F(x) \geq c_0$,

in particular $F(x) = c_0$ if $c_{i0} = 0,\ i = 1, ..., n-1$, that is if initially the unique isotope present is the n^{th} one.

Let us remark that if $\gamma_n = \min \gamma_i$, but $\mu(-\gamma_n) > 1$ then in general F is not positive everywhere. Indeed even in the case $n = 2$ we have seen that $F = \gamma_1 c_{10}$, and in general, for $n > 2$ we have, from (45) and since $\mu(-\gamma_n) = h(-\gamma_n) > 1$:

$$F = \lim_{t\to\infty} t^{-(h(-\gamma_n)-1)} e^{\gamma_n t} c = \lim_{t\to\infty} \gamma_n t^{-(h(-\gamma_n)-1)} \int_0^t e^{\gamma_n \tau} \sum_{i=1}^{n-1} c_i(\tau)\, d\tau.$$

If $c_{i0} = 0,\ i = 1, ..., n-1$ and $c_{n0} > 0$, then the initial data satisfy **H1** but $F = 0$.

case II

This case occurs when **H1** does not imply that $F(x)$ is positive in $\overline{\Omega}$, but $\mathbf{r}$ has a unique asymptotic limit for all data satisfying **H1**, as solution of an ODE. In this class we can find the example with Λ given by (42) with $\gamma_i = \gamma > 0,\ i = 1, ..., n$. Under hypothesis **H1**, in this case we have that, $\forall x \in \Omega$ and for $t > 1$:

$$\begin{aligned} &F_i(x) = \frac{\gamma^{n-1}}{(n-1)!} c_{10}(x)\delta_i^n, \quad i = 1, ..., n, \\ &0 \le r_i \le \frac{i}{\gamma t}, \qquad\qquad\qquad\quad i = 1, ..., n-1, \end{aligned} \tag{46}$$

where δ_i^n is the Kronecker symbol.
Then for any initial datum satisfying hypothesis **H1**, we have that

$$\lim_{t\to\infty} \mathbf{r} = \mathbf{0},$$

uniformly in Ω, that is asimptotically the unique isotope present is the n^{th} one, however M_∞ depends only on the 1^{st} isotope (see Theorem 3.1).

To prove (46) we remark that Λ is multiple of a Jordan normal form and the solution can be explicitly written as follows:

$$\begin{cases} e^{\gamma t} c_i = \sum_{j=1}^{i} c_{j0} \dfrac{(\gamma t)^{i-j}}{(i-j)!}, & i = 1, ..., n-1, \\ e^{\gamma t} c = \sum_{i=1}^{n-1} c_{i0} \sum_{j=1}^{n-i} \dfrac{(\gamma t)^j}{(j)!} + c_0. \end{cases} \tag{47}$$

Then, recalling that $h(-\gamma_n) = n$ and $\lim_{t\to\infty} t^{-(h(-\gamma_n)-1)} e^{\gamma t} \tilde{\mathbf{C}} = \mathbf{F}$, (46) follows.

Let us remark that for any initial data such that $c_{10} > 0$ we have for $t \to \infty$:

$$r_i \simeq \frac{(n-1)!}{(i-1)!} (\gamma t)^{-(n-i)}, \qquad i = 1, ..., n-1,$$

that is the estimate (46) is almost sharp.

case III

This case occurs when **H1** does not imply that $F(x)$ is positive, and $\mathbf{r}$, as solution of an ODE, does not have a unique asymptotic limit, for all the data satisfying **H1**.

example IIIa) The matrix Λ is diagonal, with eigenvalues not all equal. This is the case of a set of isotopes which decade out of the element with coefficients of decay not all equal.

Assuming the isotopes to be ordered with $\gamma_1 \leq \gamma_2 \leq ... \leq \gamma_n$, $\gamma_1 < \gamma_n$, from the explicit solution one can directly observe that, denoting $\mu(-\gamma_1) = j < n$:

$$F(x) = \sum_{i=1}^{j} c_{i0}(x),$$

and if $F(x_0) > 0$ then $\mathbf{r}(X(t;x_0),t)$ tends asymptotically to a limit which can depend on the initial data, if $j > 1$, but it is such that $\sum_{i=1}^{j} r_i$ tends to 1 and r_i tends to 0 for $i > j$, as $t \to \infty$.

On the other hand, we have that if $c_{10}(x_0) = ... = c_{(n-1)0}(x_0) = 0$, $c_{n0}(x_0) > 0$ then $F(x_0) = 0$ and $\mathbf{r}(X(t;x_0),t) \equiv 0$, that is a different limit from the previous one.

Then in general there does not exist a limit for $\mathbf{r}$ in the whole Ω.

example IIIb) Let us consider the example (42) assuming now that the γ_i are not all equal and that $-\gamma_n$ is not the maximum eigenvalue. Then if we choose the initial data $c_{i0} = 0$, $i = 1, ..., n-1$, $c_{n0} > 0$, satisfying hypothesis **H1**, we have the solution:

$$\tilde{\mathbf{C}} = c_0 e^{-\gamma_n t}\mathbf{v}^n, \quad \mathbf{v}^n = (0, ..., 0, 1).$$

Denoted by $\lambda_1 = -\min_{i=1,...,n} \gamma_i$ the maximum eigenvalue, say $-\gamma_k$, $k \neq n$, then, for this initial condition we have:

$$\mathbf{F} = \lim_{t\to\infty} t^{-(h(\lambda_1)-1)} e^{-\lambda_1 t}\tilde{\mathbf{C}} = \lim_{t\to\infty} t^{-(h(\lambda_1)-1)} e^{-(\gamma_n-\gamma_k)t} c_0 \mathbf{v}^n = \mathbf{0}.$$

Moreover, on any characteristic $X(t;x_0)$ with x_0 such that $c_{i0} = 0$, $i = 1, ..., n-1$, $c_{n0} > 0$, we have $\mathbf{r} = \mathbf{0}$.

On the other hand we can show, see [8, Example 2], that in this case $\mathbf{F}(x) = \beta(x)\mathbf{v}^k$, where $\mathbf{v}^k$ is given by:

$$\begin{aligned} &v^{k,i} = 0, && i = 1, ...k-1, \quad \text{if } k > 1, \\ &v^{k,i} = \textstyle\prod_{j=i}^{n-1} \dfrac{\gamma_{j+1} - \gamma_k}{\gamma_j}, && i = k, ..., n-1, \\ &v^{k,n} = 1 + \textstyle\sum_{i=1}^{n-1} \prod_{j=i}^{n-1} \dfrac{\gamma_{j+1} - \gamma_k}{\gamma_j}. \end{aligned} \tag{48}$$

Then if $F(x_0)$ is positive on the characteristic starting in x_0 we have

$$\lim_{t\to\infty} \mathbf{r} = \mathbf{r}_E \neq \mathbf{0}, \quad \mathbf{r}_{E,i} = \frac{v^{k,i}}{v^{k,n}}, \quad i = 1, ...n-1,$$

that is in general a limit for $\mathbf{r}$ does not exist in the whole Ω.

In particular if $k = 1$, all the components of $\mathbf{r_E}$ are positive and $\sum_{i=1}^{n-1} \mathbf{r}_{E,i} < 1$, that is, from a physical point of view, we have the so called secular equilibrium of all the n isotopes.

Acknowledgments. The authors wish to thank Claudio Pescatore for his helpful suggestions and comments.

References

[1] H. Amann, *Ordinary Differential Equations*, de Gruyter, Berlin, 1990.

[2] M. Bertsch, M.E. Gurtin, D. Hilhorst, *On interacting populations that disperse to avoid crowding: the case of equal dispersal velocities.* Nonlinear Anal. **11** (1987), 493–499.

[3] H.F. Bremer, E.L. Cussler, *Diffusion in the Ternary System d-Tartaric Acid c-Tartaric Acid Water at* $25^o C$. AIChE Journal, **16** (1980), 832–838.

[4] C. Bardos, A.Y. Leroux, J.C. Nedelec, *First order quasilinear equations with boundary conditions*, Comm. Partial Differential Equations **4** (1979), 1017–1034.

[5] E. Comparini, R. Dal Passo, C. Pescatore, M. Ughi, *On a model for the propagation of isotopic disequilibrium by diffusion* Math. Models Methods Appl. Sci. **19** (2009), 1277–1294.

[6] E. Comparini, A. Mancini, C. Pescatore, M. Ughi, *Numerical results for the Codiffuson of Isotopes*, Communications to SIMAI Congress, vol. 3, ISSN: 1827-9015 (2009).

[7] E. Comparini, C. Pescatore, M. Ughi, *On a quasilinear parabolic system modelling the diffusion of radioactive isotopes*, Rend. Istit. Mat. Univ. Trieste **39** (2007), 127–140.

[8] E. Comparini, M. Ughi, *Large time behaviour of the solution of a parabolic-hyperbolic system modelling the codiffusion of isotopes*, Adv. Math. Sc. Appl. **21** (2011), 305–319.

[9] *Rock matrix diffusion as a mechanism for radionuclide retardation: natural radioelement migration in relation to the microfractography and petrophysics of fractured crystalline rock*, Report EUR 15977 EN (see sections 3.7.4 and 3.7.5), European Commission, Brussels, (1995).

[10] *Rock matrix diffusion as a mechanism for radionuclide retardation: natural radioelement migration in relation to the microfractography and petrophysics of fractured crystalline rock*, Report EUR 17121 EN, European Commission, Brussels, (1997).

[11] A. Friedman, *Partial Differential Equations of Parabolic Type*, Prentice-Hall, Englewood Cliffs, 1964.

[12] J.K. Hale, *Ordinary Differential Equations*, Pure and applied mathematics, Krieger, Huntington, 1980.

[13] G.E. Hernandez, *Existence of solutions in a population dynamics problem*, Quart. Appl. Math. **43** (1986), 509–521.

[14] G.E. Hernandez, *Localization of age-dependent anti-crowding populations*, Quart. Appl. Math. **53** (1995), 35–52.
[15] T. Gimmi, H.N. Waber, A. Gautschi and A. Riibel, *Stable water isotopes in pore water of Jurassic argillaceous rocks as tracers for solute transport over large spatial and temporal scales*, Water Resources Research, 43, (2007).
[16] KASAM Nuclear Waste state of the art reports 2004, Swedish Government Official Reports SOU 2004-67, (2005).
[17] Latham, A. G. 1993. Diffusion-sorption modelling of natural U in weathered granite fractures: potential problems. Proceedings of Migration '93, 701–710.
[18] R.C. MacCamy, *A population model with nonlinear diffusion*, J. Differential Equations **39**, (1981), 52–72.
[19] Percy, E. C., Prikryl J. D., Lesley B., W. *Uranium transport through fractured silicic tuff and relative retention in areas with distinct fracture characterisics*, Applied Geochemistry **10** (1995), 685–704.
[20] C. Pescatore, *Discordance in understanding of isotope solute diffusion and elements for resolution*, Proceedings OECD/NEA "Radionuclide retention in geological media", Oskarsham, Sweden, (2002), 247–255.

Authors' addresses:

Elena Comparini
Dipartimento di Matematica "U. Dini"
Università di Firenze
V.le Morgagni 67/a, I-50134 Firenze (Italy)
E-mail: elena.comparini@unifi.it

Maura Ughi
Dipartimento di Matematica e Geoscienze
Università di Trieste
V. Valerio 12/b, I-34127 Trieste (Italy)
E-mail: ughi@units.it

Received March 19, 2012
Revised June 25, 2012

Rend. Istit. Mat. Univ. Trieste
Volume 44 (2012), 153–171

Linearizations, normalizations and isochrones of planar differential systems[1]

Marco Sabatini

A Fabio Zanolin, per i suoi primi sessant'anni.

Abstract. *In the first section we collect some unpublished results presented in [17], related to linearizations and normalizations of planar centers. In the second section we consider both the problem of finding isochrones of isochronous systems (centers or not) and its inverse, i.e. given a family of curves filling an open set, how to construct a system having such curves as isochrones. In particular, we show that for every family of curves* $y = mx + d(x)$, $m \in \mathbb{R}$, *there exists a Liénard system having such curves as isochrones.*

Keywords: planar systems, period function
MS Classification 2010: 34C25

1. Introduction

Let Ω be an open connected subset of the real plane. Let us consider a differential system

$$z' = V(z), \qquad z \equiv (x, y) \in \Omega, \tag{1}$$

$V(z) = (v_1(z), v_2(z)) \in C^\infty(\Omega, \mathbb{R}^2)$. We denote by $\phi_V(t, z)$, the local flow defined by (1). A connected subset $\mathcal{P} \subset \Omega$ covered with concentric non-trivial cycles is said to be a *period annulus*. If O is an isolated critical point of (1), we say that O is a *center* if it has a punctured neighbourhood which is a period annulus of Ω. The largest neighbourhood N_O of O such that $N_O \setminus \{O\}$ is a period annulus of Ω is said to be the its *central region*. On every period annulus one can define the *period function* $\tau(z)$, defined as the minimum positive period of the cycle starting at z. It can be proved that τ has the same regularity as the

[1]This paper was partially supported by the PRIN project *Equazioni differenziali ordinarie: sistemi dinamici, metodi topologici e applicazioni*. Symbolic and numeric computations were performed using Maple$^{\text{TM}}$ 11.

system. A period annulus is said to be *isochronous* if τ is constant. The study of τ, and in particular isochronicity, is related to boundary value problems and stability theory. In [1] several methods and results related to isochronicity theory were reviewed. One of the oldest ones is the linearization one, dating back to Poincaré. It consist in looking for a transformation that takes (1) into a linear system. Since every linear center is isochronous, if such a map exists, (1) has an isochronous center. Poincaré proved that if (1) is analytical and O is a non-degenerate critical point, then it admits a local linearization at O if and only if O is isochronous. Such a result is purely existential, giving no hints about how such a linearization could be obtained, in order to prove O actually to be isochronous. Linearizations of special classes of isochronous centers were found later by applying different techniques, as in [13].

A different method to prove isochronicity was introduced in [16, 21], based on the use of Lie brackets. Let us consider a second differential system

$$z' = W(z), \qquad z \equiv (x, y) \in \Omega, \tag{2}$$

$W(z) = (w_1(z), w_2(z)) \in C^\infty(\Omega, \mathbb{R}^2)$, $\phi_W(s, z)$ the local flow defined by (2). We say that (1) and (2) *commute*, or that V and W are *commutators*, if their Lie brackets $[V, W]$ vanish identically on Ω. A center is isochronous if and only if V it has a non-trivial (transversal at non-critical points) commutator W [16]. In several cases looking for a commutator turns out to be easier than looking for a linearization [1]. Also, as shown in [8], isochronicity is equivalent to the existence of a vector field W *normalized* by V, i.e. of a vector field W and a function μ such that $[V, W] = \mu W$. Every commutator is a normalizer, but the converse is not true, since the normalizing condition is expressed by one equality, the commutation condition by two.

Poincaré linearization theorem implies that an isochronous analytical center has a non-trivial commutator, since every linear center commutes with a transversal (at non-critical points) linear system. Conversely, if an analytical center has a non-trivial commutator, then it is isochronous, hence by Poincaré theorem it has an analytical linearization. The extension of such a relationship to non-analytical systems was studied in [22]. Procedures to get the linearization, starting form a given commutator, were studied in [5, 6, 9, 15], for several classes of analytical and non-analytical systems. In such papers it was always assumed the commutator W to have a non-degenerate critical point at O, usually having a linear part of star-node type. In the first section of this paper we present an approach, first presented in the unpublished preprint [17], where such an assumption is not required. The absence of a non-degeneracy assumption does not allow us to prove the existence of a linearizing diffeomorphism. In fact, we only prove the existence of a bijective linearizing map which fails to be a diffeomorphism at the critical point, where we lose the differentiability of the inverse map. In this section we also consider the existence of *normalizations*,

i.e. maps that take (1) into a system of the form

$$\dot{u} = v\,\varphi(u^2+v^2), \qquad \dot{v} = -u\,\varphi(u^2+v^2).$$

Such a question was considered in [12].

In the second section we are concerned with the existence of *isochrones*, or *isochronous sections*, i.e. curves met by the local flow of (1) at equal time intervals. If O is an isochronous center, then every curve meeting its cycles exactly at a single point, even if not transversal, is an isochrone. The existence of isochrones becomes less obvious when dealing with cycles, isolated (limit cycles) or not, or with rotation points, or boundaries of attraction regions [19]. The existence of isochrones in a neighbourhood U_γ of a cycle γ, in relation to the existence of commutators or normalizers, was considered in [19, 20].

Following [2], we say that a point $z^* \in U_\gamma$ has *asymptotic phase* with respect to γ if there exists a point $z_* \in \gamma$ such that $\lim_{t\to+\infty} |\phi_V(t,z_*) - \phi_V(t,z^*)| = 0$, or $\lim_{t\to-\infty} |\phi_V(t,z_*) - \phi_V(t,z^*)| = 0$. In such a case, z^* is said to be *in phase* with z_*. In [2] a cycle is said to be isochronous if it has a neighbourhood U_γ such that every point of U_γ is in phase with some point of γ. A cycle is isochronous if and only if it has an isochrone, since the set of points in phase with a given $z_* \in \gamma$ is an isochrone, and vice-versa. Every hyperbolic limit cycle is isochronous, in such a sense [11]. Even non-hyperbolic limit cycles can be isochronous, under some additional conditions on the first return time map [2, 4]. The asymptotic phase approach cannot be extended to some other situations, as attraction boundaries, since if the boundary of the attraction region of an isochronous system is unbounded, then for every z in the boundary, $\phi_V(t,z)$ does not exist for all $t \in \mathbb{R}$.

If a system has an isochrone, then it has infinitely many ones, obtained from the given one by means of the local flow ϕ_V. If a cycle $\phi_V(t,z)$ is isochronous, such curves cover a neighbourhood of $\phi_V(t,z)$. If a critical point O is isochronous, then the system's isochrones cover a punctured neighbourhood of O. If a boundary is isochronous, then the system's isochrones cover a one-sided neighbourhood of such a boundary.

Given a family of curves covering an open set, one can consider an inverse problem, consisting in finding a differential system having such curves as isochrones. In the second section we describe an elementary approach to such a problem, with special regard to Liénard systems.

2. Linearizations and normalizations

Let Ω be an open connected subset of the real plane. We assume systems (1) and (2) to have the same, isolated critical points. We denote by $\phi_V(t,z)$, $\phi_W(s,z)$ the local flows of (1) and (2). If $I \in C^\infty(\Omega,\mathbb{R})$, we denote by $\partial_V I$, $\partial_W I$, the derivatives of I along the solutions of (1), (2), respectively. Similarly

for $\partial_W I$ and for the derivative of a vector field along the solutions of (1) or (2). We write $[V,W] = \partial_V W - \partial_W V$, $A = V \wedge W = v_1 w_2 - v_2 w_1$. We say that W is a *non-trivial normalizer* of V if $A \neq 0$ at regular points and $V \wedge [V,W] = 0$. In this case, we define the function μ as follows,

$$\mu = \frac{V \wedge [V,W]}{|V|^2}.$$

If W is a normalizer of V, then the time-map $\phi_W(s,z)$ takes locally arcs of V-orbits into arcs of V-orbits. When both vector fields are non-trivial normalizers of each other we say that they are non-trivial *commutators*. By the transversality of V and W, this occurs when $[V,W] = 0$. In such a case, if $\phi_V(t, \phi_W(s,z))$ and $\phi_W(s, \phi_V(t,z))$ are defined for all $(s,t) \in J_s \times J_t$, J_s, J_t intervals containing 0, then one has the following commutativity property

$$\phi_V(t, \phi_W(s,z)) = \phi_W(s, \phi_V(t,z)).$$

We say that a function $I \in C^\infty(\Omega, \mathbb{R})$ is an *first integral* of (1), or V, if I is non-constant on any open subset of Ω, and $\partial_V I = 0$ in Ω. We say that a function $F \in C^\infty(\Omega, \mathbb{R})$ is an *integrating factor* of (1) if the divergence of the field FV vanishes in Ω. In such a case the differential form $\omega = -Fv_2 dx + Fv_1 dy$ is closed, and a potential exists on every simply connected subset of Ω. If FV does not vanish identically on any open subset of Ω, then such a potential is a first integral of (1). We say that a function $G \in C^\infty(\Omega, \mathbb{R})$, $G(z) \neq 0$ for all $z \in \Omega$, is an *inverse integrating factor* of (1) if $\dfrac{1}{G}$ is an integrating factor of (1).

If W is a normalizer of V, then $A = V \wedge W$ is an inverse integrating factor of V [7]. Similarly, if V is a normalizer of W, then $A = V \wedge W$ is an inverse integrating factor of W, so that, if V and W commute, then $A = V \wedge W$ is an inverse integrating factor both of V and W. Let us denote by $\mathcal{T}$ the set of points where V and W are transversal:

$$\mathcal{T} = \{z \in U : A(z) \neq 0\}.$$

For every $z \in \mathcal{T}$, we set $B(z) = \dfrac{1}{A(z)}$.

If W is a non-trivial normalizer of V, then for every point $z \in \mathcal{T}$ there exists a disk U_w^z and a function $S^z \in C^\infty(U_w^z, \mathbb{R})$, determined up to an additive constant κ_w^z, such that $\nabla S^z = B(-v_2, v_1)$. As a consequence, $\partial_V S^z = 0$. Similarly, if V is a non-trivial normalizer of W, then for every point $z \in \mathcal{T}$ there exists a disk U_w^z and a function $T^z \in C^\infty(U_w^z, \mathbb{R})$, determined up to an additive constant κ_w^z, such that $\nabla T^z = B(w_2, -w_1)$ and $\partial_W T^z = 0$.

If V and W commute, something more can be said, as in next lemma. We say that a map *rectifies* a vector field V if it takes (1) into a non-zero constant one. We say that a map *linearizes* a vector field V if it takes (1) into a linear

one. We say that a map *normalizes* a vector field V if it takes (1) into a system of the following form

$$\dot{u} = v\,\varphi(u^2+v^2), \qquad \dot{v} = -u\,\varphi(u^2+v^2).$$

The orbits of such a system are contained in circles centered at O. If $\varphi(u^2+v^2) \neq 0$ on a given circle, then its minimal period is $\dfrac{1}{\varphi(u^2+v^2)}$. As a consequence, if such a system is defined in a neighbourhood of O, its period function is bounded only if $\varphi(u^2+v^2)$ does not approach 0. In the following we shall take into account also bijective C^∞ maps which fail to be diffeomorphisms just at a point.

For every point $z \in \Omega \cap \mathcal{T}$, let us set $U^z = U^z_v \cap U^z_w$. Then, for every point $z \in \Omega \cap \mathcal{T}$, we can define the map $\Gamma^z = (S^z, T^z) \in C^\infty(U^z, \mathbb{R}^2)$.

LEMMA 2.1. *Let V and W be non-trivial commutators. Then, for every choice of κ^z_v, κ^z_w, Γ^z is a local diffeomorphism that rectifies locally both (1) and (2). Moreover, for every $\zeta \in U^z$, $\zeta = \phi_V(t_\zeta, \phi_W(s_\zeta, z)) = \phi_W(s_\zeta, \phi_V(t_\zeta, z))$, one has:*

$$\phi_V(t,\zeta) = (\Gamma^z)^{-1}(t + t_\zeta, s_\zeta), \qquad \phi_W(s,\zeta) = (\Gamma^z)^{-1}(t_\zeta, s + s_\zeta). \tag{3}$$

Proof. The regularity of Γ^z comes from those of S^z, T^z. The map Γ^z has jacobian matrix:

$$J_{\Gamma^z} = \begin{pmatrix} -Bv_2 & Bv_1 \\ Bw_2 & -Bw_1 \end{pmatrix}$$

whose determinant is $-B$, that does not vanish on $\mathcal{T}$. Hence Γ^z is locally invertible on all of $\mathcal{T}$, that is at every regular point. As for the transformed systems, we know from what above that $\partial_V S^z = 0$, $\partial_W T^z = 0$. Moreover,

$$\begin{cases} \partial_V T^z &= Bw_2v_1 - Bw_1v_2 = BA = 1 \\ \partial_W S^z &= -Bv_2w_1 + Bv_1w_2 = BA = 1. \end{cases}$$

This shows that Γ rectifies locally both systems.

We prove only the first equality in (3), the second one can be proved similarly. We have: $\Gamma^z(\phi_V(t,\zeta)) = \Gamma^z(\phi_V(t, \phi_V(t_\zeta, \phi_W(s_\zeta, z)))) = \Gamma^z(\phi_V(t + t_\zeta, \phi_W(s_\zeta, z)))) = (t + t_\zeta, s_\zeta)$. By the local invertibility of Γ^z we get $\phi_V(t,\zeta) = \Gamma^{z\,-1}(t + t_\zeta, s_\zeta)$. □

LEMMA 2.2. *Let $\mathcal{P}$ is an open isochronous period annulus of (1). Then, for every vector field W such that $[V,W] = 0$ on $\mathcal{P}$, there exists a map $\Lambda_W \in C^\infty(\mathcal{P}, \mathbb{R}^2)$ that linearizes both (1) and (2).*

Proof. Possibly multiplying V by $\dfrac{\tau(z)}{2\pi}$, we may assume the cycles of V to have minimal period 2π. Let us consider $z_0 \in \mathcal{P}$. The W-orbit $\phi_W(s, z_0)$ meets all the V-cycles in $\mathcal{P}$ exactly once. Let T^{z_0}, S^{z_0} be the maps of Lemma 2.1, defined in a suitable neighbourhood U_{z_0} of z_0. Let us choose the integration constants so that $T(z_0) = 0$, $S(z_0) = 0$. By Lemma 2.1, S^{z_0} and T^{z_0} coincide, respectively, with s and t of $\phi_W(s, z_0)$, $\phi_V(t, z_0)$. Hence S^{z_0} can be extended in a unique way to all of $\mathcal{P}$, by using the commutativity of the local flows ϕ_V and ϕ_W. Let us denote again by S^{z_0} and T^{z_0} the extended maps. The function T^{z_0} is not continuous at some point of every cycle, since $\phi_V(2\pi, z_0) = z_0$. Anyway, the functions $\cos T^{z_0}$, $\sin T^{z_0}$ are well-defined on all of $\mathcal{P}$. Their regularity comes from Lemma 2.1, since at every point they coincide, up to an additive constant, with some $\cos T^z$, $\sin T^z$.

Let us define Λ_W as follows,

$$\Lambda_W(z) = \Big(e^{S^{z_0}(z)} \cos\big(T^{z_0}(z)\big), e^{S^{z_0}(z)} \sin\big(T^{z_0}(z)\big)\Big) = (u, v).$$

Then Λ_W takes V-cycles into circles, and is one-to-one on cycles. This implies that Λ_W is one-to-one on all of $\mathcal{P}$.

Λ_W linearizes both (1) and (2). In fact, writing S and T for $S^{z_0}(z)$ and $T^{z_0}(z)$, one has

$$\begin{cases} \partial_V u &= e^S\ \partial_V S \cos T - e^S \sin T\ \partial_V T = -e^S \sin T = -v \\ \partial_V v &= e^S\ \partial_V S \sin T + e^S \cos T\ \partial_V T = e^S \cos T = u, \end{cases}$$

$$\begin{cases} \partial_W u &= e^S\ \partial_W S \cos T - e^S \sin T\ \partial_W T = e^S \cos T = u \\ \partial_W v &= e^S\ \partial_W S \sin T + e^S \cos T\ \partial_W T = e^S \sin T = v. \end{cases}$$

□

In next theorem we prove that starting from a commutator of (1) one can find a linearization, even without the non-degeneracy assumption on the commutator.

THEOREM 2.3. *Let O be an isochronous center of (1), with central region N_O. Then, for every vector field W such that $[V, W] = 0$ on $N_O \setminus \{O\}$, there exists a map $\Lambda^0_W \in C^\infty(N_O, R)$ that linearizes (1).*

Proof. Let z_0 be a point of $\mathcal{P} = N_O \setminus O$, and Λ be defined as in Lemma 2.2. Possibly multiplying the vector field W by -1, in order to make its orbits tend to O as $s \to -\infty$, we may assume O to be asymptotically stable for (2). Let us define the map Λ^*_W as follows,

$$\Lambda^*_W(z) = \begin{cases} O & \text{if } \ z = O, \\ \Lambda_W(z) & \text{if } \ z \neq O. \end{cases}$$

Then $\Lambda^*_W \in C^0(N_O, \mathbb{R}) \cap C^\infty(\mathcal{P}, \mathbb{R})$. Working as in [14], thm 1.3, one can prove the existence of a first integral $I \in C^\infty(N_O, \mathbb{R})$, such that $\Lambda^0_W = I\Lambda^*_W \in C^\infty(N_O, \mathbb{R})$. By Lemma 2.2, the map $w = \Lambda_W(z)$ transforms (1) into the linear system

$$\dot{u} = -v, \qquad \dot{v} = u.$$

Then, setting $\varepsilon = \Lambda^0_W(z) = I(z)\Lambda^*_W(z) = Iw$, one has

$$\dot{\varepsilon} = \dot{(Iw)} = \dot{I}w + I\dot{w} = IMw = M(Iw) = M\varepsilon,$$

hence Λ^0_W linearizes (1). □

The above theorem allows to prove the existence of a normalization for every system with a center at O.

COROLLARY 2.4. *Let O be a center of (1), with central region N_O. Then there exists a map $\Lambda_0 \in C^\infty(N_O, R)$ that normalizes (1).*

Proof. Let us consider the system

$$\dot{z} = \frac{\tau(z)}{2\pi} V(z). \tag{4}$$

Such a system is of class C^∞ in $\mathcal{P} = N_O \setminus \{O\}$, since $\tau \in C^\infty(\mathcal{P}, \mathbb{R})$. $\mathcal{P}$ is an isochronous annulus, with minimal period 2π. By Theorem 2.3, there exists a map $\Lambda_0 \in C^\infty(N_O, R)$ that linearizes (4), taking it into the system

$$\dot{u} = -v, \qquad \dot{v} = u.$$

As a consequence, system (1) is taken into the system

$$\dot{u} = -\frac{2\pi}{\tau(\Lambda_0(z))}\, v, \qquad \dot{v} = \frac{2\pi}{\tau(\Lambda_0(z))}\, u. \tag{5}$$

The function $\tau(z)$ is a first integral of (4), hence $\tau(\Lambda_0(z))$ is a first integral of (5). The orbits of (5) are circles centered at the origin, hence there exists a function $\beta \in C^\infty((0, +\infty), \mathbb{R})$ such that $\tau(\Lambda_0(z)) = \beta(u^2 + v^2)$. Then, setting

$$\varphi(u^2 + v^2) = -\frac{2\pi}{\beta(u^2 + v^2)}$$

satisfies the definition of normalized system. □

We consider now the special case of hamiltonian systems

$$\dot{x} = H_y \qquad \dot{y} = -H_x, \tag{6}$$

where $H \in C^\infty(\Omega, \mathbb{R})$. A map is said to be a *canonical transformation* if it transforms every hamiltonian system into a hamiltonian system. A diffeomorphism is a canonical transformation if and only if its jacobian determinant is a non-zero constant. The approach of Theorem 2.3 does not allow to get a canonical linearization on all of N_O, since the smoothing procedure affects the value of the jacobian determinant. On the other hand, one can characterize hamiltonian systems with commutators in terms of jacobian maps, i.e. maps wih constant non-vanishing jacobian determinant [17].

COROLLARY 2.5. *Let $H \in C^\infty(\Omega, R)$. Let z be a regular point of the hamiltonian system (6). Then (6) has a nontrivial commutator in a neighbourhood U^z of z if and only if there exist $P, Q \in C^\infty(U^z, \mathbb{R})$ such that:*

i) the map $\Lambda(z) = (P(z), Q(z))$ has jacobian determinant $\equiv 1$ in U^z;

ii) $H = \frac{P^2+Q^2}{2}$.

If (6) has an isochronous period annulus $\mathcal{P}$, then Λ can be extended to all of $\mathcal{P}$, and is a canonical linearization of (6) on $\mathcal{P}$. If (6) has a non-isochronous period annulus $\mathcal{P}$, then such a Λ is a canonical normalization of (6) on $\mathcal{P}$.

Proof. Assume that $H = \frac{P^2+Q^2}{2}$, with $P_xQ_y - P_yQ_x \equiv 1$. Then the hamiltonian system (6) has the form

$$\begin{cases} \dot{x} &= PP_y + QQ_y \\ \dot{y} &= -PP_x - QQ_x. \end{cases} \tag{7}$$

and commutes with the system:

$$\begin{cases} \dot{x} &= -PQ_y + QP_y \\ \dot{y} &= PQ_x - QP_x. \end{cases} \tag{8}$$

Conversely, assume (6) to commute with (2). Let z be a non-critical point of (6). Then the function $A = H_y w_2 + H_x w_1$ is an inverse integrating factor for both (6) and (2). Hence there exist a neighbourhood U^z of z, and functions S and T, local first integrals of (6) and (2). In particular:

$$\nabla H = A \nabla S.$$

This implies that $A_x S_y + A S_{yx} = H_{yx} = H_{xy} = A_y S_x + A S_{xy}$, so that $A_y S_x - A_x S_y = 0$. Hence the level sets of A and S coincide, so that A is a first integral of (6), too. Since the gradient of S does not vanish, there exist two scalar functions h, a such that $H = h(S)$, $A = a(S)$. We have:

$$h'(S) \nabla S = \nabla H = a(S) \nabla S.$$

that gives $h' = a$. Now let us consider the map

$$\Lambda(\zeta) = (P(\zeta), Q(\zeta)) = (\sqrt{2h(S(\zeta))}\cos T(\zeta), \sqrt{2h(S(\zeta))}\sin T(\zeta)).$$

The jacobian determinant of Λ is identically 1:

$$\begin{aligned}\det \Lambda(\zeta) &= \begin{vmatrix} \frac{h'(S)S_x}{\sqrt{2h(S)}}\cos T - \sqrt{2h(S)}T_x \sin T & \frac{h'(S)S_y}{\sqrt{2h(S)}}\cos T - \sqrt{2h(S)}T_y \sin T \\ \frac{h'(S)S_x}{\sqrt{2h(S)}}\sin T + \sqrt{2h(S)}T_x \cos T & \frac{h'(S)S_y}{\sqrt{2h(S)}}\sin T + \sqrt{2h(S)}T_y \cos T \end{vmatrix} \\ &= h'(S)\left[S_x T_y - S_y T_x\right] = h'(S)\left[\frac{H_x}{A}\frac{w_1}{A} + \frac{H_y}{A}\frac{w_2}{A}\right] \\ &= h'(S)\frac{H_x w_1 + H_y w_2}{a(S)^2} = h'(S)\frac{a(S)}{a(S)^2} = 1.\end{aligned}$$

Moreover $P^2 + Q^2 = 2h(S) = 2H$, as required.

Now, let $\mathcal{P}$ be an isochronous period annulus. Without loss of generality, we may assume the period to be 2π. Working as in Lemma 2.2, one proves that Λ can be extended to all of $\mathcal{P}$, and that it linearizes (2).

If $\mathcal{P}$ is a non-isochronous period annulus, then working as in Corollary 2.4 one obtains a new system

$$\dot{x} = \frac{\tau H_y}{2\pi}, \qquad \dot{x} = -\frac{\tau H_x}{2\pi}, \tag{9}$$

which is itself a hamiltonian system, since $\dfrac{\tau(z)}{2\pi}$ is a first integral of (6). $\mathcal{P}$ is an isochronous period annulus of (9), hence there exists a canonical map Λ that linearizes (9) on $\mathcal{P}$. As in Corollary 2.4, such a linearization is a canonical normalization of (6) on $\mathcal{P}$. □

A different, and more satisfactory approach to canonical linearizations for hamiltonian systems can be found in [12].

3. Isochrones

When dealing with centers the natural definition of isochronicity is given by requiring T to be constant. This is no longer possible when dealing with systems having non-periodic oscillations, as systems with foci. In such a case one can extend the isochronicity definition by considering *isochrones*, or *isochronous sections*, i.e. curves δ such that $\phi_V(T, \delta) \subset \delta$ for a fixed T, not necessarily positive. This in turn implies $\phi_V(nT, \delta) \subset \delta$, for every positive integer n. Usually such isochrones are taken transversal to V, but this is not necessary, in order to identify the existence of isochronous oscillations. Isochrones can

exist in a neighbourhood of a rotation point, or a cycle, or a boundary (of a central or attraction region). In a neighbourhood of a semi-stable cycle one can consider $\phi_V(T,\delta) \subset \delta$ for $T > 0$ on one side of the cycle, $\phi_V(-T,\delta) \subset \delta$ on the opposite side. If a system (1) admits a linearization Λ, then the half-lines l_θ originating at O are isochrones of the linear system, hence the curves $\Lambda^{-1}(l_\theta)$ are isochrones of (1). The linearization method can be adapted to deal with non-periodic solutions, as in the case of foci [5]. On the other hand, it cannot be applied to the study of a limit cycle's isochrones, since linear systems do not have limit cycles. The same happens for attraction boundaries, since if a linear system has an asymptotically stable point, then it is globally attractive. A different approach can be based on normalizers, since if V is a normalizer of W, then the orbits of W are isochrones of V [8]. Looking for a normalizer is an effective way both to prove a system's isochronicity, and for attacking the inverse problem, i.e. to construct an isochronous system with a given family of curves as isochrones. In fact, one can consider two problems naturally related to isochrones:

- given a system with isochronous oscillations, find a family of isochrones covering a (punctured) neighbourhood, or a one-sided neighbourhood, of a point, or cycle, or boundary;
- given a family of curves covering an open set, find a system admitting such curves as isochrones.

A related question is that of constructing an isochronous system with some prescribed dynamic properties, as centers, foci, or limit cycles. All such problems are strictly related. We first show a simple procedure to construct new isochronous systems starting from a given one.

LEMMA 3.1. *If V normalizes W on an open set U, then for every function $J \in C^\infty(U,\mathbb{R})$, and for every first integral of (2) $I^W \in C^\infty(U,\mathbb{R})$, the vector field $I^W V + JW$ normalizes W.*

Proof. Assume $[V,W] = \mu W$ on U. Then one has

$$[I^W V + JW, W] = \big(I^W \mu - \partial_W J\big) W.$$

□

If (2) is isochronous, passing from V to $I^W V + JW$ we can modify V's dynamics getting a new isochronous system with different properties. For instance we can pass form a center to a system with a focus and one or more limit cycles. In order to construct smooth vector fields, one has to consider only constant first integrals I^W. In fact, a non-constant first integral of (2) is not continuous at the critical point, since it assumes different values on different

orbits. This is not an issue if one looks for an isochronous perturbation in a neighbourhood of a cycle, neglecting the effects of such a perturbation at the critical point located inside the cycle.

One can construct several examples, starting form any couple of commuting vector fields [1]. In order to get the desired dynamics, one has to choose the proper function J, which determines the attractive or repulsive effect of JW. Starting with a jacobian map $\Lambda(x,y) = (P(x,y), Q(x,y))$, we consider the hamiltonian systems (7) and (8) of the previous section. Then we perturb (7) choosing J as a function of H, so that the limit cycles of the new system, corresponding to the zeroes of J, are cycles of (7). For example, if H assumes the value 1 in the period annulus, we can take $J(x,y) = H(x,y)^2 - 1$, obtaining the system

$$\begin{cases} \dot{x} &= PP_y + QQ_y + (H^2-1)(-PQ_y + QP_y) \\ \dot{y} &= -PP_x - QQ_x + (H^2-1)(PQ_x - QP_x), \end{cases} \tag{10}$$

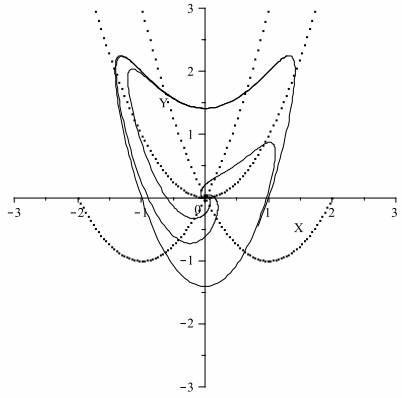
with a limit cycle coinciding with the level set $H = 1$.

If the jacobian map is $\Lambda(x,y) = (x, y - x^2)$, then system (10) has the form

$$\begin{cases} \dot{x} &= x + y - x^2 - \frac{xy^4}{4} - \frac{x^3y^2}{2} - \frac{x^5}{4} + x^3y^3 + x^5y - \frac{3y^2x^5}{2} - \frac{x^7}{2} + x^7y - \frac{x^9}{4} \\ \dot{y} &= -x + y + x^2 + 2xy - 2x^3 - \frac{x^2y^3}{2} - \frac{y^5}{4} + \frac{3x^2y^4}{4} - \frac{x^4y}{4} + \frac{x^4y^2}{2} - \frac{x^4y^3}{2} + \\ &\quad + \frac{x^4y^2}{2} - \frac{x^4y^3}{2} - \frac{x^6}{4} + \frac{x^6y}{2} - \frac{x^6y^2}{2} - \frac{x^8y}{2}x^8 + \frac{3x^8y}{4} - \frac{x^{10}}{4}. \end{cases} \tag{11}$$

Its isochrones are the curves $ax + b(y - x^2) = 0$, for $a, b \in \mathbb{R}$. In next figure we have plotted in continuous line some orbits of (11), and in dotted line the isochrones contained in the curves $y = -2x + x^2$, $y = x^2$, $y = 2x + x^2$. The system has a limit cycle contained in the level set $x^2 + (y - x^2)^2 = 1$.

Figure 1: The system (11)



By Poincaré's theorem, system (11) is linearizable at O, but its linearization is no longer Λ, which linearizes (7), but transforms (11) into the system

$$\begin{cases} \dot{u} &= v + u\Big(1 - H^2(\Lambda^{-1}(u,v))\Big) \\ \dot{v} &= -u + v\Big(1 - H^2(\Lambda^{-1}(u,v))\Big), \end{cases}$$

A normalizer can be also produced by means of a different procedure. In next statement we characterize normalizers in terms of first integrals. We do not know whether such a statement already appeared elsewhere.

THEOREM 3.2. *Let K be a first integral of (2) on an open set A. Assume W and ∇K not to vanish on A. Then V is a non-trivial normalizer of W if and only if for all $z^* \in A$ there exists a neighbourhood U^* and a function $\nu^* : U^* \to \mathbb{R}$, $\nu^* \neq 0$ such that*

$$\partial_V K = \nu^*(K).$$

Proof. Let V be a non-trivial normalizer of W. Let us choose arbitrarily a W-orbit γ^* and a point $z^* \in \gamma^*$. Every point z in a neighbourhood U^* of z^* can be written as $z = \phi_W(s, \phi_V(t, z^*))$. V is a normalizer, hence the parameter t depends only on the orbit to which z belongs. Hence the function that associates to a point $z \in U^*$ the value $t(z)$ of the parameter such that $z = \phi_V(t(z), \phi_W(s, z^*))$ is a first integral of (2). By construction, one has

$$\partial_V t(z) = 1.$$

The above formula also implies that ∇t does not vanish on A. Hence there exists a scalar function χ such that $K(z) = \chi(t(z))$, with $\chi'(t) \neq 0$ because both ∇t and ∇K do not vanish. Then

$$\partial_V K(z) = \chi'(t(z))\partial_V t(z) = \chi'(t(z)) = \chi'(\chi^{-1}(K(z))).$$

Then it is sufficient to set $\nu^*(K) = \chi'(\chi^{-1}(K))$.

Conversely, let us assume that there exists a scalar function ν^* such that $\partial_V K = \nu^*(K)$. Since ∇K does not vanish on A, locally K does not has the same value on different orbits, so that every arc of orbit in U^* can be identified as $K^{-1}(l) \cap U^*$, for some $l \in \mathbb{R}$. This establishes a one-to-one correspondence between the W-orbits of in U^* and the values of K. The relationship $\partial_V K = \nu^*(K)$ implies that $K(\phi_V(t, z))$ depends only on the initial value of K (in particular, it does not depend on the initial point z), hence the local flow $\phi_V(t, \cdot)$ takes arcs of orbits of (2) into arcs of orbits of (2), that is, V is a normalizer of W. □

Theorem 3.2 allows to construct systems with prescribed isochrones without referring to any smooth linearization. In fact, the system we consider now do not necessarily admit linearizations, since they are not regular enough.

COROLLARY 3.3. *Assume that for every non-critical point z of (2) there exist a neighbourhood $U_z \subset \Omega$ and functions $K \in C^\infty(U_z, \mathbb{R})$, $\xi \in C^0(U_z, \mathbb{R})$, $\nu \in C^0(\mathbb{R}, \mathbb{R})$, such that in U_z one has $|\nabla K| \neq 0$ and*

$$W = \left(\frac{K_x}{|\nabla K|^2} \nu(K) + \xi K_y, \frac{K_y}{|\nabla K|^2} \nu(K) - \xi K_x \right). \tag{12}$$

Then (2) is an isochronous system, whose isochrones are locally defined by the level curves of K.

Proof. On every U_z, one has $\dot{K} = \nu(K)$, hence by Lemma 3.2, system (12) normalizes the hamiltonian system having K as hamiltonian function. Hence its isochrones are the orbits of such a hamiltonian system, i.e. K's level sets. □

Corollary 3.3 provides a tool for constructing systems with pre-assigned isochrones. In this case the system's attractors depend on the function ξ. We give some examples generating rational vector fields. Let us consider a one-to-one-map $\Lambda \in C^\infty(\Omega, \mathbb{R}^2)$, such that $\Lambda(0,0) = (0,0)$. Setting $\Lambda(x,y) = (P(x,y), Q(x,y))$, we may consider polar coordinates (ρ, θ) in the (P,Q)-plane. Let us consider a strictly increasing function η, and K locally defined as follows,

$$K(x,y) = \eta(\theta(P(x,y), Q(x,y))).$$

Such a function is defined only locally, since $\theta(P(x,y), Q(x,y))$ is not a single-valued function, but the corresponding system (12), for an arbitrary choice of ν and ξ, is well defined on all of $\Omega \setminus O$. It can be extended to all of Ω by adding the origin as a stationary point. The new vector field can be discontinuous at O, but the dynamics at regular points do not change. Adapting the usual terminology, we say that O is a center if it surrounded by non-trivial cycles, or a focus if every orbit in a neighbourhood of O spirals towards O or away from O. If it has a section, then it is isochronous. The isochrones are locally contained in K's level curves, which coincide with those of $\theta(P(x,y), Q(x,y))$, i.e. half-lines starting at the origin in the (P,Q)-plane, as for system (10):

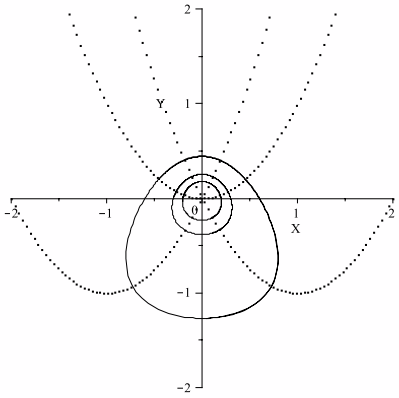
$$aP(x,y) + bQ(x,y) = 0, \qquad a, b \in \mathbb{R}.$$

If $\Lambda(x,y) = (x, y - x^2)$, $\eta(t) = t$, $\nu(t) = 1$, $\xi(x,y) = 0$, then O is a center of (12), since its orbits are symmetric with respect to the y-axis:

$$\dot{x} = -\frac{(y+x^2)(x^4 - 2yx^2 + y^2 + x^2)}{x^2 + y^2 + 2yx^2 + x^4}, \qquad \dot{y} = \frac{x(x^4 - 2yx^2 + y^2 + x^2)}{x^2 + y^2 + 2yx^2 + x^4}$$

Its isochrones are the parabolas $ax + b(y - x^2) = 0$. In Figure 2 we show three cycles and six isochrones contained in $y = -2x + x^2$, $y = x^2$, $y = 2x + x^2$. If

Figure 2: $\Lambda(x,y)=(x,y-x^2)$, $\eta(t)=t$, $\nu(t)=1$, $\xi(x,y)=0$.



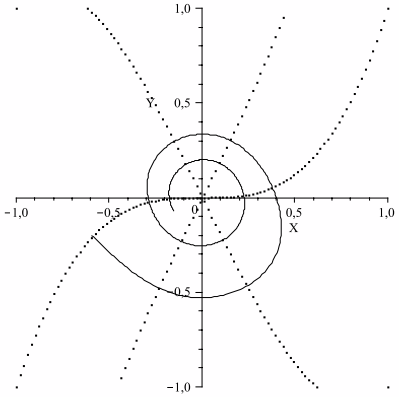
$\Lambda(x,y)=(x,y-x^3)$, $\eta(t)=t$, $\nu(t)=1$, $\xi(x,y)=\frac{x^2+y^2-1}{500}$, then O is a focus of (12):

$$\dot{x}=-\frac{(y+x^2)(x^4-2yx^2+y^2+x^2)}{x^2+y^2+2yx^2+x^4}+\frac{x(x^2+y^2-1)}{500(x^2+y^2-2x^3y+x^6)},$$

$$\dot{y}=\frac{x(x^4-2yx^2+y^2+x^2)}{x^2+y^2+2yx^2+x^4}+\frac{(2x^3+y)(x^2+y^2-1)}{500(x^2+y^2-2x^3y+x^6)},$$

Its isochrones are the cubics $ax+b(y-x^3)=0$. In Figure 3 we show a spiralling orbit and the isochrones contained in $y=-2x+x^3$, $y=x^3$, $y=2x+x^3$. The

Figure 3: $\Lambda(x,y)=(x,y-x^3)$, $\eta(t)=t$, $\nu(t)=1$, $\xi(x,y)=\frac{x^2+y^2-1}{500}$.



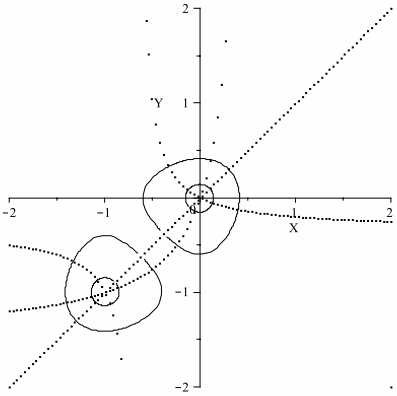
last two examples are constructed starting with globally invertible maps. This is not the case with next one, where we use the map $\Lambda(x,y) = (x+xy, y+xy)$, $\eta(t) = t$, $\nu(t) = 1$, $\xi(x,y) = 0$. Λ is only locally invertible at O, where we find a family of local isochrones defined by $a(x+xy) + b(y+xy) = 0$, $a, b \in \mathbb{R}$. Moreover, there exist other isochrones defined by the same equation, passing through the point $(-1,-1)$, where the system has another center.

$$\dot{x} = -\frac{y(1+y)(x^2+y^2+2xy^2+2x^2y+2x^2y^2)}{x^2+y^2+2x^3+2y^3+x^4+y^4},$$

$$\dot{y} = \frac{x(1+x)(x^2+y^2+2xy^2+2x^2y+2x^2y^2)}{x^2+y^2+2x^3+2y^3+x^4+y^4},$$

In Figure 4 we show both centers and the isochrones contained in the curves $x - y = 0$, $(x+xy) + 2(y+xy) = 0$, $-3(x+xy) + (y+xy) = 0$.

Figure 4: $\Lambda(x,y) = (x+xy, y+xy)$, $\eta(t) = t$, $\nu(t) = 1$, $\xi(x,y) = 0$.



The above procedure may not be the most efficient way to find a system with a given family of isochrones, in particular if one is looking for systems of a special form. In [18] some sufficient conditions for isochronicity of Liénard systems were given. In particular, it was proved that if

$$\sigma(x) = 2x^2 f(x) \int_0^x sf(s)ds - 4\left(\int_0^x sf(s)ds\right)^2 + x^3 g_n(x) - x^4 g_n'(x) \qquad (13)$$

vanishes identically, then all the oscillations around the origin of the Liénard system

$$\dot{x} = y - F(x), \qquad \dot{y} = -g(x), \qquad (14)$$

where $F'(x) = f(x)$, are isochronous. The paper [18] was concerned with centers, but its conclusions are valid for more general systems, since they are based on the properties a differential system equivalent to (14),

$$\dot{x} = y - xb(x), \qquad \dot{y} = -c(x) - yb(x), \tag{15}$$

under some additional conditions. The equivalence conditions of (14) and (15) are the following ones,

$$b(x) = \frac{\int_0^x sf(s)ds}{x^2} = \frac{I(x)}{x^2}, \qquad c(x) = g(x) - xb(x)^2.$$

Without loss of generality we may assume $g(x) = x + h.\ o.\ t.$. In this case the isochronicity condition (13) is equivalent to $c(x) = x$, so that (15) has the following form

$$\dot{x} = y - xb(x), \qquad \dot{y} = -x - yb(x). \tag{16}$$

Such a system has constant angular speed. If $b(x)$ is an odd function, then O is a center, hence an isochronous one. If $b(x)$ is not odd, the system can have a focus at O, with attraction (repulsion) region possibly bounded by a limit cycle or an unbounded orbit. Also, it is possible that several concentric limit cycles surround O. In all such cases, the half-lines starting at the origin are isochrones of (16). These allows to find isochrones for system (14), when (13) holds, since the transformation $(x, y) \mapsto (x, y + F(x) - xb(x))$ takes (15) into (14). Such a transformation is canonical, and its inverse is a canonical normalization of (14). In next theorem, we consider the converse statement. For a special class of curves filling an open region, we find a Liénard system having such curves as isochrones.

THEOREM 3.4. *For every function $d \in C^\infty(\mathcal{I}, \mathbb{R})$, $\mathcal{I}$ open interval containing 0, the Liénard system*

$$\dot{x} = y - (xd(x))', \qquad \dot{y} = -x(1 + d'(x)^2), \tag{17}$$

has the curves

$$y = mx + d(x), \qquad m \in \mathbb{R},$$

as isochrones.

Proof. The isochrones $ax + by = 0$ of (16) are taken into the curves $ax + b(y - F(x) + xb(x)) = 0$, so that the graphs of the functions

$$y = mx + F(x) - xb(x)$$

are isochrones of (14), under the condition (13). Imposing the equality $F(x) - xb(x) = d(x)$ leads to

$$d(x) = F(x) - xb(x) = F(x) - \frac{\int_0^x sf(s)ds}{x}.$$

Multiplying the first and last terms by x and differentiating, one has

$$F(x) = (xd(x))' = d(x) + xd'(x).$$

Substituting this expression into $d(x) = F(x) - xb(x)$ one obtains $b(x) = d'(x)$. In order to find an isochronous system having the curves $y = mx + d(x)$ as isochrones, we have to find $g(x)$ such that (13) holds. From [18] one has the isochronicity condition that relates $g(x)$ to $f(x)$. If $g'(0) = 1$, one has

$$g(x) = x + \frac{1}{x^3}\left(\int_0^x sf(s)ds\right)^2 = x + \frac{I(x)^2}{x^3}.$$

Since, from what above, $I(x) = x(F(x) - d(x))$, one has

$$\frac{I(x)^2}{x^3} = \frac{x^2(F(x) - d(x))^2}{x^3} = \frac{(xd'(x))^2}{x} = xd'(x)^2,$$

that gives

$$g(x) = x + xd'(x)^2.$$

□

System (17) is equivalent to the Liénard equation

$$\ddot{x} + (xd(x))''\dot{x} + x(1 + d'(x)^2) = 0.$$

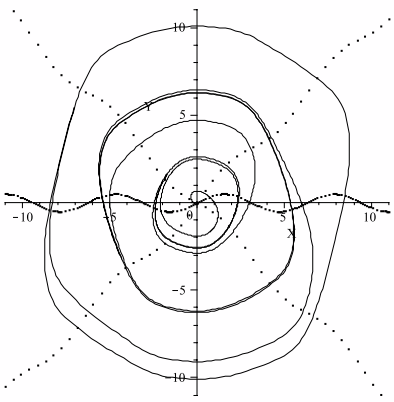
The function $d(x)$ determines the above system's dynamics. If $d(x)$ is even, then $F(x) = (xd(x))'$ is even, hence the origin is a center. If $d(x)$ is not even, then the origin is a focus.

In Figure 1 we have chosen $d(x) = \dfrac{\sin x}{2}$, and plotted the orbits of (14) as continuous lines. The dotted lines are the isochrones contained in $y = -x + \dfrac{\sin x}{2}$, $y = \dfrac{\sin x}{2}$, $y = x + \dfrac{\sin x}{2}$. The figure shows three limit cycles and six isochrones. Presumably the system has infinitely many limit cycles all meeting such isochrones.

After finding the explicit form of system (17), one can check that it normalizes a transversal system. By Lemma 3.2, it is sufficient to find two functions K and ν such that $\dot{K} = \nu(K)$. This implies that the hamiltonian system having K as hamiltonian is normalized by (17). Since the isochrones can be seen as the level sets of the function $H(x, y) = \dfrac{y - d(x)}{x}$, for $x \neq 0$, one can take $K(x, y) = \arctan\left(\dfrac{y - d(x)}{x}\right)$. The derivative of $H(x, y)$ along the solutions of (17) is

$$\dot{H} = -\frac{(y - d(x))^2 + x^2}{x^2} = -H^2 - 1,$$

Figure 5: $d(x) = \frac{\sin x}{2}$.



hence one has

$$\dot{K} = -1.$$

The hamiltonian system having K as hamiltonian function is

$$\dot{x} = \frac{x}{x^2 + (y - d(x))^2}, \qquad \dot{y} = \frac{y - d(x) + xd'(x)}{x^2 + (y - d(x))^2}.$$

Its orbits are the system's isochrones.

References

[1] J. Chavarriga and M. Sabatini, *A survey of isochronous centers*, Qual. Theory Dyn. Syst. **1** (1999), 1–70.

[2] C. Chicone and W. Liu, *Asymptotic phase revisited*, J. Differential Equations **204** (2004), 227–246.

[3] C. Chicone and R. Swanson, *Linearization via the Lie derivative*, Electronic Journal of Differential Equations. Monograph, 02. Southwest Texas State University, San Marcos, TX, 2000. Front matter + 64 pp. (electronic).

[4] F. Dumortier, *Asymptotic phase and invariant foliations near periodic orbits*, Proc. Amer. Math. Soc. **134** (2006), 2989–2996.

[5] I. Garcia, J. Giné and S. Maza, *Linearization of smooth planar vector fields around singular points via commuting flows*, Commun. Pure Appl. Anal. **7** (2008), 1415–1428.

[6] I. Garcia and M. Grau, *Linearization of analytic isochronous centers from a given commutator*, J. Math. Anal. Appl. **339** (2008), 740–745.

[7] I. Garcia and S. Maza, *A survey on the inverse integrating factor*, Qual. Theory Dyn. Syst. **9** (2010), 115–166.

[8] J. Giné and M. Grau, *Characterization of isochronous foci for planar analytic differential systems*, Proc. Roy. Soc. Edinburgh Sect. A **135** (2005), 985–998.
[9] J. Giné and S. Maza, *Lie symmetries for the orbital linearization of smooth planar vector fields around singular points*, J. Math. Anal. Appl. **345** (2008), 63–69.
[10] J. Guckenheimer, *Isochrons and phaseless sets*, J. Math. Biol. **1** (1974/75), 259–273.
[11] M. W. Hirsch, C. C. Pugh and M. Shub, *Invariant manifolds*, Lecture Notes in Mathematics, Vol. 583, Springer, Berlin, 1977.
[12] F. Mañosas and J. Villadelprat, *Area-preserving normalizations for centers of planar Hamiltonian systems*, J. Differential Equations **179** (2002), 625–646.
[13] P. Mardesic, C. Rousseau and B. Toni, *Linearization of isochronous centers*, J. Differential Equations **121** (1995), 67–108.
[14] L. Mazzi and M. Sabatini, *A characterization of centers via first integrals*, J. Differential Equations **76** (1998), 222–237.
[15] L. Mazzi and M. Sabatini, *Commutators and linearizations of isochronous centers*, Atti Accad. Naz. Lincei Cl. Sci. Fis. Mat. Natur. Rend. Lincei (9) Mat. Appl. **1** (2000), 81–98.
[16] M. Sabatini, *Characterizing isochronous centres by Lie brackets*, Differ. Equ. Dyn. Syst. **5** (1997), 91–99.
[17] M. Sabatini, *The time of commuting systems*, preprint, Univ. di Trento, 1996.
[18] M. Sabatini, *On the period function of Liénard systems*, J. Differential Equations **152** (1999), 467–487.
[19] M. Sabatini, *Non-periodic isochronous oscillations in plane differential systems*, Ann. Mat. Pura Appl. (4) **182** (2003), 487–501.
[20] M. Sabatini, *Isochronous sections via normalizers*, Internat. J. Bifur. Chaos Appl. Sci. Engrg. **15** (2005), 3031–3037.
[21] M. Villarini, *Regularity properties of the period function near a centre of a planar vector field*, Nonlinear Anal. **19** (1992), 787–803.
[22] M. Villarini, *Smooth linearizations of centres*, Ann. Fac. Sci. Toulouse Math. (6) **9** (2000), 565–570.

Author's address:

Marco Sabatini
Dipartimento di Matematica
Università di Trento
I-38050 Povo (TN) - Italy.
E-mail: marco.sabatini@unitn.it

Received April 18, 2012
Revised August 30, 2012

Rend. Istit. Mat. Univ. Trieste
Volume 44 (2012), 173–185

A global bifurcation result for a second order singular equation[1]

Anna Capietto, Walter Dambrosio
and Duccio Papini

Dedicated, with gratefulness and friendship, to Professor Fabio Zanolin on the occasion of his 60th birthday

Abstract. *We deal with a boundary value problem associated to a second order singular equation in the open interval* $(0,1]$. *We first study the eigenvalue problem in the linear case and discuss the nodal properties of the eigenfunctions. We then give a global bifurcation result for nonlinear problems.*

Keywords: self-adjoint singular operator, spectrum, nodal properties, global bifurcation
MS Classification 2010: 34C23, 34B09, 35P05

1. Introduction

We are concerned with a second order ODE of the form

$$-u'' + q(x)u = \lambda u + g(x,u)u,\ \lambda \in \mathbb{R},\ x \in (0,1], \tag{1}$$

where $q \in C((0,1])$ satisfies

$$\lim_{x\to 0^+} \frac{q(x)}{l/x^\alpha} = 1\,, \tag{2}$$

for some $l > 0$ and $\alpha \in (0, 5/4)$, and $g \in C([0,1]\times\mathbb{R})$ is such that

$$\lim_{u\to 0} g(x,u) = 0, \quad \text{uniformly in } x \in (0,1]. \tag{3}$$

The constant $5/4$ arises in a rather straightforward manner in the study of the differential operator in the left-hand side of (1) (cf. [17, p. 287-288]); details are given in Remark 2.3 below.

[1]Under the auspices of GNAMPA-I.N.d.A.M., Italy. The work has been performed in the frame of the M.I.U.R. Projects 'Topological and Variational Methods in the Study of Nonlinear Phenomena' and 'Nonlinear Control: Geometrical Methods and Applications'.

We will look for solutions u of (1) such that $u \in H_0^2(0,1)$.

When the x-variable belongs to a compact interval, problems of the form (1) have been very widely studied. A more limited number of contributions is available in the literature when the x-variable belongs to a (semi)-open interval, as it is the case in the present paper, or to an unbounded interval [7, 8].

We treat (1) in the framework of bifurcation theory. For this reason, we first discuss in Section 2 the eigenvalue problem

$$-u'' + q(x)u = \lambda u, \quad x \in (0,1],\ \lambda \in \mathbb{R}. \tag{4}$$

For such singular problems, the well-known embedding of (4) (by an elementary application of the integration by parts rule, together with the boundary condition $u(0) = 0 = u(1)$) in the setting of eigenvalue problems for compact self-adjoint operators cannot be performed. Thus, the questions of the existence of eigenvalues and of the nodal properties of the associated eigenfunctions have various delicate features. For a comprehensive account on the spectral properties of the Schrödinger operator we refer to the books [12] and [10]; for more specific results on singular problems in $(0,1)$ we refer, among many others, to [5, 14].

However, the linear spectral theory for singular problems is well-established and can be found, among others, in the classical book by Coddington and Levinson [4] and in the (relatively) more recent text by Weidmann [17]. The former monograph focuses on a generalization of the so-called "expansion theorem" valid for functions in $L^2([0,1])$ and, by doing this, a sort of "generalized shooting method" is performed. On the other hand, in [17] the singular problem is tackled from an abstract point of view; more precisely, it is considered the general question of the existence of a self-adjoint realization of the formal differential expression $\tau u = -u'' + q(x)u$ and the important Weyl alternative theorem [17, Theorem 5.6] is used. It is interesting to observe that the approach in [4] (based on more elementary ODE techniques) and the abstract one in [17] lead in different ways to the important concepts of "limit point case" and "limit circle case". The knowledge of one (or the other) case is ensured by suitable assumptions on q and leads to information on the boundary conditions to be added to (4) in order to have a self-adjoint realization of τ.

In the setting of the present paper, the operator τ is regular at $x = 1$; this implies that it is in the limit circle case. Moreover, under assumption (2), from [17, Theorem 6.4] it follows that τ is in the limit circle case also in $x = 0$. Thus, the differential operator $A : u \mapsto \tau u$ with

$$D(A) = \{u \in L^2(0,1) :\ u, u' \in AC(0,1),\ \tau u \in L^2(0,1),\ \lim_{x \to 0^+} (xu'(x) - u(x)) = 0 = u(1)\}$$

is a self-adjoint realization of τ ([17, p. 287-288]). We prove in Proposition 2.2 that in fact $D(A) = H_0^2(0,1)$; to do this, we need some knowledge of the behaviour of the solutions of (4) near zero. These estimates are developed in Proposition 2.1 by means of the classical Levinson theorem [6, Theorem 1.8.1]. Finally, at the end of Section 2 we focus on the nodal properties of a solution to (4); more precisely, in Proposition 2.4 we prove that (4) is non-oscillatory and conclude in Proposition 2.5 that the spectrum of A is purely discrete and that, for every $n \in \mathbb{N}$, the eigenfunction associated to the eigenvalue λ_n has $(n-1)$ simple zeros in $(0,1)$.

Section 3 contains a global bifurcation result (Theorem 3.2) which follows in a rather straightforward manner as an application of the celebrated Rabinowitz theorem in [11].

In order to exclude alternative (2) in Theorem 3.2, we use a technique that we already applied for Hamiltonian systems in $\mathbb{R}^{2N}$ in [2] and for planar Dirac-type systems in [3]. More precisely, we introduce a continuous integer-valued functional defined on the set of solutions to (1). Due to the singularity at $x = 0$, some care is necessary in order to prove its continuity; this is the content of Proposition 3.4. We can then state and prove our main result (Theorem 3.5).

In what follows, for a given function p we write $p(x) \sim \dfrac{m}{x^a}, x \to 0^+$, when

$$\lim_{x \to 0^+} \frac{p(x)}{m/x^a} = 1 \tag{5}$$

for some $m, a \in \mathbb{R}^+$.

Finally, we write

$$H_0^2(0,1) = \{u \in H^2(0,1) : \ u(0) = 0 = u(1)\},$$

equipped with the norm defined by

$$||u||^2 = ||u||^2_{L^2(0,1)} + ||u''||^2_{L^2(0,1)}, \quad \forall\ u \in H_0^2(0,1).$$

2. The linear equation

In this section we study a linear second order equation of the form

$$-u'' + q(x)u = \lambda u, \quad x \in (0,1],\ \lambda \in \mathbb{R}. \tag{6}$$

We will assume that $q \in C((0,1])$ and that

$$q(x) \sim \frac{l}{x^\alpha}, \quad x \to 0^+, \tag{7}$$

for some $l > 0$ and $\alpha \in (0, 5/4)$. Without loss of generality we may suppose that

$$q(x) > 0, \quad \forall\, x \in (0, 1]. \tag{8}$$

For every $u : (0, 1] \to \mathbb{R}$ we denote by τu the formal expression

$$\tau u = -u'' + q(x)u;$$

First of all, we study the asymptotic behaviour of solutions of (6) when $x \to 0^+$; to this aim, let us introduce the change of variables $t = -\log x$ and let

$$w(t) = u(e^{-t}), \quad \forall\, t > 0.$$

From the relations

$$\begin{aligned} w'(t) &= -e^{-t}u'(e^{-t}) \\ w''(t) &= e^{-t}u'(e^{-t}) + e^{-2t}u''(e^{-t}), \end{aligned} \tag{9}$$

we deduce that u is a solution of (6) on $(0, 1)$ if and only if w is a solution of

$$-w'' - w' + e^{-2t}q(e^{-t})w = \lambda e^{-2t}w \tag{10}$$

on $(0, +\infty)$. Equation (10) can be written in the form

$$Y' = (C + R(t))Y, \tag{11}$$

where $Y = (w, z)^T$ and

$$C = \begin{pmatrix} 0 & 1 \\ 0 & -1 \end{pmatrix}, \quad R(t) = \begin{pmatrix} 0 & 0 \\ e^{-2t}q(e^{-t}) - \lambda e^{-2t} & 0 \end{pmatrix}, \quad \forall\, t > 0. \tag{12}$$

Now, let us observe that C has eigenvalues $\lambda_1 = 0$, $\lambda_2 = -1$ and corresponding eigenvectors $u_1 = (1, 0)$, $u_2 = (1, -1)$ and that $R \in L^1(0, +\infty)$; therefore, an application of [6, Theorem 1.8.1] implies that (11) has two linearly independent solutions Y_1, Y_2 such that

$$\begin{aligned} Y_1(t) &= u_1 + o(1), \quad t \to +\infty, \\ Y_2(t) &= (u_2 + o(1))e^{-t}, \quad t \to +\infty. \end{aligned} \tag{13}$$

As a consequence, we obtain the following result:

PROPOSITION 2.1. *For every $\lambda \in \mathbb{R}$ the equation* (6) *has two linearly independent solutions $u_{1,\lambda}$, $u_{2,\lambda}$ such that*

$$\begin{aligned} &u_{1,\lambda}(x) = 1 + o(1),\ u'_{1,\lambda}(x) = o\left(\frac{1}{x}\right) \quad x \to 0^+, \\ &u_{2,\lambda}(x) = x + o(x),\ u'_{2,\lambda}(x) = 1 + o(1), \quad x \to 0^+, \end{aligned} \tag{14}$$

and $u_{2,\lambda} \in H^2(0,1)$.

For every $f \in L^2(0,1)$ the solutions of $\tau u = f$ are given by

$$u(x) = c_1 u_{1,0}(x) + c_2 u_{2,0}(x) + u_f(x), \quad \forall\ x \in (0,1),\ c_1, c_2 \in \mathbb{R}, \tag{15}$$

where

$$\begin{aligned} &u_f(x) = \textstyle\int_0^x G(x,t) f(t)\, dt, \quad \forall\ x \in (0,1), \\ &G(x,t) = u_{1,0}(t) u_{2,0}(x) - u_{2,0}(t) u_{1,0}(x), \quad \forall\ x \in (0,1),\ t \in (0,1) \end{aligned} \tag{16}$$

fulfill $G \in L^\infty((0,1)^2)$, $u_f(0) = 0 = u'_f(0)$ and $u_f \in H^2(0,1)$.

Proof. The estimates in (14) follow from (9) and (13), while (16) is the usual variation of constants formula. Moreover, from (14) we obtain that $u_{2,\lambda}, u'_{2,\lambda} \in L^2(0,1)$. On the other hand we have

$$q(x) u_{2,\lambda}(x) \sim x^{1-\alpha}, \quad x \to 0^+, \tag{17}$$

which implies that $q u_{2,\lambda} \in L^2(0,1)$, since $\alpha < 5/4$ (cf. Remark 2.3 for comments on this restriction); using the fact that $\tau u_{2,\lambda} = \lambda u_{2,\lambda}$, we deduce that

$$u''_{2,\lambda} = \lambda u_{2,\lambda} - q u_{2,\lambda} \in L^2(0,1).$$

From now on, we will indicate $u_i = u_{i,0}$, $i = 1, 2$. The fact that the function G defined in (16) belongs to the space $L^\infty((0,1)^2)$ is a consequence of the asymptotic estimates (14). Moreover, from (16) we also deduce that $u_f(0) = 0$ and that

$$u'_f(x) = \int_0^x (u_1(t) u'_2(x) - u_2(t) u'_1(x)) f(t)\, dt, \quad \forall\ x \in (0,1), \tag{18}$$

which implies $u'_f(0) = 0$.

Finally, the condition $u_f(0) = 0 = u'_f(0)$ guarantees that $u_f, u'_f \in L^2(0,1)$; as far as the second derivative of u_f is concerned, let us observe that we have

$$\tau u_f = f$$

and so

$$u''_f = f - q u_f. \tag{19}$$

Using the fact that $u_f(0) = 0 = u'_f(0)$ and (7), it follows that $q u_f \in L^2(0,1)$; hence $u_f \in H^2(0,1)$. □

In what follows, we study the spectral properties of suitable self-adjoint realizations of τ; to this aim, let us first observe that the differential operator τ is regular at $x = 1$. As a consequence, it is in the limit circle case at $x = 1$; moreover, from (7), according to [17, Theorem 6.4], τ is in the limit circle case also in $x = 0$.

The differential operator A defined by

$$\begin{aligned} D(A) = \{u \in L^2(0,1) : u, u' \in AC(0,1),\ \tau u \in L^2(0,1), \\ \lim_{x\to 0^+}(xu'(x) - u(x)) = 0 = u(1)\} \\ Au = \tau u, \quad \forall\ u \in D(A), \end{aligned}$$

is then a self-adjoint realization of τ ([17, p. 287-288]). We can show the validity of the following Proposition:

PROPOSITION 2.2. *The relation*

$$D(A) = H_0^2(0,1)$$

holds true. Moreover, A has a bounded inverse $A^{-1} : L^2(0,1) \to H_0^2(0,1)$.

Proof. 1. Let us start proving that $H_0^2(0,1) \subset D(A)$. It is well known that $H_0^2(0,1) \subset C^1(0,1)$; hence, for every $u \in H_0^2(0,1)$ we have $u, u' \in AC(0,1)$. Moreover, using the fact that $u(0) = 0$ we deduce that

$$u(x) = u'(0)x + o(x), \quad x \to 0^+$$

and

$$q(x)u(x) = u'(0)x^{1-\alpha} + o(x^{1-\alpha}), \quad x \to 0^+;$$

the condition $\alpha < 5/4$ guarantees again that $qu \in L^2(0,1)$ and therefore $\tau u = -u'' + qu \in L^2(0,1)$. Finally, the regularity of u and u' imply that

$$\lim_{x\to 0^+}(xu'(x) - u(x)) = 0$$

and so also the boundary condition in the definition of $D(A)$ is satisfied.

Now, let us prove that $D(A) \subset H_0^2(0,1)$; for every $u \in D(A)$ let $f = \tau u \in L^2(0,1)$. From (15) we deduce that u can be written as

$$u = c_1 u_1 + c_2 u_2 + u_f, \tag{20}$$

for some $c_1, c_2 \in \mathbb{R}$; it is easy to see that the function u_1 does not satisfy the boundary condition given in $x = 0$ in the definition of $D(A)$, while u_2 and u_f do. Hence $u \in D(A)$ if and only if $c_1 = 0$; the last statement of Proposition 2.1 implies then that $u \in H^2(0,1)$. As in the first part of the proof, the regularity

of u allows to conclude that the boundary condition in $x = 0$ given in $D(A)$ reduces to $u(0) = 0$.

2. Let us study the invertibility of A; the existence of a bounded inverse of A is equivalent to the fact that $0 \in \rho_A$, being ρ_A the resolvent of A. Since A is self-adjoint on $H_0^2(0,1)$, this follows from the surjectivity of A (cf. [16, Theorem 5.24]); hence, it is sufficient to prove that A is surjective.

To this aim, let us first observe that condition (8) guarantees that 0 cannot be an eigenvalue of A. Now, let us fix $f \in L^2(0,1)$ and let us prove that there exists $u \in H_0^2(0,1)$ such that $Au = f$, i.e. $\tau u = f$; by applying Proposition 2.1 we deduce again that (20) holds true and the same argument of the first part of the proof implies that $c_1 = 0$.

Hence we obtain $u = c_2 u_2 + u_f$; from Proposition 2.1 we deduce that this function belongs to $H^2(0,1)$ and satisfies the boundary condition $u(0) = 0$. In order to prove that the missing condition $u(1) = 0$ is fulfilled for every $f \in L^2(0,1)$, let us observe that $u_2(1) \neq 0$, otherwise u_2 would be an eigenfunction of A associated to the zero eigenvalue. Therefore, $u(1) = 0$ is satisfied if

$$c_2 = -\frac{u_f(1)}{u_2(1)},$$

for every $f \in L^2(0,1)$. □

REMARK 2.3. *As for the restriction $\alpha < 5/4$, we observe that for the proofs of Proposition 2.1 and Proposition 2.2 it is sufficient to require the milder condition $\alpha < 3/2$. The fact that $\alpha < 5/4$ is used (cf. [17, p. 287-288]) in order to obtain that $D(A)$ is the one described above. Finally, we observe that in the particular case when $\alpha < 1$ the problem is regular (cf., among others, [9]).*

The spectral properties of A are related to the oscillatory behaviour of solutions of (6). We first recall the following definition:

DEFINITION 2.4. *The differential equation* (6) *is oscillatory if every solution u has infinitely many zeros in* $(0,1)$. *It is non-oscillatory when it is not oscillatory.*

We observe that the regularity assumptions on q imply that solutions of (6) have a finite number of zeros in any interval of the form $[a,1)$, for every $0 < a < 1$. Moreover, from (7) we infer that for every $\lambda \in \mathbb{R}$ there exists $c(\lambda) \in (0,1]$ such that

$$\lambda - q(x) < 0, \quad \forall\, x \in (0, c(\lambda)).$$

An application of the Sturm comparison theorem proves that every solution of (6) has at most one zero in $(0, c(\lambda))$; as a consequence, we obtain the following result:

PROPOSITION 2.5. *For every $\lambda \in \mathbb{R}$ the differential equation* (6) *is non-oscillatory.*

Once Proposition 2.5 is obtained, we can provide in a straightforward way some useful information on the spectral properties of A; more precisely, denoting by σ_{ess} the essential spectrum of a given operator, we have:

PROPOSITION 2.6. *([17, Theorem 14.3, Theorem 14.6 and Theorem 14.9], [12, Theorem XIII.1]) The differential operator A is bounded-below and satisfies*

$$\sigma_{ess}(A) = \varnothing.$$

Moreover, there exists a sequence $\{\lambda_n\}_{n\in\mathbb{N}}$ of simple eigenvalues of A such that

$$\lim_{n\to+\infty} \lambda_n = +\infty$$

and for every $n \in \mathbb{N}$ the eigenfunction u_n of A associated to the eigenvalue λ_n has $(n-1)$ simple zeros in $(0,1)$.

REMARK 2.7. *According to [17], operators of the form τ (defined on functions whose domain is $(0,+\infty)$) arise when the time independent Schrödinger equation with spherically symmetric potential*

$$-\Delta u(x) + V(|x|)u(x) = \lambda u(x), \quad u \in L^2(\mathbb{R}^m) \tag{21}$$

is reduced to an infinite system of eigenvalue problems associated to the ordinary differential operators in $L^2(0,+\infty)$

$$\tau_i = -\frac{d^2}{dr^2} + \frac{1}{r^2}\Big[i(i+m-2) + \frac{1}{4}(m-1)(m-3)\Big] + V(r)$$

($i \in \mathbb{N}$). In Appendix 17.F of [17] it is treated the case of a potential V satisfying assumptions (which enable to consider Coulomb potentials) that lead to (7). *More precisely, it is shown that for $m = 3, i = 0$ the operator is in the limit circle case at zero and self-adjoint extensions of τ_0 are described.*

3. The main result

In this section we are interested in proving a global bifurcation result for a nonlinear eigenvalue problem of the form

$$-u'' + q(x)u = \lambda u + g(x,u)u, \ \lambda \in \mathbb{R}, \ x \in (0,1], \tag{22}$$

where $q \in C((0,1])$ satisfies (7) and $g \in C([0,1]\times\mathbb{R})$ is such that

$$\lim_{u\to 0} g(x,u) = 0, \quad \text{uniformly in } x \in [0,1]. \tag{23}$$

We will look for solutions u of (22) such that $u \in H_0^2(0,1)$. To this aim, let Σ denote the set of nontrivial solutions of (22) in $H_0^2(0,1) \times \mathbb{R}$ and let $\Sigma' = \Sigma \cup \{(0,\lambda) \in H_0^2(0,1) \times \mathbb{R} : \ \lambda \text{ is an eigenvalue of } A\}$, where A is as in Section 2.

Let M denote the Nemitskii operator associated to g, given by

$$M(u)(x) = g(x, u(x))u(x), \quad \forall\, x \in [0,1],$$

for every $u \in H_0^2(0,1)$. We can show the validity of the following:

PROPOSITION 3.1. *Assume $g \in C([0,1] \times \mathbb{R})$ and (23). Then $M : H_0^2(0,1) \longrightarrow L^2(0,1)$ is a continuous map and satisfies*

$$M(u) = o(||u||), \quad u \to 0. \tag{24}$$

Proof. 1. We first show that $Mu \in L^2(0,1)$ when $u \in H_0^2(0,1)$. When this condition holds, $u \in L^\infty(0,1)$ and the continuity of g implies that there exists $C_u > 0$ such that

$$|g(x, u(x))u(x)| \le C_u, \quad \forall\, x \in [0,1].$$

As a consequence we obtain $Mu \in L^\infty(0,1) \subset L^2(0,1)$.

2. Let us prove that M is continuous. Let us fix $u_0 \in X$ and let $u_n \in X$ such that $u_n \to u_0$ when $n \to +\infty$; the continuous embedding

$$H_0^2(0,1) \subset L^\infty(0,1)$$

and the uniform continuity of g on compact subsets of $[0,1] \times \mathbb{R}$ ensure that

$$g(x, u_n(x)) \to g(x, u_0(x)) \quad \text{in} \ \ L^\infty(0,1). \tag{25}$$

This is sufficient to conclude that $Mu_n \to Mu_0$ in $L^\infty(0,1)$ and hence $Mu_n \to Mu_0$ in $L^2(0,1)$.

3. Finally, let us prove (24): using again the fact that $H_0^2(0,1) \subset L^\infty(0,1)$, we have

$$||Mu||_{L^2(0,1)} \le ||g(x,u(x))||_{L^\infty(0,1)} ||u||_{L^2(0,1)} \le ||g(x,u(x))||_{L^\infty(0,1)} ||u||,$$

for all $u \in H_0^2(0,1)$; hence, we deduce that

$$\frac{||Mu||_{L^2(0,1)}}{||u||} \le ||g(x,u(x))||_{L^\infty(0,1)}, \quad \forall\, u \in H_0^2(0,1),\ u \neq 0.$$

Therefore the result follows from (23) and (25). □

Now, let us observe that the search of solutions $u \in H_0^2(0,1)$ of (22) is equivalent to the search of solutions of the abstract equation

$$Au = \lambda u + M(u), \quad (u,\lambda) \in H_0^2(0,1) \times \mathbb{R}; \tag{26}$$

on the other hand, (26) can be written in the form

$$w = \lambda Rw + M(Rw), \ (w,\lambda) \in L^2(0,1) \times \mathbb{R}, \tag{27}$$

where $R : L^2(0,1) \to H_0^2(0,1)$ is the inverse of A (cf. Proposition 2.2).

Now, from [17, Theorem 7.10] we deduce that R is compact; this fact and the continuity of M guarantee that the operator $MR : L^2(0,1) \to H_0^2(0,1)$ is compact. Moreover, the condition

$$M(Rw) = o(||w||_{L^2(0,1)}), \quad w \to 0, \tag{28}$$

is a consequence of (24). From an application of the global bifurcation result of Rabinowitz (cfr. [11]) to (27) we then obtain the following result:

THEOREM 3.2. *Assume* (7) *and* (23). *Then, for every eigenvalue* λ_n *of* A *there exists a continuum* C_n *of nontrivial solutions of* (22) *in* $H_0^2(0,1)\times\mathbb{R}$ *bifurcating from* $(0,\lambda_n)$ *and such that one of the following conditions holds true:*

(1) C_n *is unbounded in* $H_0^2(0,1) \times \mathbb{R}$*;*

(2) C_n *contains* $(0,\lambda_{n'}) \in \Sigma'$*, with* $n' \neq n$.

Now, let us observe that a more precise description of the bifurcating branch, eventually leading to exclude condition (2), can be obtained when there exists a continuous functional $j : \Sigma' \to \mathbb{N}$ (cf. [2, Pr. 2.1]). In order to define such a functional, we will use the fact that nontrivial solutions of (22) have a finite number of zeros in $(0,1)$; this will be a consequence of our next result.

For every $\lambda \in \mathbb{R}$ and for every nontrivial solution $u \in H_0^2(0,1)$ of (22) let us define $q_{u,\lambda} : (0,1] \to \mathbb{R}$ by $q_{u,\lambda}(x) = q(x) - \lambda - g(x,u(x))$, for every $x \in (0,1]$. The following Lemma holds true:

LEMMA 3.3. *For every* $\lambda \in \mathbb{R}$ *and for every nontrivial solution* $u \in H_0^2(0,1)$ *of* (22) *there exists a neighborhood* $U \subset H_0^2(0,1)\times\mathbb{R}$ *of* (u,λ) *and* $x_{u,\lambda} \in (0,1)$ *such that*

$$q_{v,\mu}(x) > 0, \quad \forall\ (v,\mu) \in U,\ x \in (0, x_{u,\lambda}]. \tag{29}$$

Proof. Let $(u,\lambda) \in H_0^2(0,1)\times\mathbb{R}$, $u \not\equiv 0$, be fixed and let U be the neighborhood of radius 1 of (u,λ) in $H_0^2(0,1)\times\mathbb{R}$; from the continuous embedding $L^\infty(0,1) \subset H_0^2(0,1)$ we deduce that if $(w,\mu) \in \Sigma \cap U_1$ then

$$||w||_{L^\infty(0,1)} \leq 1 + ||u||_{L^\infty(0,1)}, \quad |\mu| \leq 1 + |\lambda|$$

and

$$q(x) - \mu - g(x, w(x)) \geq q(x) - |\lambda| - 1 - \max_{\substack{x \in [0,1], \\ |s| \leq 1 + ||u||_{L^\infty(0,1)}}} |g(x,s)|, \quad \forall\, x \in (0,1).$$

From (7) we then deduce that there exists $x_{(u,\lambda)} \in (0,1)$, depending only on (u,λ), such that

$$q(x) - \mu - g(x, w(x)) > 0, \quad \forall\, x \in (0, x_{(u,\lambda)}].$$

□

Now, let us observe that for every $\lambda \in \mathbb{R}$ and for every nontrivial solution $u \in H_0^2(0,1)$ of (22) the function u is a nontrivial solution of the linear equation

$$-w'' + (q(x) - g(x, u(x)) - \lambda)w = 0. \tag{30}$$

From Lemma 3.3, with an argument similar to the one which led to Proposition 2.5, we deduce that all the nontrivial solutions of (30) (in particular u) have a finite number of zeros in $(0,1)$. We denote by $n(u)$ this number.

We are then allowed to define the functional j by setting

$$j(u,\lambda) = \begin{cases} n(u) & \text{if } u \not\equiv 0 \\ \\ n-1 & \text{if } u \equiv 0 \text{ and } \lambda = \lambda_n, \end{cases} \tag{31}$$

for every $(u,\lambda) \in \Sigma'$. Let us observe that the definition $j(0,\lambda_n) = n-1$ is suggested by Proposition 2.6.

PROPOSITION 3.4. *The function $j : \Sigma' \to \mathbb{N}$ is continuous.*

Proof. 1. As for the continuity of j in every point of the form $(0,\lambda_n)$, $n \in \mathbb{N}$, we refer to [15, Lemma 2.5].

2. Let us now fix $(u_0,\lambda_0) \in \Sigma$ and let $(u,\lambda) \in U$, with U as in Lemma 3.3; this Lemma guarantees that both u and u_0 have no zeros in $(0, x_{u_0,\lambda_0})$.

On the other hand, in the interval $[x_{u_0,\lambda_0}, 1]$ a standard continuous dependence argument (cf. also [11]) ensures that u and u_0 have the same numbers of zeros if (u,λ) is in a sufficiently small neighborhood of (u_0,λ_0). As a consequence, we obtain that there exists a neighborhood U_0 of (u_0,λ_0) such that

$$j(u,\lambda) = j(u_0,\lambda_0), \quad \forall\, (u,\lambda) \in U_0.$$

□

As a consequence, from Theorem 3.2 and Proposition 3.4 we deduce the final result:

THEOREM 3.5. *Assume* (7) *and* (23). *Then, for every eigenvalue* λ_n *of* A *there exists a continuum* C_n *of nontrivial solutions of* (22) *in* $H_0^2(0,1)\times\mathbb{R}$ *bifurcating from* $(0,\lambda_n)$ *and such that condition (1) of Theorem 3.2 holds true and*

$$j(u,\lambda) = n-1, \quad \forall\ (u,\lambda) \in C_n. \tag{32}$$

REMARK 3.6. *Theorem 3.2 can be proved as an application of Stuart's result [15, Theorem 1.2] as well. However, since in the situation considered in this paper the singularity at zero does not affect the compactness of the operator* R *defined after* (27)*, we chose to apply Rabinowitz theorem [11]. We finally mention the interesting paper [1], where global branches of solutions, with prescribed nodal properties, are obtained for a second order degenerate problem in* $(0,1)$.

REFERENCES

[1] H. BERESTYCKI AND M.J. ESTEBAN, *Existence and bifurcation of solutions for an elliptic degenerate problem*, J. Differential Equations **134** (1997), 1–25.

[2] A. CAPIETTO AND W. DAMBROSIO, *Preservation of the Maslov index along bifurcating branches of solutions of first order systems in* $\mathbb{R}^n$, J. Differential Equations **227** (2006), 692–713.

[3] A. CAPIETTO AND W. DAMBROSIO, *Planar Dirac-type systems: the eigenvalue problem and a global bifurcation result*, J. London Math. Soc. **81** (2010), 477–498.

[4] E.A. CODDINGTON AND N. LEVINSON, *Theory of Ordinary Differential Equations*, McGraw-Hill, 1955.

[5] B. ĆURGUS AND T.T. READ, *Discreteness of the spectrum of second-order differential operators and associated embedding theorems*, J. Differential Equations **184** (2002), 526–548.

[6] M.S.P. EASTHAM, *The Asymptotic Solution of Linear Differential Systems*, London Math. Society Monographs New Series, 1989.

[7] P. FELMER AND J.J. TORRES, *A nonlinear eigenvalue problem in* $\mathbb{R}$ *and multiple solutions of nonlinear Schrödinger equation*, Adv. Differential Equations **7** (2002), 1215–1234.

[8] F. HADJ SELEM, *Radial solutions with prescribed numbers of zeros for the nonlinear Schrödinger equation with harmonic potential*, Nonlinearity **24** (2011), 1795–1819.

[9] R. LEMMERT AND W. WALTER, *Singular nonlinear boundary value problems*, Appl. Anal. **72** (1999), 191–203.

[10] D.B. PEARSON, *Quantum Scattering and Spectral Theory*, Academic Press, London, 1988.

[11] P. RABINOWITZ, *Some global results for non-linear eigenvalue problems*, J. Funct. Anal. **7** (1971), 487–513.

[12] M. REED AND B. SIMON, *Methods of Modern Mathematical Physics. Vol. 4: Analysis of Operators*, Academic Press, London, 1978.

[13] H. SCHMID AND C. TRETTER, *Eigenvalue accumulation for Dirac operators with spherically symmetric potential*, J. Differential Equations **181** (2002), 511–542.

[14] I. Sim, R. Kajikiya, and Y.-H. Lee, *On a criterion for discrete or continuous spectrum of p-Laplace eigenvalue problems with singular sign-changing weights*, Nonlinear Anal. **72** (2010), 3515–3534.
[15] C. Stuart, *Global properties of components of solutions of non-linear second order differential equations on the half-line*, Ann. Scuola Norm. Sup. Pisa Cl. Sci. (4) **2** (1975), 265–286.
[16] J. Weidmann, *Linear Operators in Hilbert Spaces*, Graduate Texts in Mathematics, no. 68, Springer, Berlin, 1980.
[17] J. Weidmann, *Spectral Theory of Ordinary Differential Equations*, Lectures Notes in Mathematics, no. 1258, Springer, Berlin, 1987.

Authors' addresses:

Anna Capietto
Dipartimento di Matematica
Università di Torino
Via Carlo Alberto 10, 10123 Torino, Italy
E-mail: anna.capietto@unito.it

Walter Dambrosio
Dipartimento di Matematica
Università di Torino
Via Carlo Alberto 10, 10123 Torino, Italy
E-mail: walter.dambrosio@unito.it

Duccio Papini
Dipartimento di Ingegneria dell'Informazione e Scienze Matematiche,
Università di Siena
Via Roma 56, 53100 Siena, Italy
E-mail: papini@dii.unisi.it

Received May 28, 2012
Revised September 3, 2012

Rend. Istit. Mat. Univ. Trieste
Volume 44 (2012), 187–195

An improvement of Massera's theorem for the existence and uniqueness of a periodic solution for the Liénard equation

Gabriele Villari

A Fabio Zanolin, "amico di una vita", per i suoi splendidi 60 anni.

Abstract. *In this paper we prove the existence and uniqueness of a periodic solution for the Liénard equation*

$$\ddot{x} + f(x)\dot{x} + x = 0.$$

The classical Massera's monotonicity assumptions, which are required in the whole line, are relaxed to the interval (α, δ), where α and δ can be easily determined. In the final part of the paper a simple perturbation criterion of uniqueness is presented.

Keywords: Liénard equation, limit cycle
MS Classification 2010: 34C25

1. Preliminaries and well-known results

The problem of existence and uniqueness of a periodic solution for the Liénard equation,

$$\ddot{x} + f(x)\dot{x} + x = 0, \tag{1}$$

has been widely investigated in the literature. Among the uniqueness results, the most interesting and intriguing one is, without any doubt, the classical Massera's Theorem. This is due to the geometrical ideas and the fact that this result, despite several efforts, is in most cases no more valid for the generalized Liénard equation

$$\ddot{x} + f(x)\dot{x} + g(x) = 0. \tag{2}$$

For related results still valid for equation (2), we refer to [1], and to [3] for the equation

$$\ddot{x} + f(x, \dot{x})\dot{x} + x = 0.$$

Throughout this paper we assume that

(A) f is continuous and there exist $a < 0 < b$ such that $f(x)$ is negative for $a < x < b$, positive outside this interval. Moreover $xF(x) > 0$ for $|x|$ large.

It is well-known (see, for instance, [14, Theorem 1]), that such condition guarantees the existence of at least a stable limit cycle.

Equation (1) is equivalent to the phase-plane system

$$\begin{cases} \dot{x} = y \\ \dot{y} = -f(x)y - x. \end{cases} \tag{3}$$

We just notice that assumption **(A)** guarantees the property of uniqueness for the solutions to the Cauchy problem associated to system (3) and therefore the trajectories of such a system cannot intersect.

The phase-plane system is equivalent to the Liénard system

$$\begin{cases} \dot{x} = y - F(x) \\ \dot{y} = -x \end{cases}, \qquad \text{where } F(x) = \int_0^x f(t)\,dt. \tag{4}$$

For equation (2) system (3) becomes

$$\begin{cases} \dot{x} = y \\ \dot{y} = -f(x)y - g(x), \end{cases} \tag{5}$$

while system (4) becomes

$$\begin{cases} \dot{x} = y - F(x) \\ \dot{y} = -g(x) \end{cases}, \qquad \text{where } F(x) = \int_0^x f(t)\,dt. \tag{6}$$

It is well-known that the nonlinear transformation $(x, y + F(x))$ takes points of system (3) in points of system (4). Such a transformation preserves the x-coordinate and this will be crucial for the proof of the main result.

Now we define the property **(B)**

(B) $F(x)$ has three zeros at $\alpha < 0$, 0, $\beta > 0$. Moreover $xF(x)$ is negative for $\alpha < x < \beta$ and positive outside this interval, while F is monotone increasing for $x < \alpha$ and $x > \beta$ (see Figure 1).

We observe that property **(A)** implies property **(B)** and that property **(B)** can be assumed even if $f(x)$ changes sign several times in the interval (α, β),

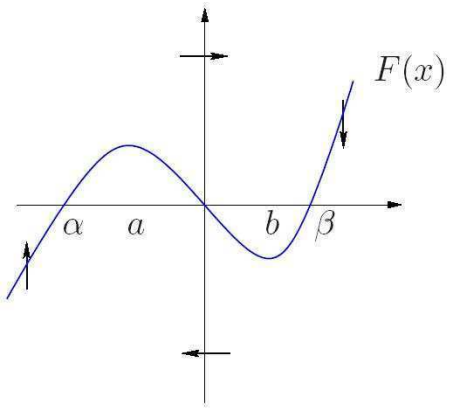


Figure 1:

which is not our case. Finally we notice that it is trivial to show that in system (4) and in system (6) the distance from the origin is increasing when $xF(x) < 0$, while is decreasing when $xF(x) > 0$.

We present the classical Massera Theorem which is a milestone among the results of limit cycles uniqueness for system (3).

THEOREM 1.1 (Massera [8]). *The system (3) has at most one limit cycle which is stable, and hence equation (1) has at most one non trivial periodic solution which is stable, provided that f is continuous and*

1. *$f(x)$ is monotone decreasing for $x < 0$,*
2. *$f(x)$ is monotone increasing for $x > 0$.*

The Theorem of Massera improved a previous result due to Sansone [12] in which there was the additional assumption $|f(x)| < 2$. This assumption comes from the fact that Sansone was using the polar coordinates. Such strong restriction on f is clearly not satisfied in the polynomial case and hence the Massera's result is much more powerful. We recall the recent paper [11] in which a discussion concerning these two results, as well as related results, may be found.

We must observe that in his paper, Massera was proving the uniqueness of limit cycles regardless the existence because only the monotonicity properties and the continuity were required. It is easy to prove that, in order to fulfill the necessary conditions for the existence of limit cycles, the only cases to be considered are

1. $f(x)$ has two zeros $a < 0 < b$. In this case property **(A)** is fulfilled and hence the existence of limit cycles is granted,

2. $f(x)$ remains negative for $x < 0$, (or for $x > 0$), while intersects the x axis once in $x > 0$ (or for $x < 0$).

In this case the existence of limit cycles is not granted. It is possible to produce examples in which, actually, there exists a unique limit cycle but, as far as we know, there is no existence result which can be applied in this situation. Moreover this case does not cover the crucial polynomial case, which is still the most important and it is related with the Lins-De Melo-Pugh conjecture [7], concerning the upper bound of limit cycles for equation (1) when $f(x)$ is a polynomial of degree n.

Now we recall another interesting result, which is due to Levinson-Smith for system (6) and to Sansone for system (4).

THEOREM 1.2 (Levinson-Smith [5] and Sansone [13], see also [15]). *If F has the property **(B)**, at most a limit cycle intersects both the lines $x = \alpha$ and $x = \beta$*

This is a very nice result, but it is abstract, because, in general, if there are no symmetry properties on f and g, such a situation is not easy to be verified. For system (6) there are sufficient conditions which guarantee that in the Liénard plane this situation actually occurs (see [2, 15] and, for more general cases, [1, 10, 17]). In the case of system (4) a sufficient condition is $|\alpha| = \beta$.

The aim of this paper is to relax the monotonicity assumptions, required by Massera, to a fixed interval given by the function f.

This will be achieved working both in the phase plane and in the Liénard plane and using property **(B)** and Theorem 1.2, together with Massera's Theorem.

Proofs are based on elementary phase plane analysis, but as far as we know, the result is original and this shows how still this classical problem deserves to be investigated.

In the final part of the paper, an existence and uniqueness result will be presented for the equation, depending on a parameter λ,

$$\ddot{x} + \lambda f(x)\dot{x} + x = 0.$$

2. The main result

We now present our result which improves the classical Massera Theorem when property **(A)** holds.

THEOREM 2.1 (Massera "revisited"). *Under the assumptions **(A)**, the Liénard system (4) has exactly one limit cycle, which is stable, provided that*

1. $|\alpha| > \beta$,

$f(x)$ *is monotone decreasing for* $\alpha < x < 0$,

$f(x)$ is monotone increasing for $0 < x < \delta$;

2. $|\alpha| < \beta$,

$f(x)$ is monotone decreasing for $\delta_1 < x < 0$,

$f(x)$ is monotone increasing for $0 < x < \beta$,

with

$$\delta = \sqrt{\left(1 + F(a) + \frac{\alpha^2}{2}\right)^2 + \beta^2}\,, \quad \delta_1 = -\sqrt{\left(-F(b) + 1 + \frac{\beta^2}{2}\right)^2 + \alpha^2},$$

where a and b are the zeros of $f(x)$ and α, β are the non trivial zeros of $F(x)$.

Proof. We preliminarly observe that, if $|\alpha| = \beta$, we can apply directly Theorem 1.2 and no monotonicity assumptions are required.

For sake of simplicity we are proving the theorem in several steps.

Step 1 We now consider the case $|\alpha| > \beta$.

*Under the assumption **(A)**, if $f(x)$ is monotone decreasing for $\alpha < x < 0$, and monotone increasing for $x > 0$, the Liénard system (4) has exactly a limit cycle, which is stable.*

In the Liénard plane any trajectory which intersects the line $x = \alpha$ in $y > 0$, also intersects the line $x = \beta$ because, as already mentioned, the distance from the origin is increasing in the strip $\alpha < 0 < \beta$.

If we keep the monotonicity properties of Massera's Theorem for $x > \alpha$, we know that, in the half plane $x > \alpha$, lies at most a stable limit cycle. This result is proved in the phase plane, but it also holds in the Liénard plane in virtue of the above mentioned property which preserves the x-coordinate, when one switches from one plane to the other. Hence in the Liénard plane there are only two possible configurations:

1. No limit cycle lies in the half plane $x > \alpha$. Hence all limit cycles must intersect both lines $x = \alpha$ and $x = \beta$ and, from Theorem 1.2, the limit cycle is unique.

2. We have a stable limit cycle in the half plane $x > \alpha$. Using again Theorem 1.2 we can have, at most, a second limit cycle intersecting both lines $x = \alpha$ and $x = \beta$. The sign conditions on f shows that such limit cycle must be semistable from his exterior. Using a perturbation argument, which may be found in [7] and [16], one can see that, with a suitable small perturbation of f near α and for $x < \alpha$, still keeping f positive and hence keeping the monotonicity properties of F required for property **(B)**, the semistable limit cycle bifurcates in two limit cycles, one

stable and one unstable, which is a contradiction because both limit cycles must intersect both the lines $x = \alpha$ and $x = \beta$. For the bifurcation from a semistable limit cycle in rotated vector fields, we refer also to the classical works of Duff [4] and Perko [9].

If $|\alpha| < \beta$ we easily get a dual result, namely:

Step 2 *Under the assumption **(A)**, if $f(x)$ is monotone decreasing for $x < 0$ and monotone increasing for $0 < x < \beta$, the Liénard system (4) has exactly a limit cycle, which is stable.*

In order to complete our proof, it is necessary to produce a fixed upper bound for the monotonicity assumptions for positive values of x.

Step 3 We consider, at first, the case $|\alpha| > \beta$.

*Under assumption **(A)**, a positive semitrajectory of the Liénard system (4), which starts at a point $P(\alpha, F(a) + 1)$, intersects the vertical isocline $y = F(x)$ in the half plane $x > 0$, at a point $S(x, F(x))$, with $x < \delta$, where $\delta = \sqrt{\left(1 + F(a) + \frac{\alpha^2}{2}\right)^2 + \beta^2}$.*

In the Liénard plane (4), the slôpe of a trajectory is given by

$$y'(x, y) = \frac{-x}{y - F(x)}.$$

At first, we observe that a positive semitrajectory, which starts at a point $P(\alpha, F(a) + 1)$, must intersect the y-axis at a point $Q(0, \bar{y})$, because the slôpe is positive, and the line $x = \beta$ at a point $R(\beta, \hat{y})$, due to the fact that, in the strip $\alpha < x < \beta$, the distance from the origin is increasing and $|\alpha| > \beta$ (see Figure 2).

$$y(Q) - y(P) = \int_{\alpha}^{0} y'(x, y)\, dx = \int_{\alpha}^{0} \frac{-x}{y - F(x)}\, dx.$$

In the strip $\alpha < x < 0$, $F(x) \leq F(a)$, the slôpe is positive and, clearly, $y - F(x) \geq y - F(a) > 1$ and therefore

$$y(Q) - y(P) < \int_{\alpha}^{0} -x\, dx = \frac{\alpha^2}{2},$$

that is

$$y(Q) = \bar{y} < 1 + F(a) + \frac{\alpha^2}{2}.$$

In the strip $0 < x < \beta$, the slôpe is negative; for this reason the positive semitrajectory intersects the β-line at a point $R(\beta, \hat{y})$, with $\hat{y} < \bar{y} < 1 + F(a) + \frac{\alpha^2}{2}$. For $x > \beta$, the distance from the origin is now decreasing. The positive semitrajectory intersects the vertical isocline $y = F(x)$ at a point $S(x, F(x))$, with

$$x < \sqrt{\left(1 + F(a) + \frac{\alpha^2}{2}\right)^2 + \beta^2} = \delta,$$

and this proves Step 3.

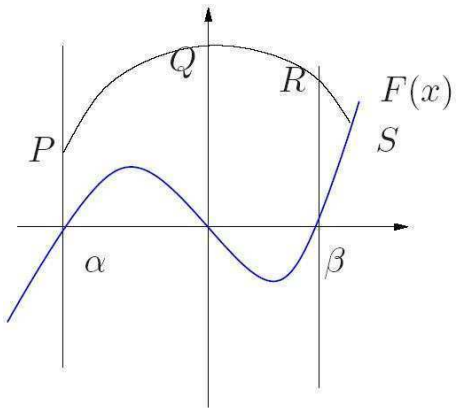


Figure 2:

From Step 3, we get that any negative semitrajectory intersecting the vertical isocline at $x > \delta$ intersects the line $x = \alpha$.

Now we require the monotonicity property of Massera Theorem just in the strip $\alpha < x < \delta$ and we can argue as in Step 1.

Again if $|\alpha| < \beta$, we can get the dual result:

Step 4 *Under assumption* ***(A)****, a positive semitrajectory of the Liénard system (4), which starts at a point* $P(\beta, F(b)-1)$*, intersects the vertical isocline* $y = F(x)$ *in the half plane* $x < 0$*, at a point* $S(x, F(x))$*, with* $x > \delta_1$*, where* $\delta_1 = -\sqrt{\left(-F(b) + 1 + \frac{\beta^2}{2}\right)^2 + \alpha^2}$.

This completes the proof of the Theorem. □

REMARK 2.1. *Observe that it is easy to see that, actually, the value* δ (δ_1) *can be improved by* $\hat{\delta} = F^{-1}\left(\sqrt{\delta^2 - x^2}\right)$ $(\hat{\delta_1} = F^{-1}\left(\sqrt{\delta_1^2 - x^2}\right))$*. However, we prefer to keep the values* δ *and* δ_1 *becuse they explicitly contain the values* a, b, α, β *and this enlights the crucial role played by the zeros of* f *and* F*.*

REMARK 2.2. *Notice that such result can also be viewed as a perturbation of the classical Massera Theorem, namely that we can perturb the function* $f(x)$ *outside the interval* $[\alpha, \delta]$ $([\delta_1, \beta])$*, keeping only the sign conditions, and still having existence and uniqueness of a stable limit cycle.*

REMARK 2.3. *Finally, as a side remark, we recall that outside the interval* $[\alpha, \delta]$ *the only restriction on* $f(x)$ *is the positivity. In the case of* f *tending, at* 0^+*, at infinity and* F *having a finite limit at infinity, still the above mentioned*

sufficient conditions for the existence of limit cycles are fulfilled [14] and the monotonicity assumptions on $[\alpha, \delta]$ *give the uniqueness.*

We already noticed that the values $F(a)$, $F(b)$ play a crucial role in order to guarantee that the trajectories of system (3) intersect both lines $x = \alpha$ and $x = \beta$.

In the light of a result in [2], proved for equation (2), which now is more powerful due to the fact that $g(x) = x$, we prove the following simple perturbation result:

THEOREM 2.2. *Under the assumption* ***(A)*** *the equation*

$$\ddot{x} + \lambda f(x)\dot{x} + x = 0$$

has a unique non trivial periodic solution for every $\lambda \geq \hat{\lambda}$, *where* $\hat{\lambda} =$

$$\sqrt{\frac{\alpha^2 - \beta^2}{F^2(b)}}, \ \textit{if} \ |\alpha| > \beta,$$

$$\sqrt{\frac{\beta^2 - \alpha^2}{F^2(a)}}, \ \textit{if} \ |\alpha| < \beta,$$

any real number if $\alpha = \beta$.

Proof. We consider only the first case, being the second one treated in the same way and the result well-known if $|\alpha| = \beta$.

As usual we consider the Liénard system

$$\begin{cases} \dot{x} = y - \lambda F(x) \\ \dot{y} = -x. \end{cases}$$

We just notice that the parameter λ does not influence the values a, b, α, β. Assumption **(A)** gives the existence of at least a limit cycle. Any positive semitrajectory which intersects the line $x = \beta$ in $y < 0$, intersects the line $x = b$ at a point $P(b, y)$, with $y < \lambda F(b)$. Recalling again the fact that, in the strip $\alpha < x < \beta$, the distance from the origin is increasing, it is straightforward to observe that if

$$\sqrt{\lambda^2 F^2(b) + b^2} \geq |\alpha|,$$

such trajectory intersects the line $x = \alpha$. Hence all limit cycles must intersect both lines $x = \alpha$ and $x = \beta$ and we can use Theorem 1.2 again. □

References

[1] T. Carletti *Uniqueness of limit cycles for a class of planar vector fields*, Qual. Theory Dyn. Syst. **6** (2005), 31–43

[2] T. Carletti and Gab. Villari, *A note on existence and uniqueness of limit cycles for Liénard systems*, J. Math. Anal. Appl. **307** (2005), 763–773.

[3] T. Carletti, L. Rosati, and Gab. Villari, *Qualitative analysis of the phase portrait for a class of planar vector fields via the comparison method*, Nonlinear Anal. **67** (2007), 39–51.

[4] G.F.D. Duff, *Limit-cycles and rotated vector fields*, Ann. of Math. (2) 57 (1953), 15–31.

[5] N. Levinson and O. Smith, *A general equation for relaxation oscillations*, Duke Math. J. **9** (1942), 382–403.

[6] A. Liénard, *Étude des oscillations entretenues*, Revue génér. de l'électr. **23** (1928), 901–902; 906–954.

[7] A. Lins, W. De Melo, and C.C. Pugh, *On Liénard's equation*, Lecture notes in Mathematics 597, Springer, New York 1976, 335–357.

[8] J.L.Massera, *Sur un Théoreme de G. Sansone sur l'equation de Liénard*, Boll. Un. Mat. Ital. (3) **9** (1954), 367–369.

[9] L.M. Perko, *Bifurcation of limit cycles: geometric theory*, Proc. Amer. Math. Soc. **114** (1992), 225–236.

[10] M. Sabatini and Gab. Villari, *About limit cycle's uniqueness for a class of generalized Liénard systems*, Appl. Math. Lett. **19** (11) (2006), 1180–1184.

[11] M. Sabatini and Gab. Villari, *On the uniqueness of limit cycles for Liénard equation: the legacy of G. Sansone*, Matematiche (Catania) **65** (2010), 201–214.

[12] G. Sansone, *Soluzioni periodiche dell'equazione di Liénard. Calcolo del periodo*, Rend. Sem. Mat. Univ. Politec. Torino **10** (1951), 155–171.

[13] G. Sansone, *Sopra l'equazione di Liénard delle oscillazioni di rilassamento*, Ann. Mat. Pura Appl. **28** (4) (1949), 153–181.

[14] Gab. Villari, *On the existence of periodic solutions for Liénard's equation*, Nonlinear Anal. **7** (1983), 71–78.

[15] Gab. Villari, *Some remarks on the uniqueness of the periodic solutions for Liénard's equation*, Boll. Un. Mat. Ital. C (6) **4** (1985), 173–182.

[16] Gab. Villari and M. Villarini, *Limit cycles and bifurcation from a separatrix for a polynomial Liénard system in the plane*, Differ. Equ. Dyn. Syst. **5** (1997), 423–437

[17] Dongmei Xiao and Zhifeng Zhang, *On the uniqueness and nonexistence of limit cycles for predator-prey systems*, Nonlinearity **16** (2003), 1185–1201.

Author's address:

Gabriele Villari
Dipartimento di Matematica "U. Dini"
Università di Firenze
viale Morgagni 67/a, 50134 Firenze, Italy
E-mail: `villari@math.unifi.it`

Received April 11, 2012
Revised September 10, 2012

Rend. Istit. Mat. Univ. Trieste
Volume 44 (2012), 197–257

On the Darboux and Birkhoff steps in the asymptotic stability of solitons

Scipio Cuccagna

Dedicated to Professor Fabio Zanolin on the occasion of his 60th birthday

Abstract. *We give a unified proof of the step to find Darboux coordinates and of the ensuing Birkhoff normal forms procedure, developed in the course of the proof of asymptotic stability of solitary waves in [4, 8, 10].*

Keywords: Hamiltonian, differential forms, flow
MS Classification 2010: 37K45

1. Introduction

The aim of this paper is to extend in a slightly more general and unified set up two important steps of the proof of the asymptotic stability of solitary waves for the Nonlinear Schrödinger equation [2, 8, 10] and the particular case of Nonlinear Dirac system treated in [4]. In both cases there is a localization at the solitary wave and a representation of the system in terms of coordinates arising from the linearization at a solitary wave. The operators $\mathcal{H}_p$ introduced later play this role. In general $\mathcal{H}_p$ has both continuous spectrum and non zero eigenvalues. The latter give rise to discrete modes which in the nonlinear problem could produce chaotic Lissaius like motions. It turns out that in [2, 3, 4, 8, 9, 10] discrete modes relax to 0 because of a mechanism of slow leaking of energy away from the discrete modes into the continuous modes, where energy disperses by linear dispersion. The idea was initiated in special situations in [5, 12, 13]. We refer to [8] for more comments and references.

The aim of this paper consists in simplifying two key steps in the proofs in [4, 8, 10]. The first step consists in searching Darboux coordinates. This allows to decrease the number of coordinates in the system and to reduce to the study of the system at an equilibrium point.

The second step consists in the implementation of the Birkhoff normal forms, to produce a simple *effective* Hamiltonian. After this, [4, 8, 10] prove

the energy leaking away from the discrete modes. In particular the key step is the proof that certain coefficients of the discrete modes equations are second powers, the *Nonlinear Fermi Golden Rule* (FGR), which generically are positive and yield discrete mode energy dissipation.

We do not discuss the FGR in this paper limiting ourselves to the search of Darboux coordinates and to the Birkhoff normal forms argument.

In this paper we avail ourselves with some ideas and notation drawn from early versions of [2] to improve the presentation in [10].

[2, 10] represent two attempts to extend the result proved in [8] for standing ground states of the NLS, to the case of moving ground states. A further goal in [2] is to develop the theory in a more abstract set up. Early versions of [2] did not encompass a Birkhoff step extendable to [4]. [2] is confined (like us here) to systems with Abelian group of symmetries.

The 1st version of the present proof was written before the 3rd version of [2] was posted on the Arxiv site. The 2nd version of [2] contained an incorrect effective Hamiltonian, see Remark 6.7 later. In the 3rd version of [2] this has been corrected, but the discussion remains sketchy and has gaps. See below at Remarks 2.10 and 6.6 and further below in this Introduction and at the beginning of Section 3.3.

We nonetheless draw from [2] a number of ideas which we list now. First of all, we draw from [2] a better choice of initial coordinates than [10]. Some of it existed also in previous literature, cf. the discussion in [11, Section 6]. We also borrow some notation, i.e. symbols $\mathcal{R}^{k,m}$ and $\mathbf{S}^{k,m}$ (which in [2] are defined incorrectly). Finally, inspired by [2] we simplify the proof in the part of the Darboux step contained in Lemma 3.6, which in [10] is more laborious.

Both here and in [10] we consider initial data in subsets of Σ_n for $n \gg 1$ which are unbounded in Σ_n and invariant for the system. We require this substantial amount of regularity and spacial decay to 0 for the classes of solutions of the system, in order to give a rigorous treatment of the flows and of the pullbacks. [2] suggests that [10] should prove decay rates in time. We do not understand the basis for this suggestion since, by the time invariance of the subsets Σ_n considered, the problem considered in [10] is very similar in this respect to the one with Σ_n replaced by H^1. Indeed time decay corresponds to bounds on norms containing time dependent weights. But if the problem is invariant by translation in time, the only information that can be derived must be invariant by translation in time, and bounds on time weighted norms do not have this property. We therefore emphasize that [10] and the present paper are very different from, say, [5, 13], which consider initial data in subsets of $H^{k,s}$ which are not invariant by the time evolution.

To find an effective Hamiltonian, we use the regularity properties of the flows, which in turn depend on the fact that we work in Σ_n for $n \gg 1$. See

Theorem 6.5 where the regularity of the flows is used to prove that the coordinate changes preserve the system. To prove for the NLS the same result in H^1, where the coordinate changes are continuous only, one needs to explain how they preserve the structure needed to make sense of the NLS. A reasonable approach to the H^1 case for the NLS is to first prove the result in our set up, then to prove the local well posedness in H^1 of the NLS within the various systems of coordinates used, and finally show that H^1 solutions of the NLS are invariant by coordinate changes, by means of a density argument and by the continuity of the coordinate changes in H^1. We do not prove here the last fact, just because everything in Section 3.2 is formulated in terms of the spaces Σ_n, but in fact for the NLS it follows by routine arguments. Since we do not provide a proof, we make no claim about H^1 solutions of the NLS, even though it seems not a far off step from what we prove here. [2] claims the result in H^1 without spelling out the proof, see Remark 6.6 below.

We discuss in some detail a key formula on the differentiation of the pullback of a differential form along a flow, see (79), which is the basis of Moser's method to find Darboux coordinates. This formula is simple in classical set ups, but in our case and in [2] its interpretation and proof are not obvious. In [2] the formula is stated and used without discussion. We treat the issue rigorously in Section 3.3, regularizing the flow, using (79) for the regularized flow, and recovering the desired equality between differential forms, by a limiting argument. Notice that we do not prove formula (79) for the non regularized flow.

We end with few remarks on the proofs.

The proof of the Darboux Theorem is a simplification of that in [10] in the part discussing the vector field. We give in Section 3.3 a detailed proof on the fact that the resulting flow transforms the symplectic form as desired. See also the introductory remarks in Section 3. Notice that parts of this discussion were skipped in [10].

The portion of our paper on the Birkhoff normal forms covers from Section 4 on and is quite different from [4, 8, 10] mainly because the pullback of the terms of the expansion of the Hamiltonian cannot be treated on a term by term basis, see Remark 5.5. What is important is to get a general structure of the pullbacks of the Hamiltonian. This is discussed in Section 4. It is likely that most of the analysis in Lemmas 4.3, 4.4 and 5.4, is not necessary to the derivation of the effective Hamiltonian, which is represented by H_2' and the null terms in $\mathbf{R}_0$ and $\mathbf{R}_1$ of the expansion in Lemma 5.4, in the final Hamiltonian. On the other hand, writing the Hamiltonian explicitly should make the arguments transparent and more clearly applicable to the part on dispersion and Fermi Golden rule.

In Section 5 we finally distinguish between discrete and continuous modes.

The present paper treats only equations whose symmetry group is Abelian. This limitation will have to be overcome to extend the theory to more general systems such for example the Dirac system without the symmetry constraints of [4].

2. Set up

- Given two vectors $u, v \in \mathbb{R}^{2N}$ we denote by $u \cdot v = \sum u_j v_j$ their inner product.
- We will consider also another quadratic form $|u|_1^2 = u \cdot_1 u$ in $\mathbb{R}^{2N}$.
- For any $n \geq 1$ we consider the space $\Sigma_n = \Sigma_n(\mathbb{R}^3, \mathbb{R}^{2N})$ defined by

$$\|u\|_{\Sigma_n}^2 := \sum_{|\alpha| \leq n} \left(\|x^\alpha u\|_{L^2(\mathbb{R}^3, \mathbb{R}^{2N})}^2 + \|\partial_x^\alpha u\|_{L^2(\mathbb{R}^3, \mathbb{R}^{2N})}^2 \right) < \infty.$$

  We set $\Sigma_0 = L^2(\mathbb{R}^3, \mathbb{R}^{2N})$. Equivalently we can define Σ_r for $r \in \mathbb{R}$ by the norm

$$\|u\|_{\Sigma_r} := \left\| (1 - \Delta + |x|^2)^{\frac{r}{2}} u \right\|_{L^2} < \infty.$$

  For $r \in \mathbb{N}$ the two definitions are equivalent, see [8]. We will not use another quite natural class of spaces denoted by $H^{k,s}$ and defined by

$$\|u\|_{H^{k,s}} := \left\| (1 + |x|^2)^{\frac{s}{2}} (1 - \Delta)^{\frac{k}{2}} u \right\|_{L^2} < \infty.$$

- $\mathcal{S}(\mathbb{R}^3, \mathbb{R}^{2N}) = \cap_{n \in \mathbb{N}} \Sigma_n(\mathbb{R}^3, \mathbb{R}^{2N})$ is the space of Schwartz functions and the space of tempered distributions is $\mathcal{S}'(\mathbb{R}^3, \mathbb{R}^{2N}) = \cup_{n \in \mathbb{N}} \Sigma_{-n}(\mathbb{R}^3, \mathbb{R}^{2N})$.
- For X and Y two Banach space, we will denote by $B(X, Y)$ the Banach space of bounded linear operators from X to Y and by $B^\ell(X, Y) = B(\prod_{j=1}^\ell X, Y)$.
- We denote by $\langle\ ,\ \rangle$ the natural inner product in $L^2(\mathbb{R}^3, \mathbb{R}^{2N})$.
- J is an invertible antisymmetric matrix in $\mathbb{R}^{2N}$. We have also $|Jy|_1 = |y|_1$ for any $y \in \mathbb{R}^{2N}$. In $L^2(\mathbb{R}^3, \mathbb{R}^{2N})$ we consider the symplectic form $\Omega = \langle J^{-1}\ ,\ \rangle$.
- We consider in $L^2(\mathbb{R}^3, \mathbb{R}^{2N})$ a linear selfadjoint elliptic differential operator $\mathcal{D}$ such that $\mathcal{D} \in B(\Sigma_r, \Sigma_{r - \mathrm{ord}\mathcal{D}})$ and $\mathcal{D} \in B(H^r, H^{r - \mathrm{ord}\mathcal{D}})$ for all r and for a fixed integer $\mathrm{ord}\mathcal{D} \geq 1$.

- We consider a Hamiltonian of the form

$$\begin{aligned} &E(U) = E_K(U) + E_P(U) \\ &E_K(U) := \frac{1}{2}\langle \mathcal{D}U, U\rangle\,, \quad E_P(U) := \int_{\mathbb{R}^3} B(|U|_1^2)dx. \end{aligned} \tag{1}$$

  Here $B \in C^\infty(\mathbb{R},\mathbb{R})$, $B(0) = B'(0) = 0$ and there exists a $p \in (2,6]$ such that for every $k \geq 0$ there is a fixed C_k with

$$\left|\nabla_\zeta^k (B(|\zeta|_1^2))\right| \leq C_k |\zeta|^{p-k-1} \quad \text{if } |\zeta| \geq 1 \text{ in } \mathbb{R}^{2N}. \tag{2}$$

Notice that $E_P \in C^5(H^1(\mathbb{R}^3,\mathbb{R}^{2N})),\mathbb{R})$. Consistently with [4, 8, 10], we focus only on *semilinear* Hamiltonians. We consider the system

$$\dot{U} = J\nabla E(U) \quad , \quad U(0) = U_0 \tag{3}$$

where for a Frechét differentiable function F the gradient $\nabla F(U)$ is defined by $\langle \nabla F(U), X\rangle = dF(U)(X)$, with $dF(U)$ the exterior differential calculated at U. We assume that

(A1) there exists d_0 such that for $d > d_0$ system (3) is locally well posed in H^d. Furthermore, the space Σ_d is invariant by this motion.

We recall the following definition.

Definition 2.1. *Given a Frechét differentiable function F, the Hamiltonian vectorfield of F with respect to a strong symplectic form ω, see [1, Chapter 9], is the field X_F such that $\omega(X_F, Y) = dF(Y)$ for any given tangent vector Y. For $\omega = \Omega$ we have $X_F = J\nabla F$.*

For F, G two scalar Frechét differentiable functions, we consider the Poisson bracket $\{F, G\} := dF(X_G)$.

If $\mathcal{G}$ has values in a given Banach space $\mathbb{E}$ and G is a scalar valued function, then we set $\{\mathcal{G}, G\} := \mathcal{G}'(X_G)$, for $\mathcal{G}'$ the Frechét derivative of $\mathcal{G}$.

We assume some symmetries in system (3). Specifically we assume what follows.

(A2) There are selfadjoint differential operators $\diamond_\ell$ for $\ell = 1, ..., n_0$ in L^2 such that $\diamond_\ell : \Sigma_n \to \Sigma_{n-d_\ell}$ for $\ell = 1, ..., n_0$. We set $\mathbf{d} = \sup_\ell d_\ell$.

(A3) We assume $[\diamond_\ell, J] = 0$ and $[\diamond_\ell, \diamond_k] = 0$.

(A4) We assume $\{\Pi_\ell, E_K\} = \{\Pi_\ell, E_P\} = 0$ for all ℓ, where $\Pi_\ell := \frac{1}{2}\langle \diamond_\ell\ ,\ \rangle$.

(A5) Set $\langle \epsilon \diamond \rangle^2 := 1 + \sum_j \epsilon^2 \diamond_j^2$. Then $\langle \epsilon \diamond \rangle^{-2} \in B(\Sigma_n, \Sigma_n)$ with

$$\|\langle \epsilon \diamond \rangle^{-2}\|_{B(\Sigma_n,\Sigma_n)} \leq C_n < \infty \text{ for any } |\epsilon| \leq 1 \text{ and } n \in \mathbb{N}. \tag{4}$$

Furthermore, for any $n \in \mathbb{Z}$ we have

$$\begin{aligned} &\textit{strong} - \lim_{\epsilon \to 0} \langle \epsilon \diamond \rangle^{-2} = 1 \text{ in } B(\Sigma_n, \Sigma_n) \\ &\lim_{\epsilon \to 0} \|\langle \epsilon \diamond \rangle^{-2} - 1\|_{B(\Sigma_n,\Sigma_{n'})} = 0 \quad \text{for any } n' \in \mathbb{Z} \text{ with } n' < n. \end{aligned} \tag{5}$$

(A6) Consider the groups $e^{J\langle \epsilon \diamond \rangle^{-2} \diamond \cdot \tau}$ defined in L^2. We assume that for any $n \in \mathbb{N}$ these groups leave Σ_n invariant and that for any $n \in \mathbb{N}$ and $c > 0$ there a C s.t. $\|e^{J\langle \epsilon \diamond \rangle^{-2} \diamond \cdot \tau}\|_{B(\Sigma_n,\Sigma_n)} \leq C$ for any $|\tau| \leq c$ and any $|\epsilon| \leq 1$.

We introduce now our *solitary* waves.

(B1) We assume that for $\mathcal{O}$ an open subset of $\mathbb{R}^{n_0}$ we have a function $p \to \Phi_p \in \mathcal{S}(\mathbb{R}^3, \mathbb{R}^{2N})$ which is in $C^\infty(\mathcal{O}, \mathcal{S})$, with $\Pi_\ell(\Phi_p) = p_\ell$, where the Φ_p are constrained critical points of E with associated Lagrange multipliers $\lambda_\ell(p)$ so that

$$\nabla E(\Phi_p) = \lambda(p) \cdot \diamond \Phi_p \tag{6}$$

(B2) We will assume that the map $p \to \lambda(p)$ is a diffeomorphism. In particular this means that the following matrix has rank n_0

$$\operatorname{rank} \left[\frac{\partial \lambda_i}{\partial p_j} \right]_{i\downarrow \ , \ j\rightarrow} = n_0. \tag{7}$$

A function $U(t) := e^{J(t\lambda(p)+\tau_0)\cdot \diamond} \Phi_p$ is a solitary wave solution of (3) for any fixed vector τ_0.

2.1. The linearization

Set $\mathcal{H}_p := J(\nabla^2 E(\Phi_p) - \lambda(p) \cdot \diamond)$. Notice that $E(e^{J\tau\cdot\diamond} U) \equiv E(U)$ for any U yields $\nabla E(e^{J\tau\cdot\diamond} U) = e^{J\tau\cdot\diamond} \nabla E(U)$ and $\nabla^2 E(e^{J\tau\cdot\diamond} U) = e^{J\tau\cdot\diamond} \nabla^2 E(U) e^{-J\tau\cdot\diamond}$. Then (6) implies $\nabla E(e^{J\tau\cdot\diamond} \Phi_p) = e^{J\tau\cdot\diamond} \lambda(p) \cdot \diamond \Phi_p$. So applying ∂_{τ_j} we obtain $(\nabla^2 E(\Phi_p) - \lambda(p) \cdot \diamond) J \diamond_j \Phi_p = 0$ and so

$$\mathcal{H}_p J \diamond_j \Phi_p = 0 \tag{8}$$

(C1) We will assume

$$\ker \mathcal{H}_p = \mathrm{Span}\{J\diamond_j \Phi_p : j = 1, ..., n_0\}. \tag{9}$$

Applying ∂_{λ_j} to (6) yields $(\nabla^2 E(\Phi_p) - \lambda(p)\cdot\diamond)\partial_{\lambda_j}\Phi_p = \diamond_j \Phi_p$. This yields

$$\mathcal{H}_p \partial_{\lambda_j}\Phi_p = J\diamond_j \Phi_p \tag{10}$$

We have

$$\langle \partial_{\lambda_j}\Phi_p, \diamond_k \Phi_p\rangle = \frac{1}{2}\partial_{\lambda_j}\langle \Phi_p, \diamond_k \Phi_p\rangle = \partial_{\lambda_j} p_k. \tag{11}$$

Necessarily, by (B2) there exists j such that $\partial_{\lambda_j} p_k \neq 0$. This implies that the *generalized kernel* is

$$N_g(\mathcal{H}_p) = \mathrm{Span}\{J\diamond_j \Phi_p, \partial_{\lambda_j}\Phi_p : j = 1, ..., n_0\}. \tag{12}$$

The map $(p, \tau) \to e^{J\tau_0\cdot\diamond}\Phi_p$ is in $C^\infty(\mathcal{O}\times\mathbb{R}^{n_0}, \mathcal{S})$.

(C2) We assume this map is a local embedding and that the image is a manifold $\mathcal{G}$.

At any given point $e^{J\tau\cdot\diamond}\Phi_p$ the tangent space of $\mathcal{G}$ is given by

$$T_{e^{J\tau\cdot\diamond}\Phi_p}\mathcal{G} = \mathrm{Span}\{e^{J\tau\cdot\diamond}\partial_{p_j}\Phi_p, e^{J\tau\cdot\diamond}\diamond_j \Phi_p : j = 1, ..., n_0\}.$$

We have $\Omega(e^{J\tau\cdot\diamond}\partial_{p_j}\Phi_p, e^{J\tau\cdot\diamond}\partial_{p_k}\Phi_p) = \Omega(\partial_{p_j}\Phi_p, \partial_{p_k}\Phi_p)$.

(C3) We assume that

$$\Omega(\partial_{p_j}\Phi_p, \partial_{p_k}\Phi_p) = 0 \text{ for all } j \text{ and } k \tag{13}$$

$$\Omega(\partial_{p_j}\Phi_p, \Phi_p) = 0 \text{ for all } j. \tag{14}$$

Notice that (14) is not required in [2] but in any case is true for the applications in [2, 4, 8, 10]. Here we use it in Lemma 3.1.

We have the following beginning of Jordan block decomposition of $\mathcal{H}_p$.

LEMMA 2.2. *Consider the operator $\mathcal{H}_p$. We have*

$$J^{-1}\mathcal{H}_p = -\mathcal{H}_p^* J^{-1}, \quad \mathcal{H}_p J = -J\mathcal{H}_p^*. \tag{15}$$

Assume (B1)–(B2) and (C1). Then we have

$$L^2 = N_g(\mathcal{H}_p) \oplus N_g^\perp(\mathcal{H}_p^*), \tag{16}$$

$$N_g(\mathcal{H}_p^*) = \mathit{Span}\{\diamond_j \Phi_p, J^{-1}\partial_{\lambda_j}\Phi_p : j = 1, ..., n_0\}. \tag{17}$$

Proof. We have $\mathcal{H}_p = JA$ for a selfadjoint operator A and with J a bounded antisymmetric operator. Then $\mathcal{H}_p^* = -AJ$ and (15) follows by direct inspection. Recall that (B1)–(B2) and (C1) imply (12). Then (15) implies (17).
The map $\psi \to \langle\ ,\psi\rangle$ establishes a map $N_g(\mathcal{H}_p^*) \to B(N_g(\mathcal{H}_p),\mathbb{R})$. By (11), formulas (12) and (17) imply that this map is an isomorphism. For any $u \in L^2$ there is exactly one $v \in N_g(\mathcal{H}_p)$ such that $\langle u,\ \rangle$ and $\langle v,\ \rangle$ coincide as elements in $B(N_g(\mathcal{H}_p^*),\mathbb{R})$. Then $u - v \in N_g^\perp(\mathcal{H}_p^*)$ and we get (16). □

Obviously Lemma 2.2 holds true only because our J is very special. For the KdV, where $J = \frac{\partial}{\partial x}$, (16)–(17) are not true.
Denote by $P_{N_g}(p) = P_{N_g(\mathcal{H}_p)}$ the projection onto $N_g(\mathcal{H}_p)$ associated to (16) and by $P(p) := 1 - P_{N_g}(p)$ the projection on $N_g^\perp(\mathcal{H}_p^*)$. We have, summing on repeated indexes,

$$P_{N_g}(p)X = -J\diamondsuit_j\Phi_p\ \langle X, J^{-1}\partial_{p_j}\Phi_p\rangle + \partial_{p_j}\Phi_p\ \langle X, \diamondsuit_j\Phi_p\rangle. \tag{18}$$

Lemma 2.3. *Assume (B1)–(B2) and (C1). Then:*

(1) $P_{N_g}(p) \in B(\mathcal{S}',\mathcal{S})$ for any $p \in \mathcal{O}$ and $P_{N_g}(p) \in C^\infty(\mathcal{O}, B(\Sigma_{-k},\Sigma_k))$ for any $k \in \mathbb{N}$.

(2) For any $p_0 \in \mathcal{O}$ and k there exists an $\varepsilon_k > 0$ such that for $|p - p_0| < \varepsilon_k$

$$P(p)P(p_0) : N_g^\perp(\mathcal{H}_{p_0}^*) \cap \Sigma_k \to N_g^\perp(\mathcal{H}_p^*) \cap \Sigma_k \tag{19}$$

is an isomorphism.

(3) For $h > k$ we have $\varepsilon_h \geq \varepsilon_k$.

Proof. Claim (1) is elementary and we skip the proof.
Consider the map $P(p)P(p_0)P(p) = 1 + P(p)(P_{N_g}(p) - P_{N_g}(p_0))P(p)$ from $N_g^\perp(\mathcal{H}_p^*) \cap \Sigma_k$ into itself. By Claim (1) and by the Fredholm alternative, this is an isomorphism for $|p - p_0| < \varepsilon_k$ with $\varepsilon_k > 0$ sufficiently small. This implies that the $P(p)P(p_0)$ in (19) is onto. For the same reasons also $P(p_0)P(p)P(p_0)$ is an isomorphism from $N_g^\perp(\mathcal{H}_{p_0}^*) \cap \Sigma_k$ into itself. Then $P(p)P(p_0)$ in (19) is one to one. This yields Claim (2).
For $h > k$ we have the commutative diagram

$$\begin{array}{ccc} N_g^\perp(\mathcal{H}_{p_0}^*) \cap \Sigma_h & \overset{P(p)P(p_0)}{\to} & N_g^\perp(\mathcal{H}_p^*) \cap \Sigma_h \\ \downarrow & & \downarrow \\ N_g^\perp(\mathcal{H}_{p_0}^*) \cap \Sigma_k & \overset{P(p)P(p_0)}{\to} & N_g^\perp(\mathcal{H}_p^*) \cap \Sigma_k \end{array}$$

with the vertical maps two embedding. This implies that for $|p - p_0| < \varepsilon_k$ we have $\ker P(p)P(p_0) = 0$ in $N_g^\perp(\mathcal{H}_{p_0}^*) \cap \Sigma_h$. To complete the proof of Claim (3),

we need to show that given $u \in N_g^{\perp}(\mathcal{H}_p^*) \cap \Sigma_h$ and the resulting $v \in N_g^{\perp}(\mathcal{H}_{p_0}^*) \cap \Sigma_k$ with $u = P(p)P(p_0)v$, we have $v \in \Sigma_h$. But this follows immediately from

$$v = u + (P_{N_g}(p) - P_{N_g}(p_0))v \text{ where } u \in \Sigma_h \text{ and } (P_{N_g}(p) - P_{N_g}(p_0))v \in \mathcal{S}.$$

□

We will denote the inverse of (19) by

$$(P(p)P(p_0))^{-1} : N_g^{\perp}(\mathcal{H}_p^*) \cap \Sigma_k \to N_g^{\perp}(\mathcal{H}_{p_0}^*) \cap \Sigma_k. \tag{20}$$

We have the following *Modulation* type lemma.

LEMMA 2.4 (Modulation). *Assume (A2), (B.1), (B.2), (C.1) and (C.3). Fix $n \in \mathbb{Z}$, $n \geq 0$ and fix $\Psi_0 = e^{J\tau_0 \cdot \diamond}\Phi_{p_0}$. Then $\exists$ a neighborhood $\mathcal{U}$ in $\Sigma_{-n}(\mathbb{R}^3, \mathbb{R}^{2N})$ of U_0 and functions $p \in C^\infty(\mathcal{U}, \mathcal{O})$ and $\tau \in C^\infty(\mathcal{U}, \mathbb{R}^{n_0})$ s.t. $p(\Psi_0) = p_0$ and $\tau(\Psi_0) = \tau_0$ and s.t. $\forall U \in \mathcal{U}$*

$$U = e^{J\tau \cdot \diamond}(\Phi_p + R) \text{ and } R \in N_g^{\perp}(\mathcal{H}_p^*). \tag{21}$$

Proof. Consider the following $2n_0$ functions:

$$\begin{aligned} \mathcal{F}_j(U, p, \tau) &:= \Omega(U - e^{J\tau \cdot \diamond}\Phi_p, e^{J\tau \cdot \diamond}\partial_{p_j}\Phi_p) \\ \mathcal{G}_j(U, p, \tau) &:= \Omega(U - e^{J\tau \cdot \diamond}\Phi_p, Je^{J\tau \cdot \diamond}\diamond_j\Phi_p). \end{aligned} \tag{22}$$

These functions belong to $C^\infty(\Sigma_{-n} \times \mathcal{O} \times \mathbb{R}^{n_0}, \mathbb{R})$. We introduce the notation $R = e^{-J\tau \cdot \diamond}U - \Phi_p$. Notice that $R = 0$ for $U = \Phi_p$. Then

$$\begin{aligned} \partial_{\tau_k}\mathcal{F}_j(U, p, \tau) &= \Omega(e^{J\tau \cdot \diamond}R, e^{J\tau \cdot \diamond}J\diamond_k\partial_{p_j}\Phi_p) - \Omega(J\diamond_k e^{J\tau \cdot \diamond}\Phi_p, e^{J\tau \cdot \diamond}\partial_{p_j}\Phi_p) \\ &= -\langle R, \diamond_k\partial_{p_j}\Phi_p\rangle - \langle \diamond_k\Phi_p, \partial_{p_j}\Phi_p\rangle \\ &= -\langle R, \diamond_k\partial_{p_j}\Phi_p\rangle - \frac{1}{2}\partial_{p_j}\langle \diamond_k\Phi_p, \Phi_p\rangle \\ &= -\langle R, \diamond_k\partial_{p_j}\Phi_p\rangle - \delta_{jk}. \end{aligned}$$

By (13) we have

$$\begin{aligned} \partial_{p_k}\mathcal{F}_j(U, p, \tau) &= \Omega(e^{J\tau \cdot \diamond}R, e^{J\tau \cdot \diamond}\partial_{p_k}\partial_{p_j}\Phi_p) - \Omega(Je^{J\tau \cdot \diamond}\partial_{p_k}\Phi_p, e^{J\tau \cdot \diamond}\partial_{p_j}\Phi_p) \\ &= \Omega(R, \partial_{p_k}\partial_{p_j}\Phi_p). \end{aligned}$$

By (A3) we have

$$\begin{aligned} \partial_{\tau_k}\mathcal{G}_j &= \Omega(e^{J\tau \cdot \diamond}R, e^{J\tau \cdot \diamond}J^2\diamond_k\diamond_j\Phi_p) - \Omega(J\diamond_k e^{J\tau \cdot \diamond}\Phi_p e^{J\tau \cdot \diamond}J\diamond_j\Phi_p) \\ &= -\langle R, J\diamond_k\diamond_j\Phi_p\rangle - \langle J\diamond_k\Phi_p, \diamond_j\Phi_p\rangle \\ &= -\langle R, J\diamond_k\diamond_j\Phi_p\rangle, \end{aligned}$$

We have

$$\begin{aligned}\partial_{p_k}\mathcal{G}_j &= \Omega(e^{J\tau\cdot\diamond}R, e^{J\tau\cdot\diamond}J\diamond_j\partial_{p_k}\Phi_p) - \Omega(e^{J\tau\cdot\diamond}\partial_{p_k}\Phi_p e^{J\tau\cdot\diamond}J\diamond_j\Phi_p)\\ &= -\langle R, \diamond_j\partial_{p_k}\Phi_p\rangle + \langle\partial_{p_k}\Phi_p, \diamond_j\Phi_p\rangle\\ &= -\langle R, \diamond_j\partial_{p_k}\Phi_p\rangle + \delta_{jk}.\end{aligned}$$

At $U=\Psi_0$, $\tau=\tau_0$ and $p=p_0$ we have $\mathcal{F}_j=\mathcal{G}_j=0$. Since in this case $R=0$ we get the desired result by the Implicit Function Theorem. □

2.2. Spectral coordinates

Lemmas 2.2–2.4 lead to a natural decomposition of (3). To write it we need further notation.
We are ready for the natural coordinates decomposition. Let $\Pi(U_0)=p_0$. We consider for $R\in N_g^\perp(\mathcal{H}^*_{p_0})$ the map

$$(\tau,p,R)\to U=e^{J\tau\cdot\diamond}(\Phi_p+P(p)R). \tag{23}$$

We have the following formulas,

$$\frac{\partial}{\partial\tau_j}=J\diamond_j U\,,\quad \frac{\partial}{\partial p_j}=e^{J\tau\cdot\diamond}(\partial_{p_j}\Phi_p+\partial_{p_j}P(p)R), \tag{24}$$

with $\frac{\partial}{\partial p_j}\in C^\infty(\mathcal{U}\cap\Sigma_k,\Sigma_{k'})$ for any pair $(k,k')\in\mathbb{N}^2$, with $\mathcal{U}\subset\Sigma_{-n}$ the neighborhood of $e^{J\tau_0\cdot\diamond}\Phi_{p_0}$ in Lemma 2.4. Similarly, $\frac{\partial}{\partial\tau_j}\in C^0(\mathcal{U}\cap\Sigma_k,\Sigma_{k-d_j})$. We have what follows.

Lemma 2.5. *Consider the $n\geq 0$ and $\mathcal{U}$ in Lemma 2.4 and fix an integer $k\geq -n$. Then the map $U\to R(U)=R$ is $C^0(\mathcal{U}\cap\Sigma_k,\Sigma_k)$. For $k\geq -n+\boldsymbol{d}$ we have $R\in C^1(\mathcal{U}\cap\Sigma_k,\Sigma_{k-\boldsymbol{d}})$. For $\mathcal{U}$ sufficiently small in Σ_{-n} the Frechét derivative $R'(U)$ of $R(U)$ is defined by the following formula, summing on the repeated index j,*

$$R'(U)=(P(p)P(p_0))^{-1}P(p)\big[e^{-J\tau\cdot\diamond}\,\mathbb{1}-J\diamond_j P(p)R\,d\tau_j-\partial_{p_j}P(p)R\,dp_j\big],$$

where $(P(p)P(p_0))^{-1}:N_g^\perp(\mathcal{H}^*_p)\cap\Sigma_{k-\boldsymbol{d}}\to N_g^\perp(\mathcal{H}^*_{p_0})\cap\Sigma_{k-\boldsymbol{d}}$ is well defined by Lemma 2.3.

Proof. The continuity of $R(U)$ follows from $R=e^{-J\tau\cdot\diamond}U-\Phi_p$ and

$$\begin{aligned}R-R' &= e^{-J\tau\cdot\diamond}U-e^{-J\tau'\cdot\diamond}U'+\Phi_{p'}-\Phi_p\\ &= \Phi_{p'}-\Phi_p+(e^{-J\tau\cdot\diamond}-e^{-J\tau'\cdot\diamond})U+e^{-J\tau'\cdot\diamond}(U-U').\end{aligned}$$

Then use $p\to\Phi_p\in C^\infty(\mathcal{O},\mathcal{S})$, the fact that $e^{J\tau\cdot\diamond}$ is strongly continuous in Σ_k and locally uniformly bounded therein. The fact that $R(U)$ has Frechét

derivative follows by the chain rule. To get the formula for $R'(U)$ notice that the equalities $R'\frac{\partial}{\partial p_j} = R'\frac{\partial}{\partial \tau_j} = 0$ and $R'e^{J\tau\cdot\diamond}P(p)P(p_0) = \mathbb{1}_{|N_g^\perp(\mathcal{H}^*_{p_0})}$ characterize R'. We claim we have

$$R' = \mathbf{a}_j d\tau_j + \mathbf{b}_j dp_j + (P(p)P(p_0))^{-1}P(p)e^{-J\tau\cdot\diamond} \tag{25}$$

for some $\mathbf{a}_j$ and $\mathbf{b}_j$. First of all, by the independence of coordinates (τ, p) from $R \in N_g^\perp(\mathcal{H}^*_{p_0})$,

$$d\tau_j \circ e^{J\tau\cdot\diamond}P(p)P(p_0) = dp_j \circ e^{J\tau\cdot\diamond}P(p)P(p_0) = 0.$$

Indeed for $g \in N_g^\perp(\mathcal{H}^*_{p_0})$ we have for instance

$$\begin{aligned} 0 = \frac{d}{dt}\tau_j(u(\tau, p, R+tg))_{|t=0} &= \frac{d}{dt}\tau_j(e^{J\tau\cdot\diamond}(\Phi_p + P(p)P(p_0)(R+tg)))_{|t=0} \\ &= d\tau_j \circ e^{J\tau\cdot\diamond}P(p)P(p_0)g. \end{aligned}$$

Secondarily, by the definition of $(P(p)P(p_0))^{-1}$,

$$(P(p)P(p_0))^{-1}P(p)e^{-J\tau\cdot\diamond} \circ e^{J\tau\cdot\diamond}P(p)P(p_0) = \mathbb{1}_{N_g^\perp(\mathcal{H}^*_{p_0})}.$$

Hence we get the claimed equality (25).
To get $\mathbf{a}_j$ and $\mathbf{b}_j$ notice that by $R'\frac{\partial}{\partial \tau_j} = 0$ and $P(p)J\diamond_j\Phi_p = 0$

$$\begin{aligned} \mathbf{a}_j &= -(P(p)P(p_0))^{-1}P(p)e^{-J\tau\cdot\diamond}\frac{\partial}{\partial \tau_j} \\ &= -(P(p)P(p_0))^{-1}P(p)e^{-J\tau\cdot\diamond}e^{J\tau\cdot\diamond}J\diamond_j(\Phi_p + P(p)R) \\ &= -(P(p)P(p_0))^{-1}P(p)J\diamond_j P(p)R. \end{aligned}$$

Similarly by $R'\frac{\partial}{\partial p_j} = 0$ and $P(p)\partial_{p_j}\Phi_p = 0$

$$\begin{aligned} \mathbf{b}_j &= -(P(p)P(p_0))^{-1}P(p)e^{-J\tau\cdot\diamond}\frac{\partial}{\partial p_j} \\ &= -(P(p)P(p_0))^{-1}P(p)(\partial_{p_j}\Phi_p + \partial_{p_j}P(p)R) \\ &= -(P(p)P(p_0))^{-1}P(p)\partial_{p_j}P(p)R. \end{aligned}$$

□

A crucial point in the stability proofs in [3, 4, 8, 10], first realized and used in [7], is the importance not to loose track of the Hamiltonian nature of (3), in whichever coordinates the system is written. Thus we have what follows.

LEMMA 2.6. *In the coordinate system* (23)*, system* (3) *can be written as*

$$\dot p = \{p, E\}\,,\ \dot\tau = \{\tau, E\}\,,\ \dot R = \{R, E\}. \tag{26}$$

Proof. The statement is not standard only for $\dot R = \{R, E\}$. Notice that (3) can be written as

$$\begin{aligned} \dot U &= J\dot\tau\cdot\diamond U + e^{J\tau\cdot\diamond}\dot p\cdot\nabla_p(\Phi_p + P(p)R) + e^{J\tau\cdot\diamond}P(p)\dot R \\ &= \sum_j \dot\tau_j\frac{\partial}{\partial\tau_j} + \dot p_j\frac{\partial}{\partial p_j} + e^{J\tau\cdot\diamond}P(p)\dot R = J\nabla E(U). \end{aligned} \tag{27}$$

When we apply the derivative $R'(U)$ to (27), all the terms in the lhs of the last line cancel except for

$$R'(U)e^{J\tau\cdot\diamond}P(p)\dot R = R'(U)J\nabla E(U) = R'(U)X_E(U) = \{R, E\},$$

from the definition of hamiltonian field and of Poisson bracket. Finally we use

$$R'(U)e^{J\tau\cdot\diamond}P(p)\dot R = \frac{d}{ds}_{|_{s=0}} R(U(\tau, p, R + s\dot R)) = \frac{d}{ds}_{|_{s=0}}(R + s\dot R) = \dot R.$$

□

2.3. Reduction of order of system (26)

The following Poisson bracket identities are useful.

LEMMA 2.7. *Consider the functions* Π_j*. Then* $X_{\Pi_j} = \frac{\partial}{\partial\tau_j}$*. In particular*

$$\{\Pi_j, \tau_k\} = -\delta_{jk}\ ,\quad \{\Pi_j, p_k\} \equiv 0\ ,\quad \{R, \Pi_j\} = 0. \tag{28}$$

Proof. (28) follows from the first claim, which is a consequence of (24):

$$X_{\Pi_j}(U) = J\nabla\Pi_j(U) = J\diamond_j U = \frac{\partial}{\partial\tau_j}.$$

□

We introduce now a new Hamiltonian:

$$K(U) := E(U) - E\left(\Phi_{p_0}\right) - \lambda_j(p(U))\left(\Pi_j(U) - \Pi_j(U_0)\right). \tag{29}$$

Notice that $K(e^{J\tau\cdot\diamond}U) \equiv K(U)$. Equivalently, $\partial_{\tau_j}K \equiv 0$. We know that for solutions of (3) we have $\Pi_j(U(t)) = \Pi_j(U_0)$ and

$$\{p_j, K\} = \{p_j, E\}\,,\ \{R, K\} = \{R, E\}\,,\ \{\tau_j, K\} = \{\tau_j, E\} + \lambda_j(p).$$

By $\partial_{\tau_j} K \equiv 0$, the evolution of the variables p, R is unchanged if we consider the following new Hamiltonian system:

$$\dot{p}_j = \{p_j, K\}, \quad \dot{\tau}_j = \{\tau_j, K\}, \quad \dot{R} = \{R, K\}. \tag{30}$$

It is elementary that the *momenta* $\Pi_j(U)$ are invariants of motion of (30).

Before exploiting the invariance of $\Pi_j(U)$ to reduce the order of the system, we introduce appropriate notation. First of all we set

$$\begin{aligned} \mathcal{P}^r &:= \mathbb{R}^{n_0} \times (\Sigma_r \cap N_g^{\perp}(\mathcal{H}_{p_0})) = \{(\tau, R)\}, \\ \widetilde{\mathcal{P}}^r &:= \mathbb{R}^{n_0} \times \mathcal{P}^r = \{(\Pi, \tau, R)\}. \end{aligned} \tag{31}$$

We set $\mathcal{P} = \mathcal{P}^0$ and $\widetilde{\mathcal{P}} = \widetilde{\mathcal{P}}^0$.

Definition 2.8. *We will say that $F(t, \varrho, R) \in C^M(I \times \mathcal{A}, \mathbb{R})$ with I a neighborhood of 0 in $\mathbb{R}$ and $\mathcal{A}$ a neighborhood of 0 in $\mathcal{P}^{-K}$ is $\mathcal{R}^{i,j}_{K,M}$ and we will write $F = \mathcal{R}^{i,j}_{K,M}$, or more specifically $F = \mathcal{R}^{i,j}_{K,M}(t, \varrho, R)$, if there exists a $C > 0$ and a smaller neighborhood $\mathcal{A}'$ of 0 s.t.*

$$|F(t, \varrho, R)| \leq C \|R\|^j_{\Sigma_{-K}} (\|R\|_{\Sigma_{-K}} + |\varrho|)^i \text{ in } I \times \mathcal{A}'. \tag{32}$$

We say $F = \mathcal{R}^{i,j}_{K,\infty}$ if $F = \mathcal{R}^{i,j}_{K,m}$ for all $m \geq M$. We say $F = \mathcal{R}^{i,j}_{\infty,M}$ if for all $k \geq K$ the above F is the restriction of an $F(t, \varrho, R) \in C^M(I \times \mathcal{A}_k, \mathbb{R})$ with $\mathcal{A}_k$ a neighborhood of 0 in $\mathcal{P}^{-k}$ and which is $F = \mathcal{R}^{i,j}_{k,M}$. Finally we say $F = \mathcal{R}^{i,j}$ if $F = \mathcal{R}^{i,j}_{k,\infty}$ for all k.

Definition 2.9. *We will say that an $T(t, \varrho, R) \in C^M(I \times \mathcal{A}, \Sigma_K(\mathbb{R}^3, \mathbb{R}^{2N}))$, with $I \times \mathcal{A}$ like above, is $\mathbf{S}^{i,j}_{K,M}$ and we will write $T = \mathbf{S}^{i,j}_{K,M}$ or more specifically $T = \mathbf{S}^{i,j}_{K,M}(t, \varrho, R)$, if there exists a $C > 0$ and a smaller neighborhood $\mathcal{A}'$ of 0 s.t.*

$$\|T(t, \varrho, R)\|_{\Sigma_K} \leq C \|R\|^j_{\Sigma_{-K}} (\|R\|_{\Sigma_{-K}} + |\varrho|)^i \text{ in } I \times \mathcal{A}'. \tag{33}$$

We use notation $T = \mathbf{S}^{i,j}$, $T = \mathbf{S}^{i,j}_{K,\infty}$ or $T = \mathbf{S}^{i,j}_{\infty,M}$ as above.

These notions will be often used also for functions $F = \mathcal{R}^{i,j}_{K,M}(\varrho, R)$ and $T = \mathbf{S}^{i,j}_{K,M}(\varrho, R)$ independent of t.

Remark 2.10. *We will see later that the coefficients of the vector fields whose flows are used to change coordinates are symbols as of Definitions 2.8 and 2.9. The definitions of the symbols $\mathcal{R}^{i,j}$ and $\mathbf{S}^{i,j}$ in [2, Definition 3.9 and 3.10] are very restrictive, since they require for the symbols to be defined in $I \times \mathcal{B}'$ with $\mathcal{B}'$ a neighborhood of the origin in $\mathcal{S}'$. The proofs in [2] at most prove that the coefficients of the vector fields in fact are symbols of the form $\mathcal{R}^{i,j}_{K,M}$ and $\mathbf{S}^{i,j}_{K,M}$*

in our sense. As an example we refer to [2, Lemmas 3.26 and 5.5]. In [2, Lemma 3.26] the fact that the b_i and the $\langle W^l; Y\rangle$ are symbols of the form $\mathcal{R}^{j,k}$ for some (j,k) in the sense of [2, Definition 3.10], requires preliminarily to show at least that they are functions of (ϱ, R) for (ϱ, R) in some neighborhood $\mathcal{U}$ of $(0,0)$ in $\mathbb{R}^{n_0}\times\mathcal{S}'$. This is not addressed in [2] and is far from trivial, since the coefficients of the linear system right above formula (3.60) are unbounded in any such $\mathcal{U}$. The justification that the coefficients $\Phi_{\mu\nu}(M)$ of χ in [2, Section 5] are in $\mathcal{S}$ is similarly inconclusive. The key step should be that the homological equation in Lemma 5.5 can be solved for all parameters k uniformly in the variable $M\in\mathbb{R}^n$, at least for $|M|<a$ for a fixed a. But the homological equations involve the perturbation of an operator and in [2] the perturbation is not fully analyzed. For example there is no discussion of the norm $\|V_M - V_0\|_{\mathcal{W}^k\to\mathcal{W}^k}$ as k grows and $|M|<a$. This norm should be expected to grow and become large, possibly breaking down the proof of $\Phi_{\mu\nu}(M)\in\mathcal{S}$. In fact it is plausible that $\Phi_{\mu\nu}(M)\in\mathcal{S}$ only for $M=0$.

From the above remarks we can see that no coordinate change in the Birkhoff or in the Darboux steps in [2] is shown to be an almost smooth transformation in the sense of [2, Definition 3.15]. Because also of the absence of a rigorous discussion on pullbacks of differential forms, we see that the proofs of the Birkhoff step, [2, Theorem 5.2], and of the Darboux step, [2, Theorem 3.21], are both inconclusive.

We proceed now to a reduction of order in (30). Write

$$\begin{aligned}\Pi_j(U) &= \Pi_j(e^{J\tau\cdot\diamond}(\Phi_p+P(p)R)) = \Pi_j(\Phi_p+P(p)R)\\ &= \tfrac{1}{2}\langle\diamond_j(\Phi_p+P(p)R),\Phi_p+P(p)R = p_j+\Pi_j(P(p)R)\\ &= p_j+\Pi_j(R)+\Pi_j((P(p)-P(p_0))R)+\langle R,\diamond_j(P(p)-P(p_0))R\rangle.\end{aligned}\tag{34}$$

We will move from variables (τ,p,R) to variables (τ,Π,R). Setting $\varrho_j=\Pi_j(R)$, we have

$$p_j=\Pi_j-\varrho_j+\widetilde{\Psi}_j(p-p_0,R)\tag{35}$$

with $\widetilde{\Psi}_j=\mathcal{R}^{0,2}(p-p_0,R)$. The implicit function theorem yields:

LEMMA 2.11. *There are functions $p_j=p_j(\Pi,\Pi(R),R)$ defined implicitly by (34), or (35), such that $p_j=\Pi_j-\varrho_j+\Psi_j(\Pi,\varrho,R)$ with $\Psi(p_0,\varrho,R)=\mathcal{R}^{0,2}(\varrho,R)$.*

We consider now (τ,Π,R) as a new coordinate system. By $\frac{\partial}{\partial\tau_k}\Pi_j(U)\equiv 0$ it follows that the vectorfields $\frac{\partial}{\partial\tau_k}$ are the same for the two systems of coordinates. In the new variables, system (30) reduces to the pair of systems

$$\dot\tau_j=\{\tau_j,K\}\,,\quad \dot\Pi_j=0\,,\tag{36}$$

$$\dot R=\{R,K\}.\tag{37}$$

System (37) is closed because of $\partial_{\tau_j} K = 0$.

3. Darboux Theorem

In this section we present one of the two main results of this paper. We seek to reproduce Moser's proof of the Darboux theorem. Specifically we look for a vector field $\mathcal{X}^t$ that will produce a flow as in (79) below. The proof of the existence and properties of $\mathcal{X}^t$ is similar to [8], but influenced by the choice of coordinates in [2]. We also add material to justify, once $\mathcal{X}^t$ has been found, the formal formula (79). Notice that for [4, 8] formula (79) does not require justification because $\mathcal{X}^t$ is a smooth vectorfield on a given manifold. But the situation in [2, 10] is different since now $\mathcal{X}^t$ is not a standard vectorfield on a manifold and Ω is not a regular differential form on the same manifold, so Lie derivative, pullbacks, push forwards and the related differentiation formulas, require justification.

Notice that, to be useful in the asymptotic stability theory, the change of variables has to be such that the new Hamiltonian equations is semilinear. This is why even in [4, 8], where we could apply the standard Darboux theorem for strong symplectic forms on Banach manifolds, see [1, Chapter 9], it is important to select $\mathcal{X}^t$ with an *ad hoc* process.

3.1. Search of a vectorfield

Recall that $\Omega = \langle J^{-1}\ ,\ \rangle$ and consider

$$\Omega_0 := d\tau_j \wedge d\Pi_j + \langle J^{-1}R', R'\rangle. \tag{38}$$

LEMMA 3.1. *At the points $e^{J\tau\cdot\diamond}\Phi_{p_0}$ for all $\tau \in \mathbb{R}^{n_0}$ we have $\Omega_0 = \Omega$. Consider the following forms:*

$$\mathrm{B}_0 := \tau_j d\Pi_j + \frac{1}{2}\langle J^{-1}R, R'\rangle; \quad \mathrm{B} := \mathrm{B}_0 + \alpha \text{ for} \tag{39}$$

$$\begin{aligned} \alpha &:= -\beta_j(p,R)d\Pi_j + \langle \Gamma(p)R + \beta_j(p,R)P^*(p)\diamond_j P(p)R, R'\rangle\ , \\ \Gamma(p) &:= \frac{1}{2}J^{-1}\left(P(p) - P(p_0)\right)\ , \\ \beta_j(p,R) &:= \frac{1}{2}\ \frac{\langle P^*(p)J^{-1}R, \partial_{p_j}P(p)R\rangle}{1 + \langle \diamond_j P(p)R, \partial_{p_j}P(p)R\rangle}\ . \end{aligned} \tag{40}$$

Then $d\mathrm{B}_0 = \Omega_0$ and $d\mathrm{B} = \Omega$.

Proof. $d\mathrm{B}_0 = \Omega_0$ follows from the definition of exterior differential. Set $\widetilde{B} := \frac{1}{2}\langle J^{-1}U,\ \rangle$. Notice that $d\widetilde{B} = \Omega$. By (23) we get:

$$\widetilde{B}(X)=\frac{1}{2}\langle J^{-1}e^{J\tau\cdot\diamond}\Phi_p,X\rangle+\frac{1}{2}\langle J^{-1}P(p)R,e^{-J\tau\cdot\diamond}X\rangle. \tag{41}$$

Set $\psi(U):=\frac{1}{2}\langle J^{-1}e^{J\tau\cdot\diamond}\Phi_p,U\rangle$. Then we claim

$$d\psi=\frac{1}{2}\langle J^{-1}e^{J\tau\cdot\diamond}\Phi_p,\ \rangle+p_jd\tau_j,$$

where in this proof we will sum on repeated indexes. The last formula implies

$$\widetilde{B}=d\psi-p_jd\tau_j+\frac{1}{2}\langle J^{-1}P(p)R,e^{-J\tau\cdot\diamond}\ \rangle. \tag{42}$$

The desired formula on $d\psi$ follows by

$$\begin{aligned}
d\psi &= \frac{1}{2}\langle J^{-1}e^{J\tau\cdot\diamond}\Phi_p,\ \rangle+\frac{1}{2}\langle e^{J\tau\cdot\diamond}\diamond_j\Phi_p,U\rangle d\tau_j+\frac{1}{2}\langle e^{J\tau\cdot\diamond}J^{-1}\partial_{p_j}\Phi_p,U\rangle dp_j\\
&= \frac{1}{2}\langle J^{-1}e^{J\tau\cdot\diamond}\Phi_p,\ \rangle+\frac{1}{2}\langle\diamond_j\Phi_p,\Phi_p+P(p)R\rangle d\tau_j\\
&\qquad\qquad +\frac{1}{2}\langle J^{-1}\partial_{p_j}\Phi_p,\Phi_p+P(p)R\rangle dp_j\\
&\overset{\text{by (17)}}{=} \frac{1}{2}\langle J^{-1}e^{J\tau\cdot\diamond}\Phi_p,\ \rangle+\underbrace{\frac{1}{2}\langle\diamond_j\Phi_p,\Phi_p\rangle}_{p_j}d\tau_j+\frac{1}{2}\underbrace{\langle J^{-1}\partial_{p_j}\Phi_p,\Phi_p\rangle}_{0\text{ by (14)}}dp_j.
\end{aligned}$$

By Lemma 2.5 and using $P(p)^*J^{-1}=J^{-1}P(p)$ we have

$$\begin{aligned}
\frac{1}{2}\langle J^{-1}P(p)R,e^{-J\tau\cdot\diamond}\ \rangle &= \frac{1}{2}\langle J^{-1}R,P(p)R'\ \rangle+\frac{1}{2}\langle J^{-1}R,P(p)J\diamond_jP(p)R\rangle d\tau_j\\
&\qquad +\frac{1}{2}\langle J^{-1}R,P(p)\partial_{p_j}P(p)R\rangle dp_j\\
&= \frac{1}{2}\langle J^{-1}R,R'\ \rangle+\frac{1}{2}\langle J^{-1}R,(P(p)-P(p_0))R'\ \rangle\\
&\qquad -\Pi_j(P(p)R)d\tau_j+\frac{1}{2}\langle J^{-1}R,P(p)\partial_{p_j}P(p)R\rangle dp_j.
\end{aligned}$$

So by (42) and using $P(p)J=JP^*(p)$ we get

$$\begin{aligned}
\widetilde{B}-d\psi &= -(\overbrace{p_j+\Pi_j(P(p)R)}^{\Pi_j})d\tau_j+\frac{1}{2}\langle J^{-1}R,R'\ \rangle\\
&\qquad +\frac{1}{2}\langle J^{-1}R,(P(p)-P(p_0))R'\ \rangle-\frac{1}{2}\langle P^*(p)J^{-1}R,\partial_{p_j}P(p)R\rangle dp_j.
\end{aligned}$$

Then $d\alpha=\Omega-\Omega_0$ for

$$\begin{aligned}
\alpha &:= \widetilde{B}-d\psi-B_0+d(\Pi_j\tau_j)\\
&= \frac{1}{2}\langle J^{-1}R,(P(p)-P(p_0))R'\ \rangle-\frac{1}{2}\langle P^*(p)J^{-1}R,\partial_{p_j}P(p)R\rangle dp_j.
\end{aligned}$$

By $p_j = \Pi_j - \Pi_j(P(p)R)$ we get

$$dp_j = d\Pi_j - \langle \diamond_j P(p)R, P(p)R' \rangle - \langle \diamond_j P(p)R, \partial_{p_j} P(p)R \rangle dp_j .$$

Then inserting the next formula in the formula for α, we obtain (40):

$$dp_j = \frac{d\Pi_j - \langle \diamond_j P(p)R, P(p)R' \rangle}{1 + \langle \diamond_j P(p)R, \partial_{p_j} P(p)R \rangle}. \tag{43}$$

□

In the Lemmas 3.2–3.6 we will initially consider the regularity of the functions in terms of the coordinates (τ, p, R).

LEMMA 3.2. *We have $\beta_j \in C^\infty(\mathcal{O} \times \Sigma_{-n}, \mathbb{R})$ for any n. For any pair (n, n') we have $\Gamma \in C^\infty(\mathcal{O}, B(\Sigma_{-n'}, \Sigma_n))$. Summing on repeated indexes, we have*

$$\begin{aligned} d\alpha \;=\; & -\partial_{p_k}\beta_j dp_k \wedge d\Pi_j - \langle \nabla_R \beta_j, R' \rangle \wedge d\Pi_j \\ & + dp_k \wedge \langle \partial_{p_k}[\Gamma(p)R + \beta_j(p,R)P^*(p)\diamond_j P(p)R], R' \rangle \\ & + \langle \nabla_R \beta_j, R' \rangle \wedge \langle P^*(p)\diamond_j P(p)R, R' \rangle + 2\langle \Gamma R', R' \rangle \,. \end{aligned} \tag{44}$$

Proof. Follows from a simple computation. In particular, for a $\mathbf{L} \in B(\Sigma_1, L^2)$ fixed, we use the formula

$$\begin{aligned} d\langle \mathbf{L}R, R' \rangle(X,Y) &:= X\langle \mathbf{L}R, R'Y \rangle - Y\langle \mathbf{L}R, R'X \rangle - \langle \mathbf{L}R, R'[X,Y] \rangle \\ &= \langle \mathbf{L}R'X, R'Y \rangle - \langle \mathbf{L}R'Y, R'X \rangle. \end{aligned}$$

□

LEMMA 3.3. *Summing on repeated indexes, we have*

$$\begin{aligned} d\alpha = {} & \widehat{\delta}_k \partial_{p_k}\beta_j d\Pi_j \wedge d\Pi_k + \langle \widehat{\Gamma}_j + (\widehat{\delta}_k \partial_{p_k}\beta_j - \widehat{\delta}_j \partial_{p_j}\beta_k)\diamond_k P(p)R, R' \rangle \wedge d\Pi_j \\ & + 2\langle \Gamma(p)R', R' \rangle + \langle \widetilde{\beta}_j, R' \rangle \wedge \langle P^*(p)\diamond_j P(p)R, R' \rangle \,, \end{aligned}$$

where we have (this time not summing on repeated indexes)

$$\begin{aligned} \widehat{\delta}_k &:= \frac{1}{1 + \langle \diamond_k P(p)R, \partial_{p_k} P(p)R \rangle} \,, \\ \widehat{\Gamma}_j &:= -\nabla_R \beta_j - \widehat{\delta}_j [\partial_{p_j}\Gamma R + \sum_{i=1}^{n_0} \beta_i \partial_{p_j} \left(P^*(p)\diamond_i P(p)\right) R] \\ &\qquad + \sum_{k=1}^{n_0} (\widehat{\delta}_k \partial_{p_k}\beta_j - \widehat{\delta}_j \partial_{p_j}\beta_k)(P^*(p) - 1)\diamond_k P(p)R \\ \widetilde{\beta}_j &:= \nabla_R \beta_j + \widehat{\delta}_j \partial_{p_j}(\Gamma + \sum_{k=1}^{n_0} \beta_k P^*(p)\diamond_k P(p))R \,. \end{aligned}$$

Proof. Follows by an elementary computation substituting (43) in (44) □

LEMMA 3.4. *For any fixed large n and for $\varepsilon_0 > 0$, consider the set $\mathcal{U}_{\mathbf{d}} \subset \widetilde{\mathcal{P}}^{\mathbf{d}} = \{(p, R)\}$ defined by $\|R\|_{\Sigma_{-n}} \leq \varepsilon_0$ and $|p - p_0| \leq \varepsilon_0$. Then for ε_0 small enough there exists a unique vectorfield $\mathcal{X}^t : \mathcal{U}_{\mathbf{d}} \to \widetilde{\mathcal{P}}$ which solves $i_{\mathcal{X}^t}\Omega_t = -\alpha$, where $\Omega_t := \Omega_0 + t(\Omega - \Omega_0)$.*

Proof. First of all we consider Y such that $i_Y\Omega_0 = -\alpha$, that is to say

$$\begin{aligned}&(Y)_{\tau_j} d\Pi_j - (Y)_{\Pi_j} d\tau_j + \langle J^{-1}(Y)_R, R'\rangle \\ &\quad = \beta_j(p, R)d\Pi_j - \langle \Gamma(p)R + \beta_j(p, R)P^*(p)\diamond_j P(p)R, R'\rangle\,.\end{aligned}$$

This yields

$$\begin{aligned}&(Y)_{\tau_j} = \beta_j(p, R) = \mathcal{R}^{0,2}(p, R)\ , \quad (Y)_{\Pi_j} = 0\ , \\ &(Y)_R = -P(p_0)J\Gamma(p)R - \beta_j(p, R)P(p_0)JP^*(p)\diamond_j P(p)R \\ &\qquad = \mathbf{S}^{1,1}(p - p_0, R) + \mathcal{R}^{0,2}(p, R)P(p_0)P(p)J\diamond_j P(p)R.\end{aligned} \tag{45}$$

Equation $i_{\mathcal{X}^t}\Omega_t = -\alpha$ is equivalent to

$$(1 + t\mathcal{K})\mathcal{X}^t = Y \tag{46}$$

where the operator $\mathcal{K}$ is defined by $i_X d\alpha = i_{\mathcal{K}X}\Omega_0$. In coordinates, (46) becomes $(\mathcal{X}^t)_{\Pi_j} = 0$ and, for $P = P(p)$,

$$(\mathcal{X}^t)_{\tau_j} + t\langle\widehat{\Gamma}_j + (\widehat{\delta}_k\partial_{p_k}\beta_j - \widehat{\delta}_j\partial_{p_j}\beta_k)\diamond_k PR, (\mathcal{X}^t)_R\rangle = -\beta_j, \tag{47}$$

$$(\mathcal{X}^t)_R + t\mathcal{L}(\mathcal{X}^t)_R = (Y)_R\ , \text{where for } X \in N_g^{\perp}(\mathcal{H}^*_{p_0}) \tag{48}$$

$$\mathcal{L}X := P(p_0)J\left[2\Gamma X + \langle\widetilde{\beta}_j, X\rangle P^*\diamond_j PR - \langle P^*\diamond_j PR, X\rangle\widetilde{\beta}_j\right]. \tag{49}$$

(49) implies the following lemma.

LEMMA 3.5. *We have, summing on repeated indexes, with i varying in some finite set,*

$$\mathcal{L}X = \mathcal{A}_j(X)J\diamond_j R + \mathcal{B}_i(X)\Psi_i \tag{50}$$

where: $\Psi_i = \mathbf{S}^{0,0}(p - p_0, R)$; for $L = \mathcal{A}_j, \mathcal{B}_i$, we have $L \in C^\infty(\mathcal{U}_{\mathbf{d}}, B(L^2, \mathbb{R}))$ with

$$L(X) = L_j\,\langle\diamond_j R, X\rangle + \langle\widetilde{L}, X\rangle, \tag{51}$$

where we have $\widetilde{L} = \mathbf{S}^{1,0}(p - p_0, R)$ and $L_j \in \mathcal{R}^{0,0}(p - p_0, R)$.

Proof. Schematically, for $\widetilde{L}_i = \mathbf{S}^{0,0}(p-p_0,R)$ and $\Psi_i = \mathbf{S}^{0,0}(p-p_0,R)$ we have

$$P(p)R = R - P_{N_g}(p)R = R + \sum_i \langle \widetilde{L}_i, R\rangle \Psi_i\,,$$
$$P^*(p)\diamond_k R = \diamond_k R - P^*_{N_g}(p)\diamond_k R = \diamond_k R + \sum_i \langle \widetilde{L}_i, R\rangle \Psi_i.$$

Then $(P^*(p)\diamond_k P(p) - \diamond_k)R = \mathbf{S}^{0,1}(p-p_0,R)$.
By the definition of $\widetilde{\beta}_j$ we have

$$\widetilde{\beta}_j = \sum_k \widehat{\delta}_j(\partial_{p_j}\beta_k)\diamond_k R + \widehat{L}$$
$$\widehat{L} := \nabla_R \beta_j + \frac{1}{2}J^{-1}\widehat{\delta}_j \partial_{p_j} P(p)R + \sum_k \beta_k \partial_{p_j}(P^*(p)\diamond_k P(p))R$$
$$- \sum_k \widehat{\delta}_j \partial_{p_j}\beta_k \left[P^*_{N_g}(p)\diamond_k P(p)R + \diamond_k P_{N_g}(p)R\right],$$

where $\widehat{L} = \mathbf{S}^{0,1}_{n,\infty}(p-p_0,R)$.
We also have $\Gamma X = \frac{1}{2}J^{-1}(P_{N_g}(p_0) - P_{N_g}(p))X = \sum_i \langle \widetilde{L}_i, X\rangle \Psi_i$ with $\widetilde{L}_i = \mathbf{S}^{1,0}(p-p_0,R)$ and $\Psi_i = \mathbf{S}^{0,0}(p-p_0,R)$. This yields the result. □

LEMMA 3.6. *System* (47)–(49) *admits exactly one solution* $\mathcal{X}^t$. *For* $\mathcal{A}_j = \mathcal{R}^{0,2}_{n,\infty}(t,p-p_0,R)$, $\mathcal{D} = \mathbf{S}^{1,1}_{n,\infty}(t,p-p_0,R)$ *with* $|t|<3$, *we have*

$$(\mathcal{X}^t)_R = \mathcal{A}_j J \diamond_j R + \mathcal{D}. \tag{52}$$

Proof. Recall Y defined by $i_Y\Omega_0 = -\alpha$. By (45) with $\widetilde{\mathcal{A}}_j = \mathcal{R}^{0,2}_{n,\infty}(p-p_0,R)$ and $\widetilde{\mathcal{D}} = \mathbf{S}^{1,1}_{n,\infty}(p-p_0,R)$ we have $(Y)_R = \widetilde{\mathcal{A}}_j J\diamond_j R + \widetilde{\mathcal{D}}$. By $(\mathcal{X}^t)_R + t\mathcal{L}(\mathcal{X}^t)_R = (Y)_R$ and Lemma 3.5 this implies for $X = (\mathcal{X}^t)_R$

$$\langle \diamond_k R, X\rangle + t\mathcal{B}_i(X)\langle \diamond_k R, \Psi_i\rangle = \langle \diamond_k R, (Y)_R\rangle$$
$$\langle \widetilde{L}, X\rangle + t\mathcal{A}_j(X)\langle \widetilde{L}, J\diamond_j R\rangle + t\mathcal{B}_i(X)\langle \widetilde{L}, \Psi_i\rangle = \langle \widetilde{L}, (Y)_R\rangle,$$

as L runs through all the $L = \mathcal{A}_j, \mathcal{B}_i$. Taking appropriate linear combinations of these equations with the coefficients L_j of $L = \mathcal{A}_j, \mathcal{B}_i$, see Lemma 3.5, for a matrix $\mathbf{R}^{0,1}(p-p_0,R)$ whose coefficients are $\mathcal{R}^{0,1}(p-p_0,R)$, we get

$$(1 + t\mathbf{R}^{0,1}(p-p_0,R)) \begin{pmatrix} \mathcal{A}_j((\mathcal{X}^t)_R) \\ \mathcal{B}_i((\mathcal{X}^t)_R) \end{pmatrix} = \begin{pmatrix} \mathcal{A}_j((Y)_R) \\ \mathcal{B}_i((Y)_R) \end{pmatrix}.$$

Then we get

$$\begin{pmatrix} \mathcal{A}_j((\mathcal{X}^t)_R) \\ \mathcal{B}_i((\mathcal{X}^t)_R) \end{pmatrix} = (1 + t\mathbf{R}^{0,1}(p-p_0,R))^{-1} \begin{pmatrix} \mathcal{A}_j((Y)_R) \\ \mathcal{B}_i((Y)_R) \end{pmatrix}. \tag{53}$$

Using the left hand side of (53) set

$$\mathcal{L}(\mathcal{X}^t)_R := \mathcal{A}_j((\mathcal{X}^t)_R)J\diamondsuit_j R + \mathcal{B}_i((\mathcal{X}^t)_R)\Psi_i. \tag{54}$$

The rhs of (54) satisfies the properties stated for the rhs of (52). Finally set $(\mathcal{X}^t)_R := (Y)_R - t\mathcal{L}(\mathcal{X}^t)_R$. This is a solution of (48). It is elementary to see from the argument that such solution is unique and that it satisfies the properties of the statement. □

With the proof of Lemma 3.6, the proof of Lemma 3.4 is completed. □

Turning to coordinates (τ, Π, R) and by Lemma 2.11 we conclude what follows.

LEMMA 3.7. *Consider the coordinate system (τ, Π, R). For G any of the $\mathcal{A}_j$, $\mathcal{D}$ in Lemma 3.6, we have $G = G(\Pi, \Pi(R), R)$, with $G(\Pi, \varrho, R)$ smooth w.r.t. $(\Pi, \varrho, R) \in \mathcal{U}_{\boldsymbol{d}}$, with $\mathcal{U}_{\boldsymbol{d}}$ formed by the $(\Pi, \varrho, R) \in \mathbb{R}^{2n_0} \times (\Sigma_{\boldsymbol{d}} \cap N_g^{\perp}(\mathcal{H}_{p_0}))$ defined by the inequalities $\|R\|_{\Sigma_{-n}} \leq \varepsilon$, $|\varrho| \leq \varepsilon$ and $|\Pi - p_0| \leq \varepsilon$ for $\varepsilon > 0$ small enough.*

3.2. Flows

The following lemma is repeatedly used in the sequel, see [2, Lemma 3.24].

LEMMA 3.8. *Below we pick $r, M, M_0, s, s', k, l \in \mathbb{N} \cup \{0\}$ with $1 \leq l \leq M$. Consider a system*

$$\begin{aligned} &\dot\tau_j = T_j(t, \Pi, \Pi(R), R)\ , \quad \dot\Pi_j = 0\ , \\ &\dot R = \mathcal{A}_j(t, \Pi, \Pi(R), R)J\diamondsuit_j R + \mathcal{D}(t, \Pi, \Pi(R), R), \end{aligned} \tag{55}$$

where we assume what follows.

- $P_{N_g(p_0)}(\mathcal{A}_j J\diamondsuit_j R + \mathcal{D}) \equiv 0$.
- *At $\Pi = p_0$, dropping the dependence on Π and for $\mathcal{U}_{-r}$ a neighborhood of 0 in $\mathcal{P}^{-r}$, we have $\mathcal{A}(t, \varrho, R) \in C^M((-3, 3) \times \mathcal{U}_{-r}, \mathbb{R}^{n_0})$ and $\mathcal{D}(t, \varrho, R) \in C^M((-3, 3) \times \mathcal{U}_{-r}, \Sigma_r)$*
- *In $(-3, 3) \times \mathcal{U}_{-r}$ for a fixed i in $\{0, 1\}$, and a fixed C_r, we have:*

$$\begin{aligned} &|\mathcal{A}(t, \varrho, R)| \leq C\|R\|_{\Sigma_{-r}}^{M_0+1}, \\ &\|\mathcal{D}(t, \varrho, R)\|_{\Sigma_r} \leq C(|\varrho| + \|R\|_{\Sigma_{-r}})^i \|R\|_{\Sigma_{-r}}^{M_0}. \end{aligned} \tag{56}$$

Let $k \in \mathbb{Z} \cap [0, r-(l+1)\boldsymbol{d}]$ and set for $s'' \geq \boldsymbol{d}$ (or $s'' \geq \boldsymbol{d}/2$ if $\boldsymbol{d}/2 \in \mathbb{N}$)

$$\mathcal{U}^{s''}_{\varepsilon_1,k} := \{(\tau, \Pi, R) \in \widetilde{\mathcal{P}}^{s''} \ : \ \Pi = p_0 \ , \ \|R\|_{\Sigma_{-k}} + |\Pi(R)| \leq \varepsilon_1\}. \tag{57}$$

Then for $\varepsilon_1 > 0$ small enough, the initial value problem associated to (55) for $\Pi = p_0$ defines a flow $\mathfrak{F}^t = (\mathfrak{F}^t_\tau, \mathfrak{F}^t_R)$ for $t \in [-2,2]$ in $\mathcal{U}^{\boldsymbol{d}}_{\varepsilon_1,k}$. In particular for $\Pi = p_0$, for R in a neighborhood $B_{\Sigma_{-k}}$ of 0 in Σ_{-k} and $\Pi(R)$ in a neighborhood $B_{\mathbb{R}^{n_0}}$ of 0 in $\mathbb{R}^{n_0}$, we have

$$\mathfrak{F}^t_R(\Pi(R), R) = e^{Jq(t,\Pi(R),R)\cdot\diamond}(R + \boldsymbol{S}(t, \Pi(R), R)), \tag{58}$$

$$\begin{gathered} \text{with } \boldsymbol{S} \in C^l((-2,2) \times B_{\mathbb{R}^{n_0}} \times B_{\Sigma_{-k}}, \Sigma_{r-(l+1)\boldsymbol{d}}) \\ q \in C^l((-2,2) \times B_{\mathbb{R}^{n_0}} \times B_{\Sigma_{-k}}, \mathbb{R}^{n_0}). \end{gathered} \tag{59}$$

For fixed $C > 0$ we have

$$\begin{aligned} &|q(t, \varrho, R)| \leq C\|R\|^{M_0+1}_{\Sigma_{(l+1)\boldsymbol{d}-r}}, \\ &\|\boldsymbol{S}(t, \varrho, R)\|_{\Sigma_{r-(l+1)\boldsymbol{d}}} \leq C(|\varrho| + \|R\|_{\Sigma_{(l+1)\boldsymbol{d}-r}})^i \|R\|^{M_0}_{\Sigma_{(l+1)\boldsymbol{d}-r}}. \end{aligned} \tag{60}$$

Furthermore we have $\boldsymbol{S} = \boldsymbol{S}_1 + \boldsymbol{S}_2$ with

$$\begin{aligned} &\boldsymbol{S}_1(t, \Pi(R), R) = \int_0^t \mathcal{D}(t', \Pi(R(t')), R(t'))dt' \\ &\|\boldsymbol{S}_2(t, \varrho, R)\|_{\Sigma_s} \leq C\|R\|^{2M_0+1}_{\Sigma_{(l+1)\boldsymbol{d}-r}}(|\varrho| + \|R\|_{\Sigma_{(l+1)\boldsymbol{d}-r}})^i. \end{aligned} \tag{61}$$

For $r-(l+1)\boldsymbol{d} \geq s' \geq s + l\mathbf{d} \geq l\mathbf{d}$ and $k \in \mathbb{Z} \cap [0, r-(l+1)\boldsymbol{d}]$ and for $\varepsilon_1 > 0$ sufficiently small, we have

$$\mathfrak{F}^t \in C^l((-2,2) \times \mathcal{U}^{s'}_{\varepsilon_1,k}, \widetilde{\mathcal{P}}^s). \tag{62}$$

Furthermore, there exists $\varepsilon_2 > 0$ such that

$$\mathfrak{F}^t(\mathcal{U}^{s'}_{\varepsilon_2,k}) \subset \mathcal{U}^{s'}_{\varepsilon_1,k} \text{ for all } |t| \leq 2 \ . \tag{63}$$

We have

$$\mathfrak{F}^t(e^{J\tau\cdot\diamond}U) \equiv e^{J\tau\cdot\diamond}\mathfrak{F}^t(U). \tag{64}$$

Proof. It is enough to focus on the equation for R. Set $S = e^{-Jq\cdot\diamond}R$ for $q \in \mathbb{R}^{n_0}$. Then consider the following system:

$$\begin{aligned} &\dot{S} = e^{-Jq\cdot\diamond}\mathcal{D}(t, \varrho, e^{Jq\cdot\diamond}S) \ , \\ &\dot{q} = \mathcal{A}(t, \varrho, e^{Jq\cdot\diamond}S) \quad , \quad q(0) = 0, \\ &\dot{\varrho}_j = \langle S, e^{-Jq\cdot\diamond}\diamond_j \mathcal{D}(t, \varrho, e^{Jq\cdot\diamond}S)\rangle \ . \end{aligned} \tag{65}$$

For $l \leq M$ and $k, s'' \in [0, r-(l+1)\mathbf{d}]$ the field in (65) is $C^l((-3,3)\times\mathcal{U}_{-k}, \Sigma_{s''}\times\mathbb{R}^{2n_0})$ with $\mathcal{U}_{-k} \subset \Sigma_{-k}\times\mathbb{R}^{2n_0}$ a neighborhood of the equilibrium 0. This follows from the fact that $(q, X) \to e^{Jq\cdot\diamond}X$ is in $C^l(\mathbb{R}^{n_0}\times\Sigma_\ell, \Sigma_{\ell-l\mathbf{d}})$ for all $\ell \in \mathbb{Z}$ and from the hypotheses on $\mathcal{A}$ and $\mathcal{D}$. For example

$$(t, q, \varrho, S) \to e^{-Jq\cdot\diamond}\diamond_j \mathcal{D}(t, \varrho, e^{Jq\cdot\diamond}S) \in C^l((-3,3)\times\mathbb{R}^{2n_0}\times\Sigma_{l\mathbf{d}-r}, \Sigma_{r-(l+1)\mathbf{d}}),$$

(more precisely for (q, ϱ, S) in a neighborhood of the origin). So

$$(t, q, \varrho, S) \to \langle S, e^{-Jq\cdot\diamond}\diamond_j \mathcal{D}(t, \varrho, e^{Jq\cdot\diamond}S)\rangle,$$

is in $C^l((-3,3)\times\mathbb{R}^{2n_0}\times\Sigma_{-k}, \mathbb{R})$ for $k \leq r-(l+1)\mathbf{d}$ (for (q, ϱ, S) near origin). For $l \geq 1$ we can apply to (65) standard theory of ODE's to conclude that there are neighborhoods of the origin $B_{\mathbb{R}^{2n_0}} \subset \mathbb{R}^{2n_0}$ and $B_{\Sigma_{-k}} \subset \Sigma_{-k}$ such that the flow is of the form

$$\begin{aligned} &S(t) = R + \mathbf{S}(t, \varrho, R)\ , \quad \mathbf{S}(0, \varrho, R) = 0\ , \\ &q(t) = q(t, \varrho, R)\ , \quad q(0, \varrho, R) = 0\ , \\ &\varrho(t) = \varrho + \overline{\varrho}(t, \varrho, R)\ , \quad \overline{\varrho}(0, \varrho, R) = 0\ , \end{aligned} \tag{66}$$

$$\begin{aligned} &\text{with } \mathbf{S} \in C^l((-2,2)\times B_{\mathbb{R}^{n_0}}\times B_{\Sigma_{-k}}, \Sigma_{r-(l+1)\mathbf{d}}) \\ &\quad \overline{\varrho}, q(t, \varrho, R) \in C^l((-2,2)\times B_{\mathbb{R}^{n_0}}\times B_{\Sigma_{-k}}, \mathbb{R}^{n_0}). \end{aligned} \tag{67}$$

For $S \in \Sigma_{\mathbf{d}} \cap B_{\Sigma_{-k}}$ and $S(0) = S$, choosing $s'' \geq \mathbf{d}$ we have $S(t) \in \Sigma_{\mathbf{d}}$ with $\Pi(S(t)) = \varrho(t)$ for $\varrho(0) = \varrho = \Pi(S)$. Then (67) yields (59) (we can replace $\Sigma_{\mathbf{d}}$ with $\Sigma_{\frac{\mathbf{d}}{2}}$ if $\frac{\mathbf{d}}{2} \in \mathbb{N}$). (58) and (59) yield (62).
We have for $R(0) = R$

$$R(t) = e^{Jq(t)\cdot\diamond}(R + \int_0^t e^{-Jq(t')\cdot\diamond}\mathcal{D}(t', \varrho(t'), R(t'))dt'). \tag{68}$$

By (A6), for $\epsilon = 0$, and by (56), for $|s''| \leq r-(l+1)\mathbf{d}$ we have

$$\begin{aligned} \|R(t)\|_{\Sigma_{s''}} &\leq C\|R\|_{\Sigma_{s''}} + C\int_0^t \|\mathcal{D}(t', \varrho(t), R(t'))\|_{\Sigma_r}dt' \\ &\leq C\|R\|_{\Sigma_{s''}} + C\int_0^t \|R(t')\|^{M_0}_{\Sigma_{-r}}(|\varrho(t')| + \|R(t')\|_{\Sigma_{-r}})^i dt' \\ &\leq C\|R\|_{\Sigma_{s''}} + C\int_0^t \|R(t')\|^{M_0}_{\Sigma_{s''}}(|\varrho(t')| + \|R(t')\|_{\Sigma_{s''}})^i dt', \end{aligned} \tag{69}$$

with the caveat that the second line is purely formal and is used to get the third

line, where the integrand is continuous. Proceeding similarly, for $\varrho(0) = \varrho$

$$\begin{aligned} |\varrho(t) - \varrho| &\leq \int_0^t |\langle R(t'), \diamond \mathcal{D}(t', R(t'), \varrho(t'))\rangle| dt' \\ &\leq \int_0^t \|R(t')\|_{\Sigma_{(l+1)\mathbf{d}-r}} \|\mathcal{D}(t', \varrho(t), R(t'))\|_{\Sigma_{r-l\mathbf{d}}} dt' \\ &\leq C \int_0^t \|R(t')\|^{M_0+1}_{\Sigma_{(l+1)\mathbf{d}-r}} (|\varrho(t')| + \|R(t')\|_{\Sigma_{(l+1)\mathbf{d}-r}})^i dt'. \end{aligned} \tag{70}$$

So for $|s''| \leq r - (l+1)\mathbf{d}$, using the continuity in t' of the integrals in the last lines of (69) and (70), by the Gronwall inequality there is a fixed C such that for all $|t| \leq 2$ we have

$$\|R(t)\|_{\Sigma_{s''}} \leq C \|R\|_{\Sigma_{s''}}, \tag{71}$$

$$|\varrho(t) - \varrho| \leq C \|R\|^{M_0+1}_{\Sigma_{(l+1)\mathbf{d}-r}} (|\varrho| + \|R\|_{\Sigma_{(l+1)\mathbf{d}-r}})^i. \tag{72}$$

By (71) for $s'' = s'$ and $s'' = -k$ and by $|\varrho(t) - \varrho| \leq C\|R\|^{M_0+1}_{\Sigma_{-k}} (|\varrho| + \|R\|_{\Sigma_{-k}})^i$, we get $\mathfrak{F}^t(\mathcal{U}^{s'}_{\varepsilon_2,k}) \subset \mathcal{U}^{s'}_{\varepsilon_1,k}$ for all $|t| \leq 2$ for $\varepsilon_1 \gg \varepsilon_2$, that is (63).
We have

$$S(t, \varrho, R) = \int_0^t e^{-Jq(t')\cdot\diamond} \mathcal{D}(t', \varrho(t'), R(t')) dt'),$$

Proceeding as for (69) and using (71)–(72) we get the estimate for $\mathbf{S}$ in (60). The estimate on q is obtained similarly integrating the second equation in (66). We have

$$\mathbf{S}_2(t, R, \varrho) = \int_0^1 dt'' \int_0^t e^{-t''q(t')\cdot\diamond} q(t') \cdot \diamond \mathcal{D}(t', \varrho(t), R(t')) dt' \tag{73}$$

Then by (71)–(72) we get

$$\begin{aligned} \|\mathbf{S}_2(t, R, \varrho)\|_{\Sigma_{r-\mathbf{d}}} &\leq C'' \int_0^t |q(t')| \|\mathcal{D}(t', \varrho(t), R(t'))\|_{\Sigma_{r-\mathbf{d}}} dt' \\ &\leq C' \int_0^t \|R(t')\|^{2M_0+1}_{\Sigma_{(l+1)\mathbf{d}-r}} (|\varrho(t')| + \|R(t')\|_{\Sigma_{(l+1)\mathbf{d}-r}})^i dt' \\ &\leq C \|R\|^{2M_0+1}_{\Sigma_{(l+1)\mathbf{d}-r}} (|\varrho| + \|R\|_{\Sigma_{(l+1)\mathbf{d}-r}})^i. \end{aligned} \tag{74}$$

This yields (61). (62) follows by (58)–(59). Finally, (64) follows immediately from (58). □

LEMMA 3.9. *Assume hypotheses and conclusions of Lemma 3.8. Consider the flow of system* (65) *for* $\Pi = p_0$ *. Denote the flow in the space with variables* $\{(\varrho, R)\}$ *by* $\mathfrak{F}^t = (\mathfrak{F}^t_\varrho, \mathfrak{F}^t_R)$*. Then we have*

$$\begin{aligned} \mathfrak{F}^t_R(\varrho, R) &= e^{Jq(t,\varrho,R)\cdot\diamond}(R + \boldsymbol{S}(t, \varrho, R)) \\ \mathfrak{F}^t_\varrho(\varrho, R) &= \varrho + \overline{\varrho}(t, \varrho, R). \end{aligned} \tag{75}$$

Furthermore, the following facts hold.

(1) Let $k \in \mathbb{Z} \cap [0, r-(l+1)\boldsymbol{d}]$ and $h \geq \max\{k + l\boldsymbol{d}, (2l+1)\boldsymbol{d} - r\}$. Then we have $\mathfrak{F}^t \in C^l((-2,2) \times \mathcal{U}_{-k}, \mathcal{P}^{-h})$ for a neighborhood of the origin $\mathcal{U}_{-k} \subset \mathcal{P}^{-k}$.

(2) Let h and k be like above with $h \leq r-(l+1)\boldsymbol{d}$. Then given a function $\mathcal{R}^{a,b}_{h,l}(\varrho, R)$, we have $\mathcal{R}^{a,b}_{h,l} \circ \mathfrak{F}^t = \mathcal{R}^{a,b}_{k,l}(t, \varrho, R)$ and given a function $\boldsymbol{S}^{a,b}_{h,l}(\varrho, R)$, we have $\boldsymbol{S}^{a,b}_{h,l} \circ \mathfrak{F}^t = \boldsymbol{S}^{a,b}_{k,l}(t, \varrho, R)$.

Proof. (75) follows by (66). By (67) we have

$$\mathbf{S} \in C^l((-2,2) \times \mathcal{U}_{-k}, \Sigma_{r-(l+1)\mathbf{d}}) \ , \quad q \text{ and } \mathfrak{F}^t_\varrho \in C^l((-2,2) \times \mathcal{U}_{-k}, \mathbb{R}^{n_0}).$$

By the above formulas we have $\mathfrak{F}^t_R \in C^l((-2,2) \times \mathcal{U}_{-k}, \Sigma_{r-(2l+1)\mathbf{d}} \cap \Sigma_{-k-l\mathbf{d}})$. This yields $\mathfrak{F}^t_R \in C^l((-2,2) \times \mathcal{U}_{-k}, \Sigma_{-h})$ and yields Claim (1).
By Claim (1), $\mathcal{R}^{a,b}_{h,l} \circ \mathfrak{F}^t \in C^l((-2,2) \times \mathcal{U}_{-k}, \mathbb{R}^{n_0})$. Let $(\varrho^t, R^t) = \mathfrak{F}^t(\varrho, R)$. Then

$$\begin{aligned} |\mathcal{R}^{a,b}_{h,l} \circ \mathfrak{F}^t(\varrho, R)| = |\mathcal{R}^{a,b}_{h,l}(\varrho^t, R^t)| &\leq C' \|R^t\|^b_{\Sigma_{-h}} (\|R^t\|_{\Sigma_{-h}} + |\varrho^t|)^a \\ &\leq C \|R\|^b_{\Sigma_{-h}} (\|R\|_{\Sigma_{-h}} + |\varrho|)^a \leq C \|R\|^b_{\Sigma_{-k}} (\|R\|_{\Sigma_{-k}} + |\varrho|)^a, \end{aligned}$$

where the first inequality uses Definition (32), the second uses (71)–(72) for $s'' = -h$ and the last is obvious. Similarly by Claim (1), $\mathbf{S}^{a,b}_{h,l} \circ \mathfrak{F}^t \in C^l((-2,2) \times \mathcal{U}_{-k}, \Sigma_h) \subset C^l((-2,2) \times \mathcal{U}_{-k}, \Sigma_k)$ and

$$\begin{aligned} \|\mathbf{S}^{a,b}_{h,l}(\varrho^t, R^t)\|_{\Sigma_k} \leq \|\mathbf{S}^{a,b}_{h,l}(\varrho^t, R^t)\|_{\Sigma_h} &\leq C' \|R^t\|^b_{\Sigma_{-h}} (\|R^t\|_{\Sigma_{-h}} + |\varrho^t|)^a \\ &\leq C \|R\|^b_{\Sigma_{-h}} (\|R\|_{\Sigma_{-h}} + |\varrho|)^a \leq C \|R\|^b_{\Sigma_{-k}} (\|R\|_{\Sigma_{-k}} + |\varrho|)^a. \end{aligned}$$

□

To prove Theorem 6.4 we will need more information on $(\Pi(R(1)), R(1))$. This is provided by the following lemma.

LEMMA 3.10. *Consider, for $\mathcal{D}$ the function in (55) at $\Pi = p_0$, the system*

$$\dot{S}(t) = \mathcal{D}(t, \Pi(R_0), S(t)) \, , \quad S(0) = R_0. \tag{76}$$

Then for $S' = S(1)$ and for $R' = R(1)$ with $R(t)$ the solution of (55) with $R(0) = R_0$, we have (same indexes of Lemma 3.8)

$$\begin{aligned} &\|R' - S'\|_{\Sigma_{-s'}} \leq C \|R_0\|^{M_0+2}_{\Sigma_{-s}} \, , \\ &\Pi(R') - \Pi(S') = \mathcal{R}^{i,2M_0+1}_{s,l}(\Pi(R_0), R_0). \end{aligned} \tag{77}$$

Proof. Recall that for $\varrho = \Pi(R)$ we have $\dot\varrho = \langle R, \diamond\mathcal{D}(t,\varrho,R)\rangle$. Similarly, for $\sigma = \Pi(S)$ we have $\dot\sigma = \langle S, \diamond\mathcal{D}(t,\varrho_0,S)\rangle$, where $\varrho_0 = \Pi(R_0)$. So we have

$$\begin{aligned}\dot\varrho - \dot\sigma &= \langle R, \diamond\mathcal{D}(t,\varrho,R)\rangle - \langle S, \diamond\mathcal{D}(t,\varrho_0,S)\rangle\\ &= \langle R-S, \diamond\mathcal{D}(t,\varrho,R)\rangle + \langle S, \diamond(\mathcal{D}(t,\varrho_0,S) - \mathcal{D}(t,\varrho,R))\rangle.\end{aligned}$$

By (56) for fixed constants and using $s' \le r - \mathbf{d}$, we have

$$\begin{aligned}|\dot\varrho - \dot\sigma| &\lesssim \|R-S\|_{\Sigma_{-s'}}\|\mathcal{D}(t,\varrho,R)\|_{\Sigma_r} + \|S\|_{\Sigma_{-s'}}\|\mathcal{D}(t,\varrho_0,S) - \mathcal{D}(t,\varrho,R)\|_{\Sigma_r}\\ &\lesssim \|R-S\|_{\Sigma_{-s'}}\|R\|^{M_0}_{\Sigma_{-s}}(|\varrho| + \|R\|_{\Sigma_{-s'}})^i + |\varrho-\varrho_0|\ \|S\|_{\Sigma_{-s'}}\|(R,S)\|^{M_0}_{\Sigma_{-s'}}\\ &\quad + \|R-S\|_{\Sigma_{-s'}}\|S\|_{\Sigma_{-s'}}\|(R,S)\|^{M_0-1}_{\Sigma_{-s'}}(|(\varrho,\varrho_0)| + \|(R,S)\|_{\Sigma_{-s'}})^i.\end{aligned}$$

We have $\dot R - \dot S = \mathcal{D}(t,\varrho,R) - \mathcal{D}(t,\varrho_0,S) + J\mathcal{A}(t,\varrho,R)(t,\varrho,R)\cdot\diamond R$ and hence for fixed constants we have, using $s \le s' - \mathbf{d}$,

$$\begin{aligned}\|R-S\|_{\Sigma_{-s'}} &\le \int_0^t [\|\mathcal{D}(\varrho,R) - \mathcal{D}(\varrho_0,S)\|_{\Sigma_{-s'}} + |\mathcal{A}|\|R\|_{\Sigma_{-s}}]dt'\\ &\lesssim \int_0^t \big[\|R-S\|_{\Sigma_{-s'}}\|(R,S)\|^{M_0-1}_{\Sigma_{-s'}}(|(\varrho,\varrho_0)| + \|(R,S)\|_{\Sigma_{-s'}})^i\\ &\qquad + |\varrho-\varrho_0|\ \|(R,S)\|^{M_0}_{\Sigma_{-s'}} + \|R\|^{M_0+2}_{\Sigma_{-s}}\big]dt'.\end{aligned}$$

Recall that $|\varrho-\varrho_0| \le C\|R_0\|^{M_0+1}_{\Sigma_{(l+1)\mathbf{d}-r}}(|\varrho_0| + \|R_0\|_{\Sigma_{(l+1)\mathbf{d}-r}})^i$ by (72), that $s < r-(l+1)\mathbf{d}$ and that we have (71) for $s'' = -s, -s'$. Then by Gronwall inequality, the above inequalities yield

$$\begin{aligned}&\|R(t) - S(t)\|_{\Sigma_{-s'}} \le C\|R_0\|^{M_0+2}_{\Sigma_{-s}}\\ &|\varrho(t) - \sigma(t)| \le C\|R_0\|^{2M_0+1}_{\Sigma_{-s}}(|\varrho_0| + \|R_0\|_{\Sigma_{-s}})^i.\end{aligned} \tag{78}$$

This yields the bounds implicit in (77). The regularity follows from Lemma 3.8. □

3.3. Darboux Theorem: end of the proof

Formally the proof should follow by $i_{\mathcal{X}^t}\Omega_t = -\alpha$, where $\Omega_t = (1-t)\Omega_0 + t\Omega$, and by

$$\frac{d}{dt}\left(\mathfrak{F}_t^*\Omega_t\right) = \mathfrak{F}_t^*\left(L_{\mathcal{X}_t}\Omega_t + \frac{d}{dt}\Omega_t\right) = \mathfrak{F}_t^*\left(di_{\mathcal{X}^t}\Omega_t + d\alpha\right) = 0. \tag{79}$$

But while for [4, 8] the above formal computation falls within the classical framework of flows, fields and differential forms, in the case of [2, 10] this is

not rigorous. In order to justify rigorously this computation, we will consider first a regularization of system (55).

LEMMA 3.11. *Consider the system*

$$\begin{aligned} &\dot\tau_j = T_j(t,\Pi,\Pi(R),R)\ , \quad \dot\Pi_j = 0\ ,\\ &\dot R = \mathcal{A}_j(t,\Pi,\Pi(R),R)J\langle\epsilon\diamondsuit\rangle^{-2}\diamondsuit_j R + \mathcal{D}_\epsilon(t,\Pi,\Pi(R),R), \end{aligned} \tag{80}$$

where $\mathcal{D}_\epsilon = \mathcal{D} + \mathcal{A}_j P_{N_g(p_0)} J\diamondsuit_j(1-\langle\epsilon\diamondsuit\rangle^{-2})R$.

(1) For $|\epsilon|\le 1$ *system* (80) *satisfies all the conclusions of Lemma 80, if we replace* $\diamondsuit$ *in* (58) *with* $\langle\epsilon\diamondsuit\rangle^{-2}\diamondsuit$ *(resp.* $\mathcal{D}$ *in* (61) *with* $\mathcal{D}_\epsilon$*), with a fixed choice of constants* ε_1, ε_2, C, *and with a fixed choice of sets* $B_{\mathbb{R}^{n_0}}$, $B_{\Sigma_{-s}}$.

(2) For $\mathcal{X}^t$ *the vector field of* (55)*, denote by* $\mathcal{X}^t_\epsilon$ *the vector field of* (80)*. Let* $n' > n + \boldsymbol{d}$ *with* $n, n' \in \mathbb{N}$. *Then for* $k \in \mathbb{Z}\cap[0,r]$ *we have*

$$\lim_{\epsilon\to 0}\mathcal{X}^t_\epsilon = \mathcal{X}^t \text{ in } C^M((-3,3)\times\mathcal{U}^{n'}_{\varepsilon_0,k},\widetilde{\mathcal{P}}^n) \text{ uniformly locally,} \tag{81}$$

that is uniformly on subsets of $(-3,3)\times\mathcal{U}^{n'}_{\varepsilon_0,k}$ *bounded in* $(-3,3)\times\widetilde{\mathcal{P}}^{n'}$.

(3) Denote by $\mathfrak{F}^t_\epsilon = (\mathfrak{F}^t_{\epsilon\tau},\mathfrak{F}^t_{\epsilon R})$ *the flow associated to* (80) *at* $\Pi = p_0$. *Let* s',s *and* k *as in the statement of Lemma 3.8. Then there is a pair* $0<\varepsilon_1<\varepsilon_0$ *such that*

$$\lim_{\varepsilon\to 0}\mathfrak{F}^t_\epsilon = \mathfrak{F}^t \text{ in } C^{l-1}([-1,1]\times\mathcal{U}^{s'}_{\varepsilon_1,k},\mathcal{U}^{s}_{\varepsilon_0,k}) \text{ uniformly locally.} \tag{82}$$

Proof. For claim (1), it is enough to check that $\mathcal{D}_\epsilon$ satisfies an estimate like the one of $\mathcal{D}$ in (60) for a fixed C for all $|\epsilon|\le 1$. Indeed, after this has been checked, the proof of Lemma 55 can be repeated verbatim, exploiting (A6) for $\epsilon\neq 0$ and with $\diamondsuit$ replaced by $\langle\epsilon\diamondsuit\rangle^{-2}\diamondsuit$.
The estimate on $\mathcal{D}_\epsilon$ needed for Claim (1) follows by the definition of $\mathcal{D}_\epsilon$, by the estimate on $\mathcal{D}$, by $P_{N_g(p_0)} = \mathbf{e}_a\langle\mathbf{e}^*_a,\ \rangle$ (sum on repeated indexes) for Schwartz functions $\mathbf{e}_a$ and $\mathbf{e}^*_a$ and, for $n\in\mathbb{N}$ with $n-1\ge s+\mathbf{d}$, and by

$$\begin{aligned} &\|P_{N_g(p_0)}J\diamondsuit_i(1-\langle\epsilon\diamondsuit\rangle^{-2})\|_{B(\Sigma_{-r},\Sigma_r)}\\ &\quad\le \|\mathbf{e}_a\langle J\diamondsuit_i(1-\langle\epsilon\diamondsuit\rangle^{-2})\mathbf{e}^*_a,\ \rangle\|_{B(\Sigma_{-r},\Sigma_r)}\\ &\qquad\le \|\mathbf{e}_a\|_{\Sigma_r}\|(1-\langle\epsilon\diamondsuit\rangle^{-2})\mathbf{e}^*_a\|_{\Sigma_{r+\mathbf{d}}} \le C(\epsilon)\|\mathbf{e}_a\|_{\Sigma_r}\ \|\mathbf{e}^*_a\|_{\Sigma_{r'}} \end{aligned} \tag{83}$$

$C(\epsilon) = \|\diamondsuit(1-\langle\epsilon\diamondsuit\rangle^{-2})\|_{B(\Sigma_{r'},\Sigma_{r+\mathbf{d}})}$ is bounded by (4) for $|\epsilon|\le 1$ for any pair (r',r) with $r' > r+\mathbf{d}$.
We consider now Claim (2). We have

$$\mathcal{X}^t - \mathcal{X}^t_\epsilon = \mathcal{A}_j(t,\varrho,R)\left(J(1-\langle\epsilon\diamondsuit\rangle^{-2})\diamondsuit_j R - P_{N_g(p_0)}J\diamondsuit_j(1-\langle\epsilon\diamondsuit\rangle^{-2})R\right).$$

We have $P_{N_g(p_0)}J\diamondsuit_j(1-\langle\epsilon\diamondsuit\rangle^{-2})R \overset{\epsilon\to 0}{\to} 0$ for $R\in\Sigma_{n'}$ for any $n'\in\mathbb{Z}$ because in fact $C(\epsilon)\overset{\epsilon\to 0}{\to} 0$ by (5), with $C(\epsilon)$ defined like above for any pair (r',r) with $r' > r+\mathbf{d}$.
Still by (5), for $n > n'+\mathbf{d}$ and for $R\in\Sigma_{n'}$ we have by (A5)

$$\begin{aligned}\|J\diamondsuit(1-\langle\epsilon\diamondsuit\rangle^{-2})R\|_{\Sigma_n} &\le \|\diamondsuit(1-\langle\epsilon\diamondsuit\rangle^{-2})\|_{B(\Sigma_{n'},\Sigma_n)}\|R\|_{\Sigma_{n'}}\\ &\le C\|(1-\langle\epsilon\diamondsuit\rangle^{-2})\|_{B(\Sigma_{n'},\Sigma_{n+\mathbf{d}})}\|R\|_{\Sigma_{n'}} \overset{\epsilon\to 0}{\to} 0\,.\end{aligned} \tag{84}$$

These facts yield (81).
We turn now to Claim (3) and to (82). By the Rellich criterion, the embedding $\Sigma_a\hookrightarrow\Sigma_b$ for $a>b$ is compact. Hence also $\mathcal{P}^a\hookrightarrow\mathcal{P}^b$ is compact. Then (82) follows by the Ascoli–Arzela Theorem by a standard argument. □

Corollary 3.12. *Consider* (55) *defined by the field* $\mathcal{X}^t$ *and consider indexes and notation of Lemma 3.8 (in particular we have* $M_0=1$ *and* $i=1$ *in* (56) *and elsewhere;* r *and* M *can be arbitrary). Consider* s'*,*s *and* k *as in 3.8. Then for the map* $\mathfrak{F}^t\in C^l(\mathcal{U}^{s'}_{\varepsilon_1,k},\widetilde{\mathcal{P}}^s)$ *derived from* (62)*, we have* $\mathfrak{F}^{1*}\Omega=\Omega_0$.

Proof. Ω_0 is constant in the coordinate system (τ,Π,R) where $R\in N_g^\perp(\mathcal{H}^*_{p_0})$, with $\Omega_0 = d\tau_j\wedge d\Pi_j + \langle J^{-1}\ ,\ \rangle$, where we apply $\langle J^{-1}\ ,\ \rangle$ only to vectors in the R space. Hence Ω_0 is C^∞ in $R\in L^2$, τ and Π, with values in $B^2(L^2,\mathbb{R})$. From Lemma 3.3 we have that $d\alpha$, so also Ω by $\Omega=\Omega_0+d\alpha$, belongs to $C^\infty(\mathcal{U}^s_{\varepsilon_0,k},B^2(\widetilde{\mathcal{P}},\mathbb{R}))$ for an $\epsilon_0>0$, and so also to $C^\infty(\mathcal{U}^s_{\varepsilon_0,k},B^2(\widetilde{\mathcal{P}}^s,\mathbb{R}))$. Let now $r-(l+1)\mathbf{d}\ge s'\ge s+l\mathbf{d}$ and $k\in\mathbb{Z}\cap[0,r-(l+1)\mathbf{d}]$. Then for a fixed $0<\varepsilon_2\ll\varepsilon_1$ and for all $|\epsilon|\le 1$ we have

$$\mathfrak{F}^t_\epsilon\in C^l((-2,2)\times\mathcal{U}^{s'}_{\varepsilon_2,k},\mathcal{U}^s_{\varepsilon_1,k}),\quad \mathfrak{F}^t_\epsilon(\mathcal{U}^{s'}_{\varepsilon_2,k})\subset\mathcal{U}^{s'}_{\varepsilon_1,k}\text{ for all }|t|\le 2 \tag{85}$$

by Lemma 3.8, for a fixed $l\ge 2$. By Lemma 3.11 we have uniformly locally

$$\lim_{\varepsilon\to 0}\mathfrak{F}^t_\epsilon=\mathfrak{F}^t\text{ in }C^l([-1,1]\times\mathcal{U}^{s'}_{\varepsilon_2,k},\mathcal{U}^s_{\varepsilon_1,k}). \tag{86}$$

Let us take $0<\varepsilon_3\ll\varepsilon_2$ s.t. $\mathfrak{F}^t_\epsilon(\mathcal{U}^{s'}_{\varepsilon_3,k})\subset\mathcal{U}^{s'}_{\varepsilon_2,k}$ for all $|t|\le 2$ and $|\epsilon|\le 1$.
In $\mathcal{U}^{s'}_{\varepsilon_3,k}$ the following computation is valid because $\mathcal{X}^t_\epsilon$ is a standard vector field in $\mathcal{U}^{s'}_{\varepsilon_1,k}$ and similarly Ω_t is a regular differential form therein:

$$\begin{aligned}\mathfrak{F}^{1*}_\epsilon\Omega-\Omega_0 &= \int_0^1\frac{d}{dt}\left(\mathfrak{F}^{t*}_\epsilon\Omega_t\right)dt = \int_0^1\mathfrak{F}^{t*}_\epsilon\left(L_{\mathcal{X}^t_\epsilon}\Omega_t+\frac{d}{dt}\Omega_t\right)dt\\ &= d\int_0^1\mathfrak{F}^{t*}_\epsilon\left(i_{\mathcal{X}^t_\epsilon}\Omega_t+\alpha\right)dt\,,\end{aligned}$$

where we recall $\Omega_t = \Omega_0 + t(\Omega - \Omega_0)$.
If we consider a ball $\mathbf{B}$ in $\mathcal{U}^{s'}_{\varepsilon_3,k}$, in the notation of Lemma 3.1, for some function $\psi_\epsilon \in C^1(\mathbf{B}, \mathbb{R})$ we can write

$$\mathfrak{F}^{1*}_\epsilon(B_0 + \alpha) - B_0 + d\psi_\epsilon = \int_0^1 \mathfrak{F}^{t*}_\epsilon \left(i_{\mathcal{X}^t_\epsilon}\Omega_t + \alpha\right) dt, \tag{87}$$

By (85)–(86) we have

$$\lim_{\epsilon\to 0}(\mathfrak{F}^{1*}_\epsilon(B_0 + \alpha) - B_0) = \mathfrak{F}^{1*}(B_0 + \alpha) - B_0 \text{ in } C^{l-1}(\mathcal{U}^{s'}_{\varepsilon_3,k}, B(\widetilde{\mathcal{P}}^{s'}, \mathbb{R})).$$

The set $\Gamma := \{\mathfrak{F}^t_\epsilon(\mathbf{B}) : |t| \le 2, |\epsilon| \le 1\}$ is a bounded subset in $\mathcal{U}^{s'}_{\varepsilon_2,k}$ because of (71)–(72). Then by (81) we have

$$\lim_{\epsilon\to 0} \mathcal{X}^t_\epsilon = \mathcal{X}^t \text{ in } C^0((-2,2)\times\Gamma, \widetilde{\mathcal{P}}^s) \text{ uniformly .}$$

Hence by $i_{\mathcal{X}^t}\Omega_t = -\alpha$ we get

$$\lim_{\epsilon\to 0}\left(i_{\mathcal{X}^t_\epsilon}\Omega_t + \alpha\right) = i_{\mathcal{X}^t}\Omega_t + \alpha = 0 \text{ in } C^0((-2,2)\times\Gamma, B(\widetilde{\mathcal{P}}^s, \mathbb{R})) \text{ uniformly.}$$

This implies

$$\begin{aligned}\lim_{\epsilon\to 0} \| \int_0^1 \mathfrak{F}^{t*}_\epsilon \left(i_{\mathcal{X}^t_\epsilon}\Omega_t + \alpha\right) dt\|_{L^\infty(\mathbf{B}, B(\widetilde{\mathcal{P}}^{s'}, \mathbb{R}))} \\ \le C \lim_{\epsilon\to 0} \|i_{\mathcal{X}^t_\epsilon}\Omega_t + \alpha\|_{L^\infty([0,1]\times\Gamma, B(\widetilde{\mathcal{P}}^s, \mathbb{R}))} = 0,\end{aligned}$$

for C an upper bound to the norms $\|(\mathfrak{F}^{t*}_\epsilon)_{|\mathfrak{F}^t_\epsilon(\upsilon)} : B(\widetilde{\mathcal{P}}^{s'}, \mathbb{R}) \to B(\widetilde{\mathcal{P}}^s, \mathbb{R})\|$ as υ varies in $\mathbf{B}$. Notice that $C < \infty$ by (82).
By (87) we conclude that uniformly

$$\lim_{\epsilon\to 0} d\psi_\epsilon = B_0 - \mathfrak{F}^{1*}(B_0 + \alpha) \text{ in } C^0(\mathbf{B}, B(\widetilde{\mathcal{P}}^{s'}, \mathbb{R})).$$

Normalizing $\psi_\epsilon(\upsilon_0) = 0$ at some given $\upsilon_0 \in \mathbf{B}$, it follows that also ψ_ϵ converges locally uniformly to a function ψ_0 with $d\psi_0 = B_0 - \mathfrak{F}^{1*}(B_0 + \alpha)$. Taking the exterior differential, we conclude that $\mathfrak{F}^{1*}\Omega = \Omega_0$ in $C^\infty(\mathcal{U}^{s'}_{\varepsilon_3,k}, B^2(\widetilde{\mathcal{P}}^{s'}, \mathbb{R}))$. □

4. Pullback of the Hamiltonian

In the somewhat abstract set up of this paper it is particularly important to have a general description of the pullbacks of the Hamiltonian K. Our main goal in this section is formula (101). This formula and its related expansion in Lemma 5.4 obtained splitting R in discrete and continuous modes, play a key role in the Birkhoff normal forms argument.

The first and quite general result is the following consequence of Lemma 3.8.

LEMMA 4.1. *Consider* $\mathfrak{F} = \mathfrak{F}_1 \circ \cdots \circ \mathfrak{F}_L$ *with* $\mathfrak{F}_j = \mathfrak{F}_j^{t=1}$ *transformations as of Lemma 3.8. Suppose that for* j *we have* $M_0 = m_j$, *with given numbers* $1 \leq m_1 \leq ... \leq m_L$. *Suppose also that all the* j *we have the same pair* r *and* M, *which we assume sufficiently large. Let* $i_j = 1$ *if* $m_j = 1$. *Fix* $0 < m' < M$

(1) Let $r > 2L(m'+1)\mathbf{d} + s'_L > 4L(m'+1)\mathbf{d} + s_1$, $s_1 \geq \boldsymbol{d}$. *Then, for any* $\varepsilon > 0$ *there exists a* $\delta > 0$ *such that* $\mathfrak{F} \in C^{m'}(\mathcal{U}^{s'_L}_{\delta,a}, \mathcal{U}^{s_1}_{\varepsilon,h})$ *for* $0 \leq a \leq h$ *and* $0 \leq h < r - (m'+1)\mathbf{d}$.

(2) Let $r > 2L(m'+1))\mathbf{d} + h > 4L(m'+1)\mathbf{d} + a$, $a \geq 0$. *The above composition, interpreting the* $\mathfrak{F}_j$*'s as maps in the* (ϱ, R) *variables as in Lemma 3.9, yields also* $\mathfrak{F} \in C^{m'}(\mathcal{U}_{-a}, \mathcal{P}^{-h})$ *for* $\mathcal{U}_{-a}$ *a sufficiently small neighborhood of the origin in* $\mathcal{P}^{-a}$.

(3) For $\mathcal{U}_{-a} \subset \mathcal{P}^{-a}$ *like above and for functions* $\mathcal{R}^{i,j}_{a,m'} \in C^{m'}(\mathcal{U}_{-a}, \mathbb{R})$ *and* $\mathbf{S}^{i,j}_{a,m'} \in C^{m'}(\mathcal{U}_{-a}, \Sigma_a)$, *the following formulas hold:*

$$\begin{aligned} &\Pi(R') := \Pi(R) \circ \mathfrak{F} = \Pi(R) + \mathcal{R}^{i_1, m_1+1}_{a,m'}(\Pi(R), R), \\ &p' := p \circ \mathfrak{F} = p + \mathcal{R}^{i_1, m_1+1}_{a,m'}(\Pi(R), R), \\ &\Phi_{p'} = \Phi_p + \mathbf{S}^{i_1, m_1+1}_{a,m'}(\Pi(R), R). \end{aligned} \tag{88}$$

(4) For a function F *such that* $F(e^{J\tau \cdot \diamond} U) \equiv F(U)$ *we have*

$$F \circ \mathfrak{F}(U) = F\left(\Phi_p + P(p)(R + \mathbf{S}^{i_1,m_1}_{k',m'}) + \boldsymbol{S}^{i_1,m_1+1}_{k',m'}\right),\ k' = r - 7L(m'+1)\mathbf{d}.$$

Proof. Recall that by (62) we have $\mathfrak{F}_j \in C^{m'}(\mathcal{U}^{s'_j}_{\varepsilon'_j,h}, \mathcal{U}^{s_j}_{\varepsilon_j,h})$ for $r - (m'+1)\mathbf{d} > s'_j \geq s_j + m'\mathbf{d}$ and appropriate choice of the $0 < \varepsilon'_j < \varepsilon_j$ and for $h \in \mathbb{Z} \cap [0, r - (m'+1)\mathbf{d}]$. So for the composition we have $\mathfrak{F} \in C^{m'}(\mathcal{U}^{\kappa}_{\varepsilon'_L,a}, \mathcal{U}^{s_1}_{\varepsilon_1,h})$ for $a \leq h$. The inequalities $r > 2L(m'+1)\mathbf{d} + s'_L > 4L(m'+1)\mathbf{d} + s_1$, $s_1 \geq \mathbf{d}$ can be accommodated since r is assumed sufficiently large. This yields claim (1).
By Lemma 3.9 we have $\mathfrak{F}_j \in C^{m'}(\mathcal{U}_{-h+jm'\mathbf{d}}, \mathcal{P}^{-h+(j-1)m'\mathbf{d}})$ with $\mathcal{U}_{-h+jm'\mathbf{d}} \subset \mathcal{P}^{-h+jm'\mathbf{d}}$ a neighborhood of the origin. So for the composition we have $\mathfrak{F} \in C^{m'}(\mathcal{U}_{-a}, \mathcal{P}^{-h})$ for $a \leq h - Lm'\mathbf{d}$. The conditions $r > 2L(m'+1)\mathbf{d} + h$, $h > 4L(m'+1)\mathbf{d} + a$ and $a \geq 0$, can be accommodated since r is assumed sufficiently large. This yields claim (2).
We now prove (88). Let first $L = 1$. By (58) we have $R' := (\mathfrak{F}_1)_R(\Pi(R), R) = e^{Jq_1 \cdot \diamond}(R + \mathbf{S}^{i_1,m_1}_{r-(m'+1)\mathbf{d},m'})$, where we use $M > m'$. Here we will omit the variables $(\Pi(R), R)$ in the $\mathbf{S}$'s and $\mathcal{R}$'s. Then we have for $a' = r - (m'+1)\mathbf{d}$

$$\Pi(R') = \Pi(R + \mathbf{S}^{i_1,m_1}_{a',m'}) = \Pi(R) + \mathcal{R}^{i_1,m_1+1}_{a'-\mathbf{d},m'}. \tag{89}$$

Here we have used

$$|\langle R, \diamond \mathbf{S}^{i_1,m_1}_{a',m'}\rangle| \leq \|R\|_{\Sigma_{-a'+\mathbf{d}}} \|\mathbf{S}^{i_1,m_1}_{a',m'}\|_{\Sigma_{a'}}.$$

By $p_j = \Pi_j - \Pi_j(R) + \mathcal{R}^{0,2}(\Pi(R), R)$ we get

$$\begin{aligned} p'_j &= \Pi_j - \Pi_j(R') + \mathcal{R}^{0,2}(\Pi(R'), R') \\ &= \Pi_j - \Pi_j(R) + \mathcal{R}^{0,2}(\Pi(R), R) + \mathcal{R}^{i_1,m_1+1}_{a'-\mathbf{d},m'} = p_j + \mathcal{R}^{i_1,m_1+1}_{a'-\mathbf{d},m'}. \end{aligned} \tag{90}$$

This yields (88) for $L = 1$ since $a \leq r - 4(m'+1)\mathbf{d} < a' - \mathbf{d}$. We extend the proof to the case $L > 1$. We write here and below $\mathfrak{F}' := \mathfrak{F}_1 \circ \cdots \circ \mathfrak{F}_{L-1}$. We suppose that $\mathfrak{F}'_R(\Pi(R), R) = e^{Jq\cdot\diamond}(R + \mathbf{S}^{i_1,m_1}_{a'_{L-1},m'})$ for $a'_{L-1} \leq r - 2(L-1)m'\mathbf{d}$, which is true for $L - 1 = 1$. Then

$$\begin{aligned} R' &= e^{J(q\circ\mathfrak{F}_L)\cdot\diamond}\left(e^{Jq_L\cdot\diamond}(R + \mathbf{S}^{i_L,m_L}_{r-(m'+1)\mathbf{d},m'}) + \mathbf{S}^{i_1,m_1}_{a'_{L-1},m'} \circ \mathfrak{F}_L\right) \\ &= e^{J(q\circ\mathfrak{F}_L+q_L)\cdot\diamond}\left(R + \mathbf{S}^{i_L,m_L}_{r-(m'+1)\mathbf{d},m'}) + e^{-Jq_L\cdot\diamond}\mathbf{S}^{i_1,m_1}_{a'_{L-1}-m'\mathbf{d},m'}\right), \end{aligned}$$

where $q_L = \mathcal{R}^{0,m_L+1}_{r-(m'+1)\mathbf{d},m'}$ and where we used the last claim in Lemma 3.9. Since $e^{-Jq_L\cdot\diamond}\mathbf{S}^{i_1,m_1}_{a'_{L-1}-m'\mathbf{d},m'} = \mathbf{S}^{i_1,m_1}_{a'_{L-1}-2m'\mathbf{d},m'}$ we conclude that there is an expansion $R' = e^{Jq\cdot\diamond}(R + \mathbf{S}^{i_1,m_1}_{a'_L,m'})$ for $a'_L \leq a'_{L-1} - 2m'\mathbf{d}$. Then

$$\mathfrak{F}_R(\Pi(R), R) = e^{Jq\cdot\diamond}(R + \mathbf{S}^{i_1,m_1}_{a'_L,m'}), \quad a'_L := r - 2Lm'\mathbf{d}. \tag{91}$$

For $a' = a'_L$ formulas (89)–(90) continue to hold. By $a < a'_L - \mathbf{d}$ this yields (88). We consider the last statement of Lemma 58. For $a' = r - (m'+1)\mathbf{d}$ we have

$$\begin{aligned} F(\mathfrak{F}_1(U)) &= F(\Phi_{p'} + P(p')e^{Jq_1\cdot\diamond}(R + \mathbf{S}^{i_1,m_1}_{a',m'})) \\ &= F(\Phi_p + P(p)e^{Jq_1\cdot\diamond}(R + \mathbf{S}^{i_1,m_1}_{a',m'}) + \mathbf{S}^{i_1,m_1+1}_{a'+\mathbf{d},m'}) \\ &= F\left(e^{Jq_1\cdot\diamond}\left(\Phi_p + P(p)(R + \mathbf{S}^{i_1,m_1}_{a',m'}) + Y\right)\right) \end{aligned}$$

with

$$Y = (e^{Jq_1\cdot\diamond} - 1)\Phi_p + [P(p), e^{Jq_1\cdot\diamond}](R + \mathbf{S}^{i_1,m_1}_{a',m'}) + e^{-Jq_1\cdot\diamond}\mathbf{S}^{i_1,m_1+1}_{a'-\mathbf{d},m'}.$$

We claim

$$Y = \mathbf{S}^{i_1,m_1+1}_{a'-2m'\mathbf{d},m'}. \tag{92}$$

To prove (92) we use $(e^{Jq_1\cdot\diamond} - 1)\Phi_p = \mathbf{S}^{i_1,m_1+1}_{r-(m'+1)\mathbf{d},m'} = \mathbf{S}^{i_1,m_1+1}_{a',m'}$. This follows from $\Phi_p \in C^\infty(\mathcal{O}, \mathcal{S})$ and

$$\left|(e^{Jq_1\cdot\diamond} - 1)\Phi_p\right|_{\Sigma_l} \leq |q_{1j}| \int_0^1 \left|e^{tJq_1\cdot\diamond}\diamond_j\Phi_p\right|_{\Sigma_l} dt \leq C_l|q_{1j}|\,|\diamond_j\Phi_p|_{\Sigma_l}. \tag{93}$$

Schematically we have, summing over repeated indexes and for $\mathbf{e}_j, \mathbf{e}_j^* \in \mathcal{S}$,

$$\begin{aligned}[P(p), e^{Jq_1\cdot\diamond}] &= [e^{Jq_1\cdot\diamond}, P_{N_g}(p)] = e^{Jq_1\cdot\diamond}\mathbf{e}_j\langle\mathbf{e}_j^*, \rangle - \mathbf{e}_j\langle e^{-Jq_1\cdot\diamond}\mathbf{e}_j^*, \rangle \\ &= (e^{Jq_1\cdot\diamond} - 1)\mathbf{e}_j\langle\mathbf{e}_j^*, \rangle - \mathbf{e}_j\langle(e^{-Jq_1\cdot\diamond} - 1)\mathbf{e}_j^*, \rangle \\ &= \mathbf{S}^{0,m_1+1}_{r-(m'+1)\mathbf{d},m'}\langle\mathbf{e}_j^*, \rangle + \mathbf{e}_j\langle\mathbf{S}^{0,m_1+1}_{r-(m'+1)\mathbf{d},m'}, \rangle.\end{aligned}$$

This yields for any $a'' \le a' = r - (m'+1)\mathbf{d}$

$$[P(p), e^{Jq_1\cdot\diamond}](R + \mathbf{S}^{i_1,m_1}_{a'',m'}) = \mathbf{S}^{i_1,m_1+2}_{a'',m'}.$$

We have $e^{-Jq_1\cdot\diamond}\mathbf{S}^{i_1,m_1+1}_{a'-\mathbf{d},m'} = \mathbf{S}^{i_1,m_1+1}_{a'-(m'+1)\mathbf{d},m'}$. Then (92) is proved. Then

$$F(\mathfrak{F}_1(U)) = F\left(\Phi_p + P(p)(R + \mathbf{S}^{i_1,m_1}_{a'-2m'\mathbf{d},m'}) + \mathbf{S}^{i_1,m_1+1}_{a'-2m'\mathbf{d},m'}\right) \tag{94}$$

for $a' = r - (m'+1)\mathbf{d}$. This proves the last sentence of our lemma for $L = 1$. For $L > 1$ set once more $\mathfrak{F}' := \mathfrak{F}_1 \circ \cdots \circ \mathfrak{F}_{L-1}$. We assume by induction that $F(\mathfrak{F}'(U))$ equals the rhs of (94) for $a' = a'_{L-1} := r - 2(L-1)m'\mathbf{d}$. Then using $\mathbf{S}^{i_1,m_1}_{l,m'} \circ \mathfrak{F}_L = \mathbf{S}^{i_1,m_1}_{l-m'\mathbf{d},m'}$ from Lemma 3.9, by (88) for $\mathfrak{F} = \mathfrak{F}_L$ and by (92) with the index 1 replaced by index L, we get

$$\begin{aligned}F(\mathfrak{F}(U)) &= F\Big(\Phi_{p'} + P(p')e^{Jq_L\cdot\diamond}(R + \mathbf{S}^{i_L,m_L}_{r-(m'+1)\mathbf{d},m'}) \\ &\qquad + P(p')\mathbf{S}^{i_1,m_1}_{a'_{L-1}-m'\mathbf{d},m'} + \mathbf{S}^{i_1,m_1+1}_{a'_{L-1}-m'\mathbf{d},m'}\Big) \\ &= F\left(e^{Jq_L\cdot\diamond}\left[\Phi_p + P(p)(R + \mathbf{S}^{i_1,m_1}_{a'_{L-1}-m'\mathbf{d},m'}) + \mathbf{S}^{i_1,m_1+1}_{a'_{L-1}-2m'\mathbf{d},m'}\right]\right).\end{aligned}$$

We conclude that $F(\mathfrak{F}(U))$ equals the rhs of (94) for $a'_L = r - 2Lm'\mathbf{d}$. In particular this proves the last sentence of our lemma for any L. □

LEMMA 4.2. *For fixed vectors* $\mathbf{u}$ *and* $\mathbf{v}$ *and for* B *sufficiently regular with* $B(0) = 0$, *we have*

$$\begin{aligned}B(|\mathbf{u}+\mathbf{v}|_1^2) &= B\left(|\mathbf{u}|_1^2\right) + B(|\mathbf{v}|_1^2) \\ &\quad + \sum_{j=0}^{3}\int_{[0,1]^2}\frac{t^j}{j!}(\partial_t^{j+1})_{|t=0}\partial_s[B(|s\mathbf{u}+t\mathbf{v}|_1^2)]\; dtds \\ &\quad + \int_{[0,1]^2} dtds\int_0^t \partial_\tau^5\partial_s[B(|s\mathbf{u}+\tau\mathbf{v}|_1^2)]\frac{(t-\tau)^3}{3!}\; d\tau.\end{aligned} \tag{95}$$

Proof. Follows by Taylor expansion in t of

$$B(|\mathbf{u}+\mathbf{v}|_1^2) = B\left(|\mathbf{u}|_1^2\right) + \int_0^1 \partial_t[B(|\mathbf{u}+t\mathbf{v}|_1^2)]dt$$
$$= B\left(|\mathbf{u}|_1^2\right) + B(|\mathbf{v}|_1^2) + \int_{[0,1]^2} dtds\ \partial_s\partial_t[B(|s\mathbf{u}+t\mathbf{v}|_1^2)].$$

□

LEMMA 4.3. *Consider a transformation $\mathfrak{F} = \mathfrak{F}_1 \circ \cdots \circ \mathfrak{F}_L$ like in Lemma 4.1 and with $m_1 = 1$, with same notations, hypotheses and conclusions. In particular we suppose r and M sufficiently large that the conclusions of Lemma 4.1 hold for preassigned sufficiently large $s = s'_L$, k' and m'. Let $k \le k' - \max\{\boldsymbol{d}, ord(\mathcal{D})\}$ and $m \le m'$. Then there are a $\underline{\psi}(\varrho) \in C^\infty$ with $\underline{\psi}(\varrho) = O(|\varrho|^2)$ near 0 and a small $\varepsilon > 0$ such that in $\mathcal{U}^s_{\varepsilon,k}$ we have the expansion*

$$K \circ \mathfrak{F} = \underline{\psi}(\Pi(R)) + \frac{1}{2}\Omega(\mathcal{H}_p P(p)R, P(p)R) + \mathcal{R}^{1,2}_{k,m} + E_P(P(p)R) + \boldsymbol{R}'' \quad (96)$$
$$\boldsymbol{R}'' := \sum_{d=2}^{4}\langle B_d(R,\Pi(R)), (P(p)R)^d\rangle + \int_{\mathbb{R}^3} B_5(x,R,R(x),\Pi(R))(P(p)R)^5(x)dx$$

with:

- $\mathcal{R}^{1,2}_{k,m} = \mathcal{R}^{1,2}_{k,m}(\Pi(R), R)$;
- $B_2(0,0) = 0$;
- $(P(p)R)^d(x)$ *represent $d-$products of components of $P(p)R$;*
- $B_d(\cdot, R, \varrho) \in C^m(\mathcal{U}_{-k}, \Sigma_k(\mathbb{R}^3, B((\mathbb{R}^{2N})^{\otimes d}, \mathbb{R})))$ *for $2 \le d \le 4$ with $\mathcal{U}_{-k} \subset \mathcal{P}^{-k}$ a neighborhood of the origin;*
- *for $\zeta \in \mathbb{R}^{2N}$ with $|\zeta| \le \varepsilon$ and $(\varrho, R) \in \mathcal{U}_{-k}$ we have for $i \le m$*

$$\|\nabla^i_{R,\zeta,\varrho} B_5(R,\zeta,\varrho)\|_{\Sigma_k(\mathbb{R}^3, B((\mathbb{R}^{2N})^{\otimes 5},\mathbb{R})} \le C_i. \quad (97)$$

Proof. Here we will omit the variables $(\Pi(R), R)$ in the $\mathbf{S}$'s and $\mathcal{R}$'s.
By Lemma 4.1 for $m \le m' \le M$, $k + \max\{\mathbf{d}, \mathrm{ord}(\mathcal{D})\} \le k' \le r - L(m'+2)\mathbf{d}$, we have

$$\begin{aligned} K(\mathfrak{F}(U)) \quad = \quad & E(\Phi_p + P(p)R + P(p)\mathbf{S}^{1,1}_{k',m'} + \mathbf{S}^{1,2}_{k',m'}) - E\left(\Phi_{p_0}\right) \\ & -(\lambda_j(p) + \mathcal{R}^{1,2}_{k,m})\left(\Pi_j(\Phi_p + P(p)R) + \mathcal{R}^{1,2}_{k,m} - \Pi_j\left(\Phi_{p_0}\right)\right), \end{aligned} \quad (98)$$

where, by (88), we have used $p' := p \circ \mathfrak{F} = p + \mathcal{R}^{1,2}_{k,m}$ and where by $k \leq k' - \mathbf{d}$

$$\Pi_j(\Phi_p + P(p)R + P(p)\mathbf{S}^{1,1}_{k',m'} + \mathbf{S}^{1,2}_{k',m'}) = \Pi_j(\Phi_p + P(p)R) + \mathcal{R}^{1,2}_{k,m}.$$

Set now $\Psi = \Phi_p + P(p)\mathbf{S}^{1,1}_{k',m'} + \mathbf{S}^{1,2}_{k',m'}$. By (95) for $\mathbf{u} = \Psi$ and $\mathbf{v} = P(p)R$

$$\begin{aligned} E_P(\Psi + P(p)R) &= E_P(\Psi) + E_P(P(p)R) \\ &+ \sum_{j=0}^{1} \int_{\mathbb{R}^3} dx \int_{[0,1]^2} \frac{t^j}{j!} (\partial_t^{j+1})_{|t=0} \partial_s [B(|s\Psi + tP(p)R|_1^2)] dtds \\ &+ \sum_{j=2}^{3} \int_{\mathbb{R}^3} dx \int_{[0,1]^2} \frac{t^j}{j!} (\partial_t^{j+1})_{|t=0} \partial_s [B(|s\Psi + tP(p)R|_1^2)] dtds \\ &+ \int_{\mathbb{R}^3} dx \int_{[0,1]^2} dtds \int_0^t \partial_\tau^5 \partial_s [B(|s\Psi + \tau P(p)R|_1^2)] \frac{(t-\tau)^3}{3!} d\tau. \end{aligned} \tag{99}$$

The last two lines can be incorporated in $\mathbf{R}''$. For example, schematically we have

$$\partial_\tau^5 \partial_s B(|s\Phi_p + \tau P(p)R|_1^2) \sim \widetilde{B}(s\Phi_p + \tau P(p)R)\ \Phi_p\ (P(p)R)^5,$$

for some $\widetilde{B}(Y) \in C^\infty(\mathbb{R}^{2N}, B^6(\mathbb{R}^{2N}, \mathbb{R}))$. This produces a term which can be absorbed in the B_5 term of $\mathbf{R}''$. In particular, (97) follows from (2). The terms in the third line of (99) can be treated similarly yielding terms which end in the B_d term of $\mathbf{R}''$ with $d = j + 1$.
The second line of (99) equals

$$\begin{aligned} &\int_{\mathbb{R}^3} dx \int_{[0,1]^2} dtds \sum_{j=0}^{1} \frac{t^j}{j!} (\partial_t^{j+1})_{|t=0} \partial_s \ \Big\{ B(|s\Phi_p + tP(p)R|_1^2) + \\ &+ \int_0^1 d\tau \partial_\tau [B(|s(\Phi_p + \tau(P(p)\mathbf{S}^{1,1}_{k',m'} + \mathbf{S}^{1,2}_{k',m'}) + tP(p)R|_1^2)] \ \Big\}. \end{aligned} \tag{100}$$

The contribution from the last line of (100) can be incorporated in $\mathbf{R}'' + \mathcal{R}^{1,2}_{k,m}$. By $k \leq k' - \mathrm{ord}(\mathcal{D})$ we have

$$\begin{aligned} E_K(\Psi + P(p)R) &= E_K(\Psi) + \langle \mathcal{D}\Phi_p, P(p)R \rangle \\ &+ \overbrace{\langle \mathcal{D}(P(p)\mathbf{S}^{1,1}_{k',m'} + \mathbf{S}^{1,2}_{k',m'}), P(p)R \rangle}^{\mathcal{R}^{1,2}_{k,m}} + E_K(P(p)R). \end{aligned}$$

Notice that from the $j = 0$ term in the first line of (100) we get

$$\begin{aligned} 2 \int_{\mathbb{R}^3} dx \int_0^1 ds \partial_s [B'(|s\Phi_p|_1^2) s\Phi_p \cdot_1 P(p)R] &= 2 \int_{\mathbb{R}^3} dx B'(|\Phi_p|_1^2) \Phi_p \cdot_1 P(p)R \\ &= \langle \nabla E_P(\Phi_p), P(p)R \rangle. \end{aligned}$$

By (6) and (16), that is $\nabla E(\Phi_p) = \lambda(p) \cdot \diamond\Phi_p \in N_g(\mathcal{H}_p^*)$, and by $P(p)R \in N_g^\perp(\mathcal{H}_p)$, we have

$$\langle \mathcal{D}\Phi_p, P(p)R\rangle + \langle \nabla E_P(\Phi_p), P(p)R\rangle = \langle \nabla E(\Phi_p), P(p)R\rangle = 0.$$

The $j = 1$ term in the first line of (100) is $\frac{1}{2}\langle \nabla^2 E_P(\Phi_p)P(p)R, P(p)R\rangle$ which summed to the $E_K(P(p)R)$ in (4) yields the $\frac{1}{2}\Omega(\mathcal{H}_p P(p)R, P(p)R)$ in (96). We have $E_K(\Psi) + E_P(\Psi) = E(\Psi)$ and

$$E(\Psi) = E(\Phi_p) + \overbrace{\langle \nabla E(\Phi_p), P(p)\mathbf{S}^{1,1}_{k',m'}\rangle}^{0} + \overbrace{\langle \nabla E(\Phi_p), \mathbf{S}^{1,2}_{k',m'}\rangle}^{\mathcal{R}^{1,2}_{k,m}} + \mathcal{R}^{1,2}_{k,m}.$$

The last term we need to analyze, for for $d(p) := E(\Phi_p) - \lambda(p) \cdot \Pi(\Phi_p)$, is

$$\begin{aligned} &E(\Phi_p) - E(\Phi_{p_0}) - \sum_j \lambda_j(p)(\Pi_j(\Phi_p) - \Pi_j(\Phi_{p_0})) \\ &\quad = d(p) - d(p_0) - \sum_j (\lambda_j(p_0) - \lambda_j(p))p_{0j} =: \widetilde{\psi}(p, p_0), \end{aligned}$$

where $\widetilde{\psi}(p, p_0) = O((p - p_0)^2)$ by $\partial_{p_j} d(p) = -p \cdot \partial_{p_j}\lambda(p)$. Notice that $\widetilde{\psi} \in C^\infty(\mathcal{O}^2, \mathbb{R})$. Now recall that in the initial system of coordinates we have $p' = \Pi - \Pi(R') + \mathcal{R}^{0,2}(\Pi(R'), R')$. Substituting p' and $\Pi(R')$ by means of (88), and R' by means of (91) we conclude that $p = p_0 - \Pi(R) + \mathcal{R}^{0,2}_{k',m'}$. Then $\widetilde{\psi}(p, p_0) = \underline{\psi}(\Pi(R)) + \mathcal{R}^{1,2}_{k,m}$ with $\underline{\psi}(\varrho) := \widetilde{\psi}(p_0 - \varrho, p_0)$ a C^∞ function with $\underline{\psi}(\varrho) = O(|\varrho|^2)$ for ϱ near 0. □

LEMMA 4.4. *Under the hypotheses and notation of Lemma 4.3, for an $\boldsymbol{R}'$ like $\boldsymbol{R}''$, for a $\psi \in C^\infty$ with $\psi(\varrho) = O(|\varrho|^2)$ near 0, we have*

$$K \circ \mathfrak{F} = \psi(\Pi(R)) + \frac{1}{2}\Omega(\mathcal{H}_{p_0}R, R) + \mathcal{R}^{1,2}_{k,m}(\Pi(R), R) + E_P(R) + \boldsymbol{R}', \quad (101)$$

$$\boldsymbol{R}' := \sum_{d=2}^{4} \langle B_d(R, \Pi(R)), R^d\rangle + \int_{\mathbb{R}^3} B_5(x, R, R(x), \Pi(R))R^5(x)dx,$$

the B_d for $d = 2, ..., 5$ with similar properties of the functions in Lemma 4.3.

Proof. We have

$$P(p)R = R + (P(p) - P(p_0))R = R + \mathbf{S}^{1,1}(p - p_0, R) = R + \mathbf{S}^{1,1}(\Pi(R), R).$$

Substituting $P(p)R = R + \mathbf{S}^{1,1}(\Pi(R), R)$ in (96) we obtain that $\mathcal{R}^{1,2}_{k,m} + \mathbf{R}''$ is absorbed in $\mathcal{R}^{1,2}_{k,m}(\Pi(R), R) + \mathbf{R}'$. This is elementary to see for the terms with $d \leq 4$. We consider the case $d = 5$.

$$
\begin{aligned}
&B_5(x,R,R(x),\Pi(R))R^i(x)(\mathbf{S}^{1,1})^{5-i}\\
&\quad=\sum_{j=0}^{5-i}\frac{1}{j!}(\partial_t^j)_{|t=0}[B_5(x,R,tR(x),\Pi(R))]R^i(x)(\mathbf{S}^{1,1})^{5-i}\\
&\qquad+\int_0^1\frac{(1-t)^{4-i}}{(4-i)!}\partial_t^{5-i}[B_5(x,R,tR(x),\Pi(R))]R^i(x)(\mathbf{S}^{1,1})^{5-i}
\end{aligned}
$$

The last term can be absorbed in the $d=5$ term of $\mathbf{R}'$. Similarly, all the other terms either are absorbed in $\mathbf{R}'$ or, like for instance the $i=j=0$ term, they are $\mathcal{R}^{1,2}$.
We write $E_P(P(p)R)=E_P(R-P_{N_g(p)}R)$ and use (95) for $\mathbf{u}=R$ and $\mathbf{v}=-P_{N_g(p)}R$. We get the sum of $E_P(R)$ with a term which can be absorbed in $\mathcal{R}^{1,2}_{k,m}(\Pi(R),R)+\mathbf{R}'$. We finally focus on

$$
\begin{aligned}
\frac{1}{2}\langle J^{-1}\mathcal{H}_pP(p)R,P(p)R\rangle &= \frac{1}{2}\langle\mathcal{D}P(p)R,P(p)R\rangle-\lambda_j(p)\Pi_j(P(p)R)\\
&\quad+\frac{1}{2}\langle\nabla^2E_P(\Phi_p)P(p)R,P(p)R\rangle.
\end{aligned}
\tag{102}
$$

We have

$$
\begin{aligned}
\langle\mathcal{D}P(p)R,P(p)R\rangle&=\langle\mathcal{D}R,R\rangle+\mathcal{R}^{1,2}_{k,m}(\Pi(R),R)\\
\langle\nabla^2E_P(\Phi_p)P(p)R,P(p)R\rangle&=\langle\nabla^2E_P(\Phi_{p_0})R,R\rangle+\mathcal{R}^{1,2}_{k,m}(\Pi(R),R)\\
&\quad+\langle(\nabla^2E_P(\Phi_p)-\nabla^2E_P(\Phi_{p_0}))R,R\rangle\\
\lambda_j(p)&=\lambda_j(p_0)+\mathcal{R}^{1,0}(\Pi(R))+\mathcal{R}^{1,2}_{k,m}(\Pi(R),R)\\
\Pi_j(P(p)R)&=\Pi_j(R)+\mathcal{R}^{1,2}_{k,m}(\Pi(R),R).
\end{aligned}
$$

Then we conclude that the right hand side of (102) is

$$
\begin{aligned}
&\overbrace{\frac{1}{2}\langle(\mathcal{D}-\lambda(p_0)\cdot\diamond+\nabla^2E_P(\Phi_{p_0}))R,R\rangle}^{\frac{1}{2}\langle J^{-1}\mathcal{H}_{p_0}R,R\rangle}+\mathcal{R}^{2,0}(\Pi(R))+\mathcal{R}^{1,2}_{k,m}(\Pi(R),R)\\
&\qquad+\frac{1}{2}\langle(\nabla^2E_P(\Phi_p)-\nabla^2E_P(\Phi_{p_0}))R,R\rangle
\end{aligned}
\tag{103}
$$

where the last term can be absorbed in the $d=2$ term of $\mathbf{R}'$ by (34). Setting $\psi(\varrho)=\underline{\psi}(\varrho)+\mathcal{R}^{2,0}(\varrho)$ with the $\mathcal{R}^{2,0}$ in (103), we get the desired result. □

We have completed the part of this paper devoted to the Darboux Theorem. The next step consists in the decomposition of R into discrete and continuous modes, and the search of a new coordinate system by an appropriate Birkhoff normal forms argument.

5. Spectral coordinates associated to $\mathcal{H}_{p_0}$

We will consider the operator $\mathcal{H}_{p_0}$, which will be central in our analysis henceforth. We will list now various hypotheses, starting with the spectrum of $\mathcal{H}_{p_0}$ thought as an operator in the natural complexification $L^2(\mathbb{R}^3, \mathbb{C}^{2N})$ of $L^2(\mathbb{R}^3, \mathbb{R}^{2N})$.

(L1) $\sigma_e(\mathcal{H}_{p_0})$ is a union of intervals in $\mathrm{i}\mathbb{R}$ with $0 \notin \sigma_e(\mathcal{H}_{p_0})$ and is symmetric with respect to 0.

(L2) $\sigma_p(\mathcal{H}_{p_0})$ is finite.

(L3) For any eigenvalue $\mathbf{e} \in \sigma_p(\mathcal{H}_{p_0})\backslash\{0\}$ the algebraic and geometric dimensions coincide and are finite.

(L4) There is a number $\mathbf{n} \geq 1$ and positive numbers $0 < \mathbf{e}'_1 \leq \mathbf{e}'_2 \leq ... \leq \mathbf{e}'_{\mathbf{n}}$ such that $\sigma_p(\mathcal{H}_{p_0})$ consists exactly of the numbers $\pm\mathrm{i}\mathbf{e}'_j$ and 0. We assume that there are fixed integers $\mathbf{n}_0 = 0 < \mathbf{n}_1 < ... < \mathbf{n}_{l_0} = \mathbf{n}$ such that $\mathbf{e}'_j = \mathbf{e}'_i$ exactly for i and j both in $(\mathbf{n}_l, \mathbf{n}_{l+1}]$ for some $l \leq l_0$. In this case $\dim\ker(\mathcal{H}_{p_0} - \mathbf{e}'_j) = \mathbf{n}_{l+1} - \mathbf{n}_l$. We assume there exist $N_j \in \mathbb{N}$ such that $N_j + 1 = \inf\{n \in \mathbb{N} : n\mathbf{e}'_j \in \sigma_e(\mathcal{H}_{p_0})\}$. We set $\mathbf{N} = \sup_j N_j$. We assume that $\mathbf{e}'_j \notin \sigma_p(\mathcal{H}_{p_0})$ for all j.

(L5) If $\mathbf{e}'_{j_1} < ... < \mathbf{e}'_{j_i}$ are i distinct λ's, and $\mu \in \mathbb{Z}^k$ satisfies $|\mu| \leq 2N+3$, then we have

$$\mu_1\mathbf{e}'_{j_1} + \cdots + \mu_k\mathbf{e}'_{j_i} = 0 \iff \mu = 0 \ .$$

The following hypothesis holds quite generally.

(L6) If $\varphi \in \ker(\mathcal{H}_{p_0} - \mathrm{i}\mathbf{e})$ for $\mathrm{i}\mathbf{e} \in \sigma_p(\mathcal{H}_{p_0})$ then $\varphi \in \mathcal{S}(\mathbb{R}^3, \mathbb{C}^{2N})$.

By (15), $\mathcal{H}_{p_0}\xi = \mathbf{e}\xi$ implies $\mathcal{H}^*_{p_0}J^{-1}\xi = -\mathbf{e}J^{-1}\xi$. Then $\sigma_p(\mathcal{H}_{p_0}) = \sigma_p(\mathcal{H}^*_{p_0})$. We denote it by σ_p.
By general argument we have:

LEMMA 5.1. *The following spectral decomposition remains determined:*

$$\begin{aligned} &N_g^{\perp}(\mathcal{H}^*_{p_0}) \otimes_{\mathbb{R}} \mathbb{C} = \big(\oplus_{\mathbf{e}\in\sigma_p\backslash\{0\}} \ker(\mathcal{H}_{p_0} - \mathbf{e})\big) \oplus X_c(p_0) \qquad (104)\\ &X_c(p_0) := \big\{N_g(\mathcal{H}^*_{p_0}) \oplus \big(\oplus_{\mathbf{e}\in\sigma_p\backslash\{0\}} \ker(\mathcal{H}^*_{p_0} - \mathbf{e})\big)\big\}^{\perp} . \end{aligned}$$

We denote by P_c the projection on $X_c(p_0)$ associated to (104). Set $\mathcal{H} := \mathcal{H}_{p_0}P_c$.
The following hypothesis is important to solve the homological equations in the Birkhoff normal forms argument.

(L7) We have $R_{\mathcal{H}} \circ \diamondsuit_j^i \in C^\omega(\rho(\mathcal{H}), B(\Sigma_n, \Sigma_n))$ for any $n \in \mathbb{N}$, any $j = 1, ..., n_0$ and for any $i = 0, 1$, where $\rho(\mathcal{H}) = \mathbb{C} \backslash \sigma_e(\mathcal{H}_{p_0})$.

For the examples in Section 7, (L7) can be checked with standard arguments. We discuss now the choice of a good frame of eigenfunctions.

LEMMA 5.2. *It is possible to choose eigenfunctions* $\xi' \in \ker(\mathcal{H}_{p_0} - \mathrm{i}\mathbf{e}'_j)$ *so that* $\Omega(\xi'_j, \overline{\xi}'_k) = 0$ *for* $j \neq k$ *and* $\Omega(\xi'_j, \overline{\xi}'_j) = -\mathrm{i}s_j$ *with* $s_j \in \{1, -1\}$. *We have* $\Omega(\xi'_j, \xi'_k) = 0$ *for all* j *and* k. *We have* $\Omega(\xi, f) = 0$ *for any eigenfunction* ξ *and any* $f \in X_c(p_0)$.

Proof. First of all, if $\lambda, \mu \in \sigma_p(\mathcal{H}_{p_0})$ are two eigenvalues with $\lambda \neq 0$ and given two associated eigenfunctions ξ_μ and ξ_λ

$$\begin{aligned} \langle J^{-1}\xi_\lambda, \overline{\xi}_\mu\rangle &= \frac{1}{\lambda}\langle J^{-1}\mathcal{H}_{p_0}\xi_\lambda, \overline{\xi}_\mu\rangle = -\frac{1}{\lambda}\langle \mathcal{H}^*_{p_0} J^{-1}\xi_\lambda, \overline{\xi}_\mu\rangle \\ &= -\frac{1}{\lambda}\langle J^{-1}\xi_\lambda, \mathcal{H}_{p_0}\overline{\xi}_\mu\rangle = -\frac{\overline{\mu}}{\lambda}\langle J^{-1}\xi_\lambda, \overline{\xi}_\mu\rangle, \end{aligned} \tag{105}$$

where for the second equality we used (15) and for the last one the fact that $\mathcal{H}_{p_0}\xi = \mu\xi$ implies $\mathcal{H}_{p_0}\overline{\xi} = \overline{\mu}\overline{\xi}$. Then, for $\mathbf{e}_j \neq \mathbf{e}_k$ and associated eigenfunctions ξ_j and ξ_k we get $\Omega(\xi_j, \overline{\xi}_k) = 0$. Notice that by a similar argument we have $\Omega(\xi_\lambda, \xi_\mu) = -\frac{\mu}{\lambda}\Omega(\xi_\lambda, \xi_\mu)$ and so $\Omega(\xi'_j, \xi'_k) \equiv 0$.
Since $\mathcal{H}_{p_0}\xi = \mathbf{e}\xi$ implies $\mathcal{H}^*_{p_0} J^{-1}\xi = -\mathbf{e}J^{-1}\xi$, for any eigenfunction ξ of $\mathcal{H}_{p_0}$ then $J^{-1}\xi$ is an eigenfunction of $\mathcal{H}^*_{p_0}$. By the definition of $X_c(p_0)$ in (104), we conclude $\Omega(\xi, f) = \langle J^{-1}\xi, f\rangle = 0$ for any $f \in X_c(p_0)$.
Let $\mathrm{i}\mathbf{e} \in \mathrm{i}\mathbb{R}\backslash\{0\}$ be an eigenvalue. By the above discussion, the Hermitian form $\langle \mathrm{i}J^{-1}\xi, \overline{\eta}\rangle$ is non degenerate in $\ker(\mathcal{H}_{p_0} - \mathrm{i}\mathbf{e})$. Then we can find a basis such that $\langle \mathrm{i}J^{-1}\eta_j, \overline{\eta}_k\rangle = -|a_j|\mathrm{sign}(a_j)\delta_{jk}$, for appropriate non zero numbers $a_j \in \mathbb{R}$. Then set $\xi' = \sqrt{|a_j|}\eta_j$. □

We set $\xi_j = \xi'_j$ and $\mathbf{e}_j = \mathbf{e}'_j$ if $s_j = 1$.
We set $\overline{\xi}_j = \xi'_j$ and $\mathbf{e}_j = -\mathbf{e}'_j$ if $s_j = -1$.
Notice that if $f \in X_c(p_0)$ then also $\overline{f} \in X_c(p_0)$. This implies that for $R \in N_g^\perp(\mathcal{H}^*_{p_0}) \otimes_{\mathbb{R}} \mathbb{C}$ with real entries, that is if $R = \overline{R}$, then we have

$$R(x) = \sum_{j=1}^{\mathbf{n}} z_j\xi_j(x) + \sum_{j=1}^{\mathbf{n}} \overline{z}_j\overline{\xi}_j(x) + f(x), \quad f \in X_c(p_0). \tag{106}$$

with $f = \overline{f}$.
By Lemma 5.2 we have, for the s_j of Lemma 5.2,

$$\frac{1}{2}\Omega(\mathcal{H}_{p_0}R, R) = \sum_{j=1}^{\mathbf{n}} \mathbf{e}_j|z_j|^2 + \frac{1}{2}\Omega(\mathcal{H}_{p_0}f, f) =: H_2. \tag{107}$$

Consider the map $R \to (z, f)$ obtained from (106). In terms of the pair (z, f), the Fréchet derivative R' can be expressed as

$$R' = \sum_{j=1}^{\mathbf{n}} (dz_j \xi_j + d\overline{z}_j \overline{\xi}_j) + f'.$$

We have

$$\Omega(R', R') = -\mathrm{i} \sum_{j=1}^{\mathbf{n}} dz_j \wedge d\overline{z}_j + \Omega(f', f'). \tag{108}$$

For a function F independent of τ and Π let us decompose X_F as of spectral decomposition (106):

$$X_F = \sum_{j=1}^{\mathbf{n}} (X_F)_{z_j} \xi_j(x) + \sum_{j=1}^{\mathbf{n}} (X_F)_{\overline{z}_j} \overline{\xi}_j(x) + (X_F)_f, \quad (X_F)_f \in X_c(p_0).$$

By $i_{X_F}\Omega = dF$ and by

$$\begin{aligned} dF &= \partial_{z_j} F dz_j + \partial_{\overline{z}_j} F d\overline{z}_j + \langle \nabla_f F, f' \rangle \\ i_{X_F}\Omega &= -\mathrm{i}(X_F)_{z_j} d\overline{z}_j + \mathrm{i}(X_F)_{\overline{z}_j} dz_j + \langle J^{-1}(X_F)_f, f' \rangle, \end{aligned}$$

we get

$$(X_F)_{z_j} = \mathrm{i}\partial_{\overline{z}_j} F \ , \quad (X_F)_{\overline{z}_j} = -\mathrm{i}\partial_{z_j} F \ , \quad (X_F)_f = J\nabla_f F.$$

This implies

$$\{F, G\} := dF(X_G) = \mathrm{i}\partial_{z_j} F \partial_{\overline{z}_j} G - \mathrm{i}\partial_{\overline{z}_j} F \partial_{z_j} G + \langle \nabla_f F, J\nabla_f G \rangle. \tag{109}$$

Hence, for H_2 defined in (107), for $z = (z_1, ..., z_{\mathbf{n}})$, using standard multi index notation and by (15), we have:

$$\{H_2, z^\mu \overline{z}^\nu\} = -\mathrm{i}\mathbf{e} \cdot (\mu - \nu) z^\mu \overline{z}^\nu \ ; \quad \{H_2, \langle J^{-1}\varphi, f \rangle\} = \langle J^{-1}\mathcal{H}\varphi, f \rangle. \tag{110}$$

5.1. Flows in spectral coordinates

We restate Lemma 3.8 for a special class of transformations.

LEMMA 5.3. *Consider*

$$\chi = \sum_{|\mu+\nu|=M_0+1} b_{\mu\nu}(\Pi(f)) z^\mu \overline{z}^\nu + \sum_{|\mu+\nu|=M_0} z^\mu \overline{z}^\nu \langle J^{-1} B_{\mu\nu}(\Pi(f)), f \rangle \tag{111}$$

with $b_{\mu\nu}(\varrho) = \mathcal{R}^{i,0}_{r,M}(\varrho)$ *and* $B_{\mu\nu}(\varrho) = \boldsymbol{S}^{i,0}_{r,M}(\varrho)$ *with* $i \in \{0, 1\}$ *fixed and* $r, M \in \mathbb{N}$ *sufficiently large and with*

$$\overline{b}_{\mu\nu} = b_{\nu\mu} \ , \quad \overline{B}_{\mu\nu} = B_{\nu\mu}, \tag{112}$$

(so that χ *is real valued for* $f = \overline{f}$*). Then we have what follows.*

(1) Consider the vectorfield X_χ defined with respect to Ω_0. Then, summing on repeated indexes (with the equalities defining the field X^{st}_χ), we have:

$$(X_\chi)_{z_j} = \mathrm{i}\partial_{\overline{z}_j}\chi =: (X^{st}_\chi)_{z_j}\,, \quad (X_\chi)_{\overline{z}_j} = -\mathrm{i}\partial_{z_j}\chi =: (X^{st}_\chi)_{\overline{z}_j}\,,$$
$$(X_\chi)_f = \partial_{\Pi_j(f)}\chi\, P_c^*(p_0)J\diamond_j f + (X^{st}_\chi)_f \text{ where } (X^{st}_\chi)_f := z^\mu \overline{z}^\nu B_{\mu\nu}(\Pi(f)).$$

(2) Denote by ϕ^t the flow of X_χ provided by Lemma 3.8 and set $(z^t, f^t) = (z, f)\circ\phi^t$. Then we have

$$z^t = z + \mathcal{Z}(t) \qquad f^t = e^{Jq(t)\cdot\diamond}(f + \boldsymbol{S}(t)) \tag{113}$$

where, for (k, m) with $k \in \mathbb{Z}\cap[0, r-(m+1)\boldsymbol{d}]$ and $1 \le m \le M$, for $B_{\Sigma_{-k}}$ a sufficiently small neighborhood of 0 in $\Sigma_{-k}\cap X_c(p_0)$ and for $B_{\mathbb{C}^{\mathbf{n}}}$ (resp.$B_{\mathbb{R}^{n_0}}$) a neighborhood of 0 in $\mathbb{C}^{\mathbf{n}}$ (resp.$\mathbb{R}^{n_0}$)

$$\begin{aligned}
&\boldsymbol{S} \in C^m((-2,2)\times B_{\mathbb{C}^{\mathbf{n}}}\times B_{\Sigma_{-k}}\times B_{\mathbb{R}^{n_0}}, \Sigma_k)\\
&q \in C^m((-2,2)\times B_{\mathbb{C}^{\mathbf{n}}}\times B_{\Sigma_{-k}}\times B_{\mathbb{R}^{n_0}}, \mathbb{R}^{n_0})\\
&\mathcal{Z} \in C^m((-2,2)\times B_{\mathbb{C}^{\mathbf{n}}}\times B_{\Sigma_{-k}}\times B_{\mathbb{R}^{n_0}}, \mathbb{C}^{\mathbf{n}}),
\end{aligned} \tag{114}$$

with for fixed C

$$\begin{aligned}
&|q(t,z,f,\varrho)| \le C(|z| + \|f\|_{\Sigma_{-k}})^{M_0+1}\\
&|\mathcal{Z}(t,z,f,\varrho)| + \|\boldsymbol{S}(t,z,f,\varrho)\|_{\Sigma_k} \le C(|z| + \|f\|_{\Sigma_{-k}})^{M_0}.
\end{aligned} \tag{115}$$

We have $\boldsymbol{S}(t,z,f,\varrho) = \boldsymbol{S}_1(t,z,f,\varrho) + \boldsymbol{S}_2(t,z,f,\varrho)$ with

$$\begin{aligned}
&\boldsymbol{S}_1(t,z,f,\varrho) = \int_0^t (X^{st}_\chi)_f\circ\phi^{t'}dt'\\
&\|\boldsymbol{S}_2(t,z,f,\varrho)\|_{\Sigma_k} \le C(|z| + \|f\|_{\Sigma_{-k}})^{2M_0+1}(|z| + \|f\|_{\Sigma_{-k}} + |\varrho|)^i.
\end{aligned} \tag{116}$$

(3) The flow ϕ^t is canonical: for s, s', k as in Lemma 3.8, the map $\phi^t \in C^l(\mathcal{U}^{s'}_{\varepsilon_1,k}, \widetilde{\mathcal{P}}^s)$ satisfies $\phi^{t*}\Omega_0 = \Omega_0$ in $C^\infty(\mathcal{U}^{s'}_{\varepsilon_2,k}, B^2(\widetilde{\mathcal{P}}^{s'}, \mathbb{R}))$ for $\varepsilon_2 > 0$ sufficiently small.

Proof. First of all notice that χ does not depend on τ and Π so that the only nonzero component of X_χ is $(X_\chi)_R = J\nabla_R\chi$. The latter is of the form indicated in claim (1) by a direct computation. Claim (2) follows now by Lemma 3.8. To prove Claim (3) we need to make rigorous the following formal computation

$$\frac{d}{dt}\phi^{t*}\Omega_0 = \phi^{t*}L_{X_\chi}\Omega_0 = \phi^{t*}di_{X_\chi}\Omega_0 = \phi^{t*}d^2\chi = 0.$$

To make sense of this we can proceed as in Corollary 3.12. We skip the proof. □

LEMMA 5.4. *Consider a transformation $\mathfrak{F} = \mathfrak{F}_1 \circ \cdots \circ \mathfrak{F}_L$ like in Lemma 4.1 and with $m_1 = 2$ and for fixed r and M sufficiently large. Denote by (k', m') the pair (k, m) of Lemma 4.4 and consider a pair (k, m) with $k \leq k'$ and $m \leq m' - (2\boldsymbol{N}+5)$. Set $H' := K \circ \mathfrak{F}$. Consider decomposition* (106). *Then on a domain $\mathcal{U}^s_{\varepsilon,k}$ like* (57) *we have*

$$H' = \psi(\Pi(f)) + H'_2 + \boldsymbol{R}\ , \tag{117}$$

for a $\psi \in C^\infty$ with $\psi(\varrho) = O(|\varrho|^2)$ near 0 and with what follows.

(1) We have

$$H'_2 = \sum_{\substack{|\mu+\nu|=2\\ \mathbf{e}\cdot(\mu-\nu)=0}} a_{\mu\nu}(\Pi(f)) z^\mu \overline{z}^\nu + \frac{1}{2}\langle J^{-1}\mathcal{H}_{p_0} f, f\rangle. \tag{118}$$

(2) We have $\boldsymbol{R} = \boldsymbol{R}_{-1} + \boldsymbol{R}_0 + \boldsymbol{R}_1 + \boldsymbol{R}_2 + \mathcal{R}^{1,2}_{k,m+2}(\Pi(f), f) + \boldsymbol{R}_3 + \boldsymbol{R}_4$, with:

$$\boldsymbol{R}_{-1} = \sum_{\substack{|\mu+\nu|=2\\ \mathbf{e}\cdot(\mu-\nu)\neq 0}} a_{\mu\nu}(\Pi(f)) z^\mu \overline{z}^\nu + \sum_{|\mu+\nu|=1} z^\mu \overline{z}^\nu \langle J^{-1} G_{\mu\nu}(\Pi(f)), f\rangle;$$

For $\mathbf{N}$ as in (L4) of this section,

$$\boldsymbol{R}_0 = \sum_{|\mu+\nu|=3}^{2\mathbf{N}+1} z^\mu \overline{z}^\nu a_{\mu\nu}(\Pi(f));$$

$$\boldsymbol{R}_1 = \sum_{|\mu+\nu|=2}^{2\mathbf{N}} z^\mu \overline{z}^\nu \langle J^{-1} G_{\mu\nu}(\Pi(f)), f\rangle;$$

$$\boldsymbol{R}_2 = \langle \mathbf{B}_2(\Pi(f)), f^2\rangle \text{ with } \mathbf{B}_2(0) = 0$$

where $f^d(x)$ represents schematically $d-$products of components of f;

$$\boldsymbol{R}_3 = \sum_{\substack{|\mu+\nu|=\\ =2N+2}} z^\mu \overline{z}^\nu a_{\mu\nu}(z, f, \Pi(f)) + \sum_{\substack{|\mu+\nu|=\\ =2N+1}} z^\mu \overline{z}^\nu \langle J^{-1} G_{\mu\nu}(z, f, \Pi(f)), f\rangle;$$

$$\boldsymbol{R}_4 = \sum_{d=2}^{4} \langle B_d(z, f, \Pi(f)), f^d\rangle + \int_{\mathbb{R}^3} B_5(x, z, f, f(x), \Pi(f)) f^5(x) dx$$
$$+ \widehat{\boldsymbol{R}}_2(z, f, \Pi(f)) + E_P(f) \text{ with } B_2(0, 0, \varrho) = 0.$$

(3) For $\delta_j := (\delta_{1j}, ..., \delta_{mj})$,

$$\begin{aligned} &a_{\mu\nu}(0) = 0 \text{ for } |\mu+\nu| = 2 \text{ with } (\mu,\nu) \neq (\delta_j, \delta_j) \text{ for all } j, \\ &a_{\delta_j\delta_j}(0) = \lambda_j(\omega_0), \\ &G_{\mu\nu}(0) = 0 \text{ for } |\mu+\nu| = 1 \ . \end{aligned} \tag{119}$$

These $a_{\mu\nu}(\varrho)$ *and* $G_{\mu\nu}(x,\varrho)$ *are* C^m *in all variables with* $G_{\mu\nu}(\cdot,\varrho) \in C^m(\mathrm{U}, \Sigma_k(\mathbb{R}^3, \mathbb{C}^{2N}))$*, for a small neighborhood* U *of* $(0,0,0)$ *in* $\mathbb{C}^{\mathbf{n}} \times (\Sigma_{-k} \cap X_c(p_0)) \times \mathbb{R}^{n_0}$ *(the space of the* (z,f,ϱ)*), and they satisfy symmetries analogous to* (112)*.*

(4) We have $a_{\mu\nu}(z,\varrho) \in C^m(\mathrm{U}, \mathbb{C})$ *.*

(5) $G_{\mu\nu}(\cdot, z, \varrho) \in C^m(\mathrm{U}, \Sigma_k(\mathbb{R}^3, \mathbb{C}^{2N}))$*.*

(6) $B_d(\cdot, z, f, \varrho) \in C^m(\mathrm{U}, \Sigma_k(\mathbb{R}^3, B((\mathbb{C}^{2N})^{\otimes d}, \mathbb{R})))$*, for* $2 \leq d \leq 4$*.* $\mathbf{B}_2(\cdot, \varrho)$ *satisfies the same property.*

(7) Let $\zeta \in \mathbb{C}^{2N}$*. Then for* $B_5(\cdot, z, f, \zeta, \varrho)$ *we have (the derivatives are not in the holomorphic sense)*

$$\text{for } |l| \leq m \ , \ \|\nabla^l_{z,f,\zeta,\varrho} B_5(z,f,\zeta,\varrho)\|_{\Sigma_k(\mathbb{R}^3, B((\mathbb{R}^{2N})^{\otimes 5}, \mathbb{R})} \leq C_l.$$

(8)

$$\begin{aligned} &\widehat{\boldsymbol{R}}_2 \in C^m(\mathrm{U}, \mathbb{C}), \\ &|\widehat{\boldsymbol{R}}_2(z,f,\varrho)| \leq C(|z| + \|f\|_{\Sigma_{-k}})\|f\|^2_{\Sigma_{-k}}; \end{aligned} \tag{120}$$

Proof. We need to express R in terms of (z,f) using (106) inside (101). We have $\Pi(R) = \Pi(f) + \mathcal{R}^{0,2}(R)$. Then, succinctly,

$$\mathcal{R}^{1,2}_{k',m'}(\Pi(R), R) = \sum_{a+b=2}^{2\mathbf{N}+1} \frac{1}{a!b!} \langle \nabla^a_\varrho \nabla^b_R \mathcal{R}^{1,2}_{k',m'}(\Pi(f), 0), (\mathcal{R}^{0,2}(R))^a R^{b\otimes} \rangle$$

$$+ \sum_{\substack{a+b \\ =2\mathbf{N}+2}} \int_0^1 \frac{(1-t)^{2\mathbf{N}+1}}{a!b!} \langle \nabla^a_\varrho \nabla^b_R \mathcal{R}^{1,2}_{k',m'}(\Pi(f) + t\mathcal{R}^{0,2}(R), tR), (\mathcal{R}^{0,2}(R))^a R^{b\otimes} \rangle dt,$$

with (k',m') the pair (k,m) of Lemma 4.4. We substitute (106), that is $R = z\cdot\xi + \overline{z}\cdot\overline{\xi} + f$. For $m \leq m' - (2\mathbf{N}+2)$ and $k \leq k'$, the terms from the $R^{b\otimes}$ of degree in f at most 1, go into $\mathbf{R}_i$ with $i = -1, 0, 1, 3$ and H'_2. For $m \leq m' - (2\mathbf{N}+4)$, the remaining terms are absorbed in $\mathcal{R}^{1,2}_{k',m+2}(\Pi(f), f) + \widehat{\mathbf{R}}_2(z, f, \Pi(f))$.
We focus now on the $d = 5$ term in (101). We substitute $R = z \cdot \xi + \overline{z} \cdot \overline{\xi} + f$. This schematically yields, for a $\widetilde{B}_5$ satisfying claim (7) with the pair (m', k'),

$$\sum_{j=0}^{5}\int_{\mathbb{R}^3}\widetilde{B}_5(x,z,f,f(x),\Pi(f))(z\cdot\xi+\overline{z}\cdot\overline{\xi})^{5-j}f^j(x)dx. \tag{121}$$

For $j=5$ we get a term that can be absorbed in the B_5 term in $\mathbf{R}_4$. Expand the $j<5$ terms in (121) as

$$\sum_{i=0}^{4-j}\int_{\mathbb{R}^3}\frac{1}{i!}(\partial_t^i)_{|t=0}\widetilde{B}_5(x,z,f,tf(x),\Pi(f))(z\cdot\xi+\overline{z}\cdot\overline{\xi})^{5-j}f^{i+j}(x)dx$$
$$+\int_{\mathbb{R}^3}\frac{1}{(4-j)!}\int_0^1\partial_t^{5-j}[\widetilde{B}_5(x,z,f,tf(x),\Pi(f))](z\cdot\xi+\overline{z}\cdot\overline{\xi})^{5-j}f^5(x)dx.$$

go into the B_d term in $\mathbf{R}_4$ The last term fits in the B_5 term in $\mathbf{R}_4$ by $m\le m'-5$. The terms in the first line go into the B_d of $\mathbf{R}_4$ for $d=i+j\ge 2$. The terms with $i+j<2$ can be treated like the $\mathcal{R}^{1,2}_{k',m'}(\Pi(R),R)$ for $m\le m'-(2\mathbf{N}+5)$ and $k\le k'$.
We focus on $E_P(R)=E_P(z\cdot\xi+\overline{z}\cdot\overline{\xi}+f)$. We use Lemma 4.2 for $\mathbf{v}=f$ and $\mathbf{u}=z\cdot\xi+\overline{z}\cdot\overline{\xi}$. Then

$$E_P(R)=E_P(f)+E_P(z\cdot\xi+\overline{z}\cdot\overline{\xi})$$
$$+\int_{\mathbb{R}^3}dx\sum_{j=0}^{3}\int_{[0,1]^2}\frac{t^j}{j!}(\partial_t^{j+1})_{|t=0}\partial_s[B(|s(z\cdot\xi+\overline{z}\cdot\overline{\xi})+tf|_1^2)]dtds$$
$$+\int_{\mathbb{R}^3}dx\int_{[0,1]^2}dtds\int_0^t\partial_\tau^5\partial_s[B(|s(z\cdot\xi+\overline{z}\cdot\overline{\xi})+\tau f|_1^2)]\frac{(t-\tau)^3}{3!}d\tau.$$

By $B(0)=B'(0)=0$, we have $E_P(z\cdot\xi+\overline{z}\cdot\overline{\xi})=\mathcal{R}^{0,4}(R)$. It is easy to conclude that this term easily fits into $\mathbf{R}_0+\mathbf{R}_3$. Similarly, the $j=0$ term fits in $\mathbf{R}_1+\mathbf{R}_3$. The $j\ge 1$ terms fit in the B_{j+1} term in $\mathbf{R}_4$. The last line fits in the B_5 term in $\mathbf{R}_4$.
The symmetries (112) for the coefficients in $H_2'+\mathbf{R}_{-1}+\mathbf{R}_0+\mathbf{R}_1$ are an elementary consequence of the fact that H' is real valued. □

Remark 5.5. *Given a Hamiltonian H' expanded as in Lemma 5.4 and given a transformation $\mathfrak{F}$, we cannot obtain the expansion of Lemma 5.4 for $H'\circ\mathfrak{F}$ analysing one by one the terms of the expansion of H'. This works in the set up of [8, 10] but not here (see in particular the discussion on the exponential under formula* (152) *later).*

6. Birkhoff normal forms

In this section we arrive at the main result of the paper.

6.1. Homological equations

We consider $a^{(\ell)}_{\mu\nu}(\varrho) \in C^{\widehat{m}}(U, C)$ for $k_0 \in \mathbb{N}$ a fixed number and U a neighborhood of 0 in $\mathbb{R}^{n_0}$. Then we set

$$H^{(\ell)}_2(\varrho) := \sum_{\substack{|\mu+\nu|=2 \\ \mathbf{e}\cdot(\mu-\nu)=0}} a^{(\ell)}_{\mu\nu}(\varrho) z^\mu \overline{z}^\nu + \frac{1}{2}\langle J^{-1}\mathcal{H}f, f\rangle. \tag{122}$$

$$\mathbf{e}_j(\varrho) := a^{(\ell)}_{\delta_j\delta_j}(\varrho), \quad \mathbf{e}(\varrho) = (\lambda_1(\varrho), \cdots, \lambda_m(\varrho)). \tag{123}$$

We assume $\mathbf{e}_j(0) = \mathbf{e}_j$ and $a^{(\ell)}_{\mu\nu}(0) = 0$ if $(\mu, \nu) \neq (\delta_j, \delta_j)$ for all j, with δ_j defined in (119).

Definition 6.1. *A function $Z(z, f, \varrho)$ is in normal form if $Z = Z_0 + Z_1$ where Z_0 and Z_1 are finite sums of the following type:*

$$Z_1 = \sum_{\mathbf{e}(0)\cdot(\nu-\mu)\in\sigma_e(\mathcal{H}_{p_0})} z^\mu \overline{z}^\nu \langle J^{-1} G_{\mu\nu}(\varrho), f\rangle \tag{124}$$

with $G_{\mu\nu}(x, \varrho) \in C^m(U, \Sigma_k(\mathbb{R}^3, \mathbb{C}^{2N}))$ for fixed $k, m \in \mathbb{N}$ and $U \subseteq \mathbb{R}^{n_0}$ a neighborhood of 0;

$$Z_0 = \sum_{\mathbf{e}(0)\cdot(\mu-\nu)=0} g_{\mu\nu}(\varrho) z^\mu \overline{z}^\nu \tag{125}$$

and $g_{\mu\nu}(\varrho) \in C^m(U, \mathbb{C})$. We assume furthermore that the above coefficients satisfy the symmetries in (112)*: that is $\overline{g}_{\mu\nu} = g_{\nu\mu}$ and $\overline{G}_{\mu\nu} = G_{\nu\mu}$.*

Lemma 6.2. *We consider $\chi = \chi(b, B)$ with*

$$\chi(b, B) = \sum_{|\mu+\nu|=M_0+1} b_{\mu\nu} z^\mu \overline{z}^\nu + \sum_{|\mu+\nu|=M_0} z^\mu \overline{z}^\nu \langle J^{-1} B_{\mu\nu}, f\rangle \tag{126}$$

for $b_{\mu\nu} \in \mathbb{C}$ and $B_{\mu\nu} \in \Sigma_{\widehat{k}}(\mathbb{R}^3, \mathbb{C}^{2N}) \cap X_c(p_0)$ with $\widehat{k} \in \mathbb{N}$, satisfying the symmetries in (112)*. Here we interpret the polynomial χ as a function with parameters $b = (b_{\mu\nu})$ and $B = (B_{\mu\nu})$. Denote by $X_{\widehat{k}}$ the space of the pairs (b, B). Let us also consider given polynomials with $K = K(\varrho)$ and $\widetilde{K} = \widetilde{K}(\varrho, b, B)$ where:*

$$K(\varrho) := \sum_{|\mu+\nu|=M_0+1} k_{\mu\nu}(\varrho) z^\mu \overline{z}^\nu + \sum_{|\mu+\nu|=M_0} z^\mu \overline{z}^\nu \langle J^{-1} K_{\mu\nu}(\varrho), f\rangle, \tag{127}$$

with $k_{\mu\nu}(\varrho) \in C^{\widehat{m}}(U, \mathbb{C})$ and $K_{\mu\nu}(\varrho) \in C^{\widehat{m}}(U, \Sigma_{\widehat{k}}(\mathbb{R}^3, \mathbb{C}^{2N}) \cap X_c(p_0))$ for U a neighborhood of 0 in $\mathbb{R}^{n_0}$, satisfying the symmetries in (112)*;*

$$\begin{aligned}\widetilde{K}(\varrho,b,B) :=& \sum_{|\mu+\nu|=M_0+1} \widetilde{k}_{\mu\nu}(\varrho,b,B) z^\mu \overline{z}^\nu \\ &+ \sum_{i=0}^{1}\sum_{j=1}^{n_0}\sum_{|\mu+\nu|=M_0} z^\mu\overline{z}^\nu \langle J^{-1}\diamondsuit_j^i K^i_{j\mu\nu}(\varrho,b,B), f\rangle,\end{aligned} \tag{128}$$

with $\widetilde{k}_{\mu\nu}\in C^{\widehat{m}}(U\times X_{\widehat{k}},\mathbb{R})$ *and* $\widetilde{K}^i_{j\mu\nu}\in C^{\widehat{m}}(U\times X_{\widehat{k}},\Sigma_{\widehat{k}}(\mathbb{R}^3,\mathbb{C}^{2N})\cap X_c(p_0))$*, satisfying the symmetries in* (112)*. Suppose also that the sums* (127) *and* (128) *do not contain terms in normal form and that* $\widetilde{K}(0,b,B)=0$*. Then there exists a neighborhood* $V\subseteq U$ *of 0 in* $\mathbb{R}^{n_0}$ *and a unique choice of functions* $(b(\varrho),B(\varrho))\in C^{\widehat{m}}(V,X_{\widehat{k}})$ *such that for* $\chi(\varrho)=\chi(b(\varrho),B(\varrho))$*,* $\widetilde{K}(\varrho)=\widetilde{K}(\varrho,b(\varrho),B(\varrho))$ *we have*

$$\left\{\chi(\varrho),H_2^{(\ell)}(\varrho)\right\}^{st}=K(\varrho)+\widetilde{K}(\varrho)+Z(\varrho) \tag{129}$$

where $\{\cdots\}^{st}$ *is the bracket* (109) *for* ϱ *fixed and where* $Z(\varrho)$ *is in normal form and homogeneous of degree* M_0+1 *in* (z,f)*.*

Proof. Summing on repeated indexes, by (110) we get

$$\begin{aligned}\{H_2^{(\ell)},\chi\}^{st} =& -\mathrm{i}\mathbf{e}(\varrho)\cdot(\mu-\nu)z^\mu\overline{z}^\nu b_{\mu\nu}(\varrho)\\ &- z^\mu\overline{z}^\nu\langle f, J^{-1}(\mathrm{i}\mathbf{e}(\varrho)\cdot(\mu-\nu)-\mathcal{H})B_{\mu\nu}(\varrho)\rangle + \widehat{K}(\varrho,b(\varrho),B(\varrho)),\end{aligned} \tag{130}$$

$$\begin{aligned}\widehat{K}(\varrho,b,B) :=& \sum_{\substack{|\mu+\nu|=2\\(\mu,\nu)\neq(\delta_j,\delta_j)\ \forall\ j}} a^{(\ell)}_{\mu\nu}(\varrho)\left[\sum_{|\mu'+\nu'|=M_0+1}\{z^\mu\overline{z}^\nu, z^{\mu'}\overline{z}^{\nu'}\}b_{\mu'\nu'}\right.\\ &\left.+\sum_{|\mu'+\nu'|=M_0}\{z^\mu\overline{z}^\nu, z^{\mu'}\overline{z}^{\nu'}\}\langle J^{-1}B_{\mu'\nu'},f\rangle\right].\end{aligned} \tag{131}$$

$\widehat{K}$ is a homogeneous polynomial of the same type of the above ones and we have $\widehat{K}(0,b,B)=0$. In particular, $\widehat{K}$ satisfies the symmetries (112) by (for $f=\overline{f}$)

$$\begin{aligned}&(a^{(\ell)}_{\mu\nu}b_{\mu'\nu'}\{z^\mu\overline{z}^\nu, z^{\mu'}\overline{z}^{\nu'}\})^* = a^{(\ell)}_{\nu\mu}b_{\nu'\mu'}\{z^\nu\overline{z}^\mu, z^{\nu'}\overline{z}^{\mu'}\}\\ &(a^{(\ell)}_{\mu\nu}\langle J^{-1}B_{\mu'\nu'},f\rangle\{z^\mu\overline{z}^\nu, z^{\mu'}\overline{z}^{\nu'}\})^* = a^{(\ell)}_{\nu\mu}\langle J^{-1}B_{\nu'\mu'},f\rangle\{z^\nu\overline{z}^\mu, z^{\nu'}\overline{z}^{\mu'}\}\end{aligned}$$

which follow by $(\mathrm{i}\partial_{z_j}F\partial_{\overline{z}_j}G-\mathrm{i}\partial_{\overline{z}_j}F\partial_{z_j}G)^*=\mathrm{i}\partial_{z_j}F^*\partial_{\overline{z}_j}G^*-\mathrm{i}\partial_{\overline{z}_j}F^*\partial_{z_j}G^*$, where in these formulas $a^*=\overline{a}$, and by the symmetries (112) for χ and for $H_2^{(\ell)}$.

Denote by $\widehat{Z}(\varrho, b, B)$ the sum of monomials in normal form of $\widetilde{K}$ and set $\mathbf{K} := \widetilde{K} + \widehat{K} - \widehat{Z}$. We look at

$$\begin{aligned} &- \mathrm{i}\mathbf{e}(\varrho)\cdot(\mu-\nu)z^\mu\overline{z}^\nu b_{\mu\nu} - z^\mu\overline{z}^\nu\langle f, J^{-1}(\mathrm{i}\mathbf{e}(\varrho)\cdot(\mu-\nu)-\mathcal{H})B_{\mu\nu}\rangle \\ &\quad + \mathbf{K}\ (\varrho, b, B) + K(\varrho) = 0 \end{aligned} \tag{132}$$

that is at

$$\begin{aligned} &k_{\mu\nu}(\varrho) + \mathbf{k}_{\mu\nu}(\varrho, b, B) - b_{\mu\nu}(\varrho)\mathrm{i}\mathbf{e}(\varrho)\cdot(\mu-\nu) = 0 \\ &B_{\mu\nu}(\varrho) = -R_{\mathcal{H}}(\mathrm{i}\mathbf{e}(\varrho)\cdot(\mu-\nu))\left[K_{\mu\nu}(\varrho) + \mathbf{K}_{\mu\nu}(\varrho, b, B)\right], \end{aligned} \tag{133}$$

with $\mathbf{k}_{\mu\nu}$ and $\mathbf{K}_{\mu\nu}$ the coefficients of $\mathbf{K}$. Notice that when $\mathbf{k}_{\mu\nu}(0, b, B) = 0$ and $\mathbf{K}_{\mu\nu}(0, b, B) = 0$, for $\varrho = 0$ there is a unique solution $(b, B) \in X_{\widehat{k}}$ given by

$$b_{\mu\nu}(0) = \frac{k_{\mu\nu}(0)}{\mathrm{i}\mathbf{e}\cdot(\mu-\nu)}, \quad B_{\mu\nu}(0) = -R_{\mathcal{H}}(\mathrm{i}\mathbf{e}\cdot(\mu-\nu))K_{\mu\nu}(0). \tag{134}$$

Lemma 6.2 is then a consequence of the Implicit Function Theorem by Hypothesis (L7) in Section 5. □

In the particular case $M_0 = 1$ we need a slight variation of Lemma 6.2.

LEMMA 6.3. *Suppose now $M_0 = 1$ and assume the notation of Lemma 6.2, assuming $K(0) = 0$, $\widetilde{K}(0,0,0) = 0$ and $\nabla_{b,B}\widetilde{K}(0,0,0) = 0$. We furthermore consider function $a_{\mu\nu}^{\mu'\nu'} \in C^{\widehat{m}}(U \times X_{\widehat{k}}, \mathbb{C})$ with $|a_{\mu\nu}^{\mu'\nu'}(\varrho, b, B)| \le C\|(b, B)\|_{X_{\widehat{k}}}$ and we set*

$$\begin{aligned} \left\{\chi(\varrho), H_2^{(\ell)}(\varrho)\right\}^{\widetilde{st}} &= \left\{\chi(\varrho), H_2^{(\ell)}(\varrho)\right\}^{st} \\ &\quad + \sum_{\substack{|\mu+\nu|=1 \\ |\mu'+\nu'|=1}} a_{\mu\nu}^{\mu'\nu'}(\varrho, b(\varrho), B(\varrho))z^\mu\overline{z}^\nu\langle\mathcal{H}B_{\mu'\nu'}(\varrho), f\rangle. \end{aligned} \tag{135}$$

Then, the same conclusions of Lemma 6.2 hold for

$$\left\{\chi(\varrho), H_2^{(\ell)}(\varrho)\right\}^{\widetilde{st}} = K(\varrho) + \widetilde{K}(\varrho) + Z(\varrho). \tag{136}$$

Proof. Like above we get to

$$\begin{aligned} &k_{\mu\nu}(\varrho) + \mathbf{k}_{\mu\nu}(\varrho, b, B) - b_{\mu\nu}\mathrm{i}\mathbf{e}(\varrho)\cdot(\mu-\nu) = 0 \\ &B_{\mu\nu} = -R_{\mathcal{H}}(\mathrm{i}\mathbf{e}(\varrho)\cdot(\mu-\nu))[K_{\mu\nu}(\varrho) + \mathbf{K}_{\mu\nu}(\varrho, b, B) + \sum_{\mu'\nu'} a_{\mu\nu}^{\mu'\nu'}(\varrho, b, B)\mathcal{H}B_{\mu'\nu'}]. \end{aligned}$$

For $(\varrho, b, B) = (0, 0, 0)$ both sides are 0. Then Lemma 6.3 follows by Implicit Function Theorem. □

6.2. The Birkhoff normal forms

Our goal in this section is to prove the following result where N is as of (L4) in Section 5.

THEOREM 6.4. *For any integer $2 \le \ell \le 2\mathbf{N}+1$ we have transformations $\mathfrak{F}^{(\ell)} = \mathfrak{F}_1 \circ \phi_2 \circ ... \circ \phi_\ell$, with $\mathfrak{F}_1$ the transformation in Corollary 3.12 the ϕ_j's like in Lemma 5.3, such that the conclusions of Lemma 5.4 hold, that is such that we have the following expansion*

$$H^{(\ell)} := K \circ \mathfrak{F}^{(\ell)} = \psi(\Pi(f)) + H_2^{(\ell)} + \mathcal{R}^{1,2}_{k,m+2}(\Pi(f), f) + \sum_{j=-1}^{4} \boldsymbol{R}_j^{(\ell)},$$

with $H_2^{(\ell)}$ of the form (118) *and with the following additional properties:*

(i) $\boldsymbol{R}_{-1}^{(\ell)} = 0$*;*

(ii) all the nonzero terms in $\boldsymbol{R}_0^{(\ell)}$ with $|\mu+\nu| \le \ell$ are in normal form, that is $\lambda \cdot (\mu - \nu) = 0$;

(iii) all the nonzero terms in $\boldsymbol{R}_1^{(\ell)}$ with $|\mu+\nu| \le \ell - 1$ are in normal form, that is $\lambda \cdot (\mu - \nu) \in \sigma_e(\mathcal{H}_{p_0})$.

Proof. The proof of Theorem 6.4 is by induction. There are two distinct parts in the proof, [2, 8, 10]. Here we follow the ordering of [2]. In the first part we assume that for some $\ell \ge 2$ the statement of the theorem is true, and we show that it continues to be true for $\ell + 1$. The proof of case $\ell = 2$, which presents some additional complications, is dealt in the second part.

In the proof we will get polynomials (111) with $M_0 = 1, ..., 2\mathbf{N}$ with decreasing (r, M) as M_0 increases. Nonetheless, in view of the fact that in Lemma 3.7 the n is arbitrarily large and that (r, M) decreases by a fixed amount at each step, these (r, M) are arbitrarily large. This is exploited in Theorem 6.5 later.

The step $\ell + 1 > 2$. We can assume that $H^{(\ell)}$ have the desired properties for indexes (k', m') (instead of (k, m)) arbitrarily large. We consider the representation (117) for $H^{(\ell)}$ and we set $\mathbf{h} = H^{(\ell)}(z, f, \varrho)$ replacing $\Pi(f)$ with ϱ in (117). Then $\mathbf{h} = H^{(\ell)}(z, f, \varrho)$ is $C^{2\mathbf{N}+2}$ near 0 in $\mathcal{P}^{s_0} = \{(\varrho, R)\}$ for $m' \ge 2\mathbf{N} + 2$ for $s_0 > \max\{\mathrm{ord}(\mathcal{H}_{p_0}), 3/2\}$ by Lemma 5.4. So we have equalities

$$a_{\mu\nu}(\varrho) = \frac{1}{\mu!\nu!}\partial_z^\mu \partial_{\bar z}^\nu \mathbf{h}_{|(z,f,\varrho)=(0,0,\varrho)}\ , \quad |\mu+\nu| \le 2\mathbf{N}+1, \tag{137}$$

$$J^{-1}G_{\mu\nu}(\varrho) = \frac{1}{\mu!\nu!}\partial_z^\mu \partial_{\bar z}^\nu \nabla_f \mathbf{h}_{|(z,f,\varrho)=(0,0,\varrho)}\ , \quad |\mu+\nu| \le 2\mathbf{N}. \tag{138}$$

We consider now a yet unknown χ as in (111) with $M_0 = \ell$, $i = 0$, $M = m'$ and $r = k'$. Set $\phi := \phi^1$, where ϕ^t is the flow of Lemma 5.3. We are seeking χ such that $H^{(\ell)} \circ \phi$ satisfies the conclusions of Theorem 6.4 for $\ell + 1$.
We know that $H^{(\ell)} \circ \phi$ satisfies the conclusions of Lemma 5.4. Therefore, to prove the induction step, all we need to do is to check that the expansion of $H^{(\ell)} \circ \phi$ satisfies $\mathbf{R}_{-1} = 0$ and that the only terms in $\mathbf{R}_0$ and $\mathbf{R}_1$ of degree $\leq \ell + 1$ are in normal form. We have

$$\begin{aligned} H_2^{(\ell)} \circ \phi &= H_2^{(\ell)} + \int_0^1 \{H_2^{(\ell)}, \chi\}^{st} \circ \phi^t dt \\ &\quad + \int_0^1 (\partial_{\varrho_j} a_{\mu\nu} z^\mu \overline{z}^\nu \{\Pi_j(f), \chi\}) \circ \phi^t dt. \end{aligned} \tag{139}$$

By (130)–(131) we have for $\varrho = \Pi(f)$

$$\begin{aligned} \{H_2^{(\ell)}, \chi\}^{st} = -\mathrm{i} &\sum_{|\mu+\nu|=\ell+1} \mathbf{e}^{(\ell)}(\varrho) \cdot (\mu - \nu) z^\mu \overline{z}^\nu b_{\mu\nu}(\varrho) \\ &- \sum_{|\mu+\nu|=\ell} z^\mu \overline{z}^\nu \langle J^{-1}(\mathrm{i}\mathbf{e}^{(\ell)}(\varrho) \cdot (\mu - \nu) - \mathcal{H}) B_{\mu\nu}(\varrho), f \rangle \\ &+ \sum_{\substack{|\mu+\nu|=2 \\ (\mu,\nu)\neq(\delta_j,\delta_j)\ \forall\ j}} a_{\mu\nu}^{(\ell)}(\varrho) \Bigg[\sum_{|\mu'+\nu'|=\ell+1} \{z^\mu \overline{z}^\nu, z^{\mu'} \overline{z}^{\nu'}\} b_{\mu'\nu'}(\varrho) \\ &\qquad + \sum_{|\mu'+\nu'|=\ell} \{z^\mu \overline{z}^\nu, z^{\mu'} \overline{z}^{\nu'}\} \langle J^{-1} B_{\mu'\nu'}(\varrho), f \rangle \Bigg]. \end{aligned} \tag{140}$$

By Lemma 5.3 for $M_0 = \ell$, $i = 0$, $M = m'$ and $r = k'$ for first and last formula and by the proof of Lemma 3.8, in particular by (72), we have

$$\begin{aligned} &z \circ \phi^t = z + \mathcal{R}^{0,\ell}_{k'',m'}(t, \Pi(f), R)\,,\ \Pi(f) \circ \phi^t = \Pi(f) + \mathcal{R}^{0,\ell+1}_{k'',m'}(t, \Pi(f), R)\,, \\ &f \circ \phi^t = e^{J\mathcal{R}^{0,\ell+1}_{k'',m'}(t,\Pi(f),R)\cdot\diamond}(f + \mathbf{S}^{0,\ell}_{k'',m'}(t, \Pi(f), R)) \end{aligned} \tag{141}$$

for $k'' \leq k' - (m'+1)\mathbf{d}$. Then, substituting (141) in (140) we get, if $k \leq k'' - \mathrm{ord}(\mathcal{H}_{p_0})$, where $\mathrm{ord}(\mathcal{H}_{p_0}) \leq \max\{\mathrm{ord}(\mathcal{D}), \mathbf{d}\}$, for $1 \leq m \leq m'$ and exploiting that an $\mathcal{R}^{0,2\ell}_{k,m}$ is also an $\mathcal{R}^{0,\ell+2}_{k,m}$ for $\ell \geq 2$,

$$\int_0^1 \{H_2^{(\ell)}, \chi\}^{st} \circ \phi^t dt = \{H_2^{(\ell)}, \chi\}^{st} + \mathcal{R}^{0,\ell+2}_{k,m}(\Pi(f), R). \tag{142}$$

We have

$$\{\Pi_j(f), \chi\} = \sum_{k=1}^{n_0} \{\Pi_j(f), \Pi_k(f)\} \partial_{\Pi_k(f)} \chi + \sum_{|\mu'+\nu'|=\ell} z^{\mu'} \overline{z}^{\nu'} \langle P_c^*(p_0) \diamond_j f, B_{\mu'\nu'} \rangle.$$

We have, for $P_d(p_0) = 1 - P_c(p_0)$ the projection on the direct sum of $N_g(\mathcal{H}_{p_0})$ and the complement of $X_c(p_0)$ in (104), and using $JP_c^*(p_0) = P_c(p_0)J$ which follows from (15),

$$\begin{aligned} \{\Pi_i(f), \Pi_j(f)\} &= \langle P_c^*(p_0)\diamond_i f, JP_c^*(p_0)\diamond_j f\rangle \\ &= \langle \diamond_i f, P_d(p_0)J\diamond_j f\rangle = \mathcal{R}^{0,2}(f). \end{aligned} \tag{143}$$

Notice also that, for $B_{\mu\nu} \in \Sigma_{k'}$ independent of $\Pi(f)$ and for $|\mu+\nu| = \ell$, we have

$$\begin{aligned} \{\Pi_i(f), z^\mu \overline{z}^\nu \langle J^{-1}B_{\mu\nu}, f\rangle\} &= z^\mu \overline{z}^\nu \langle P_c^*(p_0)\diamond_i f, B_{\mu\nu}\rangle \\ &= z^\mu \overline{z}^\nu \langle f, \diamond_i B_{\mu\nu}\rangle - z^\mu \overline{z}^\nu \langle P_d^*(p_0)\diamond_i f, B_{\mu\nu}\rangle \\ &= \mathcal{R}^{0,\ell+1}_{k'-\mathbf{d},\infty}(R) + \mathcal{R}^{0,\ell+1}(R). \end{aligned} \tag{144}$$

By (143)–(144) we conclude that $\{\Pi_j(f), \chi\} = \mathcal{R}^{0,\ell+1}_{k'-\mathbf{d},m'}(\Pi(f), R)$. By (141) we get for $m \le m'$

$$\begin{gathered} \{\Pi_j(f), \chi\} \circ \phi^t = \mathcal{R}^{0,\ell+1}_{k'-\mathbf{d},m'}\left(\Pi(f) + \mathcal{R}^{0,\ell+1}_{k'',m'}(t, \Pi(f), R), S\right), \\ \text{for } S := e^{J\mathcal{R}^{0,\ell+1}_{k'',m'}(t,\Pi(f),R)\cdot\diamond}\left(R + \mathbf{S}^{0,\ell}_{k'',m'}(t, \Pi(f), R)\right). \end{gathered}$$

Then

$$\{\Pi_j(f), \chi\} \circ \phi^t = \mathcal{R}^{0,\ell+1}_{k''-m'\mathbf{d},m'}(t, \Pi(f), R). \tag{145}$$

By (141) and (145) the last term in (139) is $\mathcal{R}^{0,\ell+2}_{k,m}(\Pi(f), R)$ for $k \le k'' - m'\mathbf{d}$. This and (142) yield for $k = \min\{k' - (2m'+1)\mathbf{d}, k' - (m'+1)\mathbf{d} - \mathrm{ord}(\mathcal{H}_{p_0})\}$

$$H_2^{(\ell)} \circ \phi = H_2^{(\ell)} + \{H_2^{(\ell)}, \chi\}^{st} + \mathcal{R}^{0,\ell+2}_{\tilde{k},m}(\Pi(f), R). \tag{146}$$

A second observation is that $\mathbf{h} = (H^{(\ell)} \circ \phi)(z, f, \varrho)$ is $C^{2\mathbf{N}+2}$ in $\mathcal{P}^{s_0} = \{(\varrho, R)\}$ for $m \ge 2\mathbf{N}+2$. We can compute again the corresponding coefficients in (137)–(138). Because of (115), for $|\mu+\nu| \le \ell$ in (137) and for $|\mu+\nu| \le \ell-1$ in (138) these coefficients are the same of $\mathbf{h} = H^{(\ell)}(z, f, \varrho)$.

A third observation is that for $j = 3, 4$ we have for $\mathbf{k} = \mathbf{R}_j^{(\ell)} \circ \phi$

$$\begin{aligned} &\partial_z^\mu \partial_{\overline{z}}^\nu \mathbf{k}_{|(0,0,\varrho)} = 0 \text{ for } |\mu| + |\nu| \le \ell + 1 \\ &\partial_z^\mu \partial_{\overline{z}}^\nu \nabla_f \mathbf{k}_{|(0,0,\varrho)} = 0 \text{ for } |\mu| + |\nu| \le \ell. \end{aligned} \tag{147}$$

By Lemma 3.10 for $l = m$, $s = k$ and $r = k'$, we have for $k \le k' - (2m+1)\mathbf{d}$

$$\Pi_j(f) \circ \phi = \Pi_j(f) \circ \phi_0 + \mathcal{R}^{0,2\ell+1}_{k,m}(\Pi(f), R), \tag{148}$$

with $\phi_0 = \phi_0^1$ and ϕ_0^t the flow defined as in Lemma 3.10 using the field X_χ^{st}. Then we have

$$\Pi_j(f) \circ \phi_0 = \Pi_j(f) + \int_0^1 \left\langle \diamond_j (X_\chi^{st})_f(\Pi(f), R \circ \phi_0^t), f \circ \phi_0^t \right\rangle dt. \tag{149}$$

By the definition of X_χ^{st} and by formulas (141) for ϕ_0^t, which are simpler because there are no phase factors, by $|\mu + \nu| = \ell$ the integrand in (149) is

$$\begin{aligned} &\left(z + \mathcal{R}^{0,\ell}_{k'',m}(t, \Pi(f), R)\right)^\mu \left(\overline{z} + \mathcal{R}^{0,\ell}_{k'',m}(t, \Pi(f), R)\right)^\nu \\ &\quad \times \left\langle \diamond_j B_{\mu,\nu}(\Pi(f)), f + \mathbf{S}^{0,\ell}_{k'',m}(t, \Pi(f), R) \right\rangle \\ &\qquad = z^\mu \overline{z}^\nu \langle \diamond_j B_{\mu,\nu}(\Pi(f)), f \rangle + \mathcal{R}^{0,2\ell}_{k'',m}(t, \Pi(f), R). \end{aligned}$$

Then for $k \le k''$ we have

$$\Pi_j(f) \circ \phi_0 = \Pi_j(f) + \langle \diamond_j (X_\chi^{st})_f, f \rangle + \mathcal{R}^{0,2\ell}_{k,m}(\Pi(f), R). \tag{150}$$

By $\ell \ge 2$ we have $2\ell \ge \ell + 2$ and so $\mathcal{R}^{0,2\ell}_{k,m}$ is an $\mathcal{R}^{0,\ell+2}_{k,m}$.
By $\psi(\varrho) = O(|\varrho|^2)$ near 0, we conclude that

$$\psi(\Pi(f)) \circ \phi = \psi(\Pi(f)) + \widetilde{K}' + \mathcal{R}^{1,\ell+2}_{k,m}(\Pi(f), R), \tag{151}$$

with $\widetilde{K}'$ a polynomial as in (128) with $M_0 = \ell$, with $\widetilde{K}'(0, b, B) = 0$ and $(\widehat{k}, \widehat{m}) = (k', m')$ satisfying. Notice that it was to get the last equality, which follows from (150), that we introduced the flow ϕ_0^t.
We now focus on $\mathbf{R}_2$. We have by (141)

$$\begin{aligned} \mathbf{R}_2 \circ \phi &= \langle \mathbf{B}_2(\Pi(f')), (f')^2 \rangle \\ &= \Big\langle \mathbf{B}_2 \left(\Pi(f) + \mathcal{R}^{0,\ell+1}_{k,m}(\Pi(f), R)\right), \\ &\qquad \left(e^{J\mathcal{R}^{0,\ell+1}_{k'',m'}(\Pi(f),R)\cdot\diamond}(f + \mathbf{S}^{0,\ell}_{k'',m'}(\Pi(f), R))\right)^2 \Big\rangle. \end{aligned} \tag{152}$$

In our present set up the exponential $e^{J\mathcal{R}^{0,\ell+1}_{k'',m'}\cdot\diamond}$ cannot be moved to the $\mathbf{B}_2$ by a change of variables in the integral as in [10]. Fortunately we know already that $H^{(\ell)} \circ \phi$ has the expansion of Lemma 5.4 and that all we need to do is to compute some derivatives of $\mathbf{R}_2 \circ \phi$.

Using the expansion in (152) and formula (116), for $i=0$ now, we set

$$\begin{aligned}\mathfrak{R}_2 &:= \langle \mathbf{B}_2(\Pi(f)), (f+\mathbf{S}^{i,\ell}_{k'',m'}(\Pi(f),R))^2\rangle\\ &= \left\langle \mathbf{B}_2(\Pi(f)), \left[f+\int_0^1 (X^{st}_\chi)_f\circ\phi^t dt+\mathbf{S}^{i,2\ell+1}_{k'',m'}(\Pi(f),R)\right]^2\right\rangle\\ &= \langle \mathbf{B}_2(\Pi(f)), f^2\rangle + 2\int_0^1 \langle \mathbf{B}_2(\Pi(f)), (X^{st}_\chi)_f\circ\phi^t\ f\rangle dt + \mathcal{R}^{i,2\ell}_{k'',m'}(\Pi(f),R).\end{aligned} \tag{153}$$

We have that $\mathbf{k}=\mathbf{R}_2\circ\phi-\mathfrak{R}_2$ is $C^{\ell+1}$ and satisfies (147). Hence the analysis of $\mathbf{R}_2\circ\phi$ reduces to that of $\mathfrak{R}_2$. By (141), for $k\le k''$, $m\le m'-1$ and $\ell>1$ we have

$$\int_0^1 X^{st}_\chi\circ\phi^t dt = X^{st}_\chi+\mathbf{S}^{0,2\ell-1}_{k'',m'-1}(\Pi(f),R) = X^{st}_\chi+\mathbf{S}^{0,\ell+1}_{k,m}(\Pi(f),R). \tag{154}$$

This implies

$$\begin{aligned}&\mathfrak{R}_2 = \langle \mathbf{B}_2(\Pi(f)), f^2\rangle + \widetilde{K}'' + \mathcal{R}^{0,\ell+2}_{k,m}(\Pi(f),R)\quad ,\\ &\widetilde{K}'' := 2\langle \mathbf{B}_2(\Pi(f)), f(X^{st}_\chi)_f\rangle.\end{aligned} \tag{155}$$

Then $\widetilde{K}''$ is a polynomial like in (128) for the pair $(\widehat{k},\widehat{m})=(k',m')$ satisfying $\widetilde{K}''(0,b,B)=0$ by $B_2(\varrho)=0$ for $\varrho=0$.
By (141) and for the pullback of the term $\mathcal{R}^{1,2}_{k',m'+2}(\Pi(f),f)$ in Lemma 5.4 we have for $\varrho=\Pi(f)$

$$\begin{aligned}\mathcal{R}^{1,2}_{k',m'+2}(\Pi(f'),f') &= \mathcal{R}^{1,2}_{k',m'+2}(\varrho,f')\\ &\quad+\int_0^1 (\nabla_\varrho \mathcal{R}^{1,2}_{k',m'+2})(\varrho+t\mathcal{R}^{0,\ell+1}_{k'',m'+2}(\varrho,f),f')\cdot\mathcal{R}^{0,\ell+1}_{k'',m'}(\varrho,f)dt\\ &= \mathcal{R}^{1,2}_{k',m'+2}(\varrho,f')+\mathcal{R}^{0,\ell+3}_{k,m}(\varrho,R)\end{aligned} \tag{156}$$

for $k\le k''-m\mathbf{d}$ and $m\le m'$, by elementary analysis of the second line.
Applying again (141) we have

$$\begin{aligned}\mathcal{R}^{1,2}_{k',m'+2}(\varrho,f') &= \mathcal{R}^{1,2}_{k',m'+2}\left(\varrho, e^{J\mathcal{R}^{0,\ell+1}_{k'',m'}(\varrho,R)\cdot\diamond}\left(f+\mathbf{S}^{0,\ell}_{k'',m'}(\varrho,R)\right)\right)\\ &= \mathcal{R}^{1,2}_{k',m'+2}\left(\varrho, f+\mathbf{S}^{0,\ell}_{k'',m'}(\varrho,R)\right)+\mathcal{R}^{1,\ell+2}_{k,m}(\varrho,R)\end{aligned} \tag{157}$$

for $k\le k''-m\mathbf{d}$ and $m\le m'-1$. Next, by Lemma 5.3, (116) and by (154),

we have $\mathbf{S}^{0,\ell}_{k'',m'}(\varrho,R) = (X^{st}_\chi)_f + \mathbf{S}^{0,\ell+1}_{k'',m'-1}(\varrho,R)$ and

$$\begin{aligned}
&\mathcal{R}^{1,2}_{k',m'+2}\left(\varrho, f + (X^{st}_\chi)_f + \mathbf{S}^{0,\ell+1}_{k'',m}(\varrho,R)\right)\\
&\quad= \mathcal{R}^{1,2}_{k',m'+2}(\varrho,f) + \int_0^1 \Big\langle \nabla_R \mathcal{R}^{1,2}_{k',m'+2}\left(\varrho, f + t(X^{st}_\chi)_f + t\mathbf{S}^{0,\ell+1}_{k'',m}(\varrho,R)\right),\\
&\qquad\qquad (X^{st}_\chi)_f + \mathbf{S}^{0,\ell+1}_{k'',m}(\varrho,R)\Big\rangle\, dt\\
&\quad= \mathcal{R}^{1,2}_{k',m'+2}(\varrho,f) + \langle \nabla_f \mathcal{R}^{1,2}_{k',m'+2}(\varrho,f), (X^{st}_\chi)_f\rangle + \mathcal{R}^{1,\ell+2}_{k,m}(\varrho,R)
\end{aligned}$$

where we have used $\ell \geq 2$, $k \leq k'' \leq k'$ and $m \leq m'-1$. Notice that we have that $\mathcal{R}^{1,2}_{k',m'+2}(\varrho,f)$ is an $\mathcal{R}^{1,2}_{k,m+2}(\varrho,f)$. Finally we have

$$\begin{aligned}
&\langle \nabla_f \mathcal{R}^{1,2}_{k',m'+2}(\varrho,f), (X^{st}_\chi)_f\rangle = \widetilde{K}''' + \overline{\mathbf{R}}_2\,,\\
&\widetilde{K}''' := \langle \nabla^2_f \mathcal{R}^{1,2}_{k',m'+2}(\varrho,0)f, (X^{st}_\chi)_f\rangle,
\end{aligned} \tag{158}$$

with $\overline{\mathbf{R}}_2$ a term we can absorb in $\widehat{\mathbf{R}}_2$ and with $\widetilde{K}'''$ like in (128) for the pair $(\widehat{k},\widehat{m}) = (k',m')$ satisfying $\widetilde{K}'''(0,b,B) = 0$.
We set

$$\mathbf{R}^{(\ell)}_0 + \mathbf{R}^{(\ell)}_1 = Z' + K + \mathbf{R}_{01}\,, \tag{159}$$

where: Z' is the sum of the monomials in normal form of degree $\leq \ell+1$; K, which is like in (127), is the sum of the the monomials of degree equal to $\ell+1$ not in normal form; $\mathbf{R}_{01}$ is the sum of the monomials of degree $> \ell+1$. By induction there are no monomials not in normal form of degree $\leq \ell$ so that each of the monomials of the lhs of (159) go into exactly one of the three terms of the rhs.
We define Z'' and $\widetilde{K}$ by setting

$$\widetilde{K}' + \widetilde{K}'' + \widetilde{K}''' = Z'' + \widetilde{K}, \tag{160}$$

collecting in Z'' all monomials of the lhs in normal form (all of degree $\ell+1$) and in $\widetilde{K}$ all monomials of the lhs not in normal form. Here $\widetilde{K}$ is like in (128) for $(\widehat{k},\widehat{m}) = (k',m')$ with $\widetilde{K}(0,b,B) = 0$.
Applying Lemma 6.2 for $(\widehat{k},\widehat{m}) = (k',m')$ we can choose χ such that for $Z = Z' + Z''$ we have

$$\{H^{(\ell)}_2, \chi\}^{st} + Z + K + \widetilde{K} = 0. \tag{161}$$

Then $H^{(\ell+1)} := H^{(\ell)} \circ \phi$ satisfies the conclusions of Theorem 6.4 for $\ell+1$.

The step $\ell + 1 = 2$. Set $H^{(1)} = K \circ \mathfrak{F}_1$. We are seeking a transformation ϕ as in the previous part such that $H^{(2)} := H^{(1)} \circ \phi$ has term $\mathbf{R}^{(2)}_{-1} = 0$ in its expansion in Lemma 5.4. The argument is similar to the previous one, but this

time χ has degree $\ell+1$ with $\ell=1$. So the steps in the previous argument where we exploited $\ell\geq 2$ need to be reframed. We know that $H^{(1)}$ satisfies Lemma 5.4 for $L=1$ for some pair that we denote by (k',m') rather than (k,m).
The proof of (142) is different from the previous one. By (77) we have for some (k,m) appropriately smaller than (k',m')

$$\{H_2^{(1)},\chi\}^{st}\circ\phi^t=\{H_2^{(1)},\chi\}^{st}\circ\phi_0^t+\mathcal{R}^{0,4}_{k,m}(\Pi(f),R). \tag{162}$$

The following linear transformation

$$(Z,\overline{Z},F)\to\begin{pmatrix}\mathrm{i}\nu_j b_{\mu\nu}(\Pi(f))\frac{Z^\mu\overline{Z}^\nu}{\overline{Z}_j}+\mathrm{i}\nu_j\frac{Z^\mu\overline{Z}^\nu}{\overline{Z}_j}\langle J^{-1}B_{\mu\nu}(\Pi(f)),F\rangle\\ -\mathrm{i}\mu_j b_{\mu\nu}(\Pi(f))\frac{Z^\mu\overline{Z}^\nu}{Z_j}-\mathrm{i}\mu_j\frac{Z^\mu\overline{Z}^\nu}{Z_j}\langle J^{-1}B_{\mu\nu}(\Pi(f)),F\rangle\\ B_{\mu\nu}(\Pi(f))Z^\mu\overline{Z}^\nu\end{pmatrix}$$

depends linearly on $(b(\varrho),B(\rho))$, for $\varrho=\Pi(f)$. Then

$$z_j\circ\phi_0^t=z_j+a_j(t,b,B)\cdot z+b_j(t,b,B)\cdot\overline{z}+\sum_{\mu\nu}c_{j\mu\nu}(t,b,B)\langle J^{-1}B_{\mu\nu},f\rangle \tag{163}$$

for $a_j,b_j\in C^\infty([0,1]\times X_{k'},\mathbb{C}^{\mathbf{n}})$ with $|a_j|+|b_j|\leq C\|(b,B)\|_{X_{k'}}$ and $c_{j\mu\nu}\in C^\infty([0,1]\times X_{k'},\mathbb{C})$. Similarly

$$f\circ\phi_0^t=f+\mathbf{a}(t,b,B)\cdot z+\ \mathbf{b}(t,b,B)\cdot\overline{z}+\sum_{\mu\nu}\mathbf{c}_{\mu\nu}(t,b,B)\langle J^{-1}B_{\mu\nu},f\rangle \tag{164}$$

with $\mathbf{a},\mathbf{b}\in C^\infty([0,1]\times X_{k'},\Sigma^{\mathbf{n}}_{k'})$ with $\|\mathbf{a}\|_{\Sigma^{\mathbf{n}}_{k'}}+\|\mathbf{b}\|_{\Sigma^{\mathbf{n}}_{k'}}\leq C\|(b,B)\|_{X_{k'}}$ and $\mathbf{c}_{\mu\nu}\in C^\infty([0,1]\times X_{k'},\Sigma_{k'})$. These coefficients satisfy appropriate symmetries that ensure $\overline{f\circ\phi_0^t}=f\circ\phi_0^t$.
We have

$$\{H_2^{(1)},\chi\}^{st}\circ\phi_0^t=\{H_2^{(1)},\chi\}^{st}(\Pi(f),R\circ\phi_0^t)+\mathcal{R}^{1,4}_{k,m}(t,\Pi(f),R). \tag{165}$$

To compute $\{H_2^{(1)},\chi\}^{st}(\Pi(f),R\circ\phi_0^t)$ we replace the R in (140) with $R\circ\phi_0^t$. The coordinates of $R\circ\phi_0^t$ can be expressed in terms of R by (163)–(164). When we substitute (z,f) in (140) using (163)–(164), by an elementary computation we obtain

$$\begin{aligned}\{H_2^{(1)},\chi\}^{st}(\varrho,R\circ\phi_0^t)&=\{H_2^{(1)},\chi\}^{st}(\varrho,R)\\ &+\sum_{\substack{|\mu+\nu|=1\\ |\mu'+\nu'|=1}}a^{\mu'\nu'}_{\mu\nu}(t,\varrho,b(\varrho),B(\varrho))z^\mu\overline{z}^\nu\langle\mathcal{H}B_{\mu\nu}(\varrho),f\rangle+A^t+\underline{\mathbf{R}}^t.\end{aligned}$$

Here:

- $a_{\mu\nu}^{\mu'\nu'}(t,\varrho,b,B) \in C^{m'}$ with $a_{\mu\nu}^{\mu'\nu'}(t,0,0,0) = 0$;
- we have

$$A^t = \sum_{|\mu+\nu|=2} \alpha_{\mu\nu}(t,\varrho,b(\varrho),B(\varrho)) z^\mu \overline{z}^\nu + \sum_{l=0}^{1} \sum_{j=1}^{n_0} \sum_{|\mu+\nu|=1} z^\mu \overline{z}^\nu \langle \diamond_j^l A_{\mu\nu}^l(t,\varrho,b(\varrho),B(\varrho)), f\rangle,$$

$\alpha_{\mu\nu}(t,\varrho,b,B)$ and $A_{\mu\nu}^l(t,\varrho,b,B)$ are $C^{m'}$ with for $i=2$

$$|\alpha_{\mu\nu}(t,\varrho,b,B)| + \|A_{\mu\nu}^l(t,\varrho,b,B)\|_{\Sigma_{k'}} \le C\|(b,B)\|_{X_{k'}}^i; \tag{166}$$

- $\underline{\mathbf{R}}^t(\varrho,z,f)$ is C^m in $(t,\varrho,z,f) \in \mathbb{R}^{n_0+1} \times \mathbb{C}^{\mathbf{n}} \times \Sigma_{-k}$ with (ϱ,z,f) near $(0,0,0)$, with for $i=2$

$$|\underline{\mathbf{R}}^t| \le C\|(b,B)\|_{X_{k'}}^2 \|f\|_{\Sigma_{-k}}^2. \tag{167}$$

Then, in the notation of Lemma 6.3

$$\int_0^1 \{H_2^{(1)},\chi\}^{st} \circ \phi_0^t dt = \{H_2^{(1)},\chi\}^{\widetilde{st}} + A + \underline{\mathbf{R}} + \mathcal{R}_{k,m}^{1,4}(\Pi(R),R), \tag{168}$$

with $A = \int_0^1 A^t dt$ and $\underline{\mathbf{R}} = \int_0^1 \underline{\mathbf{R}}^t dt$ are like A^1 and $\underline{\mathbf{R}}^1$. Then, using also (162), we get the following analogue of (146):

$$H_2^{(1)} \circ \phi = H_2^{(1)} + \{H_2^{(1)},\chi\}^{\widetilde{st}} + A + \underline{\mathbf{R}} + \mathcal{R}_{k,m}^{0,4}(\Pi(f),R). \tag{169}$$

(148) remains true also for $\ell = 1$. We consider (149) and expand

$$\langle \diamond_j (X_\chi^{st})_f(\Pi(f), R\circ\phi_0^t), f\circ\phi_0^t\rangle = \langle \diamond_j (X_\chi^{st})_f(\Pi(f),R), f\rangle + A^t + \mathbf{R}^t,$$

with A^t and $\mathbf{R}^t$ like the previous ones but such that (166)–(167) hold for $i=1$. This yields

$$\Pi_j(f)\circ\phi_0 = \Pi_j(f) + A' + \underline{\mathbf{R}}'. \tag{170}$$

Here $\underline{\mathbf{R}}'$ is like $\underline{\mathbf{R}}^1$ such that (167) holds for $i=1$. A' is like A^1 such that (166) holds for $i=1$.
By $\psi(\varrho) = O(|\varrho|^2)$ near 0 and (148) we get the first equality in

$$\begin{aligned} \psi(\Pi(f))\circ\phi &= \psi(\Pi(f))\circ\phi_0 + \mathcal{R}_{k,m}^{1,3}(\Pi(f),R) \\ &= \psi(\Pi(f)) + \widetilde{K}' + \mathcal{R}_{k',m'}^{1,2}(\Pi(f),f) + \mathcal{R}_{k,m}^{1,3}(\Pi(f),R), \end{aligned} \tag{171}$$

where $\widetilde{K}' = \mathcal{R}^{1,2}_{k',m'}(\Pi(f), R)$ is a polynomial in R as in (128) with $\widetilde{K}'(0, b, B) = 0$. The second line in (171) follows by $\psi(\varrho) = O(|\varrho|^2)$, by the fact that $\psi(\varrho)$ is smooth and by (170). Notice that by choosing $m \leq m' - 2$ we have $\mathcal{R}^{1,2}_{k',m'}(\Pi(f), f) = \mathcal{R}^{1,2}_{k,m+2}(\Pi(f), f)$.
The discussion of $\mathbf{R} \circ \phi$ is similar to the previous one after (152) . This time, though, by (77) we write

$$\int_0^1 X^{st}_\chi \circ \phi^t dt = \int_0^1 X^{st}_\chi \circ \phi^t_0 dt + \mathbf{S}^{0,3}_{k,m}(\Pi(f), R). \tag{172}$$

By (163)–(164) we get

$$\int_0^1 X^{st}_\chi \circ \phi^t_0 dt = X^{st}_\chi + \mathbf{A} \text{ in } \mathcal{P}^{k'}, \tag{173}$$

with $(z, f) \to \mathbf{A}(\varrho, z, f)$ linear, with $C^{m'}$ dependence in ϱ and with

$$\|\mathbf{A}(\varrho, z, f)\|_{\mathcal{P}^{k'}} \leq C\|(b(\varrho), B(\varrho))\|_{X_{k'}}(|z| + \|f\|_{\Sigma_{-k'}}). \tag{174}$$

This yields, for $\mathfrak{R}_2$ defined as in (153),

$$\begin{aligned}\mathfrak{R}_2 &= \left\langle \mathbf{B}_2(\Pi(f)), \left[f + \int_0^1 (X^{st}_\chi)_f \circ \phi^t_0 dt\right]^2 \right\rangle + \mathcal{R}^{1,3}_{k,m}(\Pi(f), R) \\ &= \langle \mathbf{B}_2(\Pi(f)), f^2\rangle + 2\langle \mathbf{B}_2(\Pi(f)), f\mathbf{A}\rangle + \langle \mathbf{B}_2(\Pi(f)), \mathbf{A}^2\rangle + \mathcal{R}^{1,3}_{k,m}(\Pi(f), R),\end{aligned}$$

where we have used $\mathbf{B}_2(0) = 0$ for the reminder.
We have

$$2\langle \mathbf{B}_2(\Pi(f)), f\mathbf{A}\rangle + \langle \mathbf{B}_2(\Pi(f)), \mathbf{A}^2\rangle = \widetilde{K}'' + \underline{\mathbf{R}}'',$$

with $\underline{\mathbf{R}}''$ like $\underline{\mathbf{R}}$ and with $\widetilde{K}''$ like (128) with $\widetilde{K}''(0, b, B) = 0$, by $\mathbf{B}_2(0) = 0$, and with $(\widehat{k}, \widehat{m}) = (k', m')$. Summing up, we have

$$\mathfrak{R}_2 = \langle \mathbf{B}_2(\Pi(f)), f^2\rangle + \widetilde{K}'' + \underline{\mathbf{R}}'' + \mathcal{R}^{1,3}_{k,m}(\Pi(f), R). \tag{175}$$

Notice that the reduction of $\mathbf{R}_2 \circ \phi$ to $\mathfrak{R}_2$ continues to hold also for $\ell = 1$.
We consider $\mathcal{R}^{1,2}_{k',m'+2} \circ \phi$ from the $\mathcal{R}^{1,2}_{k',m'+2}$ term in the expansion of $\mathbf{R}$ in Lemma 5.4. Then, by (156) and by (172)–(173), for $\varrho = \Pi(f)$ we have

$$\mathcal{R}^{1,2}_{k',m'+2}(\Pi(f'), f') = \mathcal{R}^{1,2}_{k',m'+2}(\varrho, f + (X^{st}_\chi)_f + \mathbf{A} + \mathbf{S}^{0,3}_{k,m}) + \mathcal{R}^{0,4}_{k,m}(\varrho, R).$$

The first term in the rhs can be expanded for $\varrho = \Pi(f)$ as

$$\mathcal{R}^{1,2}_{k',m'+2}(\varrho, f + (X^{st}_\chi)_f + \mathbf{A}) + \mathcal{R}^{1,4}_{k,m}(\varrho, R).$$

We have for $\varrho = \Pi(f)$

$$\mathcal{R}^{1,2}_{k',m'+2}(\varrho, f + (X^{st}_\chi)_f + \mathbf{A}) = \mathfrak{B}_2(\varrho)(f + (X^{st}_\chi)_f + \mathbf{A})^2 + \mathcal{R}^{1,3}_{k,m}(\varrho, R),$$

with $\mathfrak{B}_2(\varrho)$ a $C^{m'}$ function with values in $B^2(\Sigma_{-k'}, \Sigma_{k'})$ with $\mathfrak{B}_2(0) = 0$. Considering the binomial expansion we get for $\varrho = \Pi(f)$

$$\mathcal{R}^{1,2}_{k',m'+2}(\Pi(f'), f') = \mathfrak{B}_2(\varrho) f^2 + \widetilde{K}''' + \underline{\mathbf{R}}''' + \mathcal{R}^{0,3}_{k,m}(\varrho, R),$$

with $\underline{\mathbf{R}}'''$ like $\underline{\mathbf{R}}$ and with $\widetilde{K}'''$ like (128) with $\widetilde{K}'''(0, b, B) = 0$ and $(\widehat{k}, \widehat{m}) = (k', m')$.
We now set $K = \mathbf{R}^{(1)}_{-1}$ and with the A of (168) we write

$$\widetilde{K}' + \widetilde{K}'' + \widetilde{K}''' + A = Z'' + \widetilde{K}, \tag{176}$$

where in Z'' we collect the null terms of the lhs and in $\widetilde{K}$ the other terms. Now we have $K(0) = 0$, $\widetilde{K}(0, 0, 0) = 0$ and $\nabla_{b,B}\widetilde{K}(0, 0, 0) = 0$. By Lemma 6.3 for $(\widehat{k}, \widehat{m}) = (k', m')$ we can choose χ such that for we have

$$\{H^{(\ell)}_2, \chi\}^{\widetilde{st}} + Z'' + K + \widetilde{K} = 0. \tag{177}$$

Then $H^{(2)} := H^{(1)} \circ \phi$ satisfies the conclusions of Theorem 6.4 for $\ell = 2$. □

Summing up, we have proved the following result, whose proof we sketch now.

THEOREM 6.5. *For fixed $p_0 \in \mathcal{O}$ and for sufficiently large $l \in \mathbb{N}$, there are a fixed $k \in \mathbb{N}$, an $\epsilon > 0$, an $1 \ll s' \ll l$ and a $1 \ll k \ll k'$ such that for solutions $\widehat{U}(t)$ to (3) with $\Pi(U) = p_0$ with $|\Pi(\widehat{R}(t))| + \|\widehat{R}(t)\|_{\Sigma_{-k}} < \epsilon$ and $\widehat{R}(t) \in \Sigma_l$, there exists a C^0 map $\Phi : \mathcal{U}^l_{\epsilon,k} \to \mathcal{U}^{s'}_{\epsilon',k'}$ such that*

$$R := \Phi_R(\Pi(\widehat{R}), \widehat{R}) = e^{Jq(\Pi(\widehat{R}),\widehat{R})\cdot\diamond}(\widehat{R} + \boldsymbol{S}(\Pi(\widehat{R}), \widehat{R})), \tag{178}$$

$$\begin{aligned} &\text{with } \boldsymbol{S} \in C^2((-2, 2) \times B_{\mathbb{R}^{n_0}} \times B_{\Sigma_{-k}}, \Sigma_{s'}) \\ &\qquad q \in C^2((-2, 2) \times B_{\mathbb{R}^{n_0}} \times B_{\Sigma_{-k}}, \mathbb{R}^{n_0}) \end{aligned} \tag{179}$$

such that $\|\boldsymbol{S}(\Pi(\widehat{R}), \widehat{R})\|_{\Sigma_{s'}} \leq C\epsilon\|\widehat{R}\|_{\Sigma_{-k}}$ and such that splitting $R(t)$ in spectral coordinates $(z(t), f(t))$ the latter satisfy

$$\dot{z}_j = \mathrm{i}\partial_{\overline{z}_j} H \ , \quad \dot{f} = J\nabla_f H \tag{180}$$

where H is a given function satisfying the properties of $H^{(2\mathbf{N}+1)}$ in Theorem 6.4.

Proof. Since in Lemma 3.7 we can pick arbitrary n, we see by the proof of Theorem 6.4 that we can suppose that the $2\mathbf{N} + 1$ transformations ϕ_ℓ are defined by flows (55) with pair (r, M) with r and M as large as needed.
Starting with an appropriate $\mathcal{U}^s_{\varepsilon_0,\kappa_0}$, we know that there is a map $\mathfrak{F} : \mathcal{U}^{s'}_{\varepsilon_1,\kappa'} \to \mathcal{U}^s_{\varepsilon_0,\kappa_0}$ as regular as needed which satisfies the conclusions of Theorem 6.4. In

particular here we have $s' \gg s$ and $1 \ll \kappa' \ll \kappa_0$ and in $\mathcal{U}^{s'}_{\varepsilon_1,\kappa'}$ we get the system (180) by pulling back the system which exists in $\mathcal{U}^{s}_{\varepsilon_0,\kappa_0}$.
We choose now $l \gg s'$, $1 \ll k \ll \kappa'$ and sufficiently small ϵ and δ with $\mathcal{U}^{l}_{\delta,k} \subset \mathcal{U}^{s}_{\varepsilon_0,\kappa_0}$ and $\mathcal{U}^{l}_{\epsilon,k} \subset \mathcal{U}^{s'}_{\varepsilon_1,\kappa'}$. Here l and κ' can be as large as we want, thanks to our freedom to choose (r, M).
By choosing δ small we can assume $\mathcal{U}^{l}_{\delta,k} \subset \mathfrak{F}(\mathcal{U}^{s'}_{\varepsilon_1,\kappa'})$. This follows from (63) which implies $\mathfrak{F}^{-1}(\mathcal{U}^{l}_{\delta,k}) \subset \mathcal{U}^{l}_{\epsilon,k}$. Finally we set $\Phi = \mathfrak{F}^{-1}$ where $\mathfrak{F}^{-1} : \mathcal{U}^{l}_{\delta,k} \to \mathcal{U}^{s'}_{\varepsilon_1,\kappa'}$.
Formula (178) and the information on $\mathbf{S}$ has been proved in the course of the proof of Lemma 4.1. The information on the phase function q can be proved by a similar induction argument, which we skip here. □

REMARK 6.6. *The paper [2] highlights in the Introduction and states in Theorem 2.2, that it is able to treat all solutions of the NLS near ground states in* H^1. *But in fact, in [2] there is no explicit proof of this. While [2] does not state the regularity properties of the maps in [2, Theorems 3.21 and 5.2], from the context they appear to be just continuous. Even if we assume that they are almost smooth transformations (but see Remark 2.10 above), nonetheless an explanation is required on why they preserve the structure needed to make sense of the NLS. But while pullbacks of the Hamiltonian are analyzed, the question on how in [2] it is possible to pullback differential forms with maps which are continuous but non differentiable, is left unexplained in [2]. So, for example, in the statement of [2, Theorem 3.21] it is claimed that* $\mathfrak{F}^*\Omega = \Omega_0$. *It is then stated that this means that in the coordinates* ϕ' *the differential form* Ω *is* Ω_0. *The meaning of this statement is unclear though, since the chart of* ϕ' *is not differentiable and differential forms are not topological invariants. The proof of [2, Theorem 3.21] does not clarify this point since formulas such as [2, (3.42)], i.e.* (79) *here, are treated on a purely formal basis, leaving unexplained basic things such as, for example, the meaning of* $\mathfrak{F}^{t*}\Omega_t$.

REMARK 6.7. *In the 2nd version of [2] there is an incorrect effective Hamiltonian. If we use the correct definition of the symbols* $\mathbf{S}^{i,j}$ *which we give above, the functions* $\Phi_{\mu\nu}$ *used in the normal form expansion in [2] are in* $\mathcal{W}^j$ *for some large* j, *rather than in* $\cap_{j\geq 0}\mathcal{W}^j$. *In pp. 25–27 in the 2nd version of [2], the* $\mathcal{W}^j$*'s are defined using the classical pair of operators* $L_\pm$, *see [14], and are closed subspaces of* $H^{j-1}(\mathbb{R}^3)$ *of finite codimension. This last fact seems to be unnoticed in [2] and leads to the breakdown of the proof in the 2nd version of [2], as we explain below. The space* $\mathcal{W}^2$, *for example, is defined by first considering* $\langle L_+u, u\rangle$ *for* $u \in \ker^\perp L_- \cap \ker^\perp L_+ \subset L^2$. *Notice that* $\langle L_+u, u\rangle \geq 0$, *see [14, Proposition 2.7] or [11, Lemma 11.12]. Proceeding like in [11, Lemma 11.13] it can be shown that for* $u \in \ker^\perp L_- \cap \ker^\perp L_+ \subset L^2$ *with* $u \neq 0$ *we have* $\|u\|_L^2 := \langle L_+u, u\rangle > 0$. *Then consider the completion of* $\ker^\perp L_- \cap \ker^\perp L_+ \cap C_0^\infty$ *by the norm* $\|u\|_L$. *This completion is exactly* $\ker^\perp L_- \cap \ker^\perp L_+ \cap H^1(\mathbb{R}^3)$.

Then $\mathcal{W}^2$ is a closed subspace of finite codimension of the latter space. Specifically, $\mathcal{W}^2$ is in the continuous spectrum part in the spectral decomposition of the operator L_-L_+, which is selfadjoint for $\langle u,v\rangle_L := \langle L_-^{-1}u,v\rangle$ in $\ker^\perp L_-$. Notice that, under hypotheses analogous to (L1)–(L6) in Section 5, L_-L_+ has finitely many eigenvalues and its eigenfunctions are Schwartz functions. Likewise, also the other $\mathcal{W}^j$'s are closed subspaces of $H^{j-1}(\mathbb{R}^3)$ of finite codimension. Later in the 2nd version of [2], at p.41, the Strichartz estimates hinge on the false inclusion of $\mathcal{W}^j$, or of $\mathcal{W}^\infty$, in $L^{\frac{6}{5}}(\mathbb{R}^3,\mathbb{C})$. Additional mistakes appear in the justification of the Fermi Golden rule. While formulas $R^\pm_{L_0}(\rho)\Phi$ in (St.2)–(St.3) on p. 38 of the 2nd version make sense because $\Phi\in H^{k,s}$ for $s>0$ appropriate, analogous formulas $R^\pm_B(\rho)\Phi$ in (6.50) and elsewhere in Section 6.2, are undefined when we know only that $\Phi\in\mathcal{W}^\infty$. In fact even $R^\pm_{-\Delta}(\rho)\Phi$ is undefined for $\rho\geq 0$ for such Φ's. So in particular, in the 2nd version of [2], the discussion of the Fermi Golden rule is purely formal. The above ones are not simple oversights. Rather, they stem from the fact that, in the 2nd version of [2], the homological equations are solved only in these $\mathcal{W}^j$'s, while it is unclear if they can be solved in spaces with spacial weights like the $H^{k,n}$ or the Σ_n for $n>0$, as we remarked in an early version of [10]. The 3rd version of [2] credits our remark for having stimulated changes in this part of the paper. These changes are classified in the 3rd version of [2] as mere simplifications, possibly leaving the wrong impression that the proof in the 2nd version of [2], while more complicated than in the 3rd version, is still correct.

7. The NLS and the Nonlinear Dirac Equation

We give a sketchy discussion of few examples.

The Nonlinear Schrödinger equation. We consider the equation

$$\mathrm{i}U_t = -\Delta U + 2B'(|U|^2)U\ .$$

Here $N=1$, $\mathcal{D}=-\Delta$, $|\ |_1=|\ |$, $J=\begin{pmatrix}0&1\\-1&0\end{pmatrix}$. There are four invariants:

$$Q(U)=\Pi_4(U)=\frac{1}{2}\langle U,U\rangle \text{ and } \Pi_j(U)=\frac{1}{2}\langle U,J\frac{\partial}{\partial x_j}U\rangle \text{ for } j\leq 3.$$

For fixed $v\in\mathbb{R}^3$ we have

$$Q(e^{-\frac{1}{2}Jv\cdot x}U)=Q(U)\ ,\ \Pi_j(e^{-\frac{1}{2}Jv\cdot x}U)=\Pi_j(U)-\frac{v_j}{2}Q(U) \text{ for } j\leq 3 \text{ and}$$

$$E(e^{-\frac{1}{2}Jv\cdot x}U)=E(U)-\sum_{j=1}^3 v_j\Pi_j(U)+\frac{v^2}{2}Q(U).$$

There is well established theory guaranteeing under appropriate hypotheses existence of open sets $\mathcal{O} \subseteq \mathbb{R}^+$ and $(\phi_\omega, 0) \in C^\infty(\mathcal{O}, \mathcal{S}(\mathbb{R}^3, \mathbb{R}^2))$ such that

$$\Delta\phi_\omega - \omega\phi_\omega + 2B'(\phi_\omega^2)\phi_\omega = 0 \quad \text{for } x \in \mathbb{R}^3.$$

More precisely it is possible to prove exponential decay to 0 of $\phi_\omega(x)$ as $x \to \infty$. For $v \in \mathbb{R}^3$ arbitrary we get $\Phi_p(x) = e^{-\frac{1}{2}Jv\cdot x}(\phi_\omega(x), 0)$ where $p_4 = \Pi_4(\phi_\omega)$ and $p_j = -\frac{1}{2}v_j p_4$ for $j \leq 3$. We have $\lambda_4(p) = -\omega - \frac{v^2}{4}$ and $\lambda_j(p) = -v_j$ for $j \leq 3$. Notice that for $\frac{d}{d\omega}Q(\phi_\omega) \neq 0$ this yields (7). Notice that

$$\nabla^2 E(e^{-\frac{1}{2}Jv\cdot x}U) = e^{-\frac{1}{2}Jv\cdot x}\left(\nabla^2 E(U) - Jv\cdot\nabla_x + \frac{v^2}{4}\right)e^{\frac{1}{2}Jv\cdot x}$$

and that $v\cdot\nabla_x \circ e^{-\frac{1}{2}Jv\cdot x} = e^{-\frac{1}{2}Jv\cdot x} \circ (v\cdot\nabla_x - J\frac{v^2}{2})$ and

$$\begin{aligned}\nabla^2 E(\Phi_p(x)) - \lambda(p)\cdot\diamond = e^{-\frac{1}{2}Jv\cdot x}&\left(\nabla^2 E((\phi_\omega,0)) - Jv\cdot\nabla_x + \frac{v^2}{4}\right)e^{\frac{1}{2}Jv\cdot x}\\ &+ Jv\cdot\nabla_x e^{-\frac{1}{2}Jv\cdot x}e^{\frac{1}{2}Jv\cdot x} + \left(\omega + \frac{v^2}{4}\right)e^{-\frac{1}{2}Jv\cdot x}e^{\frac{1}{2}Jv\cdot x}.\end{aligned}$$

They imply

$$\mathcal{H}_p = e^{-\frac{1}{2}Jv\cdot x}\mathcal{H}_\omega e^{\frac{1}{2}Jv\cdot x}, \quad \mathcal{H}_\omega := J(\nabla^2 E((\phi_\omega,0)) + \omega). \tag{181}$$

The multiplier operator $e^{-\frac{1}{2}Jv\cdot x}$ is an isomorphism in all spaces Σ_n so all the information on the spectrum of $\mathcal{H}_p$ is obtained from the spectrum of $\mathcal{H}_\omega$. We have $\mathcal{H}_\omega = \mathcal{H}_{0\omega} + V$ where $H_{0\omega} := J(-\Delta + \omega)$ and

$$V := 4J\begin{pmatrix} -B'(\phi_\omega^2) - 2B''(\phi_\omega^2)\phi_\omega^2 & 0 \\ 0 & -B'(\phi_\omega^2)\end{pmatrix}.$$

This yields $\sigma_e(\mathcal{H}_\omega) = \sigma(H_{0\omega}) = (-\infty, -\omega] \cup [\omega, \infty)$ and that $\sigma_p(\mathcal{H}_\omega)$ is finite with finite multiplicities. The fact that $\sigma_p(\mathcal{H}_\omega)$ is in the complement of $\sigma_e(\mathcal{H}_\omega)$ is expected to be true generically. Set $\mathcal{H} = \mathcal{H}_\omega P_c(\omega)$ for $P_c(\omega)$ the projection on $X_c(\mathcal{H}_\omega)$.

LEMMA 7.1. *The statement in (A5) is true.*

Proof. Notice that Σ_n is invariant by Fourier transform so that (4) is equivalent to the fact that for the following multiplier operator (that is an operator $\psi(x)$ which maps $u \to (\psi u)(x) := \psi(x)u(x)$) we have

$$\|(1 + \epsilon^2 + \epsilon^2|x|^2)^{-2}\|_{B(\Sigma_n,\Sigma_n)} \leq C_n < \infty \ \forall\ |\epsilon| \leq 1 \text{ and } n \in \mathbb{N}. \tag{182}$$

Similarly (5) is equivalent to

$$\begin{aligned} &strong - \lim_{\epsilon\to 0}(1+\epsilon^2+\epsilon^2|x|^2)^{-2} = 1 \text{ in } B(\Sigma_n,\Sigma_n) \qquad (183)\\ &\lim_{\epsilon\to 0}\|(1+\epsilon^2+\epsilon^2|x|^2)^{-2}-1\|_{B(\Sigma_n,\Sigma_{n'})} = 0 \quad \text{for any } n'\in\mathbb{N} \text{ with } n'<n. \end{aligned}$$

Both (182)–(183) are elementary to check using the first definition of Σ_n in Section 2, computing commutators of the multiplier operators with ∂_x^α and computing elementary bounds on the derivatives of the multipliers. □

LEMMA 7.2. *The statement in (A6) is true.*

Proof. Using the Fourier transformation like in Lemma 7.1, (A6) is equivalent to the statement that for any $n\in\mathbb{N}$ and $c>0$ there a C s.t. the following multiplier operator satisfies

$$\|e^{(1+\epsilon^2+\epsilon^2|x|^2)^{-2}J(\tau_4-\sum_{j=1}^3 x_j\tau_j)}\|_{B(\Sigma_n,\Sigma_n)}\le C$$

for any $|\tau|\le c$ and any $|\epsilon|\le 1$. This too is elementary to check. □

LEMMA 7.3. *The statement in (L7) is true.*

Proof. From $\sigma(\mathcal{H})=\sigma_e(\mathcal{H}_\omega)$ we have $R_{\mathcal{H}}\in C^\omega(\rho(\mathcal{H}),B(L^2,L^2))$.
We have $R_{\mathcal{H}_{0\omega}}$ and $R_{\mathcal{H}_{0\omega}}\partial_{x_j}$ are in $C^\omega(\rho(\mathcal{H}),B(\Sigma_n,\Sigma_n))$ for any $n\in\mathbb{N}$. By conjugation by Fourier transform this is equivalent to the statement that for $z\in\rho(\mathcal{H}_{0\omega})$ and $i=0,1$, we have

$$\xi_j^i\begin{pmatrix}(|\xi|^2+\omega-z)^{-1} & 0\\ 0 & -(|\xi|^2+\omega+z)^{-1}\end{pmatrix}\in B(\Sigma_n,\Sigma_n).$$

This is elementary, using the first definition of Σ_n in Section 2.
We have for $i=0,1$

$$R_{\mathcal{H}}(z)\partial_{x_j}^i = R_{\mathcal{H}_{0\omega}}(z)P_c(\omega)\partial_{x_j}^i - R_{\mathcal{H}_{0\omega}}(z)VR_{\mathcal{H}}(z)\partial_{x_j}^i. \qquad (184)$$

From (184) we derive, for $\|\ \| = \|\ \|_{B(L^2,L^2)}$.

$$\|R_{\mathcal{H}}(z)\partial_{x_j}^i\|\le\|(1+R_{\mathcal{H}_{0\omega}}(z)V)^{-1}\|\|R_{\mathcal{H}_{0\omega}}(z)P_c(\omega)\partial_{x_j}^i\|, \qquad (185)$$

which yields the $n=0$ case.
From (184) we derive

$$\begin{aligned}\|R_{\mathcal{H}}(z)\partial_{x_j}^i\|_{B(\Sigma_n,\Sigma_n)} &\le C\|R_{\mathcal{H}_{0\omega}}(z)\partial_{x_j}^i\|_{B(\Sigma_n,\Sigma_n)}\\ &\quad + C\|R_{\mathcal{H}_{0\omega}}(z)\|_{B(\Sigma_n,\Sigma_n)}\|\langle x\rangle^n V\|_{W^{n,\infty}}\|R_{\mathcal{H}}(z)\partial_{x_j}^i\|_{B(H^n,H^n)}.\end{aligned}$$

The last factor is bounded. Indeed for $\mathbf{v}=R_{\mathcal{H}}(z)\partial_{x_j}^i\mathbf{u}$ we have

$$\partial_x^\alpha\mathbf{v} = R_{\mathcal{H}}(z)\partial_x^\alpha\partial_{x_j}^i\mathbf{u} + R_{\mathcal{H}}(z)[V,\partial_x^\alpha]\partial_{x_j}^i\mathbf{u}$$

and induction in n yields the desired bounds $\|\mathbf{v}\|_{H^n}\le C\|\mathbf{u}\|_{H^n}$. □

The Nonlinear Dirac Equation. Here the unknown U is $\mathbb{C}^4$-valued, u^* its complex conjugate and for $m > 0$

$$\mathrm{i}U_t - D_m U - Vu + 2B'(U \cdot \beta U^*)\beta U = 0 \tag{186}$$

where we assume for the moment $V = 0$ and where $D_m = -\mathrm{i}\sum_{j=1}^3 \alpha_j \partial_{x_j} + m\beta$, with for $j = 1, 2, 3$

$$\alpha_j = \begin{pmatrix} 0 & \sigma_j \\ \sigma_j & 0 \end{pmatrix}, \ \beta = \begin{pmatrix} I_{\mathbb{C}^2} & 0 \\ 0 & -I_{\mathbb{C}^2} \end{pmatrix},$$

$$\sigma_1 = \begin{pmatrix} 0 & 1 \\ 1 & 0 \end{pmatrix}, \sigma_2 = \begin{pmatrix} 0 & \mathrm{i} \\ -\mathrm{i} & 0 \end{pmatrix}, \sigma_3 = \begin{pmatrix} 1 & 0 \\ 0 & -1 \end{pmatrix}.$$

Notice that the symmetry group (186) is not Abelian. In [4] there is a symmetry restriction on the solutions considered, by looking only at functions such that for any $x \in \mathbb{R}^3$ we have $U(-x) = \beta U(x)$ and $U(-x_1, -x_2, x_3) = S_3 U(x_1, x_2, x_3)$ with $S_3 := \begin{pmatrix} \sigma_3 & 0 \\ 0 & \sigma_3 \end{pmatrix}$. We need to redefine the spaces Σ_n in the proof, introducing these symmetries. This does not affect the proof.
There is a unique invariant $Q(U) = \frac{1}{2}\|u\|_{L^2}$. In this case $\diamond_1 U = U$ for any u. Hence all the changes of variables are diffeomorphism within each space $\mathcal{P}^K$ (or $\widetilde{\mathcal{P}}^K$).
(A5)–(A6) in this case are elementary. In fact (A5) is unnecessary, (A6) is necessary only for $\epsilon = 0$, in which case is trivial. (L7) is necessary only for $i = 0$ (given that the only $\diamond_j$ is the identity) and can be proved in a way similar to Lemma 7.3.

Nonlinear Dirac Equation with a Potential. Pick $V \in \mathcal{S}(\mathbb{R}^3, B(\mathbb{C}^4))$ with $V(x)$ selfadjoint for the scalar product in $\mathbb{C}^4$ for any $x \in \mathbb{R}^3$. Then generically $\sigma_p(D_m + V) \subset (-m, m)$. Suppose $\sigma_p(D_m + V) = \{e_0, ..., e_{\mathbf{n}}\}$ with $e_0 < ... < e_{\mathbf{n}}$. Then bifurcation yields corresponding families of small standing waves $e^{-\mathrm{i}\omega t}\phi_\omega(x)$ of (186). For generic V the e_j have multiplicity 1. If we focus on e_0, for generic smooth $B'(r)$ there will be a smooth family $\omega \to \phi_\omega$ in $C^\infty(\mathcal{O}, \Sigma_n)$ for any n, with $\mathcal{O}$ an open interval one of whose endpoints is e_1. Then it can be shown that for generic V the hypotheses (L1)–(L6) in Section are true, as well as all the previous hypotheses. Indeed in this case, taking ω sufficiently close to e_0, we have eigenvalues with $\mathbf{e}'_j$ arbitrarily close to $e_j - e_0$. Generically this yields (L4)–(L5). The multiplicity of the $\mathrm{i}\mathbf{e}'_j$ is 1. We have $\sigma_e(\mathcal{H}_\omega) = (-\infty, -m + |\omega|] \cup [m - |\omega|, \infty)$. An eigenvalue λ of $\mathcal{H}_\omega$ is either $\lambda = 0$, or $\lambda = \pm\mathrm{i}\mathbf{e}'_j$ for some j. This in particular yields (L1)–(L3).

References

[1] R. Abraham, J. Marsden and T. Ratiu, *Manifolds, Tensor Analysis and Applications*, Springer, Berlin, 2000.

[2] D. Bambusi, *Asymptotic stability of ground states in some Hamiltonian PDEs with symmetry*, arXiv:1107.5835v3, version of the 24th February 2012.
[3] D. Bambusi and S. Cuccagna, *On dispersion of small energy solutions of the nonlinear Klein Gordon equation with a potential*, Amer. J. Math. **133** (2011), 1421–1468.
[4] N. Boussaid and S. Cuccagna, *On stability of standing waves of nonlinear Dirac equations*, Comm. Partial Differential Equations **37** (2012), 1001–1056.
[5] V. Buslaev and G. Perelman, *On the stability of solitary waves for nonlinear Schrödinger equations*, Nonlinear evolution equations, Amer. Math. Soc. Transl. Ser. 2 **164** (1995), 75–98.
[6] S. Cuccagna, *On asymptotic stability of ground states of NLS*, Rev. Math. Phys. **15** (2003), 877–903.
[7] S. Cuccagna, *On instability of excited states of the nonlinear Schrödinger equation*, Physica D **238** (2009), 38–54.
[8] S. Cuccagna, *The Hamiltonian structure of the nonlinear Schrödinger equation and the asymptotic stability of its ground states*, Comm. Math. Phys. **305** (2011), 279–331.
[9] S. Cuccagna, *On scattering of small energy solutions of non autonomous hamiltonian nonlinear Schrödinger equations*, J. Differential Equations **250** (2011), 2347–2371.
[10] S. Cuccagna, *On asymptotic stability of moving ground states of the nonlinear Schrödinger equation*, to appear Trans. Amer. Math. Soc.
[11] I. Rodnianski, W. Schlag and A. Soffer, *Asymptotic stability of N-soliton states of NLS*, (2003), arXiv:math/0309114v1.
[12] I. M. Sigal, *Nonlinear wave and Schrödinger equations. I. Instability of periodic and quasi-periodic solutions*, Comm. Math. Phys. **153** (1993), 297–320.
[13] A. Soffer and M. I. Weinstein, *Resonances, radiation damping and instability in Hamiltonian nonlinear wave equations*, Invent. Math. **136** (1999), 9–74.
[14] M. I. Weinstein, *Modulation stability of ground states of nonlinear Schrödinger equations*, SIAM J. Math. Anal. **16** (1985), 472–491.

Author's address:

Scipio Cuccagna
Department of Mathematics and Geosciences
University of Trieste
Via Valerio 12/1, Trieste, I-34127 Italy
E-mail: scuccagna@units.it

Received March 5, 2012
Revised September 21, 2012

Rend. Istit. Mat. Univ. Trieste
Volume 44 (2012), 259–284

Infinitely many radial solutions of a mean curvature equation in Lorentz-Minkowski space

Denis Bonheure, Colette De Coster
and Ann Derlet

To Fabio, with esteem and friendship

Abstract. *In this paper, we show that the quasilinear equation*

$$-\mathrm{div}\left(\frac{\nabla u}{\sqrt{1-|\nabla u|^2}}\right) = |u|^{\alpha-2}u, \quad \text{in } \mathbb{R}^N$$

has a positive smooth radial solution at least for any $\alpha > 2^\star = 2N/(N-2)$, $N \geq 3$. *Our approach is based on the study of the optimizers for the best constant in the inequality*

$$\int_{\mathbb{R}^N} (1-\sqrt{1-|\nabla u|^2}) \geq C\left(\int_{\mathbb{R}^N} |u|^\alpha\right)^{\frac{N}{\alpha+N}},$$

which holds true in the unit ball of $W^{1,\infty}(\mathbb{R}^N) \cap \mathcal{D}^{1;2}(\mathbb{R}^N)$ *if and only if* $\alpha \geq 2^\star$. *We also prove that the best constant is not achieved for* $\alpha = 2^\star$. *As a byproduct, our arguments combined with Lusternik-Schnirelmann category theory allow to construct a sequence of radial solutions.*

Keywords: Mean curvature equation in the Lorentz-Minkowski space, Lusternik-Schnirelmann category, multiplicity, super critical exponent
MS Classification 2010: 35J25,35J93, 58E05, 35A23, 35Q75

1. Introduction

It is well known [19] that the Lane-Emden equation

$$-\Delta u = |u|^{\alpha-2}u \ \text{ in } \mathbb{R}^N, \tag{1}$$

admits no nontrivial nonnegative solution for $2 < \alpha < 2^\star$, $N \geq 3$, while, for $\alpha = 2^\star$, any positive solution can be written in the form

$$u_{\delta,a}(x) = \beta_N \left(\frac{\delta}{\delta^2 + |x-a|^2} \right)^{\frac{N-2}{2N}},$$

as proved by Caffarelli, Gidas and Spruck [11]. For $\alpha > 2^\star$, the set of all positive radial solutions is a one-parameter family $\{u_a(r) = a u_1(a^{(\alpha-2)/2} r) : a > 0\}$, where u_1 is strictly decreasing in r (see for instance [20]). Non radial singular solutions have been constructed by Dancer, Guo and Wei [15]. We mention that it is still open whether all smooth positive solutions are radially symmetric around some point or not.

The prescribed mean curvature equation in Euclidian space

$$-\mathrm{div}\left(\frac{\nabla u}{\sqrt{1+|\nabla u|^2}} \right) = |u|^{\alpha-2}u \ \text{ in } \mathbb{R}^N,$$

has also been the object of many studies. It has been considered, among others, by Ni and Serrin [26] and del Pino and Guerra [17]. It is known that this problem has infinitely many radial positive solution if $\alpha \geq 2^\star$ and no smooth positive solutions if $\alpha \leq (2N-2)/(N-2)$. In contrast with the non-existence result for the Lane-Emden equation in the subcritical range, del Pino and Guerra proved the existence of many positive solutions when $\alpha = 2^\star - \epsilon$, for sufficiently small $\epsilon > 0$.

In this work, we aim to study the following prescribed mean curvature equation in the Lorentz-Minkowski space

$$Q(u) = |u|^{\alpha-2}u \ \text{ in } \mathbb{R}^N, \tag{2}$$

where

$$Q(u) = -\mathrm{div}\left(\frac{\nabla u}{\sqrt{1-|\nabla u|^2}} \right). \tag{3}$$

The quasilinear operator Q is a classical object in Riemannian geometry. The Lorentz-Minkowski space $\mathbb{L}^{N+1} = \{(x,t) \in \mathbb{R}^N \times \mathbb{R}\}$, with the flat metric $\sum_{j=1}^N (dx_j)^2 - (dt)^2$ is the natural framework of classical relativity. If M is an N-dimensional hypersurface of $\mathbb{L}^{N+1}$ that is the graph of a smooth function $u \in C^1(\Omega)$ with $\|\nabla u\|_{L^\infty} < 1$, the local mean curvature of M is given by $Q(u)$, see for instance [2, 12]. The determination of maximal or constant mean curvature hypersurfaces is an important issue in classical relativity. The volume integral $\int_\Omega \sqrt{1-|\nabla u|^2}$ gives the area integral in $\mathbb{L}^{N+1}$ and surfaces of maximal area (or simply maximal surfaces) solve the equation $Q(u) = 0$ in Ω.

For functions defined on the whole of $\mathbb{R}^N$, the operator Q is relevant in Maxwell-Born-Infeld field theory, see for instance [7, 8, 22, 23]. Basically, in this theory, which is fully relativistic, it is assumed that there is a maximal field strength. This lead Born and Infeld to consider the following Lagrangian density, expressed in Lorentz-Minkowski space,

$$\mathcal{L}_{BI} = b^2 \left(1 - \sqrt{1 - \frac{|\vec{E}|^2 - |\vec{B}|^2}{b^2} - \frac{(\vec{E}.\vec{B})^2}{b^4}} \right),$$

where $\vec{E}$ is the electric field, $\vec{B}$ is the magnetic field and b is the maximal admissible value of the electric field.

Up to our knowledge, the equation (2) has never been considered in the literature, at least in $\mathbb{R}^N$. We refer to [3, 4, 9, 14] for recent results on the existence of radial solutions for BVPs involving Q in the ball with either Dirichlet or Neumann conditions.

Supercritical problems are usually difficult to tackle through variational methods. For instance, concerning the Lane-Emden equation, Farina [18] has obtained a Liouville-type result for C^2 solutions of (1) with finite Morse index. Basically, if the dimension is small ($N \leq 10$), the only finite Morse index solution is 0 except at the critical exponent where the above-mentioned positive solutions arise as constrained minimizers on a manifold of codimension 1.

In contrast, we show here that the quasilinear equation (2) has a smooth positive radial solution for any $\alpha > 2^\star$, $N \geq 3$ by using simple arguments from Critical Point Theory and the Calculus of Variations. In fact, when $\alpha > 2^\star$, we have enough compactness to deal with the problem in a standard way. Indeed, we minimize the volume integral

$$\int_{\mathbb{R}^N} (1 - \sqrt{1 - |\nabla u|^2}), \tag{4}$$

truncated in a convenient way, constrained to the unit sphere of $L^\alpha(\mathbb{R}^N)$. Then we prove a gradient estimate which is uniform with respect to the truncation parameter.

Our first main result is the following.

THEOREM 1.1. *If $\alpha > 2^\star$, equation (2) has a positive radial classical solution.*

We restrict here our attention to the existence of *radially* symmetric solutions. On the one hand, we expect that all positive smooth solutions are indeed radially symmetric, though this is an open question. On the other hand, our solution arises as a constrained minimizer and its Schwarz symmetric rearrangement yields a radially symmetric minimizer (and therefore a radially symmetric solution).

Surprisingly, our approach to establish the existence of a solution of (2) fails in the critical case $\alpha = 2^\star$. Indeed, as stated in Theorem 1.2 below, the solution of Theorem 1.1 realizes the best constant in an inequality between the volume integral (4) and the L^α-norm. This inequality still holds for $\alpha = 2^\star$ but the best constant is not achieved. We emphasize that this contrasts with the Sobolev inequality.

In the sequel, we denote by $\mathcal{X}$ the functional space

$$\mathcal{X} := \Big\{ u \in \mathcal{D}^{1;2}(\mathbb{R}^N) : \nabla u \in L^\infty(\mathbb{R}^N) \text{ and } \|\nabla u\|_{L^\infty} \leq 1 \Big\},$$

endowed with the norm

$$\|u\|_{\mathcal{D}^{1;2}(\mathbb{R}^N)} := \left(\int_{\mathbb{R}^N} |\nabla u|^2 \right)^{1/2}.$$

We establish the following Sobolev-type inequality.

THEOREM 1.2. *There exists $C > 0$ such that*

$$\int_{\mathbb{R}^N} (1 - \sqrt{1 - |\nabla u|^2}) \geq C \left(\int_{\mathbb{R}^N} |u|^\alpha \right)^{\frac{N}{\alpha+N}} \tag{5}$$

for every $u \in \mathcal{X}$ if and only if $\alpha \geq 2^\star$. Moreover, the best constant

$$\inf_{u \in \mathcal{X} \setminus \{0\}} \frac{\int_{\mathbb{R}^N} (1 - \sqrt{1 - |\nabla u|^2})}{\left(\int_{\mathbb{R}^N} |u|^\alpha \right)^{\frac{N}{N+\alpha}}}$$

is achieved by a radial solution of (2) for $\alpha > 2^\star$ while it is not achieved for $\alpha = 2^\star$.

The fact that inequality (5) does not hold below the critical exponent is rather clear since the volume integral (4) is bounded from above by the Dirichlet energy. This does not mean that (2) has no non trivial nonnegative solutions for $\alpha < 2^\star$ though we conjecture that this is indeed the case. One can for instance exclude the existence of fast decaying solution but we are not able to prove a complete non-existence result for $\alpha < 2^\star$. Also the existence of a positive solution of (2) in the critical case $\alpha = 2^\star$ remains an interesting open question.

At last, as a natural extension of our existence result, we combine our previous approach with Lusternik-Schnirelmann category theory to obtain a sequence of solutions whose volume integral diverge. Namely we prove the following multiplicity result.

THEOREM 1.3. *For any $\alpha > 2^\star$, equation (2) has a sequence of radial solutions $(u^k)_{k\in\mathbb{N}}$ such that*

$$\int_{\mathbb{R}^N} (1 - \sqrt{1 - |\nabla u^k|^2}) \to +\infty \text{ as } k \to \infty.$$

Again, we first consider an auxiliary problem and conclude by a sharp uniform estimate on the gradient of our solutions. Note that we do not provide sign information on solutions though one could probably argue as in [27, 5] to obtain a sequence of sign changing solutions. We leave this, as well as the existence of infinitely many positive solutions, as open questions.

The paper is organized as follows. Section 2 contains some preliminary results on the functional spaces we will work with. In Section 3, we establish the existence of at least one classical solution of (2) (see Theorem 1.1 above). Section 4 is devoted to the proof of the inequality in Theorem 1.2 and especially to the existence of optimizers for the best constant in this inequality. Finally, in Section 5, we obtain infinitely many solutions of (2) as stated in Theorem 1.3.

With some abuse of notation, we will sometimes consider radial functions as functions of one variable, thus writing $u(|x|)$ or $u(x)$ or $u(r)$. For any set $\mathcal{A}$ of functions, $\mathcal{A}_{rad}$ is defined as the set of all radially symmetric functions of $\mathcal{A}$. Throughout the paper, C denotes a positive constant that can change from line to line.

2. Functional framework and preliminary results

Let us set $a_0(s) = (1-s)^{-1/2}$ for all $s < 1$. Equation (2) can be written as

$$-\operatorname{div}\left(a_0(|\nabla u|^2)\nabla u\right) = |u|^{\alpha-2}u \quad \text{in } \mathbb{R}^N.$$

We introduce the energy functional

$$I_0(u) := \frac{1}{2}\int_{\mathbb{R}^N} A_0(|\nabla u|^2),$$

where $A_0(t) = \int_0^t a_0(s)\,ds$ for all $t \le 1$. This functional is well defined on $\mathcal{X} = \{u \in \mathcal{D}^{1;2}(\mathbb{R}^N) : \nabla u \in L^\infty(\mathbb{R}^N) \text{ and } \|\nabla u\|_{L^\infty} \le 1\}$, because we have

$$\frac{1}{2}|\nabla u|^2 \le 1 - \sqrt{1-|\nabla u|^2} = \frac{|\nabla u|^2}{1+\sqrt{1-|\nabla u|^2}} \le |\nabla u|^2.$$

LEMMA 2.1. *Let $u \in \mathcal{X}$. Then $|\nabla u| \in L^q(\mathbb{R}^N)$ for every $q \ge 2$, and $u \in L^s(\mathbb{R}^N)$ for every $s \ge 2^\star$. Moreover, u can be assumed to be continuous and such that*

$$\lim_{|x|\to\infty} u(x) = 0.$$

Proof. Since $|\nabla u| \leq 1$ and $|\nabla u| \in L^2(\mathbb{R}^N)$, we infer that $|\nabla u| \in L^q(\mathbb{R}^N)$ for every $q \geq 2$. It then follows that $u \in L^{qN/(N-q)}(\mathbb{R}^N)$ for every $q \geq 2$, and, by interpolation, $u \in L^s(\mathbb{R}^N)$ for every $s \geq 2^\star$. Observe also that since $u \in W^{1,r}(\mathbb{R}^N)$ for some $r > N$, it can be assumed to be continuous and moreover $\lim_{|x|\to\infty} u(x) = 0$. □

Working with the functional I_0 in $\mathcal{X}$ requires some care. Since I_0 is weakly lower semi-continuous, a natural way to obtain a solution of (2) consists in minimizing I_0 constrained to the manifold

$$\mathcal{M}_0 := \left\{ u \in \mathcal{X} : \int_{\mathbb{R}^N} |u|^\alpha = 1 \right\}.$$

However, it is not clear that minimizers solve an associated Euler-Lagrange equation. Indeed, the functional I_0 is C^1 only at points $u \in \mathcal{X}$ with Lipschitz constant $Lip(u)$ strictly less than 1. Without this condition, minimizers solely solve a variational inequality.

To overcome this lack of differentiability on the boundary of $\mathcal{X}$, we will work with an auxiliary functional. This type of truncation argument has already been used in [13, 14] to deal with Dirichlet boundary condition in an interval or a ball. Here, one of the novelties is that an a priori L^∞ bound on minimizers cannot be derived from the solely boundedness of the gradient. Therefore, we truncate the volume integral in a different way than in [13, 14] and we deal with a different functional framework.

We now define our auxiliary functional. For $\theta \in]0,1[$, define $a_\theta : \mathbb{R} \to \mathbb{R}^+$ by

$$a_\theta(s) = a_0(s) \text{ for } 0 \leq s \leq 1-\theta \quad \text{and} \quad a_\theta(s) = \gamma s^p + \delta \text{ for } s > 1-\theta, \quad (6)$$

where γ and δ are chosen in such a way that a_θ is C^1. The exponent p will be chosen later according to the value of α in (2).

In the sequel, we will work with the spaces $\mathcal{D}^{1;r}_{rad}(\mathbb{R}^N)$ and $\mathcal{D}^{1;(2,q)}_{rad}(\mathbb{R}^N)$, defined respectively as the closure of the smooth compactly supported radially symmetric functions for the norms

$$\|u\|_{\mathcal{D}^{1;r}} := \left(\int_{\mathbb{R}^N} |\nabla u|^r \right)^{\frac{1}{r}}$$

and

$$\|u\|_{\mathcal{D}^{1;(2,q)}} := \left(\int_{\mathbb{R}^N} |\nabla u|^2 \right)^{\frac{1}{2}} + \left(\int_{\mathbb{R}^N} |\nabla u|^q \right)^{\frac{1}{q}},$$

with $1 < q, r < \infty$. Consider the manifold

$$\mathcal{M} := \left\{ u \in \mathcal{D}^{1;(2,q)}_{rad}(\mathbb{R}^N) : \int_{\mathbb{R}^N} |u|^\alpha = 1 \right\}.$$

We will look for critical points of I_θ constrained to $\mathcal{M}$ where

$$I_\theta : \mathcal{D}^{1;(2,q)}_{rad}(\mathbb{R}^N) \to \mathbb{R}^+$$

is defined by

$$I_\theta(u) := \frac{1}{2} \int_{\mathbb{R}^N} A_\theta(|\nabla u|^2),$$

and $A_\theta(t) = \int_0^t a_\theta(s)\, ds$.

We next recall some elementary facts. We quote them in separate lemmas for further references in the text. We do not provide the details for Lemma 2.2 which follows from standard arguments. We refer for instance to [28, 25, 6] for Lemma 2.3, whereas Lemma 2.4 can easily be deduced from [25, Corollary II-3]. Below, $q^\star := qN/(N-q)$ for $q < N$.

LEMMA 2.2. *Let $u \in \mathcal{D}^{1;(2,q)}_{rad}(\mathbb{R}^N)$. Then $u \in \mathcal{D}^{1;r}_{rad}(\mathbb{R}^N)$ for every $r \in [2, q]$. If $q < N$ then $u \in L^s(\mathbb{R}^N)$ for every $s \in [2^\star, q^\star]$; if $q = N$ then $u \in L^s(\mathbb{R}^N)$ for every $s \in [2^\star, +\infty[$; if $q > N$ then $u \in L^s(\mathbb{R}^N)$ for every $s \in [2^\star, +\infty]$. Moreover, the embeddings are continuous.*

LEMMA 2.3. *Let $r \in [2, q]$ if $q < N$, and $r \in [2, N[$ if $q \geq N$. Then there exists $C > 0$ (depending only on N and r) such that for all $u \in \mathcal{D}^{1;(2,q)}_{rad}(\mathbb{R}^N)$, there holds*

$$|u(x)| \leq C|x|^{-\frac{N-r}{r}} \|\nabla u\|_{L^r},$$

for almost all $x \in \mathbb{R}^N \setminus \{0\}$.

LEMMA 2.4. *Let $(u_n)_n \subset \mathcal{D}^{1;(2,q)}_{rad}(\mathbb{R}^N)$ be a bounded sequence. If $q < N$ then for any $s \in]2^\star, q^\star[$, there exists a subsequence which converges weakly in $\mathcal{D}^{1;(2,q)}_{rad}(\mathbb{R}^N)$ and strongly in $L^s(\mathbb{R}^N)$. If $q > N$, the same result holds for any $s \in]2^\star, +\infty[$.*

We close this section by a uniform estimate on the regularization schema. Observe that for $\theta_1 := 1/(2p+1)$, the function a_θ defined in (6) is given by

$$a_{\theta_1}(s) = 1/\sqrt{1-s} \text{ if } 0 \leq s \leq 1-\theta_1 \quad \text{and} \quad a_{\theta_1}(s) = \gamma_p s^p \text{ if } s > 1-\theta_1,$$

where $\gamma_p = \sqrt{2p+1}\,((2p+1)/2p)^p$. Therefore, for all $\theta \in]0, \theta_1]$ and $s \in \mathbb{R}^+$ we have

$$\frac{\gamma_p}{p+1} s^{p+1} \leq A_{\theta_1}(s) \leq A_\theta(s), \tag{7}$$

and

$$\begin{array}{rcll} A_\theta(s) \geq A_{\theta_1}(s) & \geq & s, & \text{if } s \leq \frac{2p}{2p+1}, \\ & \geq & \frac{\gamma_p}{p+1}\left(\frac{2p}{2p+1}\right)^p s, & \text{if } s > \frac{2p}{2p+1}. \end{array} \tag{8}$$

Inequalities (7) and (8) lead to uniform estimates (with respect to θ) in $\mathcal{D}^{1;(2,2p+2)}_{rad}(\mathbb{R}^N)$. They will be important keys in the sequel to obtain a priori bounds independent of the truncation parameter θ. As for an upper bound on I_θ, we observe that for all $u \in \mathcal{D}^{1;(2,2p+2)}(\mathbb{R}^N)$,

$$A_\theta(|\nabla u|^2) \leq C\,(|\nabla u|^{2p+2} + |\nabla u|^2). \tag{9}$$

for some constant C depending on θ. The functional I_θ is then well defined in $\mathcal{D}^{1;(2,q)}_{rad}(\mathbb{R}^N)$ with $q := 2p+2$ and it is straightforward that I_θ is C^1 on $\mathcal{D}^{1;(2,q)}_{rad}(\mathbb{R}^N)$.

The preceding lemmas suggest to choose p in the definition of a_θ such that $q = 2p+2$ satisfies $q^\star > \alpha$. Indeed, a lower bound in $\mathcal{D}^{1;(2,2p+2)}(\mathbb{R}^N)$ will follow from (7) and (8) whereas $\mathcal{M}$ is weakly closed as soon as $q^\star > \alpha$.

3. Existence of a positive solution for supercritical exponents

In this section, we prove that equation (2) has at least one positive solution.

3.1. The auxiliary problem

We will first look for a solution of the modified problem

$$-\mathrm{div}\left(a_\theta(|\nabla u_\theta|^2)\nabla u_\theta\right) = \lambda_\theta \alpha |u_\theta|^{\alpha-2}u_\theta \ \text{ in } \mathbb{R}^N,$$

where a_θ is defined in (6). It will turn out that if the parameter θ is small enough, this solution also solves the original equation (2). From now on, we assume $\theta \in]0,\theta_1]$. Recall also that $q^\star > \alpha$ (which can be written as $q > N\alpha/(N+\alpha)$), $\theta_1 = 1/(2p+1)$ and $q = 2p+2$.

PROPOSITION 3.1. *Let $\alpha > 2^\star$ and $q > \frac{N\alpha}{N+\alpha}$. Then there exists $u_\theta \in \mathcal{D}^{1;(2,q)}_{rad}(\mathbb{R}^N)$ such that*

$$c^1_\theta := \min_{v\in\mathcal{M}} I_\theta(v) = I_\theta(u_\theta) > 0. \tag{10}$$

For any minimizer u_θ of (10), there exists $\lambda_\theta \in \mathbb{R}^+$ such that u_θ is a weak solution of the equation

$$-\left(r^{N-1}a_\theta(|u_\theta'|^2)u_\theta'\right)' = \lambda_\theta \alpha r^{N-1}|u_\theta|^{\alpha-2}u_\theta, \tag{11}$$

i.e.

$$\int_0^{+\infty} r^{N-1} a_\theta(|u_\theta'|^2) u_\theta' v' = \lambda_\theta \alpha \int_0^{+\infty} r^{N-1} |u_\theta|^{\alpha-2} u_\theta v,$$

for every $v \in \mathcal{D}_{rad}^{1;(2,q)}(\mathbb{R}^N)$.

Moreover, for every $s \in [2^\star, q^\star]$, $s \in [2^\star, +\infty[$ *or* $s \in [2^\star, +\infty]$ *if* $q < N$, $q = N$ *and* $q > N$ *respectively, there exist* $C_1, M_1 > 0$ *independent of* $\theta \in]0, \theta_1]$ *such that*

$$\max\{\|u_\theta\|_{\mathcal{D}^{1;(2,q)}}, \|u_\theta\|_{L^s}\} \le C_1 \quad \textit{and} \quad c_\theta^1 \le M_1. \tag{12}$$

Proof. We proceed in several steps.

Step 1: Lower bounds on c_θ^1. The inequalities (7) and (8) imply the existence of a positive constant C depending only on p such that, for all $\theta \in]0, \theta_1]$ and all $u \in \mathcal{D}_{rad}^{1;(2,q)}(\mathbb{R}^N)$,

$$I_\theta(u) \ge C \int_{\mathbb{R}^N} |\nabla u|^2 \quad \text{and} \quad I_\theta(u) \ge C \int_{\mathbb{R}^N} |\nabla u|^q. \tag{13}$$

As $\alpha > 2^\star$, we have $2 < \frac{N\alpha}{N+\alpha} < q$ and we deduce by interpolation and Sobolev inequality that for all $u \in \mathcal{M}$,

$$I_\theta(u) \ge C \int_{\mathbb{R}^N} |\nabla u|^{\frac{N\alpha}{N+\alpha}} \ge C \left(\int_{\mathbb{R}^N} |u|^\alpha \right)^{\frac{N}{N+\alpha}} = C > 0, \tag{14}$$

for some $C > 0$ which depends only on p, α and N. This implies that $\inf_{v \in \mathcal{M}} I_\theta(v) > 0$.

Step 2: Existence of a minimizer. Let $(u_n)_n \subset \mathcal{M}$ be a minimizing sequence, i.e.

$$I_\theta(u_n) \to \inf_{v \in \mathcal{M}} I_\theta(v)$$

as $n \to \infty$. Choosing $\bar{u} \in \mathcal{M}$ a smooth function such that $|\nabla \bar{u}(x)| < 1 - \theta_1$ for all $x \in \mathbb{R}^N$, we can assume w.l.g. that

$$I_\theta(u_n) \le \int_{\mathbb{R}^N} (1 - \sqrt{1 - |\nabla \bar{u}|^2}) =: M_1 \tag{15}$$

for any $n \in \mathbb{N}$ and any $\theta \in]0, \theta_1]$.

It then follows from (13) and (15) that $(u_n)_n$ is bounded in $\mathcal{D}_{rad}^{1;(2,q)}(\mathbb{R}^N)$. Since $\alpha > 2^\star$ and $q > \frac{N\alpha}{N+\alpha}$, Lemma 2.4 implies that, up to a subsequence, $(u_n)_n$ converges weakly to u_θ in $\mathcal{D}_{rad}^{1;(2,q)}(\mathbb{R}^N)$ and strongly in $L^\alpha(\mathbb{R}^N)$ as $n \to \infty$. Obviously, $\int_{\mathbb{R}^N} |u_\theta|^\alpha = 1$ and $u_\theta \in \mathcal{M}$.

Moreover, I_θ being convex and continuous, I_θ is weakly lower semi-continuous and

$$I_\theta(u_\theta) \leq \liminf_{n\to\infty} I_\theta(u_n) = \inf_{v\in\mathcal{M}} I_\theta(v).$$

Since $u_\theta \in \mathcal{M}$, we conclude that $I_\theta(u_\theta) = \inf_{v\in\mathcal{M}} I_\theta(v)$.

Step 3: A priori bounds on the family $\{u_\theta : \theta \in]0,\theta_1]\}$. From (15), we infer that

$$c^1_\theta = I_\theta(u_\theta) \leq M_1. \tag{16}$$

By (13) and (16), u_θ is bounded in $\mathcal{D}^{1;(2,q)}$ uniformly in θ. The a priori bound in L^s follows from Lemma 2.2 according to whether $q < N$, $q = N$ or $q > N$.

Step 4: The Euler-Lagrange equation. By the Lagrange multiplier rule, there exists $\lambda_\theta \in \mathbb{R}$ such that for all $\varphi \in \mathcal{D}^{1;(2,q)}_{rad}(\mathbb{R}^N)$,

$$I'_\theta(u_\theta)(\varphi) = \lambda_\theta\alpha \int_{\mathbb{R}^N} |u_\theta|^{\alpha-2}u_\theta\ \varphi.$$

This means that

$$-\mathrm{div}\left(a_\theta(|\nabla u_\theta|^2)\nabla u_\theta\right) = \lambda_\theta\alpha|u_\theta|^{\alpha-2}u_\theta \ \text{ in } \mathbb{R}^N,$$

in the weak sense. As u_θ is radial, (11) follows. □

Observe that it is standard to prove that u_θ is a classical solution of (11) on $]0,+\infty[$. If $q > N$ then the solution is bounded and we can apply the regularity theory of Lieberman [24] to deduce that the weak solution u_θ is also $C^{1,\alpha}$ for some $0 < \alpha < 1$ in a neighborhood of the origin. We can deduce the regularity at the origin from even simpler arguments if $q < N$. Observe that for $\alpha > 2^\star$, we have $N - N/\alpha > N\alpha/(N+\alpha)$. In particular, Proposition 3.1 holds if $q > N - N/\alpha$.

LEMMA 3.2. *Let* $\alpha > 2^\star$ *and* $N - \frac{N}{\alpha} < q < N$. *If* u_θ *is a minimizer of* (10), *it is bounded in* $\mathcal{C}^1(\mathbb{R}^N)$ *and either* $u_\theta > 0$ *or* $u_\theta < 0$ *on* $\mathbb{R}^N$.

Proof. As u_θ is a solution of (11) on $]0,+\infty[$, it is standard to check that, for $r > 0$, u_θ is regular. On the other hand, one observes that $r^{N-1}a_\theta(|u'_\theta|^2)u'_\theta$ satisfies the Cauchy condition at the origin so that it has a finite limit as $r \to 0$. This limit must be zero otherwise we have

$$r^{N-1}a_\theta(|u'_\theta|^2)|u'_\theta|^2 \geq Cr^{-\frac{N-1}{q-1}}$$

near 0, which is not integrable because $q < N$. This contradicts the fact that, as u_θ is a weak solution of (11), we have

$$\int_0^{+\infty} r^{N-1}a_\theta(|u'_\theta|^2)|u'_\theta|^2 = \lambda_\theta\alpha.$$

We now claim that u'_θ is bounded. Integrating the equation, we get

$$\left|a_\theta(|u'_\theta(r)|^2)u'_\theta(r)\right| = \frac{\lambda_\theta \alpha}{r^{N-1}} \int_0^r s^{N-1}|u_\theta(s)|^{\alpha-1}\, ds,$$

for all $r \in [0,\infty[$. Using the estimate from Proposition 3.1, it follows that

$$a_\theta(|u'_\theta(r)|^2)\,|u'_\theta(r)| \le C\lambda_\theta \alpha r^{\frac{N(q^\star-\alpha+1)}{q^\star}-N+1}\|u_\theta\|_{L^{q^\star}}^{\alpha-1},$$

with $C>0$. Moreover, we have $N(q^\star-\alpha+1)/q^\star - N + 1 > 0$ since we assume $N - \frac{N}{\alpha} < q$, and therefore $u'_\theta(0) = 0$ and, for $r \le 1$, we conclude that

$$a_\theta(|u'_\theta(r)|^2)\,|u'_\theta(r)| \le C\lambda_\theta \alpha \|u_\theta\|_{L^{q^\star}}^{\alpha-1}.$$

We next deduce from Lemma 2.3 and Proposition 3.1 that for all $r > 1$,

$$\begin{aligned}\left|a_\theta(|u'_\theta(r)|^2)u'_\theta(r)\right| &= \frac{\lambda_\theta\alpha}{r^{N-1}}\left[\int_0^1 s^{N-1}|u_\theta(s)|^{\alpha-1}\,ds + \int_1^r s^{N-1}|u_\theta(s)|^{\alpha-1}\,ds\right]\\ &\le C\lambda_\theta\alpha\|u_\theta\|_{L^{q^\star}}^{\alpha-1} + \frac{\lambda_\theta\alpha}{r^{N-1}}\|u'_\theta\|_{L^2}^{\alpha-1}\int_1^r s^{N-1}s^{-\frac{(N-2)(\alpha-1)}{2}}\,ds\\ &\le C\left(1 + r^{1-\frac{(N-2)(\alpha-1)}{2}}\right),\end{aligned}$$

and since $\alpha > 2^\star$, we have

$$1 - \frac{N-2}{2}(\alpha-1) < -\frac{N}{2},$$

so that the claim follows.

As $u'_\theta(0) = 0$ one proves by standard arguments that u_θ is a classical solution.

To show that any minimizer satisfies either $u_\theta > 0$ or $u_\theta < 0$, we argue by contradiction. Indeed, if u_θ changes sign, then $|u_\theta| \in \mathcal{M}$ and $I_\theta(|u_\theta|) = I_\theta(u_\theta)$. In other words, $v = |u_\theta|$ is also a minimizer, and vanishes at some point $r_0 \in [0,\infty[$. Since v is a solution of (11) with $\min_{[0,+\infty[} v = v(r_0) = 0$ and the solutions of (11) are regular, we also have $v'(r_0) = 0$, which contradicts the local uniqueness of the solution of the Cauchy problem. This concludes the proof. □

3.2. Back to the original equation (2)

We now prove that the solution obtained in Proposition 3.1 is a solution of our original problem (2) provided the parameter θ is small enough.

In the sequel, $(u_\theta, \lambda_\theta)$ is the solution of

$$-\mathrm{div}\left(a_\theta(|\nabla u_\theta|^2)\nabla u_\theta\right) = \lambda_\theta \alpha |u_\theta|^{\alpha-2}u_\theta \ \text{ in } \mathbb{R}^N, \tag{17}$$

obtained in Proposition 3.1. We first estimate the Lagrange multiplier through an argument of the Calculus of Variations.

LEMMA 3.3. *For all $\theta \in]0, \theta_1]$, we have $0 < \lambda_\theta = \frac{N}{N+\alpha} c^1_\theta$.*

Proof. Multiplying (17) by u_θ and integrating, we obtain

$$\int_{\mathbb{R}^N} a_\theta(|\nabla u_\theta|^2)|\nabla u_\theta|^2 = \lambda_\theta \alpha \int_{\mathbb{R}^N} |u_\theta|^\alpha = \lambda_\theta \alpha. \tag{18}$$

Next, we prove that

$$\int_{\mathbb{R}^N} a_\theta(|\nabla u_\theta|^2)|\nabla u_\theta|^2 = \frac{N\alpha}{2N+2\alpha} \int_{\mathbb{R}^N} A_\theta(|\nabla u_\theta|^2). \tag{19}$$

To this end, consider the function $f : \mathbb{R}^+ \to \mathbb{R}$ defined by $f(t) := I_\theta\big(t^{\frac{N}{\alpha}} u_\theta(tx)\big)$. For all $t \in \mathbb{R}^+$, $t^{N/\alpha}u_\theta(tx) \in \mathcal{M}$, and f achieves its minimum at $t = 1$. A change of variable yields

$$f(t) = \frac{1}{2}\int_{\mathbb{R}^N} A_\theta\Big(t^{\frac{2N}{\alpha}+2}|\nabla u_\theta(tx)|^2\Big)\,dx = \frac{1}{2t^N}\int_{\mathbb{R}^N} A_\theta\Big(t^{\frac{2N}{\alpha}+2}|\nabla u_\theta(y)|^2\Big)\,dy.$$

From the last equality and Lebesgue's dominated convergence theorem, it is easy to see that f is differentiable. Hence, as $f(1)$ is a minimum, we have

$$f'(1) = \frac{1}{2}\left[\left(\frac{2N}{\alpha}+2\right)\int_{\mathbb{R}^N} a_\theta(|\nabla u_\theta|^2)|\nabla u_\theta|^2 - N\int_{\mathbb{R}^N} A_\theta(|\nabla u_\theta|^2)\right] = 0,$$

which proves (19).

Combining (19) with (18), we conclude that

$$\lambda_\theta = \frac{N}{2N+2\alpha}\int_{\mathbb{R}^N} A_\theta(|\nabla u_\theta|^2) = \frac{N}{N+\alpha}c^1_\theta > 0.$$

□

An important consequence of this lemma is that the uniform estimate on the levels c^1_θ from Proposition 3.1 yields a uniform estimate on the Lagrange multiplier. This estimate allows to deduce that, for θ small, our regularization leads to a solution of an unperturbed equation (with Lagrange multiplier).

PROPOSITION 3.4. *Assume $N - N/\alpha < q < N$. For $\alpha > 2^\star$ and θ small enough, the function u_θ obtained in Proposition* 3.1 *is a radial solution of*

$$-\mathrm{div}\left(\frac{\nabla u}{\sqrt{1-|\nabla u|^2}}\right) = \lambda |u|^{\alpha-2}u \quad \text{in } \mathbb{R}^N, \tag{20}$$

with

$$\lambda = \int_{\mathbb{R}^N} \frac{|\nabla u|^2}{\sqrt{1-|\nabla u|^2}} > 0.$$

Moreover either $u_\theta > 0$ or $u_\theta < 0$ on $\mathbb{R}^N$.

Proof. Consider the solution $(u_\theta, \lambda_\theta)$ of

$$-\mathrm{div}\left(a_\theta(|\nabla u_\theta|^2)\nabla u_\theta\right) = \lambda_\theta \alpha |u_\theta|^{\alpha-2}u_\theta \ \text{ in } \mathbb{R}^N,$$

obtained in Proposition 3.1. Let us prove the existence of a constant $E > 0$ such that for all $\theta \in]0, \theta_1]$ and all $r > 0$,

$$\left|a_\theta(|u_\theta'(r)|^2)u_\theta'(r)\right| \le E. \tag{21}$$

We argue as in Lemma 3.2 to deduce uniform estimates. First, using the uniform estimates from Proposition 3.1 and Lemma 3.3, it follows that for all $r < 1$ and all $\theta \in]0, \theta_1]$,

$$a_\theta(|u_\theta'(r)|^2)\,|u_\theta'(r)| \le C\lambda_\theta \alpha \|u_\theta\|_{L^{q^\star}}^{\alpha-1} \le C,$$

with $C > 0$ independent of $\theta \in]0, \theta_1]$. Moreover, by Lemma 2.3, Proposition 3.1 and Lemma 3.3, we have for all $r > 1$ and $\theta \in]0, \theta_1]$,

$$\begin{aligned}\left|a_\theta(|u_\theta'(r)|^2)u_\theta'(r)\right| &\le C\lambda_\theta \alpha \|u_\theta\|_{L^{q^\star}}^{\alpha-1} + \frac{\lambda_\theta \alpha}{r^{N-1}} \|\nabla u_\theta\|_{L^2}^{\alpha-1} \int_1^r s^{N-1} s^{-\frac{(N-2)(\alpha-1)}{2}} ds \\ &\le C\left(1 + r^{1-\frac{(N-2)(\alpha-1)}{2}}\right),\end{aligned}$$

where $C > 0$ is still independent of θ. As $\alpha > 2^\star$, we have

$$1 - \frac{N-2}{2}(\alpha - 1) < -\frac{N}{2}.$$

This proves (21).

Finally, by construction of a_θ, (21) implies that $|u_\theta'(r)| \le 1 - \epsilon$ for some $\epsilon > 0$, and hence u_θ solves (20) for θ small enough. More precisely, we have for all $r \ge 0$ and all $\theta < \min\{\theta_1, 1/(1+E^2)\}$,

$$|u_\theta'(r)| \le \frac{E}{\sqrt{1+E^2}},$$

and the result follows for $\theta < \min\{\theta_1, 1/(1+E^2)\}$. The fact that $\lambda := \lambda_\theta$ is bounded away from zero follows from Lemma 3.3. □

Proof of Theorem 1.1. By Proposition 3.4, we know that for θ small enough, u_θ is a radial solution of (20) i.e. u_θ is a solution of

$$-\left(r^{N-1}\frac{v'}{\sqrt{1-|v'|^2}}\right)' = \lambda r^{N-1}|v|^{\alpha-2}v \ \text{ in }]0,+\infty[.$$

Observe that w_t defined by $w_t(r) = tu_\theta(r/t)$ solves

$$-\left(r^{N-1}\frac{w'}{\sqrt{1-|w'|^2}}\right)' = \lambda\frac{1}{t^\alpha}r^{N-1}|w|^{\alpha-2}w \ \text{ in }]0,+\infty[.$$

Then w_t is a solution of the original equation (2) if $t = \lambda^{1/\alpha}$. □

REMARK 3.5. *Note that, for $\theta \in]0,\theta_1]$, w_t satisfies in fact*

$$I_\theta(|\nabla w_t|^2) = \min\left\{I_\theta(|\nabla v|^2) : v \in \mathcal{D}^{1;(2,q)}_{rad}(\mathbb{R}^N),\ \int_{\mathbb{R}^N}|v|^\alpha = \lambda^{\frac{\alpha+N}{\alpha}}\right\},$$

where $\lambda = c^1_\theta N/(N+\alpha) > 0$.

4. Optimizers in the inequality involving the volume integral

This section deals with the proof of Theorem 1.2 stated in the introduction. This theorem will follow from Proposition 4.1, Proposition 4.2, Proposition 4.4 and Proposition 4.5 below.

PROPOSITION 4.1. *Assume $\alpha \geq 2^\star$. Then there exists a constant $C > 0$, depending only on α and N, such that for all $u \in \mathcal{X}$,*

$$\int_{\mathbb{R}^N}(1-\sqrt{1-|\nabla u|^2}) \geq C\left(\int_{\mathbb{R}^N}|u|^\alpha\right)^{\frac{N}{\alpha+N}}. \tag{22}$$

Proof. If $\alpha \geq 2^\star$ then $2 \leq N\alpha/(N+\alpha)$. Hence, using the fact that $\|\nabla u\|_{L^\infty} \leq 1$ and Sobolev inequality, we have for all $u \in \mathcal{X}$,

$$\int_{\mathbb{R}^N}(1-\sqrt{1-|\nabla u|^2}) \geq \frac{1}{2}\int_{\mathbb{R}^N}|\nabla u|^2 \geq \frac{1}{2}\int_{\mathbb{R}^N}|\nabla u|^{\frac{N\alpha}{N+\alpha}} \geq C\left(\int_{\mathbb{R}^N}|u|^\alpha\right)^{\frac{N}{N+\alpha}},$$

where $C > 0$ depends only on α and N. □

Observe that the exponent $N\alpha/(\alpha+N)$ in the L^α-norm naturally arises in the proof when using Sobolev inequality. The presence of this exponent can also be explained from the invariance of the inequality (22) under the homeomorphisms $\phi_t : u(\cdot) \mapsto tu(\cdot/t)$ for $t > 0$.

We next show that the inequality (22) does not hold whatever $C > 0$ when $\alpha < 2^\star$.

PROPOSITION 4.2. *If $\alpha < 2^\star$ then*

$$\inf_{u\in\mathcal{X}\setminus\{0\}} \frac{\int_{\mathbb{R}^N}(1-\sqrt{1-|\nabla u|^2})}{\left(\int_{\mathbb{R}^N}|u|^\alpha\right)^{\frac{N}{N+\alpha}}} = 0.$$

Proof. It is straightforward to construct a sequence $(u_n)_n \subset \mathcal{D}^{1;2}(\mathbb{R}^N)$ such that $\|\nabla u_n\|_{L^\infty} \leq 1$, $\|u_n\|_{L^\alpha} = 1$, and $\int_{\mathbb{R}^N}|\nabla u_n|^2 \to 0$ as $n \to \infty$. Then we have for all $n \in \mathbb{N}$,

$$\int_{\mathbb{R}^N}(1-\sqrt{1-|\nabla u_n|^2}) = \int_{\mathbb{R}^N}\frac{|\nabla u_n|^2}{1+\sqrt{1-|\nabla u_n|^2}} \leq \int_{\mathbb{R}^N}|\nabla u_n|^2,$$

and the conclusion follows. □

We now focus on the best constant for which (22) holds when $\alpha > 2^\star$. We will use the following lemma. For the definition and basic properties of the symmetric rearrangement, the reader is referred to [21, 29] (among many others). Since we adapt a rather classical lemma from [29], we keep the notations therein. In particular, the *symmetric rearrangement $u^\bigstar$ of u* is the function whose graph is the Schwarz symmetrization of $|u|$, see for instance [29, Definition 1.C].

LEMMA 4.3. *For all $u \in \mathcal{X}$, we have the inequality*

$$\int_{\mathbb{R}^N}\left(1-\sqrt{1-|\nabla u|^2}\right) \geq \int_{\mathbb{R}^N}\left(1-\sqrt{1-|\nabla u^\bigstar|^2}\right), \tag{23}$$

where $u^\bigstar : \mathbb{R} \to \mathbb{R}$ denotes the symmetric rearrangement of u.

Proof. First we observe that $u^\bigstar$ is well defined if $u \in \mathcal{X}$ because u is Lipschitz continuous and all the level sets $\{x \in \mathbb{R}^N : u(x) > t\}$ $(t \in \mathbb{R})$ have finite measure. In addition, by the Pólya-Szegö inequality (see for instance [10, Theorem 4.7]), we have

$$\|\nabla u^\bigstar\|_{L^\infty} \leq \|\nabla u\|_{L^\infty} \leq 1$$

and

$$\int_{\mathbb{R}^N}|\nabla u^\bigstar|^2 \leq \int_{\mathbb{R}^N}|\nabla u|^2.$$

Therefore the right-hand side of (23) makes sense and both sides of the inequality are finite because ∇u is square integrable for $u \in \mathcal{X}$.

It is proven in [29, Theorem 1.C] (see also [21]) that the inequality

$$\int_{\mathbb{R}^N} \Phi(|\nabla u|) \geq \int_{\mathbb{R}^N} \Phi(|\nabla u^\star|)$$

holds for any Lipschitz-continuous u which decays at infinity and any convex, increasing function $\Phi : [0,\infty[\to [0,\infty[$ satisfying $\Phi(0) = 0$.

For all $n \in \mathbb{N}$, let us consider the functions $H_n, G_n : [0,\infty[\to [0,\infty[$ defined by

$$\begin{aligned} H_n(s) &= 1-(1-s)^{1/2}, && \text{for } s < 1-1/n^2, \\ &= 1-\frac{1}{n}+\frac{n}{2}(s-1+\frac{1}{n^2}), && \text{for } s \geq 1-1/n^2, \end{aligned}$$

and $G_n(s) = H_n(s^2)$. Observe that G_n is convex, increasing and satisfies $G_n(0) = 0$. Hence, by [29, Theorem 1.C], we know that

$$\int_{\mathbb{R}^N} G_n(|\nabla u|) \geq \int_{\mathbb{R}^N} G_n(|\nabla u^\star|). \tag{24}$$

As $u \in \mathcal{D}^{1;2}(\mathbb{R}^N)$, the measure of the set $A := \{x \in \mathbb{R}^N : |\nabla u| \geq 1/2\}$ is finite and the fact that $u \in \mathcal{X}$ implies that, for all $n \geq 2$, $|G_n(|\nabla u(x)|^2)| \leq h(x)$ with $h \in L^1(\mathbb{R}^N)$ defined by

$$\begin{aligned} h(x) &= 1, && \text{for } x \in A, \\ &= |\nabla u|^2, && \text{for } x \notin A. \end{aligned}$$

Hence, we can apply Lebesgue's dominated convergence theorem to prove that

$$\int_{\mathbb{R}^N} G_n(|\nabla u|) \to \int_{\mathbb{R}^N} \left(1-\sqrt{1-|\nabla u|^2}\right). \tag{25}$$

as n goes to infinity. We can argue in the same way to prove that

$$\int_{\mathbb{R}^N} G_n(|\nabla u^\star|) \to \int_{\mathbb{R}^N} \left(1-\sqrt{1-|\nabla u^\star|^2}\right). \tag{26}$$

We then conclude by (24), (25) and (26). □

With this lemma at hand we can prove the following proposition.

PROPOSITION 4.4. *If $\alpha > 2^\star$, the infimum*

$$C(\alpha) := \inf_{u \in \mathcal{X}\setminus\{0\}} \frac{\int_{\mathbb{R}^N}(1-\sqrt{1-|\nabla u|^2})}{\left(\int_{\mathbb{R}^N}|u|^\alpha\right)^{\frac{N}{N+\alpha}}}$$

is achieved by a radial solution of (2).

Proof. By Lemma 4.3 and the L^α-norm-preserving property of the symmetric rearrangement, we may restrict our attention to a minimizing sequence $(u_n)_n \subset \mathcal{X}$ of radial functions. Since the quotient is invariant under the family of homeomorphisms $\phi_t : v(\cdot) \mapsto tv(\cdot/t)$ $(t > 0)$, we may assume that $\int_{\mathbb{R}^N} |u_n|^\alpha = 1$. It is easily seen that $(u_n)_n$ is a priori bounded in $\mathcal{X}$. Lemma 2.1 then provides a bound in $\mathcal{D}^{1;q}(\mathbb{R}^N)$ for every $q \geq 2$. From Lemma 2.4, we deduce the required compactness to conclude that $(u_n)_n$ weakly converges in $\mathcal{D}^{1;2}(\mathbb{R}^N)$ to a function $u \in \mathcal{X}$ with $\int_{\mathbb{R}^N} |u|^\alpha = 1$. The fact that u realizes the infimum $C(\alpha)$ follows from the weak lower semi-continuity (with respect to the weak convergence in $\mathcal{D}^{1;2}(\mathbb{R}^N)$) of the volume integral.

To show that $w_t = t\,u(\cdot/t)$ solves (2) for some $t > 0$, we first prove that $|\nabla u|$ is bounded away from 1. Denoting by u_θ a minimizer of I_θ over $\mathcal{M}$ (see Proposition 3.1), we have for all $\theta > 0$,

$$I_\theta(u_\theta) \leq I_\theta(u) \leq I_0(u), \tag{27}$$

where the second inequality follows from the ordering property of the family I_θ. Moreover, we have established in Section 3 that $I_\theta(u_\theta) = I_0(u_\theta)$ for θ small enough. As u is a minimizer of I_0 this implies that the inequalities in (27) are in fact equalities. In particular, $I_\theta(u) = I_\theta(u_\theta)$, and u is a minimizer of I_θ over $\mathcal{M}$ too. The arguments of Proposition 3.4 now apply so that

$$|u'(r)| \leq \frac{E}{\sqrt{1+E^2}} < 1,$$

for some $E > 0$ and we conclude as in the proof of Theorem 1.1. □

We now turn to the case of the critical exponent $\alpha = 2^\star$.

PROPOSITION 4.5. *The infimum*

$$C(2^\star) = \inf_{u \in \mathcal{X}\setminus\{0\}} \frac{\int_{\mathbb{R}^N} (1 - \sqrt{1 - |\nabla u|^2})}{\left(\int_{\mathbb{R}^N} |u|^{2^\star}\right)^{\frac{N}{N+2^\star}}}$$

is not achieved.

Proof. Assume by contradiction that $C(2^\star)$ is achieved by some $u \in \mathcal{X}$. As above, we may suppose that u is radial. Let us prove that

$$\int_0^\infty r^{N-1} \left[\left(1 + \frac{N}{\alpha}\right) \frac{u'^2}{\sqrt{1-|u'|^2}} - N(1 - \sqrt{1-|u'|^2}) \right] \leq 0. \tag{28}$$

Define for all $t \in [0,1]$,

$$f(t) := \frac{1}{2} \int_{\mathbb{R}^N} A_0\Big(t^{\frac{2N}{\alpha}+2} |\nabla u(tx)|^2\Big)\, dx = \frac{1}{2t^N} \int_{\mathbb{R}^N} A_0\Big(t^{\frac{2N}{\alpha}+2} |\nabla u(y)|^2\Big)\, dy.$$

Let $t \in (0,1)$ be fixed. As 1 is a minimum of f, the mean value theorem yields the existence of $\tilde{t} \in (t,1)$ such that

$$f'(\tilde{t}) \leq 0. \tag{29}$$

(Note that we cannot conclude as in Lemma 3.3 that $f'(1) = 0$ because f may not be well defined for $t > 1$.) Here, the mean value theorem applies because f is continuous on $[t,1]$ and differentiable on $(t,1)$. In order to prove the differentiability of f in $\tilde{s} \in (t,1)$, observe first that from the strict inequality $\tilde{s}^{(2N/\alpha)+2} |\nabla u|^2 < 1$ a.e. in $\mathbb{R}^N$, we deduce the differentiability of the integrand. Moreover the derivative of the integrand satisfies the uniform estimate

$$\left| \left(\frac{2N}{\alpha} + 2 \right) \frac{\tilde{s}^{\frac{2N}{\alpha}+1} |\nabla u|^2}{\sqrt{1 - \tilde{s}^{\frac{2N}{\alpha}+2} |\nabla u|^2}} \right| \leq C |\nabla u|^2,$$

which holds for all s close to $\tilde{s}$ and almost every $x \in \mathbb{R}^N$. Lebesgue's dominated convergence theorem implies then that f is differentiable on $(t,1)$, and the inequality (29) is equivalent to

$$\begin{aligned} - N\tilde{t}^{-N-1} \int_{\mathbb{R}^N} \Big(1 - \sqrt{1 - \tilde{t}^{\frac{2N}{\alpha}+2} |\nabla u|^2} \Big) \\ + \tilde{t}^{-N} \int_{\mathbb{R}^N} \frac{1}{2} \left(\frac{2N}{\alpha} + 2 \right) \frac{\tilde{t}^{\frac{2N}{\alpha}+1} |\nabla u|^2}{\sqrt{1 - \tilde{t}^{\frac{2N}{\alpha}+2} |\nabla u|^2}} \leq 0. \end{aligned} \tag{30}$$

Next, we consider $(t_k) \subset (0,1)$ such that $t_k \to 1$ as $k \to \infty$. From what precedes, we infer the existence of a sequence $(\tilde{t}_k) \subset (0,1)$ still converging to 1 as k goes to ∞, and satisfying (30) with $\tilde{t} = \tilde{t}_k$ for all $k \in \mathbb{N}$. This implies that, for every $k \in \mathbb{N}$,

$$\begin{aligned} 0 &\leq \int_{\mathbb{R}^N} \left(\frac{N}{\alpha} + 1 \right) \frac{\tilde{t}_k^{\frac{2N}{\alpha}+1} |\nabla u|^2}{\sqrt{1 - \tilde{t}_k^{\frac{2N}{\alpha}+2} |\nabla u|^2}} \\ &\leq \frac{N}{\tilde{t}_k} \int_{\mathbb{R}^N} \Big(1 - \sqrt{1 - \tilde{t}_k^{\frac{2N}{\alpha}+2} |\nabla u|^2} \Big) \\ &\leq N \tilde{t}_k^{\frac{2N}{\alpha}+1} \int_{\mathbb{R}^N} |\nabla u|^2 \\ &\leq N \int_{\mathbb{R}^N} |\nabla u|^2. \end{aligned}$$

Hence, it follows from Fatou's Lemma, Lebesgue's dominated convergence the-

orem and (30) that $\frac{|\nabla u|^2}{\sqrt{1-|\nabla u|^2}} \in L^1(\mathbb{R}^N)$ and

$$\begin{aligned}\int_{\mathbb{R}^N} \frac{|\nabla u|^2}{\sqrt{1-|\nabla u|^2}} &\leq \liminf_{k\to\infty} \int_{\mathbb{R}^N} \left(\frac{N}{\alpha}+1\right) \frac{\tilde{t}_k^{\frac{2N}{\alpha}+1} |\nabla u|^2}{\sqrt{1-\tilde{t}_k^{\frac{2N}{\alpha}+2} |\nabla u|^2}} \\ &\leq \liminf_{k\to\infty} \frac{N}{\tilde{t}_k} \int_{\mathbb{R}^N} \Big(1-\sqrt{1-\tilde{t}_k^{\frac{2N}{\alpha}+2} |\nabla u|^2}\Big) \\ &= N \int_{\mathbb{R}^N} \Big(1-\sqrt{1-|\nabla u|^2}\Big).\end{aligned}$$

This implies that (28) holds.

To conclude, we define the function $g : [0,1[\to \mathbb{R}$ by $g(s) := (1+\frac{N}{\alpha}-N)s - N\sqrt{1-s} + N$ and we compute $g(0) = 0$, $g'(0) = 1+\frac{N}{\alpha}-\frac{N}{2}$ and $g''(s) = \frac{N}{4(1-s)^{3/2}}$. Therefore we have $g(s) > 0$ for $s \in]0,1[$ if and only if $1+\frac{N}{\alpha}-\frac{N}{2} \geq 0$, which is true if and only if $\alpha \leq 2^\star$. Hence, we infer that

$$\begin{aligned}0 &< \int_0^\infty r^{N-1} \frac{g(u'^2)}{\sqrt{1-|u'|^2}} \\ &= \int_0^\infty r^{N-1} \left[\left(1+\frac{N}{\alpha}\right) \frac{u'^2}{\sqrt{1-|u'|^2}} - N(1-\sqrt{1-|u'|^2})\right],\end{aligned}$$

which contradicts (28). □

5. A multiplicity result

In this section, we use again the auxiliary functional I_θ defined in Section 2.

Since the manifold $\mathcal{M}$ is symmetric and I_θ is an even functional, Lusternik-Schnirelmann category theory provides a sequence of critical values for I_θ constrained to $\mathcal{M}$. More precisely, let $\mathcal{A}$ denote the set of closed and symmetric (with respect to the origin) subsets of $\mathcal{D}_{rad}^{1;(2,q)}(\mathbb{R}^N)$. We define the usual min-max values

$$c_\theta^k := \inf_{A\in\Gamma_k} \max_{u\in A} I_\theta(u),$$

where

$$\Gamma_k := \{A \subset \mathcal{M} : A \in \mathcal{A}, A \text{ is compact and } \gamma(A) \geq k\},$$

and $\gamma(A)$ is the genus of the set A. We refer e.g. to [1] for the definition of the genus and for more details on Lusternik-Schnirelmann theory.

We first show that these levels are indeed critical levels of I_θ. It is clear that $\mathcal{M} \subset \mathcal{A}$ and $\gamma(\mathcal{M}) = +\infty$. Next we show that I_θ satisfies the Palais-Smale

condition on $\mathcal{M}$ by which we mean that every sequence $(u_n)_n \subset \mathcal{M}$ such that $I_\theta(u_n)$ is bounded and

$$I'_{\theta|\mathcal{M}}(u_n) \to 0$$

admits a converging subsequence. Here $I'_{\theta|\mathcal{M}}$ denotes the derivative of I_θ constrained to $\mathcal{M}$. Denoting by

$$T_u\mathcal{M} := \{v \in \mathcal{D}^{1;(2,q)}_{rad}(\mathbb{R}^N) : \int_{\mathbb{R}^N} |u|^{\alpha-2}uv = 0\}$$

the tangent space to $\mathcal{M}$ at u, the projection $P_u : \mathcal{D}^{1;(2,q)}_{rad}(\mathbb{R}^N) \to T_u\mathcal{M}$ is given by

$$P_u(w) = w - u\int_{\mathbb{R}^N} |u|^{\alpha-2}uw.$$

Then, for any $w \in \mathcal{D}^{1;(2,q)}_{rad}(\mathbb{R}^N)$ we have $v = P_u(w) \in T_u\mathcal{M}$ and

$$I'_{\theta|\mathcal{M}}(u)(v) = I'_{\theta|\mathcal{M}}(u)(P_u(w)) = I'_\theta(u)(w) - \lambda I'_\theta(u)(u),$$

where $\lambda = \int_{\mathbb{R}^N} |u|^{\alpha-2}uw$.

To prove the Palais-Smale condition, we will use the following convexity inequalities.

LEMMA 5.1. *There exist $\gamma_2, \gamma_q > 0$ such that for every $u, v \in \mathcal{D}^{1;(2,q)}_{rad}(\mathbb{R}^N)$,*

$$I_\theta(\frac{u+v}{2}) \le \frac{1}{2}I_\theta(u) + \frac{1}{2}I_\theta(v) - \gamma_2\int_{\mathbb{R}^N} |\nabla u - \nabla v|^2 \tag{31}$$

and

$$I_\theta(\frac{u+v}{2}) \le \frac{1}{2}I_\theta(u) + \frac{1}{2}I_\theta(v) - \gamma_q\int_{\mathbb{R}^N} |\nabla u - \nabla v|^q. \tag{32}$$

Proof. Since I_θ has a uniformly positive definite second derivative, we can apply [16, Lemma 2.3] to deduce (31). In order to prove (32), we first observe that [16, Lemma 2.1] allows to show that $s \to A_\theta(s^2)$ is strongly q-monotone. This yields, for some $\gamma_q > 0$, the inequality

$$A_\theta\left(\left[\frac{u'(r)+v'(r)}{2}\right]^2\right) \le \frac{1}{2}A_\theta(u'(r)^2) + \frac{1}{2}A_\theta(v'(r)^2) - 2\gamma_q|u'(r) - v'(r)|^q$$

where u, v are given functions in $\mathcal{D}^{1;(2,q)}_{rad}(\mathbb{R}^N)$ and $r > 0$. Multiplying by r^{N-1} and integrating from 0 to $+\infty$, we deduce (32). □

We now turn to the verification of the Palais-Smale condition.

LEMMA 5.2. *For $\alpha > 2^\star$, the functional I_θ satisfies the Palais-Smale condition on $\mathcal{M}$.*

Proof. Let $(u_n)_n \subset \mathcal{M}$ be a Palais-Smale sequence, i.e. $I_\theta(u_n)$ is bounded and

$$I'_{\theta|\mathcal{M}}(u_n) \to 0.$$

Since I_θ is coercive, it is clear that $(u_n)_n$ is bounded in $\mathcal{D}^{1;(2,q)}_{rad}(\mathbb{R}^N)$ and therefore, by Lemma 2.4, up to a subsequence, there exists $u \in \mathcal{M}$ such that u_n converges weakly to u in $\mathcal{D}^{1;(2,q)}_{rad}(\mathbb{R}^N)$ and strongly in $L^\alpha(\mathbb{R}^N)$ as $n \to \infty$. Since $(u_n)_n$ is a Palais-Smale sequence, we have, as $n \to \infty$,

$$I'_\theta(u_n)(P_{u_n}(u_n - u)) = I'_\theta(u_n)(u_n - u) - \lambda_n I'_\theta(u_n)(u_n) \to 0,$$

where we have written $\lambda_n = \int_{\mathbb{R}^N} |u_n|^{\alpha-2}u_n(u_n - u)$. Now, using the fact that $(u_n)_n$ is bounded in $\mathcal{D}^{1;(2,q)}_{rad}(\mathbb{R}^N)$ and $u_n \to u$ in $L^\alpha(\mathbb{R}^N)$, we infer $\lambda_n \to 0$ and $I'_\theta(u_n)(u_n)$ is bounded. Hence, we deduce that

$$\limsup_{n\to\infty} I'_\theta(u_n)(u_n - u) \le 0. \tag{33}$$

To complete the proof, it remains to show that $(u_n)_n$ converges strongly to u, which amounts to prove that

$$\|u_n - u\| = \left(\int_{\mathbb{R}^N} |\nabla u_n - \nabla u|^2\right)^{\frac{1}{2}} + \left(\int_{\mathbb{R}^N} |\nabla u_n - \nabla u|^q\right)^{\frac{1}{q}} \to 0,$$

as $n \to \infty$. Since I_θ is locally bounded, we may assume that $I_\theta(u_n)$ converges. By weak lower semi-continuity, we infer

$$I_\theta(u) \le \liminf_{n\to\infty} I_\theta(u_n),$$

whereas the convexity of I_θ and (33) implies

$$\limsup_{n\to\infty} I_\theta(u_n) \le I_\theta(u) + \limsup_{n\to\infty} I'_\theta(u_n)(u_n - u) \le I_\theta(u).$$

Hence $I_\theta(u_n)$ converges to $I_\theta(u)$. Using again the lower semi-continuity of I_θ, (31) and (32), we conclude that

$$I_\theta(u) \le \liminf_{n\to\infty} I_\theta\left(\frac{u_n + u}{2}\right) \le I_\theta(u) - \gamma_2 \limsup_{n\to\infty} \int_{\mathbb{R}^N} |\nabla u_n - \nabla u|^2$$

and

$$I_\theta(u) \le \liminf_{n\to\infty} I_\theta\left(\frac{u_n + u}{2}\right) \le I_\theta(u) - \gamma_q \limsup_{n\to\infty} \int_{\mathbb{R}^N} |\nabla u_n - \nabla u|^q.$$

This concludes the proof. □

Classical arguments now show that the level c_θ^k are critical values. We keep the notation $\theta_1 = 1/(2p+1)$.

PROPOSITION 5.3. *Assume $\alpha > 2^\star$ and $N - \frac{N}{\alpha} < q < N$. For every $k \geq 1$, there exists $\mu_\theta^k \in \mathbb{R}^+$ and $u_\theta^k \in \mathcal{D}_{rad}^{1;(2,q)}(\mathbb{R}^N)$ such that u_θ^k is a weak solution of*

$$-\mathrm{div}\left(a_\theta(|\nabla u_\theta^k|^2)\nabla u_\theta^k\right) = \mu_\theta^k \alpha |u_\theta^k|^{\alpha-2}u_\theta^k \quad \text{in } \mathbb{R}^N, \tag{34}$$

and $I_\theta(u_\theta^k) = c_\theta^k \to +\infty$ as $k \to \infty$. Moreover, u_θ^k is bounded in $\mathcal{C}^1(\mathbb{R}^N)$ and there exists $C_k > 0$, $M_k > 0$ such that, for all $\theta \in]0, \theta_1]$,

$$\max\{\|u_\theta^k\|_{\mathcal{D}^{1;(2,q)}}, \|u_\theta^k\|_{L^{q^\star}}\} \leq C_k \quad \text{and} \quad c_\theta^k \leq M_k.$$

Proof. The proof follows easily from [1, Theorem 10.9 and Theorem 10.10] observing also that one can bound the min-max levels taking smooth functions such that $|\nabla u(x)| < 1 - \theta_1$ as competitors in the definition of c_θ^k. Then it is enough to follow the lines of the proof of Proposition 3.1 and Lemma 3.2. □

The next step towards the proof of Theorem 1.3 consists in finding a priori bounds for the Lagrange multiplier μ_θ^k with respect to θ. The argument in Lemma 3.3 cannot be used here (except for u_θ^1 which is a global minimizer). We then go back to the equation to derive the identity (19) for the solutions u_θ^k. In fact, we just need an inequality.

LEMMA 5.4. *For all $\theta \in]0, \theta_1]$, $0 < \mu_\theta^k \leq \frac{N}{N+\alpha} c_\theta^k$.*

Proof. Multiplying (34) by u_θ^k and integrating, remembering also that $u_\theta^k \in \mathcal{M}$, we obtain

$$\int_{\mathbb{R}^N} a_\theta(|\nabla u_\theta^k|^2)|\nabla u_\theta^k|^2 = \mu_\theta^k \alpha \int_{\mathbb{R}^N} |u_\theta^k|^\alpha = \mu_\theta^k \alpha.$$

This shows $\mu_\theta^k > 0$.

Let us prove that

$$\int_{\mathbb{R}^N} a_\theta(|\nabla u_\theta^k|^2)|\nabla u_\theta^k|^2 \leq \frac{N\alpha}{2N+2\alpha} \int_{\mathbb{R}^N} A_\theta(|\nabla u_\theta^k|^2). \tag{35}$$

This implies that

$$\mu_\theta^k \leq \frac{N}{2N+2\alpha} \int_{\mathbb{R}^N} A_\theta(|\nabla u_\theta^k|^2) = \frac{N}{N+\alpha} c_\theta^k.$$

We know that u_θ^k is a solution of

$$-\left(r^{N-1} a_\theta(v'^2)v'\right)' = \mu_\theta^k \alpha r^{N-1} |v|^{\alpha-2}v,$$

bounded in $\mathcal{C}^1(\mathbb{R}^N)$ and satisfying $\int_{\mathbb{R}^N} |\nabla u_\theta^k(x)|^2 < \infty$. Let us define the function

$$\begin{aligned} F(r) &= r^N a_\theta(|v'|^2)|v'|^2 - \frac{1}{2} r^N A_\theta(|v'|^2) + \mu r^N |v|^\alpha + \frac{N}{\alpha} r^{N-1} v' v a(|v'|^2) \\ &= r\left(r^{N-1} v' a_\theta(|v'|^2)\right) v' - \frac{1}{2} r^N A_\theta(|v'|^2) + \mu r^N |v|^\alpha \\ &\qquad + \frac{N}{\alpha} v\left(r^{N-1} a_\theta(|v'|^2) v'\right), \end{aligned}$$

where for short we have written $\mu = \mu_\theta^k$ and $v = u_\theta^k$. Then, using the equation, we compute

$$\begin{aligned} F'(r) &= r^{N-1} a_\theta(|v'|^2)|v'|^2 + r[\left(r^{N-1} a_\theta(|v'|^2) v'\right)' v' + r^{N-1} a_\theta(|v'|^2) v' v''] \\ &\quad - r^N v' v'' a_\theta(|v'|^2) - \frac{N}{2} r^{N-1} A_\theta(|v'|^2) + \mu N r^{N-1} |v|^\alpha \\ &\quad + \mu\alpha r^N |v|^{\alpha-2} v v' + \frac{N}{\alpha} v (r^{N-1} a_\theta(|v'|^2) v')' + \frac{N}{\alpha} r^{N-1} a_\theta(|v'|^2)|v'|^2 \\ &= r^{N-1}\left[(1 + \frac{N}{\alpha}) a_\theta(|v'|^2)|v'|^2 - \frac{N}{2} A_\theta(|v'|^2)\right]. \end{aligned}$$

As v' and v are bounded, we have $F(0) = 0$. To estimate F at $+\infty$, we integrate the equation and we obtain

$$r^{N-1} a_\theta(|v'|^2) v' = -\mu\alpha \int_0^r (s^{N-1} |v|^{\alpha-2}(s) v(s))\, ds.$$

Using the decay estimate of Lemma 2.3, the a priori bound of Proposition 5.3 and the arguments of Lemma 3.2, we deduce that

$$\begin{aligned} r^{N-1} a_\theta(|v'|^2)|v'| &\le \mu\alpha \int_0^r (s^{N-1} |v|^{\alpha-1}(s))\, ds. \\ &\le \mu\alpha \int_0^1 (s^{N-1} |v|^{\alpha-1}(s))\, ds + C \int_1^r (s^{N-1} s^{-\frac{N-2}{2}(\alpha-1)})\, ds \\ &\le C(1 + \int_1^r (s^{\frac{3N-4}{2} - \frac{N-2}{2}\alpha})\, ds) \\ &\le C(1 + r^{\frac{3N-2}{2} - \frac{N-2}{2}\alpha}). \end{aligned}$$

Hence, we deduce that

$$a_\theta(|v'|^2)|v'| \quad \le \quad C(r^{1-N} + r^{\frac{N}{2} - \frac{N-2}{2}\alpha}),$$

and since $a_\theta(|v'|^2) \ge 1$, the same estimate holds for $|v'|$. This implies, again by

Lemma 2.3 and Proposition 5.3, that

$$\begin{aligned} F(r) &\leq r^N a_\theta(|v'|^2)|v'|^2 + \mu r^N |v|^\alpha + \frac{N}{\alpha} r^{N-1} v' v a_\theta(|v'|^2) \\ &\leq C(r^{2-N} + r^{\frac{2+N-(N-2)\alpha}{2}} + r^{2N-(N-2)\alpha} + r^{N-\frac{N-2}{2}\alpha} + r^{-\frac{N-2}{2}}). \end{aligned}$$

Since $\alpha > 2^\star$, we infer that

$$\limsup_{r\to\infty} F(r) \leq 0$$

and therefore

$$\int_0^\infty F'(r) \leq \limsup_{r\to\infty} F(r) - \lim_{r\to 0} F(r) = 0. \tag{36}$$

This completes the proof of (35). □

We are now able to complete the proof of Theorem 1.3.

PROPOSITION 5.5. *For $\alpha > 2^\star$ and θ small enough, the function $t_k u_\theta^k(r/t_k)$ where u_θ^k is given by Proposition 5.3, and*

$$t_k = \left(\int_{\mathbb{R}^N} \frac{|\nabla u_\theta^k|^2}{\sqrt{1 - |\nabla u_\theta^k|^2}} \right)^{1/\alpha},$$

is a solution of (2).

Proof. Fix $k \geq 1$. The proof follows from arguments that were used in Section 3. Indeed, since we have an estimate of the Lagrange multiplier μ_θ^k and on u_θ^k in $\mathcal{D}_{rad}^{1;(2,q)}(\mathbb{R}^N)$ as well as in $L^{q^\star}(\mathbb{R}^N)$ which are independent of θ, we infer, as in the proof of Proposition 3.4, that

$$|\nabla u_\theta^k| \leq \frac{E}{\sqrt{1+E^2}},$$

for some $E > 0$. The result then follows for $\theta < \min\{\theta_1, 1/(1+E^2)\}$ as in the proof of Theorem 1.1. □

REFERENCES

[1] A. AMBROSETTI AND A. MALCHIODI, *Nonlinear Analysis and Semilinear Elliptic Problems*, Cambridge Studies in Advanced Mathematics, vol. 104, Cambridge University Press, Cambridge, 2007.

[2] R. BARTNIK AND L. SIMON, *Spacelike hypersurfaces with prescribed boundary values and mean curvature*, Comm. Math. Phys. **87** (1982), 131–152.

[3] C. Bereanu, P. Jebelean, and J. Mawhin, *Variational methods for nonlinear perturbations of singular ϕ-laplacians*, Rend. Lincei Mat. Appl. **22** (2011), 89–111.
[4] C. Bereanu, P. Jebelean, and P.J. Torres, *Positive radial solutions for Dirichlet problems with mean curvature operators in Minkowski space*, preprint.
[5] D. Bonheure, A. Derlet, and S. de Valeriola, *On the multiplicity of nodal solutions of a prescribed mean curvature problem*, to appear in Math. Nachr.
[6] D. Bonheure, J.M. Gomes, and L. Sanchez, *Positive solutions of a second-order singular ordinary differential equation*, Nonlinear Anal. **61** (2005), 1383–1399.
[7] M. Born and L. Infeld, *Foundation of the new field theory*, Nature **132** (1933), 1004.
[8] M. Born and L. Infeld, *Foundation of the new field theory*, Proc. Roy. Soc. London A **144** (1934), 425–451.
[9] H. Brezis and J. Mawhin, *Periodic solutions of the forced relativistic pendulum*, Differential Integral Equations **23** (2010), 801–810.
[10] A. Burchard, *A short course on rearrangement inequalities*, June 2009, `http://www.math.toronto.edu/almut/rearrange.pdf`.
[11] L.A. Caffarelli, B. Gidas, and J. Spruck, *Asymptotic symmetry and local behavior of semilinear elliptic equations with critical Sobolev growth*, Comm. Pure Appl. Math. **42** (1989), 271–297.
[12] S.-Y. Cheng and S.-T. Yau, *Maximal space-like hypersurfaces in the Lorentz-Minkowski spaces*, Ann. of Math. **104** (1976), 407–419.
[13] I. Coelho, C. Corsato, F. Obersnel, and P. Omari, *Positive solutions of the Dirichlet problem for the one-dimensional Minkowski-curvature equation*, Adv. Nonlinear Stud. **12** (2012), 621–638.
[14] I. Coelho, C. Corsato, and S. Rivetti, *Positive radial solutions of the Dirichlet problem for the Minkowski-curvature equation in a ball*, preprint.
[15] E.N. Dancer, Z. Guo, and J. Wei, *Non-radial singular solutions of Lane-Emden equation in $\mathbb{R}^N$*, to appear in Indiana Univ. Math. J.
[16] P. De Nápoli and M.C. Mariani, *Mountain pass solutions to equations of p-Laplacian type*, Nonlinear Anal. **54** (2003), 1205–1219.
[17] M. del Pino and I. Guerra, *Ground states of a prescribed mean curvature equation*, J. Differential Equations **241** (2007), 112–129.
[18] A. Farina, *On the classification of solutions of the Lane-Emden equation on unbounded domains of $\mathbb{R}^N$*, J. Math. Pures Appl. **87** (2007), 537–561.
[19] B. Gidas and J. Spruck, *Global and local behavior of positive solutions of nonlinear elliptic equations*, Comm. Pure Appl. Math. **24** (1981), 525–598.
[20] C. Gui, W.-M. Ni, and X. Wang, *On the stability and instability of positive steady states of a semilinear heat equation in $\mathbb{R}^n$*, Comm. Pure Appl. Math. **45** (1992), 1153–1181.
[21] B. Kawohl, *Rearrangements and Convexity of Level Sets in PDE*, Lecture Notes in Mathematics, vol. 1150, Springer-Verlag, Berlin, 1985.
[22] M. K.-H. Kiessling, *Some uniqueness results for stationary solutions to the Maxwell-Born-Infeld field equations and their physical consequence*, Phys. Lett. A **375** (2011), 3925–3930.

[23] M. K.-H. Kiessling, *On the quasi-linear elliptic PDE* $-\nabla\cdot(\nabla u/\sqrt{1-|\nabla u|^2}) = 4\pi\sum_k a_k\delta_{s_k}$ *in physics and geometry*, Comm. Math. Phys. **314** (2012), 509–523.
[24] G. Lieberman, *Boundary regularity for solutions of degenerate elliptic equations*, Nonlinear Anal. TMA **12** (1988), 1203–1219.
[25] P.-L. Lions, *Symétrie et compacité dans les espaces de Sobolev*, J. Funct. Anal. **49** (1982), 315–334.
[26] W.-M. Ni and J. Serrin, *Existence and non-existence theorems for ground states for quasilinear partial differential equations*, Att. Convegni Lincei **77** (1985), 231–257.
[27] M. Ramos, H. Tavares, and W. Zou, *A Bahri-Lions theorem revisited*, Adv. Math. **222** (2009), 2173–2195.
[28] W.A. Strauss, *Existence of solitary waves in higher dimensions*, Comm. Math. Phys. **55** (1977), 149–162.
[29] G. Talenti, *Inequalities in rearrangement invariant function spaces*, Nonlinear analysis, function spaces and applications, Vol. 5 (Prague, 1994), Prometheus, Prague, 1994, pp. 177–230.

Authors' addresses:

Denis Bonheure
Département de Mathématique
Université libre de Bruxelles
CP 214 Boulevard du Triomphe, B-1050 Bruxelles, Belgium
E-mail: denis.bonheure@ulb.ac.be

Colette De Coster
Université de Valenciennes et du Hainaut Cambrésis, LAMAV, FR CNRS 2956
Institut des Sciences et Techniques de Valenciennes
F-59313 Valenciennes Cedex 9, France
E-mail: Colette.DeCoster@univ-valenciennes.fr

Ann Derlet
Institut de mathématique de Toulouse, CeReMath
Université de Toulouse
21 Allée de Brienne, F-31000 Toulouse, France
E-mail: aderlet@univ-tlse1.fr

Received July 30, 2012
Revised September 26, 2012

Rend. Istit. Mat. Univ. Trieste
Volume 44 (2012), 285–296

From probability to sequences and back

Roman Frič

Dedicated to Fabio Zanolin

Abstract. *This is a survey covering sequential structures and their applications to the foundations of probability theory. Sequential convergence, convergence groups and the extension of sequentially continuous maps belong to general topology and Trieste for long has been a center of sequential topology. We begin with some personal reflections, continue with topological problems motivated by the extension of probability measures, and close with some recent results related to the categorical foundations of probability theory.*

Keywords: Convergence of sequences, sequentially continuous map, field of sets, extension of probability measures, convergence group, free group, completion, categorical approach to probability, bold algebra, MV-algebra, D-poset of fuzzy sets, state, extension of states, epireflection
MS Classification 2010: 06D35, 54C20, 60A10, 04A72, 28C99, 54B30

1. Introduction

My PhD advisor Professor Josef Novák (1905 - 1999) and Professor Mario Dolcher (1920 - 1997), the PhD advisor of Fabio Zanolin, have had a common interest in sequential convergence and sequential topology (cf. [3]). Fabio has solved some problems posed by Novák related to sequential convergence spaces and groups ([37, 38]) and our personal meeting at the Prague Topological Symposium in 1982 resulted in friendship, fruitful cooperation, and a series of joint papers ([2, 12, 13, 14, 15, 16, 17, 18, 19]).

During my first visit of Italy in 1986, my homeland Slovakia (part of Czechoslovakia until 1993) and Italy have been separated by the Iron Curtain. That time, due to the Helsinky Agreement in 1975, scientific contacts and even joint research have been more easy and, thanks to a generous support by the Consiglio Nazionale delle Ricerche, I had both honor and pleasure to spend few fantastic weeks within the mathematical community in Trieste. Besides intensive joint research on convergence groups with Fabio, my plan was to present some results of Novák and members of his research team. The topic was "topological (sequential) aspects of the extension of measure". While working on my colloquium presentation, I have solved the "product problem for

sequential envelopes" (the product of sequential envelopes is equal to the sequential envelope of product, cf. [5]). The theory of sequential envelopes and its applications to probability has been a big theme for people around Novák ([5, 6, 9, 10, 25, 26, 29, 30, 31, 34]). Indeed, sequential envelopes are epireflections similar to the Čech-Stone β-compactification, the Hewitt υ-realcompactification, and the E-compactifications of S. Mrówka, for which the product problems and their solutions are really "hard mathematics" (cf. [21]). I remember being so happy, that even the bad news about Chernobyl looked unimportant to me (that time the information was very limited).

At this point, let me provide some background information about Josef Novák and his interest in the relationship between (sequential) topology and probability. He was a student of Eduard Čech and hence a topologist by faith. During WWII, Czech universities have been closed by the Nazi authorities and Novák became involved in statistical applications. Continuity in applications usually means sequential continuity, while the "real topology" means ultrafilters, compactness, and the like... The idea of Novák was to utilize sequences in general topology as much as possible (remember his construction of a regular topological space every continuous function on which is constant). The extension of probability measures (in fact bounded sequentially continuous functions) from a field $\mathbb{A}$ of subsets to the generated σ-field $\sigma(\mathbb{A})$ served as a canonical example in three directions.

1. Operations in $\mathbb{A}$ are sequentially continuous, hence we can study $\mathbb{A}$ as a sequential convergence algebra (group) and $\sigma(\mathbb{A})$ can be considered as its sequential completion.

2. The sequential convergence in a field of sets is determined by probability measures (a sequence $\{A_n\}_{n=1}^{\infty}$ converges to A iff the sequence $\{p(A_n)\}_{n=1}^{\infty}$ converges to $p(A)$ for all probability measures p) - a sequential version of complete regularity of a topological space. The problem is to find suitable sequential absolute properties of $\sigma(\mathbb{A})$ analogous to absolute properties like compactness or realcompactness.

3. Sequential convergence structures do not belong to the mainstream of general topology, hence there was a need to develop a suitable classification of such structures and to introduce characteristic properties guaranteeing relevant constructions in the realm of sequential structures. Observe that sequences are "short and meager", so that analogous topological and sequential constructions usually have different properties, for example, unlike βX and υX, the extension of bounded sequentially continuous functions and unbounded sequentially continuous functions are equivalent constructions ([5]).

An interested reader can find more detailed information about sequential structures in [6] and references therein.

In the present paper I will concentrate on the outcome of research related

to the second of the three directions. Most of our joint research with Fabio Zanolin concerned the other two directions. Here I mention two main themes related to sequential convergence groups, also known as $\mathcal{L}$-groups.

1. Free convergence groups. Beside being a natural construction, the free group serves as a vehicle to transport properties of sequential convergence spaces to $\mathcal{L}$-groups (cf. [12, 14, 15, 16]).

2. Coarse convergence groups. To define a compatible sequential convergence (we assume unique limits) for a given group G, it is the same as to define a suitable subgroup of G^N (the group of all sequences converging to the neutral element of G). This relates algebraic properties of G, resp. G^N, and certain properties of the convergence in question. Coarse convergence means that it cannot be enlarged without ruining the compatibility (e.g. the uniqueness of limits). The coarseness can be characterized by an algebraic condition, which results in an nice interplay between algebra and sequential topology. Coarse groups have interesting nontrivial properties (cf. [2, 13, 17, 19, 35]).

2. Measure extension theorem and more

In this section we outline the basic ideas of Josef Novák related to the extension of probability measures and leading to the notion of sequential envelope (cf. [8]).

THEOREM 2.1 (METHM – classical). *Let $\mathbb{A}$ be a field of sets, let $\sigma(\mathbb{A})$ be the generated σ-field, and let p be a probability measure on $\mathbb{A}$. Then there exists a unique probability measure $\overline{p}$ on $\sigma(\mathbb{A})$ such that $\overline{p}(A) = p(A)$ for all $A \in \mathbb{A}$.*

The proof (usually based on the outer measure) can be found in any treatise on measure. However, additional properties of $\sigma(\mathbb{A})$ are usually not mentioned there. J. Novák pointed out that from the "topological viewpoint" $\sigma(\mathbb{A})$ can be viewed as a maximal object over which all probability measures on $\mathbb{A}$ can be extended.

In order to make the text more self-contained, we recall some facts about fields of sets. Let X be a set. Then each subset $A \subseteq X$ can be viewed as the indicator function $\chi_A \in \{0,1\}^X$, $\chi_A(x) = 1$ if $x \in A$ and $\chi_A(x) = 0$ otherwise. Moreover, a sequence $\{A_n\}_{n=1}^{\infty}$ converges to A (i.e. $A = \limsup A_n = \liminf A_n$) iff the sequence $\{\chi_{A_n}\}_{n=1}^{\infty}$ converges pointwise to χ_A. If $\mathbb{A}$ is a field of subsets of X, then the generated σ-field $\sigma(\mathbb{A})$ is the smallest sequentially closed subset of $\{0,1\}^X$ containing $\mathbb{A}$ and $\mathbb{A}$ is sequentially dense in $\sigma(\mathbb{A})$ (i.e. each $A \in \sigma(\mathbb{A})$ can be reached by iterations, up to ω_1 times, of adding sequential limits, starting with sequences from $\mathbb{A}$). Observe that if two probability measures on $\sigma(\mathbb{A})$ coincide on $\mathbb{A}$, then a topological argument guarantees that they are identical. Let $\mathbb{A}, \mathbb{B}$ be fields of subsets of X and let $\mathbb{A} \subseteq \mathbb{B}$. A sequence $\{A_n\}_{n=1}^{\infty}$ of sets in $\mathbb{A}$ is said to be *P-Cauchy* if for each probability measure p on $\mathbb{A}$ the sequence $\{p(A_n)\}_{n=1}^{\infty}$ is a Cauchy sequence of real numbers. If for

each probability measure p on $\mathbb{A}$ there exists a probability measure $\overline{p}$ on $\mathbb{B}$ such that $\overline{p}(A) = p(A)$ for all $A \in \mathbb{A}$, then $\mathbb{A}$ is said to be *P-embedded* in $\mathbb{B}$.

THEOREM 2.2. *The following are equivalent*

(i) $\mathbb{A} = \sigma(\mathbb{A})$*;*

(ii) Each P-Cauchy sequence converges in $\mathbb{A}$*;*

(iii) $\mathbb{A}$ *is sequentially closed in each field of sets* $\mathbb{B}$ *in which* $\mathbb{A}$ *is P-embedded.*

Proof. (i) implies (ii). Assume (i) and let $\{A_n\}_{n=1}^{\infty}$ be a P-Cauchy sequence in $\mathbb{A}$. Since each $x \in X$ represents a point-probability, the sequence $\{A_n\}_{n=1}^{\infty}$ (pointwise) converges in $\{0,1\}^X$. From $\mathbb{A} = \sigma(\mathbb{A})$ it follows that $\mathbb{A}$ is sequentially closed and hence $\{A_n\}_{n=1}^{\infty}$ converges in $\mathbb{A}$.

(ii) implies (iii). Let $\mathbb{A}$ be P-embedded in $\mathbb{B}$ and let $\{A_n\}_{n=1}^{\infty}$ be a sequence in $\mathbb{A}$ which converges in $\mathbb{B}$. Since each $\overline{p} \in P(\mathbb{B})$ is sequentially continuous, $\{A_n\}_{n=1}^{\infty}$ is P-Cauchy and hence converges in $\mathbb{A}$.

(iii) implies (i). From the classical METHM it follows that $\mathbb{A}$ is P-embedded in $\sigma(\mathbb{A})$. Thus (iii) implies that $\mathbb{A}$ sequentially closed in $\sigma(\mathbb{A})$ and hence $\mathbb{A} = \sigma(\mathbb{A})$. This completes the proof. □

THEOREM 2.3 (METHM – Novák). *Let* $\mathbb{A}$ *be a field of subsets of* X *and let* $\sigma(\mathbb{A})$ *be the generated* σ*-field. Then* $\sigma(\mathbb{A})$ *is a maximal field of subsets of* X *in which* $\mathbb{A}$ *is P-embedded and sequentially dense.*

Proof. The assertion follows from the preceding theorem. Let $\mathbb{A}$ be a field of subsets of X. Assume that $\mathbb{A}$ is P-embedded and sequentially dense in a field $\mathbb{B}$. Clearly, $\mathbb{A}$ is P-embedded and sequentially dense in $\sigma(\mathbb{B})$. Since the generated σ-field of a field of subsets of X is the smallest sequentially closed system in $\{0,1\}^X$ containing the field in question, necessarily $\sigma(\mathbb{B}) = \sigma(\mathbb{A})$. Thus $\sigma(\mathbb{A})$ is maximal. This completes the proof. □

Observe that σ-fields form a special class of fields of subsets. Indeed, $\mathbb{A} = \sigma(\mathbb{A})$ means that $\mathbb{A}$ has the following absolute property with respect to the extension of probability measures (cf. [7]): $\mathbb{A}$ is sequentially closed in each field of subsets in which it is P-embedded (in this respect, this absolute property is similar to the compactness).

J. Novák showed that each bounded σ-additive measure on a ring of sets $\mathbb{A}$ is sequentially continuous ([28]) and pointed out the topological aspects of the extension of such measures on $\mathbb{A}$ over the generated σ-ring $\sigma(\mathbb{A})$: it is of a similar nature as the extension of bounded continuous functions on a completely regular topological space X over its Čech-Stone compactification βX (or as the extension of continuous functions on X over its Hewitt realcompactification υX). He developed a theory of sequential envelopes and (exploiting the Measure Extension Theorem) he proved that $\sigma(\mathbb{A})$ is the sequential envelope of $\mathbb{A}$ with respect to the probabilities. However, the sequential continuity

does not capture other properties (e.g. additivity) of probability measures. We show that in the category ID of D-posets of fuzzy sets (such D-posets generalize both fields of subsets and their fuzzy counterparts called bold algebras) probabilities are morphisms and the extension of probabilities on $\mathbb{A}$ over $\sigma(\mathbb{A})$ is a completely categorical construction (an epireflection, see [1]).

OBSERVATION 2.4. *Novák's original construction of the sequential envelope of a space X (a set carrying sequential convergence and the corresponding convergence closure) with respect to a given class $\mathcal{C}_0$ of sequentially continuous functions into $[0,1]$ follows the usual construction of β-compactification: embedding X into the power $[0,1]^{\mathcal{C}_0}$ and taking the closure (instead of the product topology, $[0,1]^{\mathcal{C}_0}$ carries the pointwise convergence, i.e. the categorical product convergence, and instead of the topological closure we take the smallest sequentially closed set containing the embedded X). In fact, this is a categorical construction of an epireflection of X, belonging to the category of space embeddable into powers $[0,1]^S$, into the subcategory of spaces embeddable as sequentially closed subspaces of powers $[0,1]^S$ (cf. [5]).*

OBSERVATION 2.5. *In the realm of sequential convergence spaces, the sequentially closed subspaces of categorical convergence powers $[0,1]^S$ possess the quality of being absolutely sequentially closed with respect to the extension of sequentially continuous functions of a given class, i.e., sequentially closed in every larger space to which sequentially continuous functions of a given class can be extended.*

OBSERVATION 2.6. *The category ID of D-posets of fuzzy sets is the result of a quest for a natural domain of generalized random events in which "all goes well":*

1. Both the classical Kolmogorovian probability theory, or CPT, and the fuzzy probability theory, or FPT, initiated by A. L. Zadeh ([36]) "live as minimal models having simple characteristic properties".

2. Probability measures, observables (i.e. preimages of random variables) and their fuzzy counterparts are morphisms.

3. Basic probability notions and constructions are categorical.

3. Notes on probability

In this section we present some notes about the foundations of probability. We will put into a perspective CPT and FPT and show why in the category ID "all goes well".

A. N. Kolmogorov in his famous "Grundbegriffe" ([22]) has "mathematized" probability via set-theoretic and maesure-theoretic constructions. Roughly, random events are "measurable" subsets of the outcomes, and probability is a measure (normed and σ-additive) on the random events. Observe that

- Random events form a σ-complete lattice of sets;
- In fact, every random event, as a subset of Ω, is a propositional function (Boolean logic).

In 1968 L. A. Zadeh ([36]) proposed to extend the classical probability to the realm of fuzzy mathematics. His idea was to extend classical random events, i.e. measurable $\{0,1\}$-valued (propositional) functions, to fuzzy random events, i.e. measurable $[0,1]$-valued (propositional) functions, and the probability measure to the integral with respect to a probability measure.

There are conceptual and theoretical differences and similarities between randomness and fuzziness (cf. [24]).

- Both systems describe uncertainty with numbers in the unit interval $[0,1]$ and both systems combine sets and propositions associatively, commutatively, and distributively;
- The key distinction concerns how the systems deal with a thing A and its opposite A^c;
- Classical logic and set theory assume that the law of noncontradiction (the law of excluded middle) is never violated. That is what makes the classical theory black or white;
- Fuzziness begins where Western logic ends. Fuzziness describes *event ambiguity*. It measures the degree to which an event occurs, not whether it occurs;
- Randomness describes the uncertainty of *event occurrence*. An event occurs or not;
- At issue is the nature of the occurring event: whether it is uncertain in any way, in particular whether it can be unambiguously distinguished from its opposite.

In order to represent a *classical object* o

- We choose a set X of attributes;
- We identify o and the set $A_o = \{x \in X;\ o \quad \text{does have} \quad x\}$.

Observe that, in fact, o can be viewed as a propositional function $o \in \{0,1\}^X$ and $x \in A_o$ iff the proposition $o(x)$ is true. Clearly, x cannot be at the same time in A_o and in its complement.

In order to represent a *fuzzy object* o

- We choose a set X of attributes;

- We identify o and the fuzzy set $o \in [0,1]^X$, where $o(x)$ is the degree to which o possesses the attribute x.

Observe that, in fact, o can be viewed as a "fuzzy propositional function" $o \in \{0,1\}^X$ and $o(x)$ tells us how much o is true at x. It can happen, that at some x both o and its complement $o^c = 1_X - o$ are "partially true", i.e., both $o(x)$ and $o^c(x) = 1 - o(x)$ are positive numbers.

QUESTION: Is it possible to build a generalized probability so that the CPT and FPT are special cases?
ANSWER: YES.

- We start with a set X of attributes and the system of potential generalized random events $[0,1]^X$ carrying the natural pointwise partial order;
- Any minimal model of generalized random events $\mathcal{X} \subseteq [0,1]^X$ has to contain the maximal and minimal random events (constant functions 0_X, 1_X) and has to be closed with respect to the relative complementation: if $u, v \in \mathcal{X}$ and $v \leq u$, i.e. $v(x) \leq u(x)$ for all $x \in X$, then $u - v \in \mathcal{X}$;
- If we assume that it is a σ-complete lattice (defined pointwise), then there exists a σ-field $\mathbf{A}$ of subsets of X such that $\mathbf{A} \subseteq \mathcal{X} \subseteq \mathcal{M}(\mathbf{A})$, where $\mathcal{M}(\mathbf{A})$ is the family of all measurable functions ranging in $[0,1]$;
- If we assume that $\mathcal{X}$ is divisible, i.e., for each $u \in \mathcal{X}$ and each natural number n there exists $v \in \mathcal{X}$ such that $nv = u$, and a σ-complete lattice, then $\mathcal{X} = \mathcal{M}(\mathbf{A})$.

The last two items are in fact deep results about the structure of "fuzzy random events" (cf. [27, Theorem 5.1]). To sum up, random events in CPT and random events in FPT are the minimal models of random events in a reasonable generalized probability; divisibility characterizes the transition from random events in CPT to random events in FPT.

4. From extension to epireflection

This section is devoted to bold algebras, distinguished domains of generalized probability (cf. [33]). First, we recall some notions used in the sequel.

D-posets have been introduced in [23] in order to model events in quantum probability. They generalize Boolean algebras, MV-algebras and other probability domains (cf. [4]) and provide a category in which generalized probability measures, called states, become morphisms. Recall that a D-poset is a partially ordered set X with the greatest element 1_X, the least element 0_X, and a partial binary operation called *difference*, such that $a \ominus b$ is defined iff $b \leq a$, and the following axioms are assumed:

(D1) $a \ominus 0_X = a$ for each $a \in X$;

(D2) If $c \leq b \leq a$, then $a \ominus b \leq a \ominus c$ and $(a \ominus c) \ominus (a \ominus b) = b \ominus c$.

A map h of a D-poset X into a D-poset Y which preserves the D-structure is said to be a *D-homomorphism*. Consider the unit interval $I = [0, 1]$ carrying the natural order, algebraic operations and convergence. Define a partial operation "$\ominus$" as follows: for $a, b \in I, b \leq a$, put $a \ominus b = a - b$. Then I carrying the natural (total) order, together with the partial operation is a D-poset. A sequentially continuous D- homomorphism of $\mathcal{X}$ into I is said to be a *state*.

Fundamental to applications are *D-posets of fuzzy sets*, i.e. systems $\mathcal{X} \subseteq [0, 1]^X$ carrying the coordinatewise partial order, coordinatewise convergence of sequences, containing the top and bottom elements of I^X, and closed with respect to the partial operation difference defined coordinatewise. We always assume that $\mathcal{X}$ is *reduced*, i.e., for $x, y \in X$, $x \neq y$, there exists $u \in \mathcal{X}$ such that $u(x) \neq u(y)$. Denote ID the category having (reduced) D-posets of fuzzy sets as objects and having sequentially continuous D-homomorphisms as morphisms. Objects of ID are subobjects of the powers I^X.

Recall ([4, 7]) that a *bold algebra* is a system $\mathcal{X} \subseteq [0, 1]^X$ containing the constant functions 0_X, 1_X and closed with respect to the usual Łukasiewicz operations: for $u, v \in \mathcal{X}$ put $(u \oplus v)(x) = u(x) \oplus v(x) = \min\{1, u(x) + v(x)\}$, $u^*(x) = 1 - u(x)$, $x \in X$. Bold algebras are MV-algebras representable as $[0, 1]$-valued functions, MV-algebras generalize Boolean algebras and bold algebras generalize in a natural way fields of sets (viewed as indicator functions). More information concerning MV-algebras and probability on MV-algebras can be found in [33]. If a bold algebra $\mathcal{X} \subseteq [0, 1]^X$ is sequentially closed in $[0, 1]^X$ (with respect to the coordinatewise sequential convergence), then $\mathcal{X}$ is a *Łukasiewicz tribe* ($\mathcal{X}$ is closed not only with respect to finite, but also with respect to countable Łukasiewicz sums, cf. [7, Corollary 2.8]). Let $\mathcal{X} \subseteq [0, 1]^X$ be a bold algebra. Then $[0, 1]^X$ is a Łukasiewicz tribe containing $\mathcal{X}$ and the intersection of all Łukasiewicz tribes $\mathcal{Y} \subseteq [0, 1]^X$ such that $\mathcal{X} \subseteq \mathcal{Y}$ is a Łukasiewicz tribe; it will be called the *induced* Łukasiewicz tribe and denoted by $\sigma(\mathcal{X})$. Each bold algebra can be considered as on object of ID. Finally, each bold algebra $\mathcal{X} \subseteq [0, 1]^X$ is a lattice, where for $u, v \in \mathcal{X}$ we have $(u \vee v)(x) = u(x) \vee v(x)$ and $(u \wedge v)(x) = u(x) \wedge v(x)$, $x \in X$.

Denote FSD the full subcategory of ID the objects of which are fields of sets and $CFSD$ its full subcategory consisting of σ-fields. It is known (cf. [32]) that sequentially continuous D-homomorphisms of a field of sets ranging in I are exactly σ-additive probability measures.

Denote BID the full subcategory of ID whose objects are bold algebras (the morphisms are exactly sequentially continuous D-morphisms). Let $CBID$ be

the subcategory of BID consisting of Łukasiewicz tribes (remember, a bold algebra $\mathcal{X} \subseteq I^X$ is a tribe iff $\mathcal{X}$ is a sequentially closed in I^X).

THEOREM 4.1. *Let $\mathcal{X} \subseteq I^X$ be a bold algebra and let $\sigma(\mathcal{X}) \subseteq I^X$ be the induced Łukasiewicz tribe. Let h be a sequentially continuous D-homomorphism of $\mathcal{X}$ into a Łukasiewicz tribe $\mathcal{Y}$. Then h can be uniquely extended to a sequentially continuous D-homomorphism h_σ of $\sigma(\mathcal{X})$ into $\mathcal{Y}$.*

Proof. Let $\mathcal{Y} = \sigma(\mathcal{Y}) \subseteq I^Y$. For each $y \in Y$, let pr_y be the y-th projection of I^Y to the factor space $I^{\{y\}}$. Then each composition $pr_y \circ h$ is a state on $\mathcal{X}$ and (cf. [7, Proposition 2.1]) it can be uniquely extended to a state $\overline{pr_y \circ h}$ on $\sigma(\mathcal{X})$. Since I^Y is a categorical product, there is a unique ID-morphism h_σ of $\sigma(\mathcal{X})$ into I^Y such that $pr_y \circ h_\sigma = \overline{pr_y \circ h}$. Clearly, for each $u \in \mathcal{X}$ and each $y \in Y$ we have $\overline{pr_y \circ h}(u) = (pr_y \circ h)(u)$. Hence $h_\sigma(u) = h(u)$ for each $u \in \mathcal{X}$. A topological argument shows that h_σ maps $\sigma(\mathcal{X})$ into $\mathcal{Y} = \sigma(\mathcal{Y})$ and that h_σ is uniquely determined (indeed, the pointwise convergence has unique limits, $\mathcal{X}$ is sequentially dense in $\sigma(\mathcal{X})$, h_σ is sequentially continuous and hence $h_\sigma(\sigma(\mathcal{X})) \subseteq \sigma((h(\mathcal{X})) \subseteq \sigma(\mathcal{Y}) = \mathcal{Y}$, (cf. [30]). □

REMARK 4.2. *If $\mathcal{Y}$ is the unit interval* [0,1] *carrying the canonical D-structure, then the previous theorem becomes the usual "State Extension Theorem" for bold algebras.*

REMARK 4.3. *Note that the embedding of a bold algebra $\mathcal{X}$ into $\sigma(\mathcal{X})$ is an epimorphism (two morphisms on $\sigma(\mathcal{X})$ agreeing on $\mathcal{X}$ are identical). This is a standard topological fact following from the uniqueness of limits, sequential continuity of morphisms, and the sequential density of $\mathcal{X}$ in $\sigma(\mathcal{X})$* (cf. [30]).

COROLLARY 4.4. *The subcategory $CBID$ is an epireflective subcategory of BID.*

Observe ([1]) that an epireflector is (roughly) a nice functor sending each object having some fundamental properties to the unique object in the subcategory of objects having some extreme properties, its epireflection, and sending each morphism to the unique morphism of the epireflection of its domain into the epireflection of its range (e.g. the completion of a metric space is an epireflection into complete metric spaces).

COROLLARY 4.5. *The subcategory $CFSD$ is an epireflective subcategory of FSD.*

Proof. Let $\mathbb{A} \subseteq \{0,1\}^X$ be a field of subset of X and let $\sigma(\mathbb{A})$ be the generated σ-field. Let h be an ID-morphism of $\mathbb{A}$ into a σ-field $\mathbb{B} = \sigma(\mathbb{B})$. Clearly, it suffices to prove that h can be uniquely extended to an ID-morphism h_σ of $\sigma(\mathbb{A})$ into $\mathbb{B}$. But $\sigma(\mathbb{A})$ and $\mathbb{B}$ are the induced Łukasiewicz tribes and the assertion follows from Theorem 4.1. □

As stated earlier, in the category ID the extension of probability measures on a field of subsets over the generated σ-field becomes a purely categorical construction. Moreover, the categorical approach leads to a better understanding of the foundations of probability theory (cf. [11, 20, 27]). Finally, observe that the sequential continuity of morphisms plays an an important role.

Acknowledgement: This work was supported by VEGA 2/0046/11.

References

[1] J. Adámek, *Theory of Mathematical Structures*, Reidel, Dordrecht, 1983.
[2] D. Dikranjan, R. Frič, and F. Zanolin, *On convergence groups with dense coarse subgroups*, Czechoslovak Math. J. **37** (1987), 471–479.
[3] M. Dolcher, *Topologie e strutture di convergenza*, Ann. Scuola Norm. Sup. Pisa **14** (1960), 63–92.
[4] A. Dvurečenskij and S. Pulmannová, *New trends in quantum structures*, Kluwer Academic Publ. and Ister Science, Dordrecht and Bratislava, 2000.
[5] R. Frič, *Remarks on sequential envelopes*, Rend. Istit. Mat. Univ. Trieste **26** (1988), 604–612.
[6] R. Frič, *History of sequential convergence spaces*, Handbook of the History of General Topology (Amsterdam), Progr. Math., vol. 1, Kluwer Academic Publishers, 1997, pp. 343–355.
[7] R. Frič, *Łukasiewicz tribes are absolutely sequentially closed bold algebras*, Czechoslovak Math. J. **52** (2002), 861–874.
[8] R. Frič, *Extension of measures: a categorical approach*, Math. Bohemica **130** (2005), 397–407.
[9] R. Frič and M. Hušek, *Projectively generated convergence of sequences*, Czechoslovak Math. J. **33** (1983), 525–536.
[10] R. Frič, K. McKennon, and S. D. Richardson, *Sequential convergence in c(x)*, Convergence structures and applications to analysis (Frankfurt/Oder, 1978) (Berlin), Abh. Akad. Wiss. DDR, Abt. Math.-Naturwiss.-Technik, 1979, vol. 4N, Akademie-Verlag, 1980, pp. 57–65.
[11] R. Frič and M. Papčo, *On probability domains II*, Internat. J. Theoret. Phys. **50** (2011), 3778–3786.
[12] R. Frič and F. Zanolin, *A convergence group having no completion*, Convergence Structures and Applications II (Schwerin, 1983) (Berlin), Abh. Akad. Wiss. DDR, Abt. Math.-Naturwiss.-Technik, 1984, vol. 2N, Akademie-Verlag, 1984, pp. 47–48.
[13] R. Frič and F. Zanolin, *Coarse convergence groups*, Convergence Structures 1984 (Proc. Conf. on Convergence, Bechyne, 1984) (Berlin), Mathematical Research/Mathematische Forschung Bd., vol. 24, Akademie-Verlag, 1985, pp. 107–114.
[14] R. Frič and F. Zanolin, *Fine convergence in free groups*, Czechoslovak Math. J. **36** (1986), 134–139.

[15] R. Frič and F. Zanolin, *Remarks on sequential convergence in free groups*, Topology and Applications (Eger, 1983) (Amsterdam), Colloq. Math. Soc. Janos Bolyai, vol. 41, North-Holland, 1986, pp. 283–291.
[16] R. Frič and F. Zanolin, *Sequential convergence in free groups*, Rend. Istit. Mat. Univ. Trieste **18** (1986), 200–218.
[17] R. Frič and F. Zanolin, *Coarse sequential convergence in groups, etc.*, Czechoslovak Math. J. **40** (1990), 459–467.
[18] R. Frič and F. Zanolin, *Strict completions of $\mathcal{L}_0^*$-groups*, Czechoslovak Math. J. **42** (1992), 589–598.
[19] R. Frič and F. Zanolin, *Relatively coarse sequential convergence*, Czechoslovak Math. J. **47** (1997), 395–408.
[20] S. Gudder, *Fuzzy probability theory*, Demonstratio Math. **31** (1998), 235–254.
[21] H. Herrlich and M. Hušek, *Some open categorical problems in top*, Appl. Categ. Structures **1** (1993), 1–19.
[22] A. N. Kolmogorov, *Grundbegriffe der wahrscheinlichkeitsrechnung*, Springer, Berlin, 1933.
[23] F. Kôpka and Chovanec, *D-posets*, Math. Slovaca **44** (1994), 21–34.
[24] B. Kosko, *Fuzziness vs. probability*, Int. J. Gen. Syst. **17** (1990), 211–240.
[25] V. Koutník, *On sequentially regular convergence spaces*, Czechoslovak Math. J. **17** (1967), 232–247.
[26] P. Kratochvíl, *Multisequences and measure*, General Topology and its Relations to Modern Analysis and Algeba IV (Proc. Fourth Prague Topological Sympos., 1976) (Praha), Part B Contributed Papers, Society of Czechoslovak Mathematicians and Physicists, pp. 237–244.
[27] R. Mesiar, *Fuzzy sets and probability theory*, Tatra Mt. Math. Publ. **1** (1992), 105–123.
[28] J. Novák, *Ueber die eindeutigen stetigen erweiterungen stetiger funktionen*, Czechoslovak Math. J. **8** (1958), 344–355.
[29] J. Novák, *On the sequential envelope*, General Topology and its Relations to Modern Analysis and Algebra (I) (Proc. (First) Prague Topological Sympos., 1961) (Prague), Part B Contributed Papers, Publishing House of the Czechoslovak Academy of Sciences, 1962, pp. 292–294.
[30] J. Novák, *On convergence spaces and their sequential envelopes*, Czechoslovak Math. J. **15** (1965), 74–100.
[31] J. Novák, *On sequential envelopes defined by means of certain classes of functions*, Czechoslovak Math. J. **18** (1968), 450–456.
[32] M. Papčo, *On measurable spaces and measurable maps*, Tatra Mt. Math. Publ. **28** (2004), 125–140.
[33] B. Riečan and D. Mundici, *Probability on MV-algebras*, Handbook of Measure Theory (Amsterdam), vol. II, North-Holland, 2002, pp. 869–910.
[34] M. Schroder, *Arrows in the "finite product theorem for certain epireflections" of R. Frič and D. C. Kent*, Math. Slovaca **45** (1995), 171–191.
[35] P. Simon and F. Zanolin, *A coarse convergence group need not be precompact*, Czechoslovak Math. J. **37** (1987), 480–486.
[36] L. A. Zadeh, *Probability measures of fuzzy events*, J. Math. Anal. Appl. **23** (1968), 421–427.

[37] F. Zanolin, *Solution of a problem of J. Novák about convergence groups*, Boll. Un. Mat. Ital. A **14** (1977), 375–381.
[38] F. Zanolin, *Example of a convergence commutative group which is not separated*, Czechoslovak Math. J. **34** (1984), 169–171.

Author's address:

Roman Frič
Mathematical Institute,
Slovak Academy of Sciences,
Grešákova 6, 040 01 Košice, Slovak Republic
and
Catholic University in Ružomberok,
Hrabovská cesta 1, 034 01 Ružomberok, Slovak Republic
E-mail: fric@saske.sk

Received June 13, 2012
Revised September 28, 2012

Rend. Istit. Mat. Univ. Trieste
Volume 44 (2012), 297–312

Limit free computation of entropy

Dikran Dikranjan and Anna Giordano Bruno

Dedicated to the sixtieth birthday of Fabio Zanolin

Abstract. *Various limit-free formulas are given for the computation of the algebraic and the topological entropy, respectively in the settings of endomorphisms of locally finite discrete groups and of continuous endomorphisms of totally disconnected compact groups. As applications we give new proofs of the connection between the algebraic and the topological entropy in the abelian case and of the connection of the topological entropy with the finite depth for topological automorphisms.*

Keywords: topological entropy, algebraic entropy, totally disconnected compact group, finite depth
MS Classification 2010: 37B40, 22C05, 54H11, 54H20, 54C70, 20K30

1. Introduction

In this paper we are concerned with the topological and the algebraic entropy respectively in the setting of continuous endomorphisms of totally disconnected compact groups and of endomorphisms of locally finite groups. In the abelian case the correspondence between these two settings - that is between continuous endomorphisms of totally disconnected compact abelian groups and endomorphisms of torsion abelian groups - is given by Pontryagin duality.

In [1] Adler, Konheim and McAndrew introduced the topological entropy for continuous selfmaps of compact spaces, while later on Bowen in [2] introduced it for uniformly continuous selfmaps of metric spaces, and this definition was extended to uniformly continuous selfmaps of uniform spaces by Hood in [9]. As explained in detail in [4], for continuous endomorphisms of totally disconnected compact groups the topological entropy can be introduced as follows. It is worth recalling that a totally disconnected compact group K has as a local base at 1 the family $\mathcal{B}(K)$ of all open subgroups of K, as proved by van Dantzig in [14].

Let K be a totally disconnected compact group and $\psi : K \to K$ a continuous endomorphism. For every open subgroup U of K and every positive integer n let

$$C_n(\psi, U) = U \cap \psi^{-1}(U) \cap \ldots \cap \psi^{-n+1}(U)$$

be the *n-th ψ-cotrajectory* of U, and the *ψ-cotrajectory* of U is

$$C(\psi, U) = \bigcap_{n=0}^{\infty} \psi^{-n}(U) = \bigcap_{n=1}^{\infty} C_n(\psi, U).$$

Note that this is the greatest ψ-invariant subgroup of K contained in U.

The *topological entropy of ψ with respect to U* is given by the following limit, which is proved to exist (see also Lemma 3.1 below),

$$H_{top}(\psi, U) = \lim_{n\to\infty} \frac{\log[K : C_n(\psi, U)]}{n}.$$

The *topological entropy* of ψ is

$$h_{top}(\psi) = \sup\{H_{top}(\psi, U) : U \in \mathcal{B}(K)\}.$$

Using ideas briefly sketched in [1], Weiss developed in [15] the definition of algebraic entropy for endomorphisms of torsion abelian groups. Moreover, Peters modified this definition in [12] for automorphisms of abelian groups, and this approach was extended to all endomorphisms of abelian groups in [3]; in [4] also the hypothesis of commutativity of the groups was removed. Following [4] we give here the definition of algebraic entropy for endomorphisms of locally finite groups, which coincides with the definition given in [1] in the abelian case.

Let G be a locally finite group and $\phi : G \to G$ an endomorphism. Denote by $\mathcal{F}(G)$ the family of all finite subgroups of G. For every finite subgroup F of G and every positive integer n let

$$T_n(\phi, F) = F \cdot \phi(F) \cdot \ldots \cdot \phi^{n-1}(F)$$

be the *n-th ϕ-trajectory* of F, and the *ϕ-trajectory* of F is

$$T(\phi, F) = \bigcup_{n=1}^{\infty} T_n(\phi, F).$$

If G is abelian, then $T(\phi, F)$ is the smallest ϕ-invariant subgroup of G containing F.

The *algebraic entropy of ϕ with respect to F* is the following limit, which exists as proved in [4],

$$H_{alg}(\phi, F) = \lim_{n\to\infty} \frac{\log |T_n(\phi, F)|}{n}.$$

The *algebraic entropy* of ϕ is

$$h_{alg}(\phi) = \sup\{H_{alg}(\phi, F) : F \in \mathcal{F}(G)\}.$$

Every locally finite group is obviously torsion, while the converse holds true under the hypothesis that the group is abelian; on the other hand, the solution of Burnside's problem shows that even groups of finite exponent may fail to be locally finite.

Yuzvinski claims at the end of his paper [17], that for every torsion abelian group G and every endomorphism $\phi : G \to G$ one has

$$h_{alg}(\phi) = \sup\left\{\log\left|\frac{T(\phi, F)}{\phi(T(\phi, F))}\right| : F \in \mathcal{F}(G)\right\}. \tag{1}$$

This formula is *false* without the assumption that ϕ is injective, as shown by Example 2.1 below (see also [7]). The huge gap in Example 2.1 is due to the special choice of the zero endomorphism. In fact, as noted in [7], Yuzvinski's claim is true for *injective* endomorphisms. A proof of this theorem, based on a much more general result on multiplicities, was given in [7]. Here we offer a short multiplicity-free proof of the following more general and precise formula that obviously implies the theorem. Note that if G is a torsion abelian group and $\phi : G \to G$ an endomorphism, then the hypothesis that $\ker\phi \cap T(\phi, F)$ is finite in the following formula is automatically satisfied (see Lemma 4.1).

Algebraic Formula. *Let G be a locally finite group, $\phi : G \to G$ an endomorphism and F a finite normal subgroup of G such that $\ker\phi \cap T(\phi, F)$ is finite. Then*

$$H_{alg}(\phi, F) = \log\left|\frac{T(\phi, F)}{\phi(T(\phi, F))}\right| - \log|\ker\phi \cap T(\phi, F)|.$$

The next corollary shows that Yuzvinski's claim holds true for injective endomorphisms.

COROLLARY 1.1. *Let G be a locally finite group, $\phi : G \to G$ an injective endomorphism and F a finite normal subgroup of G. Then*

$$H_{alg}(\phi, F) = \log\left|\frac{T(\phi, F)}{\phi(T(\phi, F))}\right|.$$

Therefore (1) *holds true whenever ϕ is injective.*

This formula suggests a similar approach for the topological entropy. Indeed it is possible to prove the following limit-free formula for the topological entropy. Also in this case, if the totally disconnected compact group K is

abelian and $\psi : K \to K$ is a continuous endomorphism, then the condition that $K/(\mathrm{Im}\psi + C(\psi, U))$ is finite is automatically satisfied.

Topological Formula. *Let K be a totally disconnected compact group, $\psi : K \to K$ a continuous endomorphism and U an open normal subgroup of K such that $K/(\mathrm{Im}\psi \cdot C(\psi, U))$ is finite. Then*

$$H_{top}(\psi, U) = \log \left| \frac{\psi^{-1}(C(\psi, U))}{C(\psi, U)} \right| - \log \left| \frac{K}{\mathrm{Im}\psi \cdot C(\psi, U)} \right|.$$

Stoyanov in [13] proved that in the compact case for the computation of the topological entropy one can reduce to surjective endomorphisms ψ, for which the quotient $K/(\mathrm{Im}\psi \cdot C(\psi, U))$ is obviously trivial. The much simpler formula in this case (practically, the topological counterpart of Corollary 1.1) is given in Corollary 3.6.

In Section 2 we give a proof of the Algebraic Formula, while in Section 3 we verify the Topological Formula. Moreover, we note how these two results give immediately a new proof of Weiss Bridge Theorem connecting the algebraic and the topological entropy by Pontryagin duality. Note that the Pontryagin dual of a torsion abelian group is a totally disconnected compact abelian group.

Weiss Bridge Theorem. *Let G be a torsion abelian group and $\phi : G \to G$ an endomorphism. Let $K = \widehat{G}$ be the Pontryagin dual of G and let $\psi = \widehat{\phi} : K \to K$ be the dual of ϕ. Then*

$$h_{alg}(\phi) = h_{top}(\psi).$$

Let K be a totally disconnected compact group and $\psi : K \to K$ a topological automorphism. In [16] Willis defined the pair (K, ψ) to have *finite depth* if there exists $U \in \mathcal{B}(K)$ such that

$$\bigcap_{n \in \mathbb{Z}} \psi^n(U) = \{1\}; \tag{2}$$

we call a subgroup U with this property *ϕ-antistable*. One can show that K must necessarily be metrizable and totally disconnected (see Section 5 for more details). For a pair (K, ψ) of finite depth, the *depth* of ψ is

$$\mathrm{depth}(\psi) = [\psi(C(\psi^{-1}, U)) : C(\psi^{-1}, U)]; \tag{3}$$

as noted in [16] this index is finite and does not depend on the choice of the ϕ-antistable subgroup $U \in \mathcal{B}(K)$.

In Section 5 an application of the Topological Formula is given in Theorem 5.2, stating that in case (K, ψ) is a pair of finite depth, then

$$h_{top}(\psi) = \log \operatorname{depth}(\psi).$$

A similar result for the measure-theoretic entropy, going into a somewhat different direction, can be found in [10, Theorem 2]. According to Halmos [8], surjective continuous endomorphisms of compact groups are measure preserving, and in this case the measure theoretic entropy coincides with the topological entropy as proved by Stoyanov [13].

2. Algebraic entropy

The following example shows that Yuzvinski's claim (1) is false without the assumption that the considered endomorphism is injective.

EXAMPLE 2.1. *Let G be a non-zero torsion abelian group, and $\phi : G \to G$ be the zero endomorphism. Take any non-zero finite subgroup F of G; then $T(\phi, F) = F$ and $\phi(T(\phi, F)) = 0$. Then*

$$\sup\left\{\log\left|\frac{T(\phi,F)}{\phi(T(\phi,F))}\right| : F \in \mathcal{F}(G)\right\} \geq \log\left|\frac{T(\phi,F)}{\phi(T(\phi,F))}\right| = \log|F|,$$

while $h_{alg}(\phi) = 0$.

In particular, when G is an infinite torsion abelian group, for every $n > 0$ we can pick a finite subgroup F_n of G of size $\geq n$. Then

$$\sup\left\{\log\left|\frac{T(\phi,F)}{\phi(T(\phi,F))}\right| : F \in \mathcal{F}(G)\right\} = \infty,$$

so in this case one has $0 = h_{alg}(\phi) \neq \infty$ in (1).

We are now in position to prove the Algebraic Formula.

THEOREM 2.2 (Algebraic Formula). *Let G be a locally finite group, $\phi : G \to G$ an endomorphism and F a finite normal subgroup of G such that $\ker\phi \cap T(\phi, F)$ is finite. Then*

$$H_{alg}(\phi, F) = \log\left|\frac{T(\phi,F)}{\phi(T(\phi,F))}\right| - \log|\ker\phi \cap T(\phi, F)|.$$

Proof. Let $K = \ker\phi \cap T(\phi, F)$, which is finite by hypothesis. We show that one can assume without loss of generality that F contains K. Indeed let $F' = FK \subseteq T(\phi, F)$. Then $T(\phi, F') = T(\phi, F)$, and so $H_{alg}(\phi, F') = H_{alg}(\phi, F)$;

moreover, $K = \ker\phi \cap F' \subseteq F'$. So we assume without loss of generality that $K \subseteq F$ and we verify that

$$H_{alg}(\phi, F) = \log\left|\frac{T(\phi, F)}{\phi(T(\phi, F))}\right| - \log|K|.$$

For the sake of brevity, we write in the sequel T_n and T, for $T_n(\phi, F)$ and $T(\phi, F)$ respectively.

Arguing as in [6, Lemma 1.1] we have that the index $|T_{n+1}/T_n|$ stabilizes, i.e., there exists $n_0 > 0$ such that for all $n > n_0$ one has $|T_{n+1}/T_n| = \alpha$, consequently $H_{alg}(\phi, F) = \log\alpha$. Our aim is to show that also

$$\left|\frac{T}{\phi(T)}\right| = \alpha \cdot |K|; \tag{4}$$

obviously this proves the theorem. Since $T = F \cdot \phi(T)$ and $(F \cdot \phi(T))/\phi(T) \cong F/(F \cap \phi(T))$, it follows that (4) is equivalent to

$$\left|\frac{F}{F \cap \phi(T)}\right| = \alpha \cdot |K|. \tag{5}$$

The increasing chain $F \cap \phi(T_n)$ of finite subgroups of F stabilizes, so there exists $n_1 > 0$ such that $F \cap \phi(T) = F \cap \phi(T_n)$ for all $n \geq n_1$. Hence (5) is equivalent to

$$\left|\frac{F}{F \cap \phi(T_n)}\right| = \alpha \cdot |K|$$

for all $n \geq n_1$.

As $F/(F \cap \phi(T_n)) \cong (F \cdot \phi(T_n))/\phi(T_n) = T_{n+1}/\phi(T_n)$, we conclude that

$$\left|\frac{F}{F \cap \phi(T_n)}\right| = \left|\frac{T_{n+1}}{\phi(T_n)}\right|. \tag{6}$$

Since $\phi(T_n) \cong T_n/(\ker\phi \cap T_n) = T_n/K$, we have $|\phi(T_n)| \cdot |K| = |T_n|$. Hence Lagrange Theorem applied to the group T_{n+1} and its subgroups T_n and $\phi(T_n)$ gives

$$\left|\frac{T_{n+1}}{\phi(T_n)}\right| = \frac{|T_{n+1}|}{|\phi(T_n)|} = \frac{|T_{n+1}| \cdot |K|}{|T_n|} = \left|\frac{T_{n+1}}{T_n}\right| \cdot |K| = \alpha \cdot |K|, \tag{7}$$

provided $n \geq \max\{n_0, n_1\}$. From (6) and (7) we get (5), and this concludes the proof. □

The next corollary is dedicated to the case of a finite normal subgroup F with $G = T(\phi, F)$. As noted in [6] this condition is not restrictive for the computation of the algebraic entropy, since

$$H_{alg}(\phi, F) = H_{alg}(\phi \restriction_{T(\phi,F)}, F) = h_{alg}(\phi \restriction_{T(\phi,F)}).$$

So in the particular case when $G = T(\phi, F)$ we have $h_{alg}(\phi) = H_{alg}(\phi, F)$.

COROLLARY 2.3. *Let G be a locally finite group, $\phi : G \to G$ an endomorphism and F a finite normal subgroup of G such that $G = T(\phi, F)$. If $\ker \phi$ is finite, then*

$$H_{alg}(\phi, F) = \log |\operatorname{coker} \phi| - \log |\ker \phi|.$$

In particular, $|\operatorname{coker} \phi| \geq |\ker \phi|$.

3. Topological entropy

The following is the counterpart of [6, Lemma 1.1] for the topological entropy. Its proof follows the one of [5, Lemma 2.2].

LEMMA 3.1. *Let K be a compact group, $\psi : K \to K$ a continuous endomorphism and U an open normal subgroup of K. For every positive integer n let $c_n := |K/C_n(\psi, U)|$. Then*

(a) c_n divides c_{n+1} for every $n > 0$.

For every $n > 0$ let $\alpha_n := c_{n+1}/c_n = |C_n(\psi, U)/C_{n+1}(\psi, U)|$. Then

(b) α_{n+1} divides α_n for every $n > 0$.

(c) Consequently the sequence $\{\alpha_n\}_{n>0}$ stabilizes, i.e., there exist integers $n_0 > 0$ and $\alpha > 0$ such that $\alpha_n = \alpha$ for every $n \geq n_0$.

(d) Moreover, $H_{top}(\psi, U) = \log \alpha$.

(e) If ψ is a topological automorphism, $H_{top}(\psi^{-1}, U) = H_{top}(\psi, U)$.

Proof. Let $n > 0$. Since there is no possibility of confusion we denote $C_n(\psi, U)$ simply by C_n.

(a) Since K/C_n is isomorphic to $(K/C_{n+1})/(C_n/C_{n+1})$, it follows that $c_{n+1}/c_n = |C_n/C_{n+1}|$ and in particular c_n divides c_{n+1}.

(b) We prove that C_n/C_{n+1} is isomorphic to a subgroup of C_{n-1}/C_n, and this gives immediately the thesis. First note that

$$\frac{C_n}{C_{n+1}} = \frac{C_n}{C_n \cap \psi^{-n}(U)} \cong \frac{C_n \cdot \psi^{-n}(U)}{\psi^{-n}(U)}.$$

Now $(C_n\psi^{-n}(U))/\psi^{-n}(U)$ is a subgroup of the quotient $(\psi^{-1}(C_{n-1})\psi^{-n}(U))/\psi^{-n}(U)$. The homomorphism $K/\psi^{-n}(U) \to K/\psi^{-n+1}(U)$ induced by ψ is injective, therefore the quotient $(\psi^{-1}(C_{n-1}) \cdot \psi^{-n}(U))/\psi^{-n}(U)$ is isomorphic to its image

$$\frac{C_{n-1} \cdot \psi^{-n+1}(U)}{\psi^{-n+1}(U)} \cong \frac{C_{n-1}}{C_{n-1} \cap \psi^{-n+1}(U)} = \frac{C_{n-1}}{C_n}.$$

(c) follows immediately from (b).

(d) By item (c) for $n_0 > 0$ we have $c_{n_0+n} = \alpha^n c_{n_0}$ for every $n \geq 0$, and by the definition of topological entropy

$$H_{top}(\psi, U) = \lim_{n\to\infty} \frac{\log c_n}{n} = \lim_{n\to\infty} \frac{\log(\alpha^n c_{n_0})}{n} = \log \alpha.$$

(e) Assume that ψ is a topological automorphism. For every positive integer n let $c_n^* := |K/C_n(\psi^{-1}, U)|$. According to (a)–(c) applied to ψ^{-1}, $H_{top}(\psi^{-1}, U) = \log \alpha^*$, where α^* is the value at which stabilizes the sequence $\alpha_n^* := c_{n+1}^*/c_n^*$. Hence it suffices to see that $c_n^* = c_n$ for all $n > 0$ and this is obvious since $\psi^{n-1}(C_n(\psi, U)) = C_n(\psi^{-1}, U)$. □

For the proof of the Topological Formula we need the following folklore fact that we give with a proof for reader's convenience.

LEMMA 3.2. *Let G be a topological group and let T be a closed subset of G. Then for every descending chain $B_1 \supseteq B_2 \supseteq \ldots \supseteq B_n \supseteq \ldots$ of closed subsets of G the intersection $B = \bigcap_{n=1}^{\infty} B_n$ is non-empty and $\bigcap_{n=1}^{\infty}(B_nT) = BT$, whenever B_1 is countably compact.*

Proof. That $B \neq \emptyset$ is a direct consequence of the countable compactness of B_1.

The inclusion $\bigcap_{n=1}^{\infty}(B_nT) \supseteq BT$ is obvious. To verify the converse inclusion pick an element $x \in \bigcap_{n=1}^{\infty}(B_nT)$. Then there exist elements $b_n \in B_n$, $t_n \in T$ such that $x = b_nt_n$ for every $n > 0$. Let D_n be the closure of the set $\{b_n, b_{n+1}, \ldots\}$ for $n > 0$. Then

$$D_n \subseteq B_n \quad \text{for each} \quad n > 0. \tag{8}$$

The countable compactness of B_1 yields that $\bigcap_{n=1}^{\infty} D_n \neq \emptyset$. Fix an element b of this intersection and note that $b \in B$ due to (8). It suffices to prove that $b^{-1}x \in T$. Since T is closed it suffices to check that $b^{-1}x$ belongs to the closure of T. To this end let $V = V^{-1}$ be a symmetric neighborhood of the neutral element of G. Then bV is a neighborhood of $b \in D_1$, so $bV \ni b_m$ for some $m > 0$. This yields $Vb^{-1} \ni b_m^{-1}$, and consequently $Vb^{-1}x \ni b_m^{-1}x = t_m$. Therefore $Vb^{-1}x \cap T \neq \emptyset$, and so $b^{-1}x$ belongs to the closure of T. □

THEOREM 3.3 (Topological Formula). *Let K be a totally disconnected compact group, $\psi : K \to K$ a continuous endomorphism and U an open normal subgroup of K such that $K/(\mathrm{Im}\psi \cdot C(\psi, U))$ is finite. Then*

$$H_{top}(\psi, U) = \log \left| \frac{\psi^{-1}(C(\psi, U))}{C(\psi, U)} \right| - \log \left| \frac{K}{\mathrm{Im}\psi \cdot C(\psi, U)} \right|.$$

Proof. Since there is no possibility of confusion we denote $C_n(\psi,U)$ and $C(\psi,U)$ simply by C_n and C respectively. Let $L = \mathrm{Im}\psi \cdot C$. We can assume without loss of generality that $U \subseteq L$. Indeed otherwise one can take $U' = U \cap L$. Then U' is open since L is open, being a closed subgroup of K of finite index by hypothesis; moreover, $C(\psi, U) = C(\psi, U')$ as $\psi^{-1}(U) = \psi^{-1}(U')$ and so $H_{top}(\psi, U) = H_{top}(\psi, U')$.

Let us note that our assumption $U \subseteq L$ and the inclusion $C \subseteq U$ imply

$$L = \mathrm{Im}\psi \cdot C \subseteq \mathrm{Im}\psi \cdot C_n \subseteq \mathrm{Im}\psi \cdot U \subseteq \mathrm{Im}\psi \cdot C \cdot U \subseteq L \cdot U = L. \tag{9}$$

The homomorphism $K/\psi^{-1}(C_n) \to K/C_n$ induced by ψ is injective and the image of $K/\psi^{-1}(C_n)$ is $\mathrm{Im}\psi \cdot C_n/C_n$. As $\mathrm{Im}\psi \cdot C_n = L$ by (9), we get

$$\left|\frac{K}{\psi^{-1}(C_n)}\right| = \left|\frac{L}{C_n}\right|. \tag{10}$$

By Lemma 3.1 there exist integers $n_0 > 0$ and $\alpha > 0$ such that $\alpha_n = \alpha$ for every $n \geq n_0$ and $H_{top}(\psi, U) = \log \alpha$. So it suffices to prove that

$$\left|\frac{\psi^{-1}(C)}{C}\right| = \alpha \cdot |L|. \tag{11}$$

We start noting that

$$\frac{\psi^{-1}(C)}{C} = \frac{\psi^{-1}(C)}{\psi^{-1}(C) \cap U} \cong \frac{\psi^{-1}(C) \cdot U}{U}. \tag{12}$$

The quotient K/U is finite and $\{(\psi^{-1}(C_n) \cdot U)/U\}_{n>0}$ is a descending chain of subgroups of K/U, hence it stabilizes, that is there exists $n_1 > 0$ such that

$$\psi^{-1}(C_n) \cdot U = \psi^{-1}(C_{n_1}) \cdot U \text{ for every } n \geq n_1;$$

in other words $\bigcap_{n=1}^{\infty}(\psi^{-1}(C_n) \cdot U) = \psi^{-1}(C_{n_1}) \cdot U$. Lemma 3.2 gives

$$\bigcap_{n=1}^{\infty}(\psi^{-1}(C_n) \cdot U) = \left(\bigcap_{n=1}^{\infty} \psi^{-1}(C_n)\right) \cdot U,$$

and $\bigcap_{n=1}^{\infty} \psi^{-1}(C_n) = \psi^{-1}(C)$, therefore

$$\psi^{-1}(C) \cdot U = \psi^{-1}(C_n) \cdot U \text{ for every } n \geq n_1. \tag{13}$$

Let $n \geq \max\{n_0, n_1\}$. Then (12) and (13) give

$$\frac{\psi^{-1}(C)}{C} \cong \frac{\psi^{-1}(C) \cdot U}{U} = \frac{\psi^{-1}(C_n) \cdot U}{U} \cong \frac{\psi^{-1}(C_n)}{\psi^{-1}(C_n) \cap U} = \frac{\psi^{-1}(C_n)}{C_{n+1}}.$$

Consequently

$$\left|\frac{\psi^{-1}(C)}{C}\right| = \left|\frac{\psi^{-1}(C_n)}{C_{n+1}}\right| = \frac{|K/C_{n+1}|}{|K/\psi^{-1}(C_n)|}. \tag{14}$$

As $|K/C_n| = |K/L| \cdot |L/C_n|$, (10) gives

$$\left|\frac{K}{\psi^{-1}(C_n)}\right| = \frac{|K/C_n|}{|K/L|}. \tag{15}$$

So using (15) in (14), and recalling that $n \geq n_0$, we can conclude that

$$\left|\frac{\psi^{-1}(C)}{C}\right| = \frac{|K/C_{n+1}|}{|K/C_n|}\left|\frac{K}{L}\right| = \left|\frac{C_n}{C_{n+1}}\right|\left|\frac{K}{L}\right| = \alpha\left|\frac{K}{L}\right|. \tag{16}$$

i.e., the wanted equality announced in (11). □

As noted in the Introduction, if K is a totally disconnected compact group, $\mathcal{B}(K)$ is a local base at 1. In this case also the subfamily $\mathcal{B}_\triangleleft(K)$ of $\mathcal{B}(K)$ of all normal open subgroups of K is a local base at 1. Indeed we have the following property, where for $U \in \mathcal{B}(K)$, the *heart* U_K of U in K is the greatest normal subgroup of K contained in U.

LEMMA 3.4. *Let K be a compact group. If $U \in \mathcal{B}(K)$, then $U_K \in \mathcal{B}_\triangleleft(K)$.*

Since for any $U, V \in \mathcal{B}(K)$, if $U \subseteq V$, then $H_{top}(\psi, V) \leq H_{top}(\psi, U)$, by the definition of topological entropy we immediately derive that it suffices to take the supremum when U ranges in a local base at 1 of K:

LEMMA 3.5. *Let K be a totally disconnected compact group, $\psi : K \to K$ a continuous endomorphism and $\mathcal{B} \subseteq \mathcal{B}(K)$ a local base at 1. Then $h_{top}(\psi) = \sup\{H_{top}(\psi, U) : U \in \mathcal{B}\}$.*

In particular, $h_{top}(\psi) = \sup\{H_{top}(\psi, U) : U \in \mathcal{B}_\triangleleft(K)\}$, so we immediately get Corollary 3.6 of the Topological Formula for continuous surjective endomorphism (in particular, for topological automorphisms).

Following Willis [16], when ψ is clear, we denote $C(\psi, U)$ also by the shorter and more suggestive U_-, and we leave U_+ denote the ψ^{-1}-cotrajectory $C(\psi^{-1}, U) = \bigcap_{n=0}^{\infty} \psi^n(U)$. We start using this notation from (17), where the first equality follows from Theorem 3.3, while the second one follows from Lemma 3.1(e) and the first equality.

COROLLARY 3.6. *Let K be a totally disconnected compact group and $\psi : K \to K$ a continuous surjective endomorphism. Then*

$$H_{top}(\psi, U) = \log\left|\frac{\psi^{-1}(U_-)}{U_-}\right| \text{ and } \left|\frac{\psi^{-1}(U_-)}{U_-}\right| = \left|\frac{\psi(U_+)}{U_+}\right| \tag{17}$$

for every $U \in \mathcal{B}_\triangleleft(K)$. In particular,

$$h_{top}(\psi) = \sup\left\{\log\left|\frac{\psi^{-1}(U_-)}{U_-}\right| : U \in \mathcal{B}_\triangleleft(K)\right\}.$$

The next corollary is dedicated to the case of an open normal subgroup U with trivial ψ-cotrajectory $U_- = C(\psi, U)$.

COROLLARY 3.7. *Let K be a totally disconnected compact group, $\psi : K \to K$ a continuous endomorphism and $U \in \mathcal{B}_\triangleleft(K)$ with trivial ψ-cotrajectory U_-. If* $\operatorname{coker}\psi = K/\mathrm{Im}\psi$ *is finite, then*

$$H_{top}(\psi, U) = \log|\ker\psi| - \log|\operatorname{coker}\psi|.$$

In particular, $|\ker\psi| \geq |\operatorname{coker}\psi|$.

The aim of the next remark is to clarify the significance of the hypothesis $|K/(\mathrm{Im}\psi \cdot U_-)| < \infty$ in Theorem 3.3. See also Remark 5.4 for an interesting consequence of Corollary 3.7.

REMARK 3.8. *Let $\psi : K \to K$ a continuous endomorphism of a totally disconnected compact group K and let U an open normal subgroup of K.*

(a) Let $K_U = K/U_-$, let $q_U : K \to K_U$ be the canonical homomorphism and let $\psi_U : K_U \to K_U$ be the induced endomorphism. Clearly, U_- is ψ-invariant (but need not be stabilized by ψ) and $q_U(U)$ is ψ_U-antistable (actually, $q(U)_- = C(\psi_U, q(U))$ is trivial). Moreover, $K/(\mathrm{Im}\psi \cdot U_-)$ is finite precisely when $\mathrm{Im}\psi_U$ has finite index in K_U. More precisely,

$$K/(\mathrm{Im}\psi \cdot U_-) \cong K_U/(\mathrm{Im}\psi_U \cdot q(U)_-) = K_U/\mathrm{Im}\psi_U = \operatorname{coker}\psi_U,$$

as $q(U)_-$ is trivial. So

$$\left|\frac{K}{\mathrm{Im}\psi \cdot U_-}\right| = \left|\frac{K_U}{\mathrm{Im}\psi_U}\right| = |\operatorname{coker}\psi_U|.$$

By Corollary 3.7 the triviality of $q(U)_-$ gives

$$H_{top}(\psi_U, q_U(U)) = \log|\ker\psi_U| - \log|\operatorname{coker}\psi_U|.$$

(b) Now let $N = U_- \cap U_+$, $K_{(U)} = K/N_U$ and let $p_U : K \to K_{(U)}$ be the canonical homomorphism. Clearly, N_U is stabilized by ψ, the induced endomorphism $\psi_{(U)}$ of $K_{(U)}$ is injective and $p_U(U)$ is $\psi_{(U)}$-antistable (one can see as before that $K/(\mathrm{Im}\psi \cdot U_-)$ is finite precisely when $\operatorname{coker}\psi_{(U)} = K_{(U)}/\mathrm{Im}\psi_{(U)}$ *is finite, etc.). One can use the pairs $(K_{(U)}, \psi_{(U)})$ of finite depth to present the pair (K, ψ) as an inverse limit of the pairs $(K_{(U)}, \psi_{(U)})$ with $U \in \mathcal{B}_\triangleleft(K)$ of finite depth (see [16, Proposition 5.3]).*

4. The abelian case

When the groups are abelian the finiteness conditions in the Algebraic Formula and in the Topological Formula are automatically satisfied. Indeed we have the following result, which applies directly for the Algebraic Formula and together with Lemma 4.3 for the Topological Formula.

LEMMA 4.1. *Let G be a torsion abelian group, $\phi : G \to G$ an endomorphism and F a finite subgroup of G. Then* $\ker\phi \cap T(\phi, F)$ *is finite.*

Proof. Since $T(\phi, F)$ is ϕ-invariant, we can assume without loss of generality that $G = T(\phi, F)$. This is a finitely generated $\mathbb{Z}[X]$-module. Therefore $\ker\phi$ is a finitely generated $\mathbb{Z}[X]$-module as well. Since the action of ϕ on $\ker\phi$ sends $\ker\phi$ to 0, we have that $\ker\phi$ is a finitely generated $\mathbb{Z}$-module. Hence $\ker\phi$ is finite. □

We recall now some definitions and results from Pontryagin duality. For a topological abelian group G the Pontryagin dual $\widehat{G}$ of G is the group $\mathrm{Chom}(G,\mathbb{T})$ of the continuous characters of G endowed with the compact-open topology [11]. The Pontryagin dual of a discrete Abelian group is always compact. Moreover, we recall that a finite abelian group is isomorphic to its dual, and the dual of a torsion abelian group is a totally disconnected compact abelian group. If $\phi : G \to G$ is an endomorphism, its Pontryagin dual $\widehat{\phi} : \widehat{G} \to \widehat{G}$ is defined by $\widehat{\phi}(\chi) = \chi \circ \phi$ for every $\chi \in \widehat{G}$. For a subset H of G, the *annihilator* of H in $\widehat{G}$ is $H^\perp = \{\chi \in \widehat{G} : \chi(H) = 0\}$.

LEMMA 4.2. *Let G be an abelian group.*

(a) If $\{H_n\}_{n>0}$ are subgroups of G, then $(\sum_{n=1}^{\infty} H_n)^\perp \cong \bigcap_{n=1}^{\infty} H_n^\perp$.

If H a subgroup of G and $\phi : G \to G$ an endomorphism, then:

(b) $\widehat{H} \cong \widehat{G}/H^\perp$;

(c) $(\phi^n(H))^\perp = (\widehat{\phi})^{-n}(H^\perp)$ for every $n \geq 0$;

(d) $\ker\phi^\perp = \mathrm{Im}\widehat{\phi}$.

(e) If $H \subseteq L$ are subgroups of G, then $H^\perp/L^\perp \cong \widehat{L/H}$.

LEMMA 4.3. *Let G be a torsion abelian group, $\phi : G \to G$ an endomorphism and F a finite subgroup of G. Let $K = \widehat{G}$, $\psi = \widehat{\phi}$ and $U = F^\perp$. Then $U \in \mathcal{B}(K)$ and*

(a) $T_n(\phi, F)^\perp = C_n(\psi, U)$ for every $n > 0$, and $T(\phi, F)^\perp = C(\psi, U)$;

(b) $\ker\phi \cap T(\phi, F) \cong K/(\mathrm{Im}\psi + C(\psi, U))$;

(c) $T(\phi,F)/\phi(T(\phi,F)) \cong \psi^{-1}(C(\psi,U))/C(\psi,U)$.

Proof. The conclusions follow from Lemma 4.2. □

Applying this lemma, the Algebraic Formula and the Topological Formula, we can now give a short proof of Weiss Bridge Theorem connecting the algebraic and the topological entropy.

COROLLARY 4.4 (Weiss Bridge Theorem). *Let G be a torsion abelian group and $\phi : G \to G$ an endomorphism, let $K = \widehat{G}$ and $\psi = \widehat{\phi}$. Then*

$$h_{alg}(\phi) = h_{top}(\psi).$$

Proof. Let $K = \widehat{G}$ and $\psi = \widehat{\phi}$. Let U be an open subgroup of K. Then F is a finite subgroup of G. By Theorem 2.2 and Theorem 3.3 and by Lemma 4.3 we can conclude that $H_{alg}(\phi,F) = H_{top}(\psi,U)$, hence $h_{alg}(\phi) = h_{top}(\psi)$. □

REMARK 4.5. *Applying Pontryagin duality in the abelian case one can also derive the Topological Formula from the Algebraic Formula. Indeed, let K be a totally disconnected compact abelian group and $\psi : K \to K$ a continuous endomorphism. Let $G = \widehat{K}$, $\phi = \widehat{\psi}$ and $F = U^{\perp}$. Then F is a finite subgroup of G. By Lemma 4.3 we have that $K/C_n(\psi,U) \cong T_n(\phi,F)$ and so*

$$H_{top}(\psi,U) = H_{alg}(\phi,F).$$

By Theorem 2.2 $H_{alg}(\phi,F) = \log|T(\phi,F)/\phi(T(\phi,F))| - \log|\ker\phi \cap T(\phi,F)|$ and again Lemma 4.3 gives

$$|T(\phi,F)/\phi(T(\phi,F))| = |\psi^{-1}(C(\psi,U))/C(\psi,U)|$$
$$\text{and } |\ker\phi \cap T(\phi,F)| = |K/(\mathrm{Im}\psi + C(\psi,U))|.$$

Therefore $H_{top}(\psi,U) = \log|\psi^{-1}(C(\psi,U))/C(\psi,U)| - \log|K/(\mathrm{Im}\psi + C(\psi,U))|$, that is the Topological Formula.

5. An application: finite depth and topological entropy

Let K be a totally disconnected compact group and $\psi : K \to K$ a topological automorphism. As recalled in the Introduction, the pair (K,ψ) has *finite depth* if there exists a ϕ-antistable $U \in \mathcal{B}(K)$ (see (3)). By Lemma 3.4 we can assume without loss of generality that U is also normal, that is $U \in \mathcal{B}_{\triangleleft}(K)$. This definition implies that

$$\text{the family } \mathcal{B}_U = \{U_n : n > 0\}, \quad \text{where } U_n := C_n(\psi,U) \cap C_n(\psi^{-1},U), \text{ is a local base at 1.} \tag{18}$$

In particular, K turns out to be necessarily metrizable and totally disconnected. Moreover, K is isomorphic to a subgroup G_1 of $F^{\mathbb{Z}}$, where F is a finite group; if σ denotes the left Bernoulli shift of $F^{\mathbb{Z}}$, then G_1 is stabilized by σ and under the identification of G with G_1 one has $\psi = \sigma \restriction_{G_1}$ (see also [10, Proposition 2]).

PROPOSITION 5.1 ([16, Proposition 5.5]). *Let (K,ψ) be a pair of finite depth. If $U, W \in \mathcal{B}_\triangleleft(K)$ are ϕ-antistable, then $[\psi(U_+) : U_+] = [\psi(W_+) : W_+]$.*

In view of this result one defines the *depth* of a pair (K,ψ) of finite depth as

$$\operatorname{depth}(\psi) = \left| \frac{\psi(U_+)}{U_+} \right|$$

for any ϕ-antistable $U \in \mathcal{B}_\triangleleft(K)$. Moreover, since

$$\left| \frac{\psi(U_+)}{U_+} \right| = \left| \frac{\psi^{-1}(U_-)}{U_-} \right| \tag{19}$$

according to (17), one can extend this definition to

$$\operatorname{depth}(\psi) = \left| \frac{\psi(U_+)}{U_+} \right| = \left| \frac{\psi^{-1}(U_-)}{U_-} \right|,$$

where $U \in \mathcal{B}_\triangleleft(K)$ is any ϕ-antistable $U \in \mathcal{B}_\triangleleft(K)$.

THEOREM 5.2. *Let (K,ψ) be a pair of finite depth. Then*

$$h_{top}(\psi) = \log \operatorname{depth}(\psi).$$

Proof. Let $U \in \mathcal{B}_\triangleleft(K)$ be ϕ-antistable. By (18) the family $\mathcal{B}_U$ is a local base at 1. Moreover, for any $n > 0$ we have $H_{top}(\psi,U_n) = \log |\psi^{-1}(C(\psi,U_n))/C(\psi,U_n)|$ by Theorem 3.3, therefore (19) gives

$$H_{top}(\psi, U_n) = \log \operatorname{depth}(\psi).$$

Hence $h_{top}(\psi) = \log \operatorname{depth}(\psi)$ by Lemma 3.5. □

The equality $h_{top}(\psi) = h_{top}(\psi^{-1})$ from Lemma 3.1(e) is well known for the topological entropy of automorphisms of compact groups, we obtain as a by-product the following fact.

COROLLARY 5.3. *Let (K,ψ) be a pair of finite depth. Then*

$$\operatorname{depth}(\psi) = \operatorname{depth}(\psi^{-1}).$$

Theorem 5.2 and Corollary 3.7 have the following consequence. According to [16, Proposition 5.5], if K is infinite, then $\operatorname{depth}(\psi) > 1$.

REMARK 5.4. *Let K be a totally disconnected compact group, $\psi : K \to K$ a continuous endomorphism and $U \in \mathcal{B}_{\lhd}(K)$ with trivial ψ-cotrajectory $C(\psi, U)$. The triviality of $C(\psi, U)$ implies that U is ψ-antistable. This yields that the pair (K, ψ) has finite depth, so if K is infinite, we have $H_{top}(\psi, U) = \log \operatorname{depth}(\psi) > 0$ by Theorem 5.2. In particular, Corollary 3.7 gives the non-obvious inequality $\log |\ker \psi| - \log |\operatorname{coker} \psi| > 0$, i.e., ψ is necessarily non-injective and $|\ker \psi| > |\operatorname{coker} \psi|$.*

Acknowledgements

It is a pleasure to thank George Willis for sending us his preprint [16] and for inspiring us to prove Theorem 5.2. We thank also the members of our Seminar on Dynamical Systems at the University of Udine for the useful discussions on this topic.

REFERENCES

[1] R. L. ADLER, A. G. KONHEIM AND M. H. MCANDREW, *Topological entropy*, Trans. Amer. Math. Soc. **114** (1965), 309–319.
[2] R. BOWEN, *Entropy for group endomorphisms and homogeneous spaces*, Trans. Amer. Math. Soc. **153** (1971), 401–414.
[3] D. DIKRANJAN AND A. GIORDANO BRUNO, *Entropy on abelian groups*, preprint http://arxiv.org/abs/1007.0533.
[4] D. DIKRANJAN AND A. GIORDANO BRUNO, *Topological entropy and algebraic entropy for group endomorphisms*, Proceedings ICTA2011 Islamabad, Pakistan July 4-10 2011 Cambridge Scientific Publishers (2012), 133–214.
[5] D. DIKRANJAN, A. GIORDANO BRUNO AND L. SALCE, *Adjoint algebraic entropy*, J. Algebra **324** (2010), 442–463.
[6] D. DIKRANJAN, B. GOLDSMITH, L. SALCE AND P. ZANARDO, *Algebraic entropy of endomorphisms of abelian groups*, Trans. Amer. Math. Soc. **361** (2009), 3401–3434.
[7] D. DIKRANJAN, M. SANCHIS AND S. VIRILI, *New and old facts about entropy on uniform spaces and topological groups*, Topology Appl. **159** (2012), 1916–1942.
[8] P. HALMOS, *On automorphisms of compact groups*, Bull. Amer. Math. Soc. **49** (1943), 619–624.
[9] B. M. HOOD, *Topological entropy and uniform spaces*, J. London Math. Soc. **8** (1974), 633–641.
[10] B. KITCHENS, *Expansive dynamics of zero-dimensional groups*, Ergodic Theory Dynam. Systems **7** (1987), 249–261.
[11] L. S. PONTRYAGIN, *Topological Groups*, Gordon and Breach, New York, 1966.
[12] J. PETERS, *Entropy on discrete Abelian groups*, Adv. Math. **33** (1979), 1–13.
[13] L. N. STOYANOV, *Uniqueness of topological entropy for endomorphisms on compact groups*, Boll. Un. Mat. Ital. B (7) **1** (1987), 829–847.

[14] D. van Dantzig, *Studien over topologische Algebra*, Dissertation, Amsterdam 1931.
[15] M. D. Weiss, *Algebraic and other entropies of group endomorphisms*, Math. Systems Theory **8** (1974/75), 243–248.
[16] G. A. Willis, *The nub of an automorphism of a totally disconnected locally compact group*, submitted http://arxiv.org/abs/1112.4239.
[17] S. Yuzvinski, *Metric properties of endomorphisms of compact groups*, Izv. Acad. Nauk SSSR, Ser. Mat. **29** (1965), 1295–1328 (in Russian). English Translation: Amer. Math. Soc. Transl. (2) **66** (1968), 63–98.

Authors' addresses:

Dikran Dikranjan
Dipartimento di Matematica e Informatica
Università di Udine
Via delle Scienze, 206 - 33100 Udine, Italy
E-mail: dikran.dikranjan@uniud.it

Anna Giordano Bruno
Dipartimento di Matematica e Informatica
Università di Udine
Via delle Scienze, 206 - 33100 Udine, Italy
E-mail: anna.giordanobruno@uniud.it

Received May 22, 2012
Revised October 2, 2012

Rend. Istit. Mat. Univ. Trieste
Volume 44 (2012), 313–348

Solvable (and unsolvable) cases of the decision problem for fragments of analysis

Domenico Cantone, Eugenio G. Omodeo
and Gaetano T. Spartà

Dedicated to Fabio Zanolin on the occasion of his 60th birthday

Abstract. *We survey two series of results concerning the decidability of fragments of Tarksi's elementary algebra extended with one-argument functions which meet significant properties such as continuity, differentiability, or analyticity. One series of results regards the initial levels of a hierarchy of prenex sentences involving a single function symbol: in a number of cases, the decision problem for these sentences was solved in the positive by H. Friedman and Á. Seress, who also proved that beyond two quantifier alternations decidability gets lost. The second series of results refers to merely existential sentences, but it brings into play an arbitrary number of functions, which are requested to be, over specified closed intervals, monotone increasing or decreasing, concave, or convex; any two such functions can be compared, and in one case, where each function is supposed to own continuous first derivative, their derivatives can be compared with real constants.*

Keywords: decidable theories, Tarski's elementary algebra, one-variable functions
MS Classification 2010: 03B25, 26A06

Introduction

We will address the decidability issue for various fragments of real analysis.

In the background, we have the fundamental decidability result proved by Tarski in [17] about the theory, named *elementary algebra*, where real *numbers* only—not functions—come into play. This result refers to the entire first-order language whose signature consists of the numerical constants 0, 1, -1, the operators $+, -, \cdot$, and the comparators $>, <, =$. As usual, an adequate basis of propositional connectives (e.g., $\wedge, \vee, \neg$) is also available, together with a

denumerable infinity of variables: these are assumed to range over the reals and can be quantified by means of the symbols $\exists, \forall$, without restraints. Tarski produced an algorithm which, given any formula Φ devoid of free variables in this language, provides the yes/no answer as whether Φ is true or false.

Note that in elementary algebra each variable represents a generic real number. If there were means to impose that some variables range over integers, then one would be able to recast in elementary algebra all sentences of elementary arithmetic, and could thereby decide which of these sentences are true: an impossible situation, as shown by Church in [4].

A decision algorithm for elementary algebra could become part of a *proof assistant*, to wit, of a computerized system offering support to scholars either by way of autonomous theorem-proving abilities or through verification that proposed proofs are impeccable [9]. Anyway, for applications of this nature one must necessarily take into account the computational cost of the algorithm.

It turns out, in particular, that although the procedure proposed by Collins [5] has doubly exponential complexity relative to the number of variables occurring in the sentence (or just exponential, if the endowment of variables is finite and fixed), its computational cost is considerably lower than in case of Tarski's algorithm. A refinement of this result is achieved with Grigoriev's algorithm [12] applicable to sentences in prenex normal form, whose complexity is doubly exponential relative to the number of quantifier alternations.

Even when we merely consider the *existential* theory of reals,[1] consisting of those sentences $\exists x_1 \cdots \exists x_n \vartheta$ in Tarksi's algebra, where ϑ is a quantifier-free formula (involving no variables distinct from $x_1, \ldots, x_n$), the known decision algorithms have a complexity at best exponential relative to the number n of variables [8]; however, if one fixes the number of variables that can be used, then an algorithm of polynomial complexity becomes available [14].

As observed by Tarski himself [17], the decidability of elementary algebra entails decidability of various other first-order theories regarding complex numbers or n-dimensional vectors, as well as decidability of elementary geometries of the plane, of 3-, or of n-dimensional space; of analogous non-Euclidean geometries, and of projective geometry. It is in fact possible to translate statements of these systems into statements about real numbers, thereby reducing their decision problems to the analogous problem for elementary algebra.

For instance, a first-order system of *elementary plane geometry* can be instructed [17, 18] over a language endowed with a denumerable infinity of variables (ranging over the points of Euclidean space), with the familiar dyadic sign $=$ (*identity* of points in the plane), with the 3-adic *betweenness* predicate

[1] As seen here, we are taking the liberty of calling 'theory' a fragment of the language of a theory proper (cf. [7])—usually of a complete one, so that the distinction between valid and true sentence becomes immaterial. Such a fragment, to wit, a syntactically delimited family Θ, does not comprise exclusively true sentences; so, when saying that a 'theory' is decidable, we will actually mean that its true sentences form a decidable subset of Θ.

symbol $B(x, y, z)$, interpreted as "y lies between x and z on the straight line xz", and with the 4-adic *equidistance* predicate symbol $D(x, y; z, t)$, interpreted as "the distance from x to y equals the distance from z to t". To get a decision method for this system:

- one associates with each sentence Φ of elementary plane geometry a sentence Φ^* of elementary algebra, by mapping each variable x of Φ into two real-valued variables $\overline{x}, \overline{\overline{x}}$ which represent its coordinates, so that to any two distinct point-variables x and y there correspond four distinct real variables $\overline{x}, \overline{\overline{x}}, \overline{y}, \overline{\overline{y}}$;

- one translates $B(\text{-}, \text{-}, \text{-})$ and $D(\text{-}, \text{-}; \text{-}, \text{-})$, inside Φ^*, into algebraic relations involving the coordinates of points.

One can achieve that the sentence Φ be true if and only if Φ^* is true; thus a decision problem for geometry gets reduced to elementary algebra. (Tarski proposes also a complete axiomatization for elementary plane geometry and, more generally, for n-dimensional Euclidean geometries [18, 19]).

A first limitation to extensions of Tarski's theories by real functions stems from the fact that by extending elementary algebra with the function $\sin x$ one disrupts its decidability [17] (in fact, by resorting to the periodicity of that function, one can define within Tarski's theory the predicate "x is an integer").

The existential theory of reals, extended with the numbers $\log 2$, π and with the functions e^x and $\sin x$ turns out to be, by itself, undecidable (Richardson, [15]).

In fact, let E^* be a set of real-valued functions (at least partially defined) of one real argument, which is closed relative to addition, subtraction, multiplication, and function composition, and which contains the identity function and all rational numbers (seen, here, as constant functions). Moreover, let E be a set of formal expressions, each one representing a function belonging to E^* so that every function in E^* is represented by at least one expression in E (if $A \in E$, we indicate by $A(x)$ the corresponding function in E^*). Suppose, also, that through an effective procedure one can, given expressions A and B in E, find expressions in E which represent the functions $A(x) + B(x)$, $A(x) - B(x)$, $A(x) \cdot B(x)$, and $A(B(x))$. Richardson proves that if E^* comprises the functions e^x, $\sin x$ and the constant functions $\log 2$, π, then the *negative value* problem "given an expression A in E, determine whether or not there is a real number $\boldsymbol{x}$ such that $A(\boldsymbol{x}) < 0$" is undecidable. Let us suppose, for the sake of contradiction, that the existential theory of reals extended with the numbers $\log 2$, π and with the functions e^x, $\sin x$ is decidable. Then, in particular, one could decide of any given sentence $(\exists x)\vartheta$, where ϑ is a quantifier-free formula of elementary algebra extended with the numbers $\log 2$, π and with the functions e^x, $\sin x$, whether $(\exists x)\vartheta$ is true or false. This could be done, in particular, for sentences of the form $(\exists x) f(x) < 0$, where f is a real function

of the real variable x, built from $x, \log 2, \pi, e^x, \sin x$ and rational constants, by means of addition, subtraction, multiplication, and function composition. In other words, the negative-value problem would be decidable that refers to the smallest collection E^* including $\{x, \log 2, \pi, e^x, \sin x\} \cup \mathbb{Q}$ and closed relative to addition, subtraction, multiplication, and function composition; but this would conflict with what was stated earlier.

Richardson also proves, under suitable assumptions about E^*, that the *identity* problem "given an expression A in E, establish whether or not $A(x) \equiv \mathbf{0}$" (where $\mathbf{0}$ is the everywhere null function over $\mathbb{R}$) and the *integration* problem "given an expression A in E, establish whether or not there is a function f in E^* such that $f'(x) \equiv A(x)$" are undecidable (the symbol $\equiv$ indicates that the functions coincide, i.e., they share the same domain, over which they take, corresponding to the same value for the argument, equal value).

In order to prove the undecidability of these problems, Richardson exploits the existence [6] of a function of type

$$P(y, x_1, \ldots, x_n) = ay + b_1 x_1 + \ldots + b_n x_n + c_1 2^{x_1} + \ldots + c_n 2^{x_n} + d\,,$$

with $a, b_1, \ldots, b_n, c_1, \ldots, c_n, d \in \mathbb{Z}$, such that the problem "given $y \in \mathbb{N}$, establish whether or not there exist $x_1, \ldots, x_n \in \mathbb{N}$ such that $P(y, x_1, \ldots, x_n) = 0$" turns out to be undecidable. In fact, arguing by contradiction, he shows that if the negative value problem, the identity problem, or the integration problem were decidable, then through the construction of suitable "intermediate problems" the said problem could be decided too.

In what follows we will present two series of decidability (and undecidability) results about fragments of real analysis, one series having been obtained by Friedman and Seress [10, 11] (concerning what we will simply designate as FS theory), and the other by Cantone, Cincotti, Ferro, Gallo, Omodeo, and Schwartz in [2, 3] (RMCF, RMCF$^+$, and RDF theories).

The FS theory consists of sentences of type $(\forall f \in \mathscr{F})\varphi$, where $\mathscr{F}$ is a family of monadic functions from $\mathbb{R}$ to $\mathbb{R}$ (respectively, from $\mathbb{I} = [0, 1]$ to $\mathbb{I}$) and φ is a first-order sentence involving, besides the function symbol f, variables ranging over $\mathbb{R}$ (resp., over $\mathbb{I}$), the comparison signs $>, <$, and $=$, the usual connectives $\wedge, \vee, \neg$, and $\exists/\forall$–quantifiers.

As for RMCF, RMCF$^+$, and RDF, these are unquantified theories involving real-valued variables (and constants), additional variables (and constants) to be interpreted as real-valued functions of a real argument, also involving operations between numbers and between functions, the ordering relations and predicate symbols for comparing functions, for comparing function derivatives and real numbers, predicates stating (strict and non-strict) function monotonicity, and predicates stating (strict and non-strict) convexity and concavity of functions over real intervals.

The style of our presentation will be rather casual; in the sense that it will privilege conceptual aspects over technical ones—without neglecting the

latter whenever deemed necessary. We will strive to bring into evidence the expressiveness of the theories presented by casting inside them various theorems of elementary analysis; thus, in the case of decidable theories, our examples will entail the possibility of proving certain theorems automatically.

1. The FS theory

To begin our discussion on the FS theory, we must recall a common classification of quantified sentences (i.e., formulae devoid of free variables) in a first-order theory. One defines a sentence φ to be Σ_k when it is either of the *prenex* type

$$(\exists x_{1,1}\cdots\exists x_{1,m_1})(\forall x_{2,1}\cdots\forall x_{2,m_2})\cdots \\ \cdots(\forall x_{k-1,1}\cdots\forall x_{k-1,m_{k-1}})(\exists x_{k,1}\cdots\exists x_{k,m_k})\varphi_0$$

(where φ_0 is quantifier-free) with k an odd number, or of the prenex type

$$(\exists x_{1,1}\cdots\exists x_{1,m_1})(\forall x_{2,1}\cdots\forall x_{2,m_2})\cdots \\ \cdots(\exists x_{k-1,1}\cdots\exists x_{k-1,m_{k-1}})(\forall x_{k,1}\cdots\forall x_{k,m_k})\varphi_0$$

(where φ_0 is devoid of quantifiers again) with k an even number; that is, if the prenex normal form of φ, in which all quantifiers have been brought to the beginning, alternates $k-1$ times between batches of existential and universal quantifiers and shows an $\exists$-quantifier at its very start. The definition of Π_k sentences is analogous, but in this case a $\forall$-quantifier occurs first.

1.1. Decidability of Σ_1 sentences, of Π_1 sentences, and of Π_2 separated sentences of FS

As already recalled, the sentences in the FS theory are of type

$$(\forall f \in \mathscr{F})\varphi\,,$$

where $\mathscr{F}$ is a family of functions from $\mathbb{R}$ to $\mathbb{R}$ (respectively, from $\mathbb{I}$ to $\mathbb{I}$) and φ is a first-order sentence involving the monadic function symbol f, individual variables ranging over $\mathbb{R}$ (resp., over $\mathbb{I}$), the dyadic comparators $>$, $<$, $=$, the propositional connectives $\wedge, \vee, \neg$, and $\exists/\forall$–quantifiers.

In our study on decidability, we first address the case in which φ is Σ_1 (to wit, φ is of type $\exists x_1\cdots\exists x_n\varphi_0$, where φ_0 is quantifier-free). We will see, in particular, that if $\mathscr{F}$ is formed by all continuous functions from $\mathbb{R}$ to $\mathbb{R}$ (or from $\mathbb{I}$ to $\mathbb{I}$), then the Σ_1 sentences are decidable; but the same is known to hold for the family of all functions from $\mathbb{R}$ to $\mathbb{R}$ (or from $\mathbb{I}$ to $\mathbb{I}$) which are differentiable, for those which are of class C^∞, and for the analytic functions.

Observe, in the first place, that the Σ_1 sentences admit an equivalent normalized form, according to the following lemma:

LEMMA 1.1 ([10, Section 1, Lemma 1.1]). *Let φ be the Σ_1 sentence $\exists x_1 \cdots \exists x_n \varphi_0$, where φ_0 is quantifier-free. Then φ is equivalent to a sentence ψ of the form*

$$\exists x_1 \cdots \exists x_p \left(\bigvee_{i=1}^{m} \left[\left(\bigwedge_{j=1}^{k_i - 1} (x_j < x_{j+1}) \right) \wedge \psi_i \right] \right) ,$$

where each ψ_i has the form $\bigwedge_{j=1}^{\ell_i} (f(x_{a_j}) = x_{b_j})$ with

(a) $1 \leqslant a_j \leqslant k_i$ and $1 \leqslant b_j \leqslant k_i$ for each j,

(b) every variable x_c ($1 \leqslant c \leqslant k_i$) occurs at least once as either x_{a_j} or x_{b_j},

(c) every variable x_c occurs at most once as x_{a_j}.

Moreover, by means of a suitable algorithm it is possible to get ψ from φ in a finite number of steps. The case $m = 0$ reflects the impossibility of having a coherent ordering for the variables of φ.

The algorithm is based on techniques such as transformation into disjunctive normal form, introduction of new variables, review of all possible orderings of the variables, and renumbering of variables.

As regards complexity, let us observe that, at least in principle, the application of this lemma could lead to a combinatorial explosion. Suffice it to say that, given r variables $x_1, ..., x_r$, the number of possible chains with the ordering $<$, with possible identifications of some variables through the equivalence relation $=$, is of order $r! \cdot r \cdot e^r$ ([2, p. 775]).

The following holds for the sentences on which we are focusing, when $\mathscr{F}$ is the family of all continuous functions from $\mathbb{R}$ to $\mathbb{R}$:

PROPOSITION 1.2 (Characterization theorem, cf. [10, Section 1, Theorem 1.3]). *Let $\mathscr{F}$ be the set of all continuous functions from $\mathbb{R}$ to $\mathbb{R}$ and let φ be a Σ_1 sentence. Let, moreover, ψ be a Σ_1 sentence, equivalent to φ, of the form*

$$\exists x_1 \cdots \exists x_p \left(\bigvee_{i=1}^{m} \left[\left(\bigwedge_{j=1}^{k_i - 1} (x_j < x_{j+1}) \right) \wedge \psi_i \right] \right)$$

meeting all conditions stated in Lemma 1.1. Then $(\forall f \in \mathscr{F})\varphi$ is true if and only if each one of the following types of formula occurs among the ψ_i's:

(1) $\bigwedge_{j=1}^{k} (f(x_j) = x_j)$;

(2) a subset of $\bigwedge_{j=1}^{k} (f(x_j) = x_{k+1-j})$ meeting condition (b) of Lemma 1.1 (here and below, if Y is a conjunction of literals, by the locution "subset of Y" we informally refer to a conjunction of some of the literals in Y);

(3) $\bigwedge_{j=1}^{\ell}(f(x_{a_j}) = x_{b_j})$ *meeting, in addition to (b) and (c) of Lemma 1.1, the conditions*

(3.a) if $f(x_{a_j}) = x_{b_j}$ *then* $x_{a_j} < x_{b_j}$,

(3.b) if $f(x_{a_j}) = x_{b_j}$, $f(x_{a_{j'}}) = x_{b_{j'}}$, *and* $x_{a_j} < x_{a_{j'}}$, *then* $x_{b_j} < x_{b_{j'}}$;

(4) $\bigwedge_{j=1}^{\ell}(f(x_{a_j}) = x_{b_j})$ *meeting, in addition to (b) and (c) of Lemma 1.1, the conditions*

(4.a) if $f(x_{a_j}) = x_{b_j}$ *then* $x_{a_j} > x_{b_j}$,

(4.b) if $f(x_{a_j}) = x_{b_j}$, $f(x_{a_{j'}}) = x_{b_{j'}}$, *and* $x_{a_j} < x_{a_{j'}}$, *then* $x_{b_j} < x_{b_{j'}}$;

(5) either one of the types $\bigwedge_{j=1}^{k}(f(x_j) = x_n)$, $\bigwedge_{j=1,j\neq n}^{k}(f(x_j) = x_n)$, *for some* n *with* $1 \leqslant n \leqslant k$;

(6) a subset of $\bigwedge_{j=1}^{k}(f(x_j) = x_{g_j})$ *meeting condition (b) of Lemma 1.1 along with the following conditions: for some* n, *with* $1 \leqslant n \leqslant k$,

(6.a) either $g_n = n$ *and*

$$\forall j[((1 \leqslant j \leqslant n-1) \Rightarrow (n+1 \leqslant g_j \leqslant k)) \wedge ((n+1 \leqslant j \leqslant k) \Rightarrow (1 \leqslant g_j \leqslant n-1))]$$

hold, or

$$\forall j[((1 \leqslant j \leqslant n) \Rightarrow (n+1 \leqslant g_j \leqslant k)) \wedge ((n+1 \leqslant j \leqslant k) \Rightarrow (1 \leqslant g_j \leqslant n))]$$

holds,

(6.b) if $1 \leqslant j < h \leqslant k$ *then* $g_h < g_j$,

(6.c) if $1 \leqslant j \leqslant n < s \leqslant g_j$ *and* $f(x_s) = x_\ell$, *then* $j < \ell$,

(6.d) if $g_j \leqslant s \leqslant n < j \leqslant k$ *and* $f(x_s) = x_\ell$, *then* $j > \ell$;

(7) a subset of $\bigwedge_{j=1}^{k}(f(x_j) = x_{g_j})$ *meeting condition (b) of Lemma 1.1 along with the following conditions: for some* n, *with* $1 \leqslant n \leqslant k$,

(7.a) either $g_n = n$ *and*

$$\forall j[((1 \leqslant j \leqslant n-1) \Rightarrow (n+1 \leqslant g_j \leqslant k)) \wedge ((n+1 \leqslant j \leqslant k) \Rightarrow (1 \leqslant g_j \leqslant n-1))]$$

hold, or

$$\forall j[((1 \leqslant j \leqslant n) \Rightarrow (n+1 \leqslant g_j \leqslant k)) \\ \wedge ((n+1 \leqslant j \leqslant k) \Rightarrow (1 \leqslant g_j \leqslant n))]$$

holds,

(7.b) *if* $1 \leqslant j < h \leqslant k$ *then* $g_h < g_j$,

(7.c) *if* $1 \leqslant j \leqslant n$, $g_j \leqslant s \leqslant k$, *and* $f(x_s) = x_\ell$, *then* $j \geqslant \ell$ *(where equality can hold only if* $j = g_j = s = \ell = n$*),*

(7.d) *if* $n+1 \leqslant j \leqslant k$, $1 \leqslant s \leqslant g_j$, *and* $f(x_s) = x_\ell$, *then* $j < \ell$*;*

(8) for some n *with* $1 \leqslant n \leqslant k$, *a subset of*

$$\bigwedge_{j=1}^{n} (f(x_j) = x_n) \wedge \bigwedge_{j=n+1}^{k} (f(x_j) = x_{g_j})$$

meeting condition (b) of Lemma 1.1 along with the conditions

(8.a) *if* $n+1 \leqslant j \leqslant k$ *then* $1 \leqslant g_j < n$,

(8.b) *if* $n+1 \leqslant j < h \leqslant k$ *then* $g_j > g_h$*;*

(9) for some n *with* $1 \leqslant n \leqslant k$, *a subset of*

$$\bigwedge_{j=1}^{n-1} (f(x_j) = x_{g_j}) \wedge \bigwedge_{j=n}^{k} (f(x_j) = x_n)$$

meeting condition (b) of Lemma 1.1 along with the conditions

(9.a) *if* $1 \leqslant j \leqslant n-1$ *then* $n < g_j \leqslant k$,

(9.b) *if* $1 \leqslant j < h \leqslant n-1$ *then* $g_j > g_h$.

Notice that a ψ_i can belong to more than one type. For instance, the formula $f(x_1) = x_1$ is of types (1), (2), (5), (6), (7), (8), (9).

Here we offer some clues about the necessity of the above conditions. If φ is true of all continuous functions from $\mathbb{R}$ to $\mathbb{R}$, then, since ψ is equivalent to φ, ψ is true of all continuous functions from $\mathbb{R}$ to $\mathbb{R}$. Therefore ψ will be satisfied, in particular, by the function $f(x) = x$; this implies that there must be a ψ_i of type (1). Likewise ψ must be true, in particular, of the function $f(x) = -x$; this implies that there must be a ψ_i of type (2). By choosing suitable functions for the remaining types, in the same fashion, one proves that the ψ_i's must include at least one formula of each type.

The proof that the above conditions are also sufficient is more intricate. To show that if ψ encompasses all nine types then ψ is true of all continuous f (from $\mathbb{R}$ to $\mathbb{R}$), one takes into account all possibilities about the number of fixpoints which a given f can own (none, exactly one, a finite number greater than one, infinitely many). One proves that in each case f falls under at least one of the nine types, and hence it satisfies ψ. Consider, e.g., the simplest case, namely the one of an f with infinitely many fixpoints: then, given a positive integer k, there must exist $x_1, \cdots, x_k \in \mathbb{R}$ such that $x_1 < \cdots < x_k$ and $f(x_1) = x_1, \ldots, f(x_k) = x_k$; therefore f satisfies the ψ_i's of type (1) and the sentence ψ.

Let us observe that through application of the preceding lemma and proposition one can decide by means of an algorithm whether each given sentence $(\forall f \in \mathscr{F})\varphi$ is true or false; otherwise stated, these results provide an automatic proof-procedure for statements of this nature.

To illustrate application of the preceding proposition, let us examine a simple example:

EXAMPLE 1.3. *Consider the sentence*

$$(\forall f \in \mathscr{F})\exists x\, \exists y(f(x) = y)\,,$$

which can be interpreted as claiming "for every continuous function f from $\mathbb{R}$ to $\mathbb{R}$ there exist $x, y \in \mathbb{R}$ such that $f(x) = y$". In this case φ is the Σ_1 sentence

$$\exists x\, \exists y(f(x) = y)\,,$$

equivalent to

$$\exists x_1\, \exists x_2[(x_1 < x_2 \wedge f(x_1) = x_2) \vee (f(x_1) = x_1) \vee (x_1 < x_2 \wedge f(x_2) = x_1)]\,.$$

The formula $(x_1 < x_2 \wedge f(x_1) = x_2)$ matches type (3), the formula $(f(x_1) = x_1)$ matches types (1), (2), (5), (6), (7), (8), (9), and the formula $(x_1 < x_2 \wedge f(x_2) = x_1)$ matches type (4). Hence all of the nine types are encompassed, which amounts to saying that the sentence $(\forall f \in \mathscr{F})\exists x, y(f(x) = y)$ is true.

The following example formalizes another lemma expressible by means of a Σ_1 sentence.

EXAMPLE 1.4. *Consider the claim "for each continuous function f from $\mathbb{R}$ to $\mathbb{R}$ there exist $x, y, z \in \mathbb{R}$, with $x < y < z$, such that either $f(x) \leqslant f(y) \leqslant f(z)$ or $f(x) \geqslant f(y) \geqslant f(z)$ holds". This can be formalized as*

$$(\forall f \in \mathscr{F})\exists x \exists y \exists z(x < y < z \wedge (f(x) \leqslant f(y) \leqslant f(z) \vee f(x) \geqslant f(y) \geqslant f(z)))$$

and hence it can be proved automatically thanks to the preceding results.

The above-seen characterization theorem concerning the family of the continuous functions (from $\mathbb{R}$ to $\mathbb{R}$) holds, with the same conditions (1) through (9), for the family of the differentiable functions (from $\mathbb{R}$ to $\mathbb{R}$), as well as for the ones of class C^∞ (from $\mathbb{R}$ to $\mathbb{R}$); this tells us, as a consequence, that if a Σ_1 sentence holds for all functions of class C^∞ from $\mathbb{R}$ to $\mathbb{R}$ then it holds, more generally, for all continuous functions from $\mathbb{R}$ to $\mathbb{R}$.

A similar characterization theorem holds for the analytic functions from $\mathbb{R}$ to $\mathbb{R}$; but in this case the claim involves only conditions (1) through (7).

Yet an analogous theorem holds for the functions (continuous, differentiable, of class C^∞, or analytic) from $\mathbb{I}$ to $\mathbb{I}$. In this case the characterization is exactly the same for all of the four collections of functions; consequently, if a Σ_1 sentence holds for all analytic functions from $\mathbb{I}$ to $\mathbb{I}$ then it holds, more generally, for all continuous functions from $\mathbb{I}$ to $\mathbb{I}$.

What said so far enables us to state the following decidability result:

PROPOSITION 1.5 (Decidability of the Σ_1 sentences of FS, cf. [10, Section 1, Theorems 1.3 through 1.6]). *The validity problem for Σ_1 sentences is solvable, relative to each one of the following families of functions from $\mathbb{R}$ to $\mathbb{R}$: continuous, differentiable, C^∞, and analytic. The same holds for the corresponding families of functions from $\mathbb{I}$ to $\mathbb{I}$.*

Otherwise stated: let $\mathscr{F}$ be the family of all continuous functions (or the one of the differentiable functions, or of the functions of class C^∞, or of the analytic functions) from $\mathbb{R}$ to $\mathbb{R}$. Then an algorithm exists which, given any sentence $(\forall f \in \mathscr{F})\varphi$, where φ is Σ_1, establishes whether it is true or false. The same holds about $\mathbb{I}$.

Let us now address the decidability problem for the $(\forall f \in \mathscr{F})\varphi$ sentences of FS where φ is a Π_1 sentence (namely, φ is of the form $\forall x_1 \cdots \forall x_n \varphi_0$, with φ_0 quantifier-free). Focusing, for the time being, on the case when $\mathscr{F}$ is the family of all continuous functions from $\mathbb{R}$ to $\mathbb{R}$, we have:

$$(\forall f \in \mathscr{F})\forall x_1 \cdots \forall x_n \, \varphi_0$$

is true if and only if its negation

$$(\exists f \in \mathscr{F})\exists x_1 \cdots \exists x_n \, \chi_0 \,,$$

where $\chi_0 = \neg\varphi_0$, is false. This happens if and only if the sentence, to be referred below as γ,

$$(\exists f \in \mathscr{F})\exists x_1 \cdots \exists x_n \left(\bigvee_{i=1}^{m} \left[\left(\bigwedge_{j=1}^{k_i - 1} (x_j < x_{j+1}) \right) \wedge \psi_i \right] \right) ,$$

which results from application of Lemma 1.1 to χ_0, is false. This happens if and only if $m = 0$. In fact, if $m = 0$ then, as already said in the claim of Lemma 1.1,

the variables of χ_0 do not admit a coherent ordering, and therefore γ is false. If, on the opposite, $m \geqslant 1$ holds, then it is possible (by assigning suitable values to the variables and by choosing a suitable interpolation polynomial $\boldsymbol{f}$ as f) to determine $\boldsymbol{x}_1, \ldots, \boldsymbol{x}_n$ and $\boldsymbol{f}$ so that they satisfy $(\bigwedge_{j=1}^{k_1-1}(x_j < x_{j+1})) \wedge \psi_1$; in particular, it suffices to assign values $\boldsymbol{x}_i = i$ $(i = 1, \ldots, n)$ to the variables and to choose as f a polynomial $\boldsymbol{f}$ such that $\boldsymbol{f}(a_j) = b_j$ whenever $f(x_{a_j}) = x_{b_j}$ occurs in ψ_1. Therefore, if $m \geqslant 1$, then γ is true.

What said so far entails a decision procedure for the case of the Π_1 sentences. Analogous considerations can be made if $\mathscr{F}$, instead of being the family of all continuous functions from $\mathbb{R}$ to $\mathbb{R}$, is either the family of all differentiable functions (from $\mathbb{R}$ to $\mathbb{R}$), the one of all functions of class C^∞ (from $\mathbb{R}$ to $\mathbb{R}$), or the one of all analytic functions (from $\mathbb{R}$ to $\mathbb{R}$). The same considerations can be made again for the corresponding families of functions from $\mathbb{I}$ to $\mathbb{I}$.

We hence get the following decidability result:

PROPOSITION 1.6 (Decidability of the Π_1 sentences of FS, cf. [10, Section 1, Theorem 1.7]). *The validity problem for Π_1 sentences is solvable, relative to each one of the following families of functions: continuous, differentiable, C^∞, and analytic. The same holds for the corresponding families of functions from $\mathbb{I}$ to $\mathbb{I}$.*

Otherwise stated: let $\mathscr{F}$ be the family of all continuous functions (or the one of the differentiable functions, or of the functions of class C^∞, or of the analytic functions) from $\mathbb{R}$ to $\mathbb{R}$. Then an algorithm exists which, given any sentence $(\forall f \in \mathscr{F})\varphi$, where φ is Π_1, establishes whether it is true or false. The same holds for $\mathbb{I}$.

Notice also that, since the characterization for all of them is the same ($m = 0$ in the sentence obtained from $\neg\varphi_0$ through application of Lemma 1.1), it turns out that these families of functions are indistinguishable relative to the Π_1 sentences; among others, a Π_1 sentence is true for all continuous functions from $\mathbb{R}$ to $\mathbb{R}$ if and only if it is true for all analytic functions from $\mathbb{I}$ to $\mathbb{I}$.

The following example formalizes a lemma (good definition of a function) expressible by means of a Π_1 sentence.

EXAMPLE 1.7. *Consider the theorem "let f be a continuous function from $\mathbb{R}$ to $\mathbb{R}$ and let $x, y, z \in \mathbb{R}$; if $f(x) = y$ and $f(x) = z$, then $y = z$". This can be formalized as*

$$(\forall f \in \mathscr{F})\forall x \forall y \forall z((f(x) = y \wedge f(x) = z) \rightarrow y = z)$$

and therefore it can be proved automatically, thanks to the preceding results (recall that the derived connective $\rightarrow$, exploited in the formalization of this sentence, can be eliminated, e.g., through the rewriting $a \rightarrow b \equiv \neg(a \wedge \neg b)$).

Let us now introduce the notion of *separated formula*. Intuitively speaking, we are talking about formulae in which the elements of the domain of f are not compared with those of its range. To state this more accurately:

DEFINITION 1.8. *Let φ_0 be a quantifier-free formula involving a monadic function f along with variables ranging over $\mathbb{R}$ (resp., over $\mathbb{I}$), the comparators $>$, $<$, $=$, and the usual connectives $\wedge, \vee, \neg$.*

We will say that φ_0 is a SEPARATED FORMULA *if it meets the following conditions:*

(a) The terms of φ_0 are of either the form x or the form $f(x)$, where x is a variable (i.e., no composition of f with itself occurs in φ_0).

(b) There are two sets, formed by variables of φ_0 and to be called set of the DOMAIN VARIABLES *and of the* RANGE VARIABLES*, respectively, such that:*

- *(b1) every variable of φ_0 belongs to exactly one of the two sets;*
- *(b2) if the term $f(x)$ occurs in φ_0, then x is a domain variable;*
- *(b3) when $f(x) > y$, $f(x) < y$, or $f(x) = y$ occurs as a subformula in φ_0, then y is a range variable;*
- *(b4) when $x > y$, $x < y$, or $x = y$ occurs as a subformula in φ_0, then x and y are either both domain variables or both range variables (that is, a domain variable is never compared with a range variable).*

To end, we will say that a sentence φ in prenex form is SEPARATED *when its unquantified part is a separated formula.*

For instance, the sentence $\exists x(f(x) = x)$ is not separated (if it were such then, due to the conditions (b2) and (b3), x would be both domain variable and range variable, which would conflict with condition (b1)).

The sentence $\exists x \exists y(f(x) = y)$ is, instead, separated (with x domain variable and y range variable).

For the $(\forall f \in \mathscr{F})\varphi$ sentences of the theory FS, when φ is a Π_2 separated sentence (i.e., a sentence of the form $\forall x_1 \cdots \forall x_n \exists x_{n+1} \cdots \exists x_m \varphi_0$, with φ_0 devoid of quantifiers and separated), then the following decidability result holds:

PROPOSITION 1.9 (Decidability of the separated Π_2 sentences of FS, cf. [10, Section 2]). *The validity problem for separated Π_2 sentences is solvable, relative to the following families of functions from $\mathbb{R}$ to $\mathbb{R}$: continuous, differentiable, C^∞, and analytic. The same holds for the corresponding families of functions from $\mathbb{I}$ to $\mathbb{I}$.*

Otherwise stated: let $\mathscr{F}$ be the family of all continuous functions (or the one of all differentiable functions, or the one of all functions of class C^∞, or the one of all analytic functions) from $\mathbb{R}$ to $\mathbb{R}$. Then there is an algorithm

which, given any sentence $(\forall f \in \mathscr{F})\varphi$*, where* φ *be a separated* Π_2 *sentence, establishes whether it is true or false. The same holds for* $\mathbb{I}$*.*

Also in this case, the decidability of sentences is obtained through a normalization lemma with the aid of characterization theorems.

The following example shows how the *intermediate value theorem* can be formalized by means of a separated Π_2 sentence.

EXAMPLE 1.10. *Consider the (intermediate value) theorem:*
"Let f *be a continuous function from* $\mathbb{R}$ *to* $\mathbb{R}$ *and let* $x_1, x_2, y_1, y_2, t \in \mathbb{R}$ *be such that* $f(x_1) = y_1$*,* $f(x_2) = y_2$ *and* $y_1 \leqslant t \leqslant y_2$*. Then there is a* $z \in \mathbb{R}$ *such that* $x_1 \leqslant z \leqslant x_2$ *and* $f(z) = t$*".*
This claim can be formalized as

$$(\forall f \in \mathscr{F})\forall x_1\forall x_2\forall y_1\forall y_2\forall t\exists z\big((f(x_1) = y_1 \wedge f(x_2) = y_2 \wedge y_1 \leqslant t \leqslant y_2) \rightarrow (x_1 \leqslant z \leqslant x_2 \wedge f(z) = t)\big)$$

and hence it can be proved automatically.

1.2. Undecidability of Σ_4 sentences

Indicate, as usual, by $\omega = \{0, 1, \ldots, n, n+1, \ldots\}$ the set of all finite ordinal numbers (where $0 = \emptyset$ and $n + 1 = \{0, \ldots, n\}$); also let $n \in \omega$. A dyadic antireflexive and symmetrical relation (on n) is a subset R of $n \times n$ which meets the following conditions (where aRb stands for $(a, b) \in R$):

antireflexivity if aRb then $a \neq b$;

symmetry if aRb then bRa.

The first-order theory of antireflexive and symmetrical relations with finite models (finite graph theory, to be indicated as GSF) is the set of all sentences φ_R, constructed from the variables (now ranging over natural numbers), by means of the dyadic predicate symbol R (to be interpreted as an antireflexive and symmetric relation), the identity relator $=$, the propositional connectives $\wedge, \vee, \neg$, and the $\exists/\forall$–quantifiers.

The validity problem for the Σ_2 sentences of this theory is undecidable [13]. Specifically, there cannot be any algorithm which, given a generic sentence of type $(\forall R)\varphi_R$ (where φ_R is a Σ_2 sentence of the GSF theory), establishes whether it is true or false.

As a matter of fact, there is an algorithm which associates with every Σ_2 sentence φ_R of the GSF theory a separated Σ_4 sentence φ of the FS theory so that $(\forall R)\varphi_R$ is true if and only if $(\forall f \in \mathscr{F})\varphi$ is true about the family $\mathscr{F}$ of all continuous functions from $\mathbb{R}$ to $\mathbb{R}$.

Consequently, if the truth problem for $(\forall f \in \mathscr{F})\varphi$ sentences (where φ is a separated Σ_4 sentence in FS and $\mathscr{F}$ is the family of all continuous functions from $\mathbb{R}$ to $\mathbb{R}$) were decidable, then the analogous problem for $(\forall R)\varphi_R$ sentences (where φ_R is a Σ_2 sentence of GSF) would also be decidable, which is not the case as just recalled above.

Therefore the truth problem for $(\forall f \in \mathscr{F})\varphi$ sentences, where φ is a separated Σ_4 sentence of FS and $\mathscr{F}$ is the family of all continuous functions from $\mathbb{R}$ to $\mathbb{R}$, turns out to be undecidable. This result can be generalized, much by the same method, into the following theorem:

PROPOSITION 1.11 (Undecidability of separated Π_4 sentences of FS, cf. [10, Section 4, Theorem 4.2] and [11, Section 4, Theorem 4.2]). *The set* $\{\varphi|(\forall f \in \mathscr{F})\varphi$ *is true*$\}$ *of sentences turns out to be undecidable in the following cases (where we say that a separated sentence of* FS *is* WEAK *if it has no subformulae of type* $f(x) < y$, $y < f(x)$, $f(x) < f(t)$, *or* $y < z$, *with* y, z *range variables; that is, if the ordering relation is not used, in it, to compare elements of the range of* f*).*

(a) $\mathscr{F}$ *is the family of all continuous functions from* $\mathbb{R}$ *to* $\mathbb{R}$ *and* φ *ranges over all separated* Σ_4 *sentences of* FS*;*

(b) more generally, $\mathscr{F}$ *is a family of functions from* $\mathbb{R}$ *to* $\mathbb{R}$ *comprising all analytic functions and* φ *ranges over the separated, weak* Σ_4 *sentences of* FS*;*

(c) $\mathscr{F}$ *is the family of all continuous functions from* $\mathbb{I}$ *to* $\mathbb{I}$ *and* φ *ranges over all separated* Σ_4 *sentences of* FS*;*

(d) more generally, $\mathscr{F}$ *is a family of functions from* $\mathbb{I}$ *to* $\mathbb{I}$ *comprising all polynomials and* φ *ranges over all separated, weak* Σ_4 *sentences of* FS*.*

On the other hand, the said set $\{\varphi|(\forall f \in \mathscr{F})\varphi$ is true$\}$ of sentences, where $\mathscr{F}$ is the family of all polynomials from $\mathbb{R}$ to $\mathbb{R}$ (resp., from $\mathbb{I}$ to $\mathbb{I}$) and φ ranges over all sentences of FS, turns out to be *co-recursively enumerable* (cf. [11, Section 4, Theorem 4.7]). Otherwise stated, there exists a computing procedure which eventually halts if and only if a sentence of the said type is submitted to it which happens to be false.

1.3. Decidability and undecidability of sentences about families of monotone functions

Let us now consider the sentences $(\forall f \in \mathscr{F})\varphi$ of the FS theory, where $\mathscr{F}$ is the family of all functions (from $\mathbb{R}$ to $\mathbb{R}$) which are continuous, monotone strictly increasing, and unlimited below as well as above. The following lemma reduces

the decidability issue for sentences of this type to the analogous issue regarding sentences of type $(\forall A_1, A_2, A_3, A_4, A_5)\varphi^+$, where $A_1, A_2, A_3, A_4, A_5 \subseteq \mathbb{R}$ and φ^+ is a sentence involving real-valued variables, the comparators $<, =$, the usual connectives $\wedge, \vee, \neg$, $\exists/\forall$–quantifiers, and predicates of type $x \in A_i$. The latter was solved in the positive, cf. [1].

LEMMA 1.12 ([10, Section 3, Lemma 3.5]). *To each sentence φ there corresponds a sentence φ^+ for which the following sentences are logically equivalent.*

(a) $(\forall f \in \mathscr{F})\varphi$, *where $\mathscr{F}$ is the family of all functions (from $\mathbb{R}$ to $\mathbb{R}$) which are continuous, monotone strictly increasing and unlimited below as well as above.*

(b) $(\forall A_1, A_2, A_3, A_4, A_5)\varphi^+$, *where $A_1, A_2, A_3, A_4, A_5 \subseteq \mathbb{R}$ and φ^+ is a sentence that involves variables ranging over $\mathbb{R}$, the comparators $<, =$, the propositional connectives $\wedge, \vee, \neg$, $\exists/\forall$–quantifiers, and predicates of type $x \in A_i$.*

Such a φ^+ can be obtained from φ through a suitable algorithm.

Here we will content ourselves with providing the intuitive idea, lying behind this lemma, that the first-order properties of a function f (which is continuous, monotone strictly increasing, and unlimited below as well as above) can be expressed as properties of sets, which are defined starting from the function (for instance, the set $\alpha(f)$ of all fixpoints of f and the set $\beta(f)$ of all left endpoints of the intervals of $\mathbb{R}\backslash\alpha(f)$).

This lemma yields, in view of the decidability of $(\forall A_1, A_2, A_3, A_4, A_5)\varphi^+$ sentences, decidability of the $(\forall f \in \mathscr{F})\varphi$ sentences of the FS theory (where $\mathscr{F}$ is the family of all functions from $\mathbb{R}$ to $\mathbb{R}$ which are continuous, monotone strictly increasing and unlimited below as well as above). This decidability result can be enhanced, much by the same method, into the following proposition:

PROPOSITION 1.13 ([10, Section 3]; [11, Sections 2 and 3]). *The set $\{\varphi|(\forall f \in \mathscr{F})\varphi$ is true$\}$ of sentences turns out to be decidable in the following cases.*

(a) $\mathscr{F}$ is the family of all functions from $\mathbb{R}$ to $\mathbb{R}$ which are continuous, monotone strictly increasing and unlimited below as well as above.

(b) $\mathscr{F}$ is the family of all functions from $\mathbb{R}$ to $\mathbb{R}$ which are continuous and monotone strictly increasing.

(c) $\mathscr{F}$ is the family of all functions from $\mathbb{R}$ to $\mathbb{R}$ which are continuous and monotone strictly decreasing.

(d) $\mathscr{F}$ is the family of all functions from $\mathbb{R}$ to $\mathbb{R}$ which are continuous and strictly monotone.

(e) *$\mathscr{F}$ is the family of all functions from $\mathbb{R}$ to $\mathbb{R}$ which are monotone nondecreasing, such that there are at most n intervals on which each of them is constant, and each of them has at most n discontinuity points (where n is a fixed number in $\mathbb{N}$).*

(f) *$\mathscr{F}$ is the family of all functions from $\mathbb{I}$ to $\mathbb{I}$ which are monotone nondecreasing, such that there are at most n intervals on which each of them is constant, and each of them has at most n discontinuity points (where n is a fixed number in $\mathbb{N}$).*

(g) *$\mathscr{F}$ is the family of all functions from $\mathbb{R}$ to $\mathbb{R}$ which are monotone and φ is a separated sentence (as by the definition seen earlier).*

The following example formalizes the property of a function from $\mathbb{R}$ to $\mathbb{R}$, continuous and monotone strictly decreasing, of having exactly one fixpoint.

EXAMPLE 1.14. *Consider the claim:*
"Let f be a function from $\mathbb{R}$ to $\mathbb{R}$, continuous and monotone strictly decreasing. Then there exists exactly one $x \in \mathbb{R}$ such that $f(x) = x$."
This claim can be formalized as

$$(\forall f \in \mathscr{F})\exists x \forall y[f(x) = x \wedge (f(y) = y \rightarrow y = x)]\,,$$

where $\mathscr{F}$ is the family of all functions from $\mathbb{R}$ to $\mathbb{R}$ which are continuous and monotone strictly decreasing. Therefore this theorem can be proved automatically.

On the opposite, decidability gets lost if one takes, as $\mathscr{F}$, the family of all functions from $\mathbb{R}$ to $\mathbb{R}$ which are continuous and monotone and have an arbitrarily large finite number of intervals on which they are constant.

As a matter of fact, given a Turing machine T endowed with symbols $\{a_0, \ldots, a_h\}$ (where a_0 stands for the *blank*) and states $\{q_0, q_1, \ldots, q_k\}$ (where q_0 is the initial state and q_1 is the final state), it is possible to construct a sentence $\varphi(T)$ such that $(\exists f \in \mathscr{F})\varphi(T)$ is true if and only if the machine T, starting with an empty tape, halts after a finite number of steps. Since $(\exists f \in \mathscr{F})\varphi(T)$ is true if and only if $(\forall f \in \mathscr{F})\neg\varphi(T)$ is false, if the truth of the $(\forall f \in \mathscr{F})\varphi$ sentences were decidable, then the truth of the $(\exists f \in \mathscr{F})\varphi(T)$ sentences would also be decidable, and therefore the problem "T will halt" would turn out to be such; however, as is well-known, the halting problem is undecidable [20].

This argument can be adjusted to all families of functions $\mathscr{F}$ (either from $\mathbb{R}$ to $\mathbb{R}$ or from $\mathbb{I}$ to $\mathbb{I}$) which include all nondecreasing monotone functions of class C^∞ and have any finite number of intervals where they are constant. The same holds for the family of all functions from $\mathbb{R}$ to $\mathbb{R}$ which are monotone, continuous on the left, and have an arbitrary finite number of discontinuity points. Hence we have the following undecidability result:

PROPOSITION 1.15 ([11, Section 1]). *The set of all* $\{\varphi|(\forall f \in \mathscr{F})\varphi$ *is true*$\}$ *sentences has, in each of the following cases, an unsolvable decision problem (case (b) generalizes case (a)).*

(a) $\mathscr{F}$ *is the family of all functions from* $\mathbb{R}$ *to* $\mathbb{R}$ *(resp., from* $\mathbb{I}$ *to* $\mathbb{I}$*) which are continuous and monotone and have an arbitrary, though finite, number of intervals over which they are constant.*

(b) $\mathscr{F}$ *is a family of functions from* $\mathbb{R}$ *to* $\mathbb{R}$ *(resp., from* $\mathbb{I}$ *to* $\mathbb{I}$*) containing all nondecreasing monotone functions of class* C^∞ *which have an arbitrary finite number of intervals over which they are constant.*

(c) $\mathscr{F}$ *is the family of all functions from* $\mathbb{R}$ *to* $\mathbb{R}$ *which are strictly monotone, continuous on the left, and have an arbitrary finite number of discontinuity points.*[2]

Nevertheless, the set $\{\varphi|(\forall f \in \mathscr{F})\varphi$ is true$\}$ of sentences, where $\mathscr{F}$ is the family of all functions from $\mathbb{R}$ to $\mathbb{R}$ (resp., from $\mathbb{I}$ to $\mathbb{I}$) which are monotone nondecreasing and have an arbitrary finite number of intervals over which they are constant and an arbitrary finite number of discontinuity points, turns out to be *co-recursively enumerable* (cf. [11, Section 3, Corollary 3.6]). In other words, there exists a computing procedure which eventually halts if and only if a sentence of the said type is initially submitted to it which happens to be false.

2. The theories RMCF, RMCF$^+$, and RDF

As said in the introduction, Tarski's elementary algebra is decidable; i.e., there is an algorithm telling one, of any given closed formula Φ of this theory, whether Φ is true or false. As recalled there, Tarski's elementary algebra is the first-order theory supplying a denumerable infinity of real-valued variables, the numerical constants 0, 1, -1 (interpreted as the corresponding real numbers), the operations $+$, $-$, and $\cdot$ (designating the familiar arithmetic operations over $\mathbb{R}$), the standard comparators $>$, $<$, and $=$, the propositional connectives $\wedge$, $\vee$, and $\neg$, and the quantifiers $\exists$ and $\forall$.

The decidability of Tarski's elementary algebra readily entails the decidability of its own existential sub-theory, consisting of all statements of the form

$$\exists x_1 \exists x_2 \cdots \exists x_n \vartheta\,,$$

where ϑ is quantifier-free and involves only variables from among $x_1, x_2, \ldots, x_n$.

---

[2] With regard to item (c), [11] does not discuss the case of functions from $\mathbb{I}$ to $\mathbb{I}$.

The existential theory of reals can be thought of as a quantifier-free language. For, a prenex sentence $\exists x_1 \exists x_2 \cdots \exists x_n \vartheta$ is true if and only if its unquantified matrix ϑ is satisfiable, and hence any truth-decision algorithm for the existential theory of reals can be used also to solve the satisfiability problem for the corresponding theory devoid of quantifiers.

The fragments of real analysis RMCF, RMCF$^+$, and RDF, which will be reviewed in this section, are in quantifier-free form. They extend the quantifier-free theory of reals with various predicates over real functions of a real variable. More specifically, the theories RMCF and RMCF$^+$ deal with continuous functions, whereas the theory RDF refers to differentiable functions with a continuous derivative.

We begin with a brief description of the theory RMCF. Later we will review in some detail RMCF$^+$, and will also give a brief outline of the theory RDF.

The theory RMCF (of Reals with Monotone and Convex Functions) [3] involves predicates for function comparison, and predicates about monotonicity of functions (strict and non-strict), and about concavity and convexity of functions (only non-strict). The atomic formulae of RMCF are of these forms:

$$\begin{array}{ll} t_1 = t_2\,, & t_1 > t_2\,, \\ F_1 = F_2\,, & F_1 > F_2\,, \\ \mathsf{Up}(F)_{[t_1,t_2]}\,, & \mathsf{Strict\_Up}(F)_{[t_1,t_2]}\,, \\ \mathsf{Down}(F)_{[t_1,t_2]}\,, & \mathsf{Strict\_Down}(F)_{[t_1,t_2]}\,, \\ \mathsf{Convex}(F)_{[t_1,t_2]}\,, & \mathsf{Concave}(F)_{[t_1,t_2]}\,. \end{array}$$

Here t_1, t_2 are numerical expression (involving real variables, the real constants 0, 1, function images of numerical expressions, and the arithmetic operations) and F_1, F_2 are functional expressions (involving function variables and constants and the operations of sum and difference between functional expressions). The functional constants are $\mathbf{0}$, $\mathbf{1}$, interpreted as the functions with fixed values 0 and 1, respectively. Function symbols are interpreted as continuous real functions of a real variable having as their domain the whole real axis $\mathbb{R}$. The predicate $F_1 = F_2$ (resp., $F_1 > F_2$) states that the real functions $\boldsymbol{f}_1$ and $\boldsymbol{f}_2$ interpreting the expressions F_1 and F_2 coincide over the whole real axis (resp., $\boldsymbol{f}_1(x) > \boldsymbol{f}_2(x)$ holds for all $x \in \mathbb{R}$). The predicate symbols express monotonicity (strict or non-strict), non-strict convexity, and non-strict concavity of functions; each of them refers to a closed bounded interval $[t_1, t_2]$. The formulae of RMCF result from propositional combinations of atomic formulae by means of the connectives $\neg$, $\wedge$, $\vee$, $\rightarrow$, $\leftrightarrow$. As said, explicit quantification is not allowed in RMCF formulae.

The above considerations could easily be formalized in a definition of the (RMCF) interpretations of formulae of RMCF. We say that an RMCF formula ϑ is *satisfiable* if there exists an RMCF interpretation (*real model*) of the symbols of ϑ which makes ϑ true. We say that an RMCF formula ϑ is *valid* (or is a *theorem*) if ϑ is true in all RMCF interpretations.

As shown in [3], there is a decision procedure which determines, for any given RMCF formula, whether it is satisfiable or not. Such a procedure is achieved through satisfiability-preserving transformations which reduce the satisfiability problem for RMCF to the satisfiability problem for Tarski's theory of reals.[3] To prove the correctness of these formula transformations, function variables are interpreted as piecewise linear functions. In addition, since a formula is valid if and only if its negation is unsatisfiable, the same algorithm tells one whether a given RMCF formula is valid or not; hence one can fully mechanize recognition of any theorem expressible in RMCF.

In [3], a variant of the theory RMCF in which function variables are interpreted as multivariate continuous real functions is also studied and a decision procedure is provided for it.

As an ending remark, note that Proposition 1.6 about the Π_1-decidability of FS, to the extent to which it refers to continuous real functions of one real variable defined all over $\mathbb{R}$, readily follows from the decidability of RMCF.

2.1. The theory RMCF$^+$

The theory RMCF$^+$ [2] (cf. also [16, pp. 165–177]) is an extension of RMCF with predicates on strict convexity and concavity of real continuous functions of a real variable. In addition, most of the predicates on functions apply both to bounded and unbounded intervals.

2.1.1. Syntax of RMCF$^+$

The language of RMCF$^+$ contains

- a denumerable infinity of individual variables, called *numerical variables*, which are denoted by $x, y, z, \ldots$;
- two *numerical constants* $0, 1$;
- a denumerable infinity of *function variables*, denoted by $f, g, h, \ldots$;
- two *functional constants* $\mathbf{0}$, $\mathbf{1}$.

The language of RMCF$^+$ also includes two distinguished symbols, $-\infty, +\infty$, which are restricted to occur only within range defining parameters, as stated in the definition of atomic RMCF$^+$-formulae below.

Numerical terms are recursively defined as follows:

(a) numerical variables and the constants $0, 1$ are numerical terms;

---

[3] We will be a bit more specific on this, and also about syntax and semantics matters, in the next section, in the context of the extension RMCF$^+$ of RMCF.

(b) if t_1, t_2 are numerical terms, so are $(t_1 + t_2), (t_1 - t_2)$, and $(t_1 \cdot t_2)$;

(c) if t is a numerical term and f is a function variable, then $f(t)$ is a numerical term.

Functional terms are recursively defined as follows:

(a) function variables and the functional constants $\mathbf{0}$, $\mathbf{1}$ are functional terms;

(b) if F_1, F_2 are functional terms, so are $(F_1 + F_2)$ and $(F_1 - F_2)$.

In the following, the expression *numerical variable* will be used also to denote the constants $0, 1$. Likewise, the expression *function variable* will be used also to denote the functional constants $\mathbf{0}$, $\mathbf{1}$

By *extended numerical variable* we mean a numerical variable or one of the symbols $-\infty, +\infty$. Likewise, by *extended numerical term* we mean a numerical term or one of the symbols $-\infty, +\infty$.

An atomic RMCF^+-formula is an expression having one of the following forms:

$$\begin{array}{ll} t_1 = t_2\,, & t_1 > t_2\,, \\ (F_1 = F_2)_A\,, & (F_1 > F_2)_{[t_1,t_2]}\,, \\ \mathsf{Up}(F)_A\,, & \mathsf{Strict\_Up}(F)_A\,, \\ \mathsf{Down}(F)_A\,, & \mathsf{Strict\_Down}(F)_A\,, \\ \mathsf{Convex}(F)_A\,, & \mathsf{Strict\_Convex}(F)_A\,, \\ \mathsf{Concave}(F)_A\,, & \mathsf{Strict\_Concave}(F)_A\,, \end{array}$$

where A stands for any of the following interval terms

$$[t_1, t_2], \quad [t_1, +\infty[, \quad]-\infty, t_2], \quad]-\infty, +\infty[\,,$$

t_1, t_2 are numerical terms, and F, F_1, F_2 are functional terms.[4]

The formulae of RMCF^+ are propositional combinations of atomic formulae by means of the usual connectives $\neg, \wedge, \vee, \rightarrow, \leftrightarrow$. Let us stress again that explicit quantification is not admitted.

To ease readability, occasionally we will use abbreviations. For instance, if t_1, t_2, t_3 are numerical terms, then $t_1 = t_2/t_3$ is a shorthand for the conjunction

$$(t_2 = t_1 \cdot t_3) \wedge (\neg(t_3 = 0))\,.$$

---

[4]Notice that literals of type $F_1 > F_2$ are admitted in RMCF^+-formulae only if restricted to finite closed intervals, rather than to possibly infinite closed intervals, as is the case for all remaining literals involving functional terms. This is due to the facts that (a) the satisfiability test for RMCF^+-formulae is based on the property that any satisfiable RMCF^+-formula admits a *canonical* model M sending function variables to piecewise linear functions with *small* quadratic perturbations on finite internal intervals and *small* exponential perturbations on the two external infinite intervals; (b) there are problems in satisfying literals of type $F_1 > F_2$ on the two external infinite intervals using linear functions with exponential perturbations in the presence of literals of the remaining types, involving functional terms.

Likewise, $t_1 > t_2/t_3$ is a shorthand for the formula

$$\Big((t_1 \cdot t_3 > t_2) \wedge (t_3 > 0)\Big) \vee \Big((t_2 > t_1 \cdot t_3) \wedge (0 > t_3)\Big) .$$

And so on.

2.1.2. Semantics of RMCF^+

An RMCF *interpretation* for the language RMCF^+ is a map M defined over terms and formulae of RMCF^+ as follows:

(a) for every numerical variable x distinct from $0, 1$, Mx is a real number;

(b) the numerical constants $0, 1$ are interpreted as the real numbers $0, 1$, respectively;

(c) the functional constants $\mathbf{0}$, $\mathbf{1}$ are interpreted as the constant functions with values 0 and 1, respectively, defined over the whole real axis $\mathbb{R}$;

(d) for each function variable f distinct from $\mathbf{0}$, $\mathbf{1}$, Mf is a continuous real function of a real variable over the whole axis $\mathbb{R}$;

(e) for each numerical term $t_1 \otimes t_2$, with $\otimes \in \{+, -, \cdot\}$, $M(t_1 \otimes t_2)$ is the real number $Mt_1 \otimes Mt_2$;

(f) for each numerical term $f(t)$, $M(f(t))$ is the real number $(Mf)(Mt)$;

(g) for each functional term $F_1 \oplus F_2$, with $\oplus \in \{+, -\}$, $M(F_1 \otimes F_2)$ is the function $MF_1 \oplus MF_2$;

(h) let t_1, t_2 be numerical terms, F, G functional terms, and A an interval term of the form

$$[t_1, t_2], \quad [t_1, +\infty[, \quad]-\infty, t_2], \quad]-\infty, +\infty[.$$

Let MA be the interpretation of the interval term A, namely

$$MA = \begin{cases} [Mt_1, Mt_2] & \text{if } A = [t_1, t_2] , \\ [Mt_1, +\infty[& \text{if } A = [t_1, +\infty] , \\ ]-\infty, Mt_2] & \text{if } A =]-\infty, t_2] , \\ ]-\infty, +\infty[& \text{if } A =]-\infty, +\infty] . \end{cases}$$

(h.1) $M(t_1 = t_2)$ (resp., $M(t_1 > t_2)$) is true if and only if $Mt_1 = Mt_2$ (resp., $Mt_1 > Mt_2$);

(h.2) $M\big((F > G)_{[t_1,t_2]}\big)$ is true if and only if $(MF)(x) > (MG)(x)$ for all $x \in [Mt_1, Mt_2]$ (thus $M\big((F > G)_{[t_1,t_2]}\big)$ is vacuously true whenever $Mt_1 > Mt_2$; a similar observation applies to the cases below);

(h.3) $M((F = G)_A)$ is true if and only if $(MF)(x) = (MG)(x)$ for all $x \in MA$;

(h.4) $M(\mathsf{Up}(F)_A)$ (resp., $M(\mathsf{Strict\_Up}(F)_A)$) is true if and only if the function MF is monotonically nondecreasing (resp., strictly increasing) in the interval MA;

(h.5) $M(\mathsf{Down}(F)_A)$ (resp., $M(\mathsf{Strict\_Down}(F)_A)$) is true if and only if the function MF is monotonically nonincreasing (resp., strictly decreasing) in the interval MA;

(h.6) $M(\mathsf{Convex}(F)_A)$ (resp., $M(\mathsf{Strict\_Convex}(F)_A)$) is true if and only if the function MF is convex (resp., strictly convex) in the interval MA;

(h.7) $M(\mathsf{Concave}(F)_A)$ (resp., $M(\mathsf{Strict\_Concave}(F)_A)$) is true if and only if the function MF is concave (resp., strictly concave) in the interval MA.

2.1.3. A decision procedure for RMCF^+ formulae: an overview

We briefly review below a decision procedure for the satisfiability problem for RMCF^+ formulae, namely an algorithm which given any RMCF^+ formula φ tells one whether or not φ is satisfiable by a real model.

Phase 1: The first phase of the algorithm consists in transforming the input formula φ into an equisatisfiable formula of the form $\bigvee_{i=1}^{n} \varphi_i$, where each φ_i, for $i = 1, \ldots, n$, is in *standard ordered form*, i.e.,

(a) φ_i is a conjunction of literals of the following simple types

$$\begin{array}{ll} x = y + w\,, & x = y \cdot w\,, \\ x > y\,, & y = f(x)\,, \\ (f = g + h)_A\,, & (f > g)_{[x_1,x_2]}\,, \\ \mathsf{Up}(f)_A\,, & \mathsf{Strict\_Up}(f)_A\,, \\ \mathsf{Convex}(f)_A\,, & \mathsf{Strict\_Convex}(f)_A\,, \end{array} \tag{1}$$

where A is an interval term of any of the following types

$$[x_1, x_2]\,, \quad [x_1, +\infty[\,, \quad]-\infty, x_2]\,, \quad]-\infty, +\infty[\,,$$

x, y, w, x_1, x_2 are numerical variables, and f, g, h are function variables.

(b) Let $x_1, \ldots, x_n$ be the *domain variables* of φ_i, namely the numerical variables x which appear in φ_i either within a functional term of the form $f(x)$ or as one of the two extremes w_b (other than $\pm\infty$) in an interval term of the form $[w_1, w_2]$. Then there exists a permutation π of $\langle 1, \ldots, n\rangle$ such that φ_i contains the literals $x_{\pi(j+1)} > x_{\pi(j)}$, for $j = 1, \ldots, n-1$ (the conjunction of such literals yields a strict ordering of the domain variables).

For instance, the formula

$$\mathsf{Down}(f)_{[x,y]} \wedge y = f(x)$$

is transformed into the equisatisfiable formula

$$\begin{aligned}\Big((\mathbf{0} = f + g)_{[x,y]} \wedge \mathsf{Up}(g)_{[x,y]} \wedge y = f(x) \wedge x > y\Big) \\ \vee \Big((\mathbf{0} = f + g)_{[x,y]} \wedge \mathsf{Up}(g)_{[x,y]} \wedge y = f(x) \wedge (x = y + 0)\Big) \\ \vee \Big((\mathbf{0} = f + g)_{[x,y]} \wedge \mathsf{Up}(g)_{[x,y]} \wedge y = f(x) \wedge y > x\Big)\,.\end{aligned}$$

Since φ is satisfiable if and only if at least one of the φ_i is satisfiable, Phase 1 allows one to reduce the satisfiability problem for general RMCF^+ formulae to the satisfiability problem for RMCF^+ conjunctions of simple atomic formulae of the types (1) in standard ordered form.

As we have noted for Lemma 1.1, in this phase a combinatorial explosion can take place, which should be counteracted by suitable measures in the implementation of the algorithm (cf. [2, p. 775]).

The subsequent phases of the algorithm will therefore address the satisfiability problem for RMCF^+ conjunctions in standard ordered form.

Thus, let φ_i be a RMCF^+ conjunction in standard ordered form (for instance, one of the conjuncts resulting from Phase 1).

Phase 2: In this phase all function variables present in φ_i are evaluated over the domain variables of φ_i. In other words, for each domain variable v_j of φ_i and each function variable f occurring in φ_i, the conjunct

$$y_j^f = f(v_j)\,,$$

where y_j^f is a freshly introduced numerical variable, is added to φ_i.

In addition, for each literal $x = f(v_j)$ initially present in φ_i, the literal

$$x = y_j^f$$

is added to φ_i.

Let ψ be the resulting formula. Plainly, ψ and φ_i are equisatisfiable.

For instance, the formula

$$\mathsf{Convex}(f)_{[x,y]} \wedge y > x$$

is transformed into the equisatisfiable formula

$$\mathsf{Convex}(f)_{[x,y]} \wedge y > x \wedge z = f(x) \wedge t = f(y)\,.$$

Phase 3: During this phase, all literals involving function variables, namely those of the form

$$\begin{array}{lll} y = f(x)\,, & (f = g + h)_A\,, & (f > g)_{[x_1,x_2]}\,, \\ \mathsf{Up}(f)_A\,, & \mathsf{Convex}(f)_A\,, & \\ \mathsf{Strict\_Up}(f)_A\,, & \mathsf{Strict\_Convex}(f)_A\,, & \end{array}$$

are removed from the formula ψ resulting from Phase 2 and are replaced by suitable RMCF^+ conjuncts not involving function variables. Thus, the resulting conjunction is a quantifier-free formula, which can be readily tested for satisfiability by any decider for Tarski's theory of reals.

This is the most critical phase of the algorithm, from the correctness point of view. Indeed, while it is not difficult to eliminate function symbols from ψ in such a way that the resulting RMCF^+ formula ψ_1 is satisfiable whenever so is the input formula ψ, particular care must be taken in order that the reverse implication holds too, namely that ψ is satisfiable whenever so is ψ_1.

Let us see in detail the steps of Phase 3. Let $V = \{v_1, \ldots, v_r\}$ be the collection of the domain variables of ψ and assume that ψ contains the literals $v_{i+1} > v_i$, for $i = 1, \ldots, r-1$ (see (b) in Phase 1). Let $\mathit{ind} : V \cup \{-\infty, +\infty\} \longrightarrow \{1, 2, \ldots, r\}$ be the *index function* of V, where

- $\mathit{ind}(v_i) = i$, for $i = 1, \ldots, r$,
- $\mathit{ind}(-\infty) = 1$ and $\mathit{ind}(+\infty) = r$.

Also, for each function variable f in ψ, let us introduce the new numerical variables γ_0^f, γ_r^f, and α_j^f, for $j = 0, 1, \ldots, r$.

We perform the following six transformation steps (five addition steps and one, the last, elimination step).

1. For each literal of the type $(f = g + h)_{[w_1,w_2]}$ in ψ, where f, g, h are function variables and w_1, w_2 are extended numerical variables, we add the following literals:

   $$y_i^f = y_i^g + y_i^h\,, \qquad \alpha_j^f = \alpha_j^g + \alpha_j^h\,,$$

   for every i such that $\mathit{ind}(w_1) \leqslant i \leqslant \mathit{ind}(w_2)$ and for every j such that $\mathit{ind}(w_1) \leqslant j \leqslant \mathit{ind}(w_2) - 1$.

In addition, if $w_1 = -\infty$, we add also the following two literals:

$$\alpha_0^f = \alpha_0^g + \alpha_0^h\,, \qquad \gamma_0^f = \gamma_0^g + \gamma_0^h\,.$$

Likewise, if $w_2 = +\infty$, we add also the following two literals:

$$\alpha_r^f = \alpha_r^g + \alpha_r^h\,, \qquad \gamma_r^f = \gamma_r^g + \gamma_r^h\,.$$

2. For each literal of the type $(f > g)_{[w_1,w_2]}$ present in ψ, where f, g are function variables and w_1, w_2 are numerical variables, we add the following literals:

$$y_j^f - y_j^g > |\alpha_j^f| + |\alpha_j^g|\,, \qquad y_{j+1}^f - y_{j+1}^g > |\alpha_j^f| + |\alpha_j^g|\,,$$

for every j such that $\mathit{ind}(w_1) \leqslant j \leqslant \mathit{ind}(w_2)$ (here and in the following it is to be understood that literals containing the absolute value function are to be considered as shorthands for equivalent RMCF^+ formulae with no occurrence of the absolute value).

3. For each literal of the form $\mathsf{Up}(f)_{[w_1,w_2]}$ in ψ, where f is a function variable and w_1, w_2 are extended numerical variables, we add the following literals:

$$y_{j+1}^f - y_j^f \geqslant 4|\alpha_j^f|\,,$$

for every j such that $\mathit{ind}(w_1) \leqslant j \leqslant \mathit{ind}(w_2) - 1$.

In addition, if $w_1 = -\infty$, we add also the following two literals:

$$\gamma_0^f \geqslant 0\,, \qquad \gamma_0^f \geqslant \alpha_0^f\,.$$

Likewise, if $w_2 = +\infty$, we add also the following two literals:

$$\gamma_r^f \geqslant 0\,, \qquad \alpha_r^f + \gamma_r^f \geqslant 0\,.$$

For literals of the form $\mathsf{Strict\_Up}(f)$, we proceed much in the same way, but using the strict inequality $>$ in place of $\geqslant$.

4. For each literal of the type $\mathsf{Convex}(f)_{[w_1,w_2]}$ in ψ, where f is a function variable and w_1, w_2 are extended numerical variables, we add the following literals:

$$0 \geqslant \alpha_i^f\,, \qquad \alpha_j^f \geqslant \frac{1}{4}\Big[y_j^f - y_{j+1}^f + (y_j^f - y_{j-1}^f - 4\alpha_{j-1}^f)\frac{v_{j+1} - v_j}{v_j - v_{j-1}}\Big]\,,$$

for every i such that $\mathit{ind}(w_1) \leqslant i \leqslant \mathit{ind}(w_2) - 1$ and every j such that $\mathit{ind}(w_1) < j < \mathit{ind}(w_2)$.

In addition, if $w_1 = -\infty$, we add also the following literal

$$0 \geqslant \alpha_0^f$$

and, provided that $w_2 \neq v_1$, also the literal

$$\frac{y_2^f - y_1^f + 4\alpha_1^f}{v_2 - v_1} \geqslant \gamma_0^f - \alpha_0^f \,.$$

Likewise, if $w_2 = +\infty$, we add also the following literal

$$0 \geqslant \alpha_r^f$$

and, provided that $w_1 \neq v_r$, also the literal

$$\alpha_r^f + \gamma_r^f \geqslant \frac{y_r^f - y_{r-1}^f - 4\alpha_{r-1}^f}{v_r - v_{r-1}} \,.$$

5. For each literal of the type $\mathsf{Strict\_Convex}(f)_{[w_1,w_2]}$ in ψ, where f is a function variable and w_1, w_2 are extended numerical variables, we add the following literals:

$$0 > \alpha_i^f \,, \qquad \alpha_j^f \geqslant \frac{1}{4}\Big[y_j^f - y_{j+1}^f + (y_j^f - y_{j-1}^f - 4\alpha_{j-1}^f)\frac{v_{j+1} - v_j}{v_j - v_{j-1}}\Big] \,,$$

for every i such that $\mathit{ind}(w_1) \leqslant i \leqslant \mathit{ind}(w_2) - 1$ and every j such that $\mathit{ind}(w_1) < j < \mathit{ind}(w_2)$.

In addition, if $w_1 = -\infty$, we add also the following literal

$$0 > \alpha_0^f$$

and, provided that $w_2 \neq v_1$, also the literal

$$\frac{y_2^f - y_1^f + 4\alpha_1^f}{v_2 - v_1} \geqslant \gamma_0^f - \alpha_0^f \,.$$

Likewise, if $w_2 = +\infty$, we add also the following literal

$$0 > \alpha_r^f$$

and, provided that $w_1 \neq v_r$, also the literal

$$\alpha_r^f + \gamma_r^f \geqslant \frac{y_r^f - y_{r-1}^f - 4\alpha_{r-1}^f}{v_r - v_{r-1}} \,.$$

6. Finally, we drop from ψ all literals involving function variables.

For instance, the formula

$$(f = g + h)_{[x,y]} \wedge y > x \wedge z_1 = f(x) \wedge z_2 = f(y) \\ \wedge t_1 = g(x) \wedge t_2 = g(y) \wedge s_1 = h(x) \wedge s_2 = h(y)$$

is transformed into the equisatisfiable formula

$$y > x \wedge z_1 = f(x) \wedge z_2 = f(y) \\ \wedge t_1 = g(x) \wedge t_2 = g(y) \wedge s_1 = h(x) \wedge s_2 = h(y) \wedge \\ (z_1 = t_1 + s_1) \wedge (z_2 = t_2 + s_2) \wedge (\alpha_f = \alpha_g + \alpha_h)\,.$$

Let ψ_1 be the resulting formula, after the execution of the steps 1–6 above. As already remarked, it can easily be shown that if ψ is satisfiable, so is ψ_1. On the other hand, if ψ_1 is satisfied by a real model M, then for each function variable f thanks to the constraints introduced during the first five addition steps above, it can be shown that there exists a function Mf which can be obtained by perturbing quadratically and exponentially a piecewise linear function through the points (Mv_j, My_j^f), for $j = 1, \ldots, r$. It turns out that the real assignment M so extended over the function variables of ψ is a model for all literals of ψ. Since ψ_1 is a quantifier-free formula of Tarski's theory of reals, its satisfiability can be tested algorithmically.

As a universally closed RMCF^+ statement is valid if and only if its negation is unsatisfiable, the satisfiability test for RMCF^+ outlined above can also be used to test the validity (i.e., theoremhood) of the universal closure of formulae of RMCF^+. Thus we have the following result.

PROPOSITION 2.1 ([2, Section 3, Theorem 1]). *The validity problem for universally closed* RMCF^+ *statements is decidable. In other words, one can test algorithmically whether any universally closed* RMCF^+ *statement is a theorem or not.*

2.1.4. Formalization in RMCF^+ of elementary lemmas in real analysis

We show by way of some examples that the theory RMCF^+ is expressive enough to allow the formulation of some elementary lemmas in real analysis, which can be proved automatically by the decision procedure outlined above.

EXAMPLE 2.2. *Consider the claim:*

"Let f and g be two real functions defined over a closed bounded interval $[a, b]$, such that $f(a) = g(a)$ and $f(b) = g(b)$. If f is strictly convex and g is concave, then $f(x) < g(x)$ for each $x \in]a, b[$."

This can be formalized by the universal closure of the RMCF^+ *formula*

$$\Big(\mathsf{Strict\_Convex}(f)_{[a,b]} \wedge \mathsf{Concave}(g)_{[a,b]} \wedge f(a) = g(a) \\ \wedge f(b) = g(b) \wedge b > x \wedge x > a\Big) \rightarrow (g(x) > f(x))\,. \quad (2)$$

To show that (2) is valid, it is sufficient to prove that its negation

$$\mathsf{Strict\_Convex}(f)_{[a,b]} \wedge \mathsf{Concave}(g)_{[a,b]} \wedge f(a) = g(a) \\ \wedge f(b) = g(b) \wedge b > x \wedge x > a \wedge \neg(g(x) > f(x))$$

is unsatisfiable. After the normalization phase (Phase 1), we obtain

$$\Big[\mathsf{Strict\_Convex}(f)_{[a,b]} \wedge \mathsf{Convex}(h)_{[a,b]} \wedge (\mathbf{0} = g + h)_{[a,b]} \\ \wedge f(a) = g(a) \wedge f(b) = g(b) \wedge b > x \wedge x > a \wedge (f(x) > g(x))\Big] \\ \vee \Big[\mathsf{Strict\_Convex}(f)_{[a,b]} \wedge \mathsf{Convex}(h)_{[a,b]} \wedge (\mathbf{0} = g + h)_{[a,b]} \\ \wedge f(a) = g(a) \wedge f(b) = g(b) \wedge b > x \wedge x > a \wedge (f(x) = g(x))\Big]\,.$$

Then, after executing the subsequent phases of the decision algorithm, we obtain the inequalities

$$(f(b) - f(x)) \cdot (x - x_1) > (f(x) - f(a)) \cdot (x_2 - x_1) \\ (-g(b) + g(x)) \cdot (x - x_1) \geqslant (-g(x) + g(a)) \cdot (x_2 - x_1)$$

which, together with $f(a) = g(a)$ *e* $f(b) = g(b)$*, imply* $f(x) < g(x)$*, contradicting both* $f(x) > g(x)$ *and* $f(x) = g(x)$*.*

Having proved that the negation of (2) is unsatisfiable, it follows that (2) is valid, thus proving that our claim expresses a theorem.

A second example is the following.

EXAMPLE 2.3. *Consider the claim:*

"Let f *and* g *be two real functions defined over a closed bounded interval* $[a, b]$*, such that* f *is strictly convex and* g *is concave in* $[a, b]$*. Then there exist at most two distinct points* $x, y \in [a, b]$ *such that* $f(x) = g(x)$ *and* $f(y) = g(y)$ *(i.e., the graphs of* f *and* g *meet in at most two points in* $[a, b]$*)."*

Observe that it can be formalized as the universal closure of the following

RMCF$^+$ *formula*

$$\Big[\mathsf{Strict\_Convex}(f)_{[a,b]} \wedge \mathsf{Concave}(g)_{[a,b]} \\ \wedge (a \leqslant x_1 \leqslant b) \wedge (a \leqslant x_2 \leqslant b) \wedge (a \leqslant x_3 \leqslant b) \\ \wedge f(x_1) = g(x_1) \wedge f(x_2) = g(x_2) \wedge f(x_3) = g(x_3)\Big] \\ \rightarrow \Big[(x_1 = x_2) \vee (x_1 = x_3) \vee (x_2 = x_3)\Big]$$

and therefore it can be proved automatically.

2.2. An overview of the theory RDF

The theory RDF (of Reals with Differentiable Functions) is an unquantified first-order theory involving various predicates on real functions of class C^1 of one real variable, namely functions with continuous first derivative. Predicates of RDF concern comparison of functions, strict and non-strict monotonicity, strict and non-strict convexity (and concavity), and comparison of first derivatives with real constants. Specifically, the atomic formulae of RDF are:

$t_1 = t_2$,	$t_1 > t_2$,
$(f = g)_A$,	$(f > g)_{[t_1,t_2]}$,
$\mathsf{Up}(f)_A$,	$\mathsf{Strict\_Up}(f)_A$,
$\mathsf{Down}(f)_A$,	$\mathsf{Strict\_Down}(f)_A$,
$\mathsf{Convex}(f)_A$,	$\mathsf{Strict\_Convex}(f)_A$,
$\mathsf{Concave}(f)_A$,	$\mathsf{Strict\_Concave}(f)_A$,
$(D[f] \geqslant t)_A$,	$(D[f] > t)_A$,
$(D[f] \leqslant t)_A$,	$(D[f] < t)_A$,
$(D[f] = t)_A$,	

where A stands for any of the following interval terms

$$[t_1, t_2], \quad [t_1, +\infty[, \quad]-\infty, t_2], \quad]-\infty, +\infty[,$$

t_1, t_2 are numerical terms, and f, g stand for function variables or the functional constants **0** and **1**. Numerical terms are arithmetic expressions involving real variables, the real constants 0, 1, functional expressions of the form $f(t)$, and the arithmetic operators.

Formulae of RDF are propositional combinations of atomic RDF-formulae with the usual logical connectives $\neg, \wedge, \vee, \rightarrow, \leftrightarrow$. Again, explicit quantification is not allowed.

Function variables are interpreted by real functions of a real variable, defined on the whole real axis $\mathbb{R}$, differentiable over $\mathbb{R}$ and with continuous derivative. The functional constants **0** and **1** are interpreted as the constant functions

with values 0 and 1, respectively. Predicates of type $(f > g)_{[t_1,t_2]}$ assert that the function f strictly dominates g in the closed bounded interval $[t_1, t_2]$. The remaining atomic formulae on functions can refer also to closed half-bounded intervals $[t_1, +\infty[$ and $]-\infty, t_2]$ and to the whole real axis $]-\infty, +\infty[$.

Based on the above indications and in analogy with what has been done in the preceding section, one can give a precise definition of RDF-interpretations. Then, satisfiable RDF-formulae are those which admit at least one satisfying interpretation (real model), and valid RDF-formulae (RDF-theorems) are those which are satisfied by all interpretations.

Domenico Cantone and Gianluca Cincotti have proved in recent years that:

- An RDF-formula φ is satisfiable if and only if it admits a *canonical* real model M which interprets the function variables of φ as piecewise linear real functions with *small* quadratic and exponential perturbations.
- Canonical models can be encoded by finitely many parameters satisfying suitable arithmetical conditions. These can be tested for satisfiability by any decision procedure for the existential Tarski's theory of reals.
- Thereby one gets the solvability of the satisfiability problem for RDF-formulae; consequently, solvability of the validity problem for RDF-formulae, because a formula is valid if and only if its negation is unsatisfiable.

The results on which we are reporting can be summarized as follows:

PROPOSITION 2.4. RDF *has solvable satisfiability and validity problems.*[5]

Before outlining the decision algorithm for RDF, we illustrate the expressiveness of this theory by formalizing in it some simple lemmas of elementary real analysis.

EXAMPLE 2.5. *Consider the claim:*

"Let f be a real function of class C^1 on the closed interval $[a, b]$, with constant first derivative. Then f is linear in $[a, b]$."

Plainly, this claim can be formalized by the RDF*-formula*

$$(D[f] = t)_{[a,b]} \rightarrow \Big(\mathsf{Convex}(f)_{[a,b]} \wedge \mathsf{Concave}(f)_{[a,b]}\Big)$$

and therefore it can be verified automatically by a decision procedure for RDF*.*

---

[5] A communication—as yet unpublished—of these results, "*Decision algorithms for fragments of real analysis. II. A theory of differentiable functions with convexity and concavity predicates*" was offered by D. Cantone and G. Cincotti at the Italian conference "Convegno italiano di Logica Computazionale" (CILC'07), 21–22 June 2007, Messina.

A continuation, due to D. Cantone and G.T. Spartà, of that study is in progress: "*Decision algorithms for fragments of real analysis. III. A theory of differentiable functions with (semi-) open intervals*". Motivations for extending RDF so as to overcome some of its expressive limitations will emerge from the discussion of Examples 2.6 and 2.7 below.

Another example is the following.

EXAMPLE 2.6 (Weak form of Rolle's theorem). *Consider the claim:*
"Let f be a real function of class C^1 on the closed interval $[a,b]$ such that $f(a) = f(b)$, $f'(a) \neq 0$, and $f'(b) \neq 0$. Then there exists $c \in]a,b[$ such that $f'(c) = 0$."
In view of the continuity of the first derivative f', this claim can be formalized by the following RDF*-formula*

$$\Big(a < b \wedge f(a) = f(b) \wedge D[f](a) \neq 0 \wedge D[f](b) \neq 0\Big) \\ \rightarrow \neg\Big((D[f] > 0)_{[a,b]} \vee (D[f] < 0)_{[a,b]}\Big)$$

and therefore it can be verified automatically by a decision procedure for RDF*.*

A final example is the following.

EXAMPLE 2.7 (Weak form of the mean-value theorem). *Consider the claim:*
"Let f be a real function of class C^1 on the closed interval $[a,b]$ such that $f'(a) \neq \dfrac{f(b)-f(a)}{b-a}$, and $f'(b) \neq \dfrac{f(b)-f(a)}{b-a}$. Then there exists $c \in]a,b[$ such that $f'(c) = \dfrac{f(b)-f(a)}{b-a}$."
Note that this claim generalizes that of the preceding example. Thus, again by the continuity of the first derivative f', it can be formalized in RDF *as follows:*

$$\Big(a < b \wedge x = \frac{f(b)-f(a)}{b-a} \wedge D[f](a) \neq x \wedge D[f](b) \neq x\Big) \\ \rightarrow \neg\Big((D[f] > x)_{[a,b]} \vee (D[f] < x)_{[a,b]}\Big)\,.$$

In Example 2.6 we had to exclude the cases in which either $f'(a) = 0$ or $f'(b) = 0$, because $(D[f] > 0)_{[a,b]} \vee (D[f] < 0)_{[a,b]}$ expresses that $D[f]$ is nonzero in the closed interval $[a,b]$, rather than in the open interval $]a,b[$. A similar remark applies to Example 2.7, where we had to assume the extra assumptions $f'(a) \neq \frac{f(b)-f(a)}{b-a}$, and $f'(b) \neq \frac{f(b)-f(a)}{b-a}$. If we could express literals of the forms $(D[f] < t)_{]a,b[}$ and $(D[f] > t)_{]a,b[}$, relative to open intervals, in both cases we could get rid of those extra assumptions.

Such remarks have motivated the study—just mentioned in a footnote—of the extension RDF^+ of RDF with literals of any of the forms

$$(f > g)_A\,, \quad (D[f] > t)_B\,, \quad (D[f] < t)_B\,, \quad (D[f] \neq t)_B\,,$$

where A stands for an open or semi-open bounded interval and B stands for an open or semi-open interval which is not necessarily bounded.

2.2.1. The decision algorithm for RDF, in outline

Much like the decision algorithm for RMCF^+, the one for RDF begins with a *normalization phase* which transforms the input formula φ into an equisatisfiable disjunction $\bigvee_{i=1}^{n} \varphi_i$, where each φ_i is a conjunction in *standard ordered form*. While the ordering condition concerning the *domain variables* of each φ_i is as before (but here we include among the domain variables also every x appearing in a term $D[f](x)$ within φ_i), the forms of the literals constituting φ_i are, for the theory at hand:

$$\begin{array}{ll} x = y + w\,, & x = y \cdot w\,, \\ x > y\,, & y = f(x)\,, \\ (f = g)_A\,, & (f > g)_{[x_1,x_2]}\,, \\ y = D[f](x)\,, & (D[f] \bowtie y)_A\,, \\ \mathsf{Strict\_Up}(f)_A\,, & \mathsf{Strict\_Down}(f)_A\,, \\ \mathsf{Convex}(f)_A\,, & \mathsf{Strict\_Convex}(f)_A\,, \\ \mathsf{Concave}(f)_A\,, & \mathsf{Strict\_Concave}(f)_A\,, \end{array} \tag{3}$$

where $\bowtie \in \{=, >, \geqslant, <, \leqslant\}$, A is an interval term of any of the following types

$$[x_1, x_2]\,, \qquad [x_1, +\infty[\,, \qquad]-\infty, x_2]\,, \qquad]-\infty, +\infty[\,,$$

x, y, w, x_1, x_2 are numerical variables, and f, g are function variables. Notice that all negative literals are eliminated by the transformation rules exploited in this phase (all of which are, conceptually, rather simple).

In order to determine whether or not φ is satisfiable, we must check one by one its disjuncts φ_i until either one of them turns out to be satisfiable, or all disjuncts have been examined without success. In preparation for this, we explicitly evaluate all function variables present in each φ_i over the domain variables of φ_i. The way to do this is closely analogous to the one discussed earlier for RMCF^+: we associate new variables y_j^f, t_j^f with each combination of a domain variable v_j of φ_i with a function variable f also appearing in φ_i, and conjoin the literals

$$y_j^f = f(v_j)\,, \qquad t_j^f = D[f](v_j)$$

with φ_i. For each literal $x = f(v_j)$ occurring in φ_i, we then insert the literal $x = y_j^f$ into φ_i; likewise, for each literal $x = D[f](v_j)$ in φ_i, we introduce the equality $x = t_j^f$. Each φ_i produced by the normalization phase is thereby transformed by the present phase into an equisatisfiable conjunction ψ_i.

We will now describe the main phase, which eliminates from each ψ_i all literals that involve function variables.

Let $V = \{v_1, v_2, \ldots, v_r\}$ be the collection of the domain variables of ψ_i with their implicit ordering, and let the index function $ind : V \cup \{-\infty, +\infty\} \longrightarrow$

$\{1, 2, \ldots, r\}$ be defined as follows:

$$ind(x) =_{\text{Def}} \begin{cases} 1 & \text{if } x = -\infty, \\ l & \text{if } x = v_l, \text{ for some } l \in \{1, 2, \ldots, r\}, \\ r & \text{if } x = +\infty. \end{cases}$$

For each function symbol f occurring in ψ_i, introduce new numerical variables γ_0^f, γ_r^f and proceed as follows:

1. For each literal of type $(f{=}g)_{[z_1,z_2]}$ occurring in ψ_i, add the literals:

$$y_i^f = y_i^g, \qquad t_i^f = t_i^g,$$

for $i \in \{ind(z_1), \ldots, ind(z_2)\}$; moreover, if $z_1 = -\infty$, add the literal:

$$\gamma_0^f = \gamma_0^g\,;$$

likewise, if $z_2 = +\infty$, add the literal:

$$\gamma_r^f = \gamma_r^g\,.$$

2. For each literal of type $(f{>}g)_{[w_1,w_2]}$ occurring in ψ_i, add the literal:

$$y_i^f > y_i^g\,,$$

for $i \in \{ind(w_1), \ldots, ind(w_2)\}$.

3. For each literal of type $(D[f]{\bowtie}y)_{[z_1,z_2]}$ occurring in ψ_i, where $\bowtie \in \{=, <, \leqslant, >, \geqslant\}$, add the formulae:

$$t_i^f \bowtie y\,,$$

$$\frac{y_{j+1}^f - y_j^f}{v_{j+1} - v_j} \bowtie y,$$

for $i, j \in \{ind(z_1), \ldots, ind(z_2)\}$, $j \neq ind(z_2)$. Moreover, if $\bowtie \in \{\leqslant, \geqslant\}$ also add the implication:

$$\left(\frac{y_{j+1}^f - y_j^f}{v_{j+1} - v_j} = y\right) \longrightarrow (t_j^f = y \wedge t_{j+1}^f = y);$$

moreover, if $z_1 = -\infty$, add the formula:

$$\gamma_0^f \bowtie y,$$

and if $z_2 = +\infty$, add the formula:

$$\gamma_r^f \bowtie y.$$

4. For each literal of type $\mathsf{Strict\_Up}(f)_{[z_1,z_2]}$ (resp. $\mathsf{Strict\_Down}(f)_{[z_1,z_2]}$) occurring in ψ_i, add the formulae:
$$t_i^f \geqslant 0 \qquad (\text{resp. } t_i^f \leqslant 0),$$
$$y_{j+1}^f > y_j^f \qquad (\text{resp. } y_{j+1}^f < y_j^f),$$
for $i, j \in \{ind(z_1), \ldots, ind(z_2)\}$, $j \neq ind(z_2)$. Moreover, if $z_1 = -\infty$, add the formula:
$$\gamma_0^f > 0 \qquad (\text{resp. } \gamma_0^f < 0),$$
and if $z_2 = +\infty$, add the formula:
$$\gamma_r^f > 0 \qquad (\text{resp. } \gamma_r^f < 0).$$

5. For each literal of type $\mathsf{Convex}(f)_{[z_1,z_2]}$ (resp. $\mathsf{Concave}(f)_{[z_1,z_2]}$) occurring in ψ_i, add the following formulae:[6]
$$t_i^f \leqslant \frac{y_{i+1}^f - y_i^f}{v_{i+1} - v_i} \leqslant t_{i+1}^f \qquad (\text{resp. } \geqslant),$$
$$\left(\frac{y_{i+1}^f - y_i^f}{v_{i+1} - v_i} = t_i^f \ \vee \ \frac{y_{i+1}^f - y_i^f}{v_{i+1} - v_i} = t_{i+1}^f \right) \longrightarrow (t_i^f = t_{i+1}^f),$$
for $i \in \{ind(z_1), \ldots, ind(z_2) - 1\}$; moreover, if $z_1 = -\infty$, add the formula:
$$\gamma_0^f \leqslant t_1^f \qquad (\text{resp. } \gamma_0^f \geqslant t_1^f),$$
and if $z_2 = +\infty$, add the formula:
$$\gamma_r^f \geqslant t_r^f \qquad (\text{resp. } \gamma_r^f \leqslant t_r^f).$$

6. For each literal of type $\mathsf{Strict\_Convex}(f)_{[z_1,z_2]}$ (resp. $\mathsf{Strict\_Concave}(f)_{[z_1,z_2]}$) occurring in ψ_i, add the following formulae:
$$t_i^f < \frac{y_{i+1}^f - y_i^f}{v_{i+1} - v_i} < t_{i+1}^f \qquad (\text{resp. } >),$$
for $i \in \{ind(z_1), \ldots, ind(z_2) - 1\}$; moreover, if $z_1 = -\infty$, add the formula:
$$\gamma_0^f < t_1^f \qquad (\text{resp. } \gamma_0^f > t_1^f),$$
and if $z_2 = +\infty$, add the formula:
$$\gamma_r^f > t_r^f \qquad (\text{resp. } \gamma_r^f < t_r^f).$$

---

[6] Observe that this group of formulae implicitly forces the relations $\frac{y_j^f - y_{j-1}^f}{v_j - v_{j-1}} \leqslant \frac{y_{j+1}^f - y_j^f}{v_{j+1} - v_j}$ for each $j \in \{ind(z_1) + 1, \ldots, ind(z_2) - 1\}$. Geometrically, the point of coordinates (v_j, y_j^f) does not lie above (resp. lies below) the straight line joining the two points (v_{j-1}, y_{j-1}^f) and (v_{j+1}, y_{j+1}^f).

7. Withdraw all literals where function variables appear.

In conclusion, the formula χ_i resulting from ψ_i through the function variable removal phase just described only involves literals of the following types:

$$t_1 \leqslant t_2\,, \qquad t_1 < t_2\,, \qquad t_1 = t_2\,,$$

where t_1 and t_2 are terms involving only real variables, the real constants 0 and 1, and the arithmetic operators $+$ and $\cdot$ (and their counterparts $-$ and $/$), so that the formula χ_i belongs to the decidable (existential) Tarski's theory of reals. Showing that our theory RDF has a solvable satisfiability problem simply amounts to showing that the main phase leading from ψ_i to χ_i preserves satisfiability. The proof of this fact, albeit not particularly deep, requires a somewhat technical and lengthy proof, which we omit here.

References

[1] J. P. Burgess and Y. Gurevich, *The decision problem for linear temporal logic*, Notre Dame J. Formal Logic **26** (1985), 115–128.

[2] D. Cantone, G. Cincotti and G. Gallo, *Decision algorithms for fragments of real analysis. I. Continuous functions with strict convexity and concavity predicates*, J. Symbolic Comput. **41** (2006), 763–789.

[3] D. Cantone, A. Ferro, E. G. Omodeo and J. T. Schwartz, *Decision algorithms for some fragments of analysis and related areas*, Comm. Pure Appl. Math. **40** (1987), 281–300.

[4] A. Church, *An unsolvable problem of elementary number theory*, Amer. J. Math. **58** (1936), 345–363.

[5] G. Collins, *Quantifier elimination for real closed fields by cylindrical algebraic decomposition*, in "Automata Theory and Formal Languages 2nd GI Conference Kaiserslautern", May 20-23, 1975 (H. Brakhage, ed.), Lecture Notes in Computer Science, vol. 33, Springer Berlin (1975), pp. 134–183.

[6] M. Davis, H. Putnam and J. Robinson, *The decision problem for exponential Diophantine equations*, Ann. of Math. **74** (1961), 425–436.

[7] H. B. Enderton, *A Mathematical Introduction to Logic*, Academic Press (1972).

[8] M. J. Fisher and M. O. Rabin, *Super-exponential complexity of Presburger arithmetic*. Complexity and Computation, Vol. VII, SIAM-AMS, Philadelphia (1974), 27–41. (Cf. http://publications.csail.mit.edu/lcs/specpub.php?id=42)

[9] A. Formisano and E. Omodeo, *Theory-specific automated reasoning*, in "A 25-year perspective on Logic Programming: Achievements of the Italian Association for Logic Programming, GULP" (A. Dovier and E. Pontelli eds.), Lecture Notes in Computer Science, vol. 6125, Springer (2010), pp. 37–63.

[10] H. Friedman and À. Seress, *Decidability in elementary analysis. I*, Adv. Math. **76** (1989), no. 1, 94–115.

[11] H. Friedman and À. Seress, *Decidability in elementary analysis. II*, Adv. Math. **79** (1990), no. 1, 1–17.

[12] D. Grigoriev, *Complexity of deciding Tarski algebra*, J. Symbolic Comput. **5** (1988), 65–108.
[13] Y. Gurevich, *Existential interpretation. II*, Arch. Math. Logic **22** (1982), 103–120.
[14] J. Renegar, *A faster PSPACE algorithm for deciding the existential theory of the reals*, 29th Annual Symposium on Foundations of Computer Science (FOCS 1988, Los Angeles, Ca., USA), IEEE Computer Society Press, Los Alamitos (1988), pp. 291–295.
[15] D. Richardson, *Some undecidable problems involving elementary functions of a real variable*, J. Symbolic Logic **33** (1968), 514–520.
[16] J. T. Schwartz, D. Cantone, and E. G. Omodeo, *Computational Logic and Set Theory*, Springer (2011). Foreword by Martin Davis.
[17] A. Tarski, *A decision method for elementary algebra and geometry*, Tech. report, RAND Corporation, Santa Monica, CA (1951), Prepared for publication with the assistance of J.C.C. McKinsey. Available at http://www.rand.org/pubs/reports/R109.
[18] A. Tarski, *What is elementary geometry?, The Axiomatic Method with Special Reference to Geometry and Physics*, North-Holland, Amsterdam (1959), pp. 16–29.
[19] A. Tarski and S. Givant, *Tarski's system of geometry*, Bull. Symbolic Logic **5** (1999), no. 2, 175–214.
[20] A. M. Turing, *On computable numbers, with an application to the Entscheidungsproblem*, Proc. London Math. Soc. **42** (1936), 230–265, a correction appeared on Proc. London Math. Soc. **43** (1937), 544–546.

Authors' addresses:

Domenico Cantone
Dipartimento di Matematica e Informatica, University of Catania
Viale Andrea Doria 6, I-95125 Catania, Italy
E-mail: cantone@dmi.unict.it

Eugenio G. Omodeo
Dipartimento di Matematica e Geoscienze, DMI, University of Trieste
Via Alfonso Valerio 12/1, I-34127 Trieste, Italy
E-mail: eomodeo@units.it

Gaetano T. Spartà
Dipartimento di Metodi e Modelli per l'Economia, il Territorio e la Finanza (MEMOTEF), University of Roma "La Sapienza"
Via Del Castro Laurenziano 9, I-00161 Roma, Italy
E-mail: gaetanosparta@virgilio.it

Received June 4, 2012
Revised October 10, 2012

Rend. Istit. Mat. Univ. Trieste
Volume 44 (2012), 349–369

On the supports for cohomology classes of complex manifolds[1]

Dario Portelli

Dedicated to Fabio Zanolin on the occasion of his sixtieth birthday

Abstract. *Let X be a compact, connected complex manifold, and let $\xi \in H^i(X, \mathbb{Q})$ be a non-trivial class. The paper deals with the possibility to construct a topological cycle Γ on X, whose class is the Poincaré dual of ξ, which is closely related in a precise sense to the complex structure of X. The desired properties of Γ allow to define a differentiable relation into a suitable space of 1-jets. This relation shows that there is a preliminary topological obstruction to construct such a Γ. The main result of the paper is that, in a relevant particular case, this obstruction disappears.*

Keywords: cohomology class, support, complex manifold, differential relation
MS Classification 2010: 32Q55

1. Introduction

Throughout the paper X will denote a compact, connected complex manifold of dimension n.

Let $\xi \in H^i(X, \mathbb{Q})$ be non zero. By a classical theorem of Thom [5] there is an integer $N > 0$ such that the Poincaré dual $PD(N\xi) \in H_k(X, \mathbb{Q})$ is the fundamental class of an oriented differentiable submanifold $\Gamma \subset X$, of dimension $k = 2n-i$ (by the way, the symbol $\subset$ will denote nonstrict inclusion throughout the paper). The set Γ is closed in X, hence compact. For our purposes the relevant property is

$$\xi|_{X-\Gamma} = 0\,. \tag{1}$$

To prove this, let T be an open tubolar neighborhood of Γ inside X. Then $Z := X - T$ is a deformation retract of $X - \Gamma$, and it is sufficient to prove that

[1]Dario Portelli was supported by MIUR funds, PRIN project "Geometria delle varietà algebriche e dei loro spazi di moduli" (cofin 2008), and by Università di Trieste - Finanziamento di Ateneo per progetti di ricerca scientifica - FRA 2011.

$\xi|_{_Z} = 0$. Denote the inclusion $Z \subset X$ by h, and assume that $h^*(\xi) = \xi|_{_Z} \neq 0$. Therefore, since the Kronecker pairing

$$\langle\ ,\ \rangle : H^{\,i}(Z, \mathbb{Q}) \times H_i(Z, \mathbb{Q}) \to \mathbb{Q}$$

is non degenerate, there is $u \in H_i(Z, \mathbb{Q})$ such that $\langle\, h^*(\xi)\,,\, u\,\rangle \neq 0$. But

$$\langle\, h^*(\xi)\,,\, u\,\rangle \;=\; \langle\, \xi\,,\, h_*(u)\,\rangle$$

and it is well known that the right hand side agrees with the intersection number of the k-cycles $PD(\xi) = [\Gamma]$ and $h_*(u)$ on X. Since these cycles can be represented by disjoint chains, we conclude $\langle\, \xi\,,\, h_*(u)\,\rangle = 0$, contradiction.

Relation (1) implies also that for any subset S of X containing Γ we have

$$\xi|_{_{X-S}} \;=\; 0\,.$$

We will say that such a subset of X is a *support* for ξ. Actually, we are interested to the possibility that Γ is contained into a *complex subspace* $Y \subset X$, i.e. that ξ has supports which are of some interest from the point of view of the Complex Geometry. Let us give a necessary condition for this.

By restricting the scalars, the complex n-dimensional vector space $T_{_P}X$ can be thought as a real $2n$-dimensional vector space. This real vector space is nothing but the tangent space at P of the differentiable manifold underlying X. Recall that multiplication by $i = \sqrt{-1}$ defines on $T_{_P}X$ a complex structure $J : T_{_P}X \to T_{_P}X$, and a real subspace of $T_{_P}X$ corresponds to a *complex* subspace of the complex space $T_{_P}X$ if and only if it is left invariant by J.

Now assume that Γ is contained into some complex subspace $Y \subsetneq X$. For any point $P \in \Gamma$ which is smooth for Y there is a chain of real tangent vector spaces

$$T_{_P}\Gamma \;\subset\; T_{_P}Y \;\subsetneq\; T_{_P}X\,.$$

But $T_{_P}Y$ is a complex subspace of $T_{_P}X$, hence

$$T_{_P}\Gamma + J(\,T_{_P}\Gamma\,) \;\subset\; T_{_P}Y \;\subsetneq\; T_{_P}X\,.$$

Note that $T_{_P}\Gamma + J(\,T_{_P}\Gamma\,)$ is in any case the *smallest* complex subspace of $T_{_P}X$ containing $T_{_P}\Gamma$. If the codimension of Y into X is assumed to be $\geq p$, then at any point $P \in \Gamma \cap Y_{\rm sm}$ we have

$$dim_{_{\mathbb{C}}}\big(\,T_{_P}\Gamma + J(\,T_{_P}\Gamma\,)\big) \;\leq\; n - p\,. \tag{2}$$

Notice that by semi-continuity this relation is actually satisfied at every point of Γ.

To try to construct such a support Y for ξ, the idea is to start from a Γ obtained by Thom's theorem, and then to deform somehow the inclusion $i : \Gamma \to X$ to get, say, an immersion $f : \Gamma \to X$, which satisfies condition (2) at any point, and moreover satisfies

$$f_*\mu_\Gamma \,=\, i_*\mu_\Gamma \,=\, [\Gamma] \,=\, PD(N\xi) \,\in\, H_k(X, \mathbb{Q})\,, \tag{3}$$

where $\mu_\Gamma \in H_k(\Gamma, \mathbb{Q})$ is the fundamental class of Γ (recall that Γ is oriented, see [5, p. 28], where, however, this assumption is implicit). Since (2) involves tangent spaces to Γ and X, the natural ambient to study how to deform the inclusion $\Gamma \subset X$ is the space $\mathscr{J}^1(\Gamma, X)$ of 1-jets of germs of maps $\Gamma \to X$, of class $\mathscr{C}^1$ at least. This space consists of all linear maps $L : T_c\Gamma \to T_xX$ for all possible choices of $c \in \Gamma$ and of $x \in X$. There are canonical maps $s : \mathscr{J}^1(\Gamma, X) \to \Gamma$ and $b : \mathscr{J}^1(\Gamma, X) \to X$ defined respectively by

$$s(L) \,:=\, c \qquad \text{and} \qquad b(L) \,:=\, x\,.$$

Moreover, every map $f : \Gamma \to X$, of class $\mathscr{C}^k$ with $k \geq 1$, lifts to the map

$$\begin{aligned} \mathscr{J}^1_f : \Gamma &\longrightarrow \mathscr{J}^1(\Gamma, X) \\ c &\mapsto d\,f_c \end{aligned} \qquad\qquad \begin{array}{ccc} & & \mathscr{J}^1(\Gamma, X) \\ & {\scriptstyle \mathscr{J}^1_f} \nearrow & \big\downarrow {\scriptstyle b} \\ \Gamma & \xrightarrow[f]{} & X \end{array} \tag{4}$$

of class $\mathscr{C}^{k-1}$, which makes the diagram on the right commutative. Note that $\mathscr{J}^1_f$ is always an embedding when $k \geq 2$, even if f is not. We set

$$\mathscr{R} \,:=\, \{\, L \in \mathscr{J}^1(\Gamma, X) \,|\, dim_{\mathbb{C}}\big(\, L(T_c\Gamma) + J(\, L(T_c\Gamma))\,\big) \,\leq\, n - p\,\}\,. \tag{5}$$

In Gromov language (see e.g. [3]) such a $\mathscr{R}$ is called a *differential relation*. Condition (2) translates nicely into this new set-up, because, if $f : \Gamma \to X$ is an immersion, it amounts to require that $\mathscr{J}^1_f(\Gamma) \subset \mathscr{R}$.

All this makes apparent that *there is a priori a topological obstruction in order to find a deformation $f : \Gamma \to X$ of the inclusion $i : \Gamma \to X$ which satisfies* (2) *and* (3). In fact, assume that there is such a f, and let us simply denote by φ its lifting to $\mathscr{J}^1(\Gamma, X)$; then $\varphi(\Gamma) \subset \mathscr{R}$. Hence, formally the map φ factorizes through the inclusion $u : \mathscr{R} \subset \mathscr{J}^1(\Gamma, X)$, namely we have the commutative diagram of topological spaces and continuous maps

$$\begin{array}{ccc} \mathscr{R} & \overset{u}{\hookrightarrow} & \mathscr{J}^1(\Gamma, X) \\ \big\uparrow {\scriptstyle \psi} & \nearrow {\scriptstyle \varphi} & \big\downarrow {\scriptstyle b} \\ \Gamma & \xrightarrow[f]{} & X \end{array}$$

which yields in homology (the fundamental class μ_Γ of Γ was already introduced above)

$$PD(N\,\xi) \,=\, [\Gamma] \,=\, f_*\,\mu_\Gamma \,=\, b_*(\varphi_*\,\mu_\Gamma) \,=\, b_*u_*\psi_*\,\mu_\Gamma \,=\, b_*\circ[u_*(\psi_*\,\mu_\Gamma)]\,.$$

Therefore, *in order that the inclusion* $\Gamma\subset X$ *can be deformed to satisfy* (2), *a necessary condition is that the class* $[\Gamma]$ *is the image via* b_* *of a class supported on* $\mathscr{R}$.

In this paper we discuss this topological obstruction in the simplest possible case, namely when $p=1$ (recall that p was introduced as the codimension into X of a complex subspace Y of X containing Γ). In this case condition (2) specializes to

$$dim_{\mathbb{C}}\big(\,L(T_c\Gamma)+J(\,L(T_c\Gamma)\,)\big)\,\leq\,n-1 \tag{6}$$

and the differential relation $\mathscr{R}$ involved becomes

$$\mathscr{R}\,=\,\{\,L\in\mathscr{J}^1(\Gamma,X)\,|\,dim_{\mathbb{C}}\big(\,L(T_c\Gamma)+J(\,L(T_c\Gamma))\,\big)\,\leq\,n-1\,\}\,.$$

To justify a further restriction in the statement of the main theorem below, let me say that the paper arose from an attempt to understand from a differential geometric point of view some aspects of the Hodge Conjecture. It is well known that Hodge (p,p)-conjecture can be reduced to the case when $dim(X)=2p$. Therefore, it was natural for a first exploration to consider only the case when $i=dim(X)=k$.

The main result of the paper is that in the particular case when $p=1$ and $i\,=\,dim(X)\,=\,k$, the topological obstruction mentioned above disappears. More precisely, we have

THEOREM 1.1. *For* X *of arbitrary dimension* n, *let* $\mathscr{R}\subset\mathscr{J}^1(\Gamma,X)$ *be defined by (6) in the particular case* $i=dim(X)=k$. *Then* $\mathscr{R}$ *is a deformation retract of* $\mathscr{J}^1(\Gamma,X)$.

Following some pioneering work of Thom [6], Gromov, Eliashberg and several other people developed the theory of differential relations (see e.g. [3]). This theory provides technical tools which should allow, in principle, to decide whether the inclusion $\Gamma\subset X$ can be deformed as desired, or not.

However, it is well known that on a general smooth, projective hypersurface $X\subset\mathbb{P}^4$, of degree 5, there are non-trivial $\xi\in H^3(X,\mathbb{Q})$ which are not supported by a divisor of X (see e.g. [7], Ch. 18). It would be of the highest interest to understand from the point of view of the differential relations why a 3-cycle Γ corresponding to such a class ξ cannot be deformed in the desired way in this case.

Theorem 1.1 is proved in § 4. The few, elementary facts about jets we will need are recalled for the reader's convenience in the second section. The study

of the basic properties of $\mathscr{R}$ used to prove Theorem 1.1 is the content of § 3. Finally, the last section contains some details on the restriction of $\mathscr{R}$ to the fibres of $(s,b) : \mathscr{J}^1(\Gamma, X) \to \Gamma \times X$, which perhaps are of independent interest.

From now on we will assume without further mention that $k = dim(X) = n$.

2. Some basic fact on 1-jets

We will consider only 1-jets, so we will always write in the sequel $\mathscr{J}$ for $\mathscr{J}^1(\Gamma, X)$, and $\mathscr{J}(U,V)$ for $\mathscr{J}^1(U,V)$. For the basic definitions and properties of the spaces of jets the interested reader is referred e.g. to [2].

Let Γ and X be differentiable varieties of class $\mathscr{C}^r$, where $r \geq 1$ is an integer, or $r = \omega$, namely Γ and X are real analytic varieties; we will mantain this convention about r throughout the paper.

A structure of differential variety on the set $\mathscr{J}(\Gamma, X)$ is given by the following atlas. Let $(U, u^1, u^2, \dots, u^n)$ and $(V, x^1, x^2, \dots, x^{2n})$ be as above; then we can represent L by a $2n \times n$ matrix with respect to the bases

$$\frac{\partial}{\partial u^1}, \frac{\partial}{\partial u^2}, \dots, \frac{\partial}{\partial u^n} \quad \text{of } T_{c}\Gamma \qquad \text{and} \qquad \frac{\partial}{\partial x^1}, \dots, \frac{\partial}{\partial x^{2n}} \quad \text{of } T_{x}X$$

canonically associated to the given coordinate charts. To represent the entries of this matrix we introduce new coordinates p_{ij}, where $1 \leq i \leq 2n$ and $1 \leq j \leq n$. Therefore, if we consider the canonical map

$$(s,b) : \mathscr{J}(\Gamma, X) \to \Gamma \times X \tag{7}$$

on the subset $\mathscr{J}(U,V) := (s,b)^{-1}(U \times V)$ of $\mathscr{J}(\Gamma, X)$ we have the local coordinates

$$u^1, u^2, \dots, u^n, x^1, x^2, \dots, x^{2n}, p_{ij}, \ 1 \leq i \leq 2n, \ 1 \leq j \leq n. \tag{8}$$

We will need in the sequel the explicit expression for the change of local coordinates in $\mathscr{J}$. For this, consider coordinate charts $(U', v^1, v^2, \dots, v^n)$ on Γ and $(V', y^1, \dots, y^{2n})$ on X, such that $U \cap U' \neq \varnothing$ and $V \cap V' \neq \varnothing$. It is clear than that

$$\mathscr{J}(U,V) \cap \mathscr{J}(U',V') = \mathscr{J}(U \cap U', V \cap V') \neq \varnothing.$$

On $\mathscr{J}(U',V')$ the local coordinates are

$$v^1, \dots, v^n, y^1, \dots, y^{2n}, q_{hk}, \ 1 \leq h \leq 2n, \ 1 \leq k \leq n,$$

and the change of local coordinates is given by the maps

$$v^k = v^k(u^1, \dots, u^n), \qquad 1 \le k \le n, \tag{9}$$

$$y^h = y^h(x^1, \dots, x^{2n}), \qquad 1 \le h \le 2n, \tag{10}$$

$$q_{hk} = \sum_{\substack{1 \le i \le 2n \\ 1 \le j \le n}} \frac{\partial y^h}{\partial x^i} \frac{\partial u^j}{\partial v^k} p_{ij}, \qquad 1 \le h \le 2n, 1 \le k \le n. \tag{11}$$

In particular, notice that, for fixed $c \in U \cap U'$ and $x \in V \cap V'$, relations (11) define a linear map. This implies that the map (7) realizes $\mathscr{J}(\Gamma, X)$ as a real vector bundle over $\Gamma \times X$, of rank $2n^2$ (by the way, if we consider higher order jets, i.e. $\mathscr{J}^r(\Gamma, X)$ with $r > 1$, we can only say that $(s, b) : \mathscr{J}^r(\Gamma, X) \to \Gamma \times X$ is an *affine* bundle). It is clear how this vector bundle trivializes; in fact, if M denotes the real vector space of $2n \times n$ matrices, then $\mathscr{J}(U, V)$ can be identified with $U \times V \times M$, and then $(s, b) : \mathscr{J}(U, V) \to U \times V$ corresponds to the projection $U \times V \times M \to U \times V$.

Define the rank of the 1-jet (c, x, L) as the rank of L. The map ρ which associates to every 1-jet its rank is easily seen to be lower semicontinuous. Hence, for any integer r, with $0 \le r \le n$, the set $\mathscr{J}_r := \{ j \in \mathscr{J} \mid \rho(j) \le r \}$ is closed in $\mathscr{J}$. We will mostly *restrict in the sequel to work on the open subset* $\mathscr{Y}$ *of* $\mathscr{J}$ *of the jets of rank* n.

3. The differential relation $\mathscr{R}$

Let us now introduce some more standard notation which will be used freely throughout the paper.

Consider coordinate charts $(U, u^1, u^2, \dots, u^n)$ for Γ and $(V, x^1, x^2, \dots, x^{2n})$ for X. More precisely, we will always assume that V is a domain of holomorphic coordinates $(z^1, \dots, z^n) \in \mathbb{C}^n$ on X, and that $z^h = x^h + ix^{n+h}$ is the decomposition of z^h into its real and imaginary parts. Then the complex structure J is given by

$$J : (x^1, \dots, x^n, x^{n+1}, \dots, x^{2n}) \mapsto (-x^{n+1}, -x^{n+2}, \dots, -x^{2n}, x^1, \dots, x^n). \tag{12}$$

Now assume that we have an immersion $f : U \to V$; we can write it in coordinates. For any $c \in U$ the image $T_c := d f_c(T_c \Gamma)$ of the differential map $d f_c$ is generated inside $T_{f(c)} X$ by the columns of the jacobian matrix

$$J_c = \frac{\partial (x^1, x^2, \dots, x^{2n})}{\partial (u^1, u^2, \dots, u^n)}(c),$$

which is a $2n \times n$ matrix. We write J_c in block form

$$J_c = \begin{pmatrix} A \\ B \end{pmatrix}, \tag{13}$$

where both A, B are $n \times n$ real matrices, whose entries depend on c. Then by (12) the subspace $T_c + J(T_c)$ of $T_P X$ is generated by the columns of the matrix

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

and relation (6) is verified at all points of U if and only if on U

$$\det \begin{pmatrix} A & -B \\ B & A \end{pmatrix} \equiv 0\,. \tag{14}$$

(13) and (14) suggest to organize the matrix (p_{ij}) in block form

$$(p_{ij})_{i,j} = \begin{pmatrix} A \\ B \end{pmatrix} \tag{15}$$

and to set

$$\mathscr{M} := \begin{pmatrix} A & -B \\ B & A \end{pmatrix}. \tag{16}$$

The determinant D_{UV} of $\mathscr{M}$ is a homogeneous polynomial, with coefficients in $\mathbb{Z}$, in the indeterminates p_{ij}, of degree $2n$.

We will check now that the loci defined on the various charts $\mathscr{J}(U, V)$ by the corresponding equations $D_{UV} = 0$ patch toghether to define a closed subset of $\mathscr{J}(\Gamma, X)$, which is the differential relation $\mathscr{R}$.

The key point is to understand how the various maps D_{UV} behave under a change of coordinates. So, let $U' \subseteq \Gamma$ and $V' \subseteq X$ denote as usual coordinate charts such that $U \cap U' \neq \varnothing$ and $V \cap V' \neq \varnothing$. Then on $\mathscr{J}(U \cap U', V \cap V')$ we have the restrictions of both D_{UV} and $D_{U'V'}$.
To simplify notations we will denote the jacobian matrices involved by

$$\mathscr{U} = \frac{\partial(y^1, \ldots, y^{2n})}{\partial(x^1, \ldots, x^{2n})} \qquad \text{and} \qquad \mathscr{V} = \frac{\partial(u^1, \ldots, u^n)}{\partial(v^1, \ldots, v^n)}\,.$$

Moreover, let us write the matrix $\mathscr{M}$ in block form as

$$\mathscr{M} = (\mathscr{P} \,|\, S\mathscr{P})\,, \tag{17}$$

where the size of each block is $2n \times n$, and

$$S := \begin{pmatrix} 0 & -I_n \\ I_n & 0 \end{pmatrix}$$

is the matrix of the complex structure J (note that this matrix is the same on every chart of X). Finally, arrange the various q_{hk} appearing in (11) in a $2n \times n$ matrix $\mathscr{Q}$. Equations (11) tell us that $\mathscr{P}$ and $\mathscr{Q}$ are related by

$$\mathscr{Q} = \mathscr{U}\,\mathscr{P}\,\mathscr{V}\,.$$

Then

$$\big(\,\mathscr{Q}\,|\,S\,\mathscr{Q}\,\big)\;=\;\big(\,\mathscr{U}\;\mathscr{P}\,\mathscr{V}\,|\,S\,\mathscr{U}\;\mathscr{P}\,\mathscr{V}\,\big)\,.$$

Since X is a complex manifold, we can restrict to the case when all the changes of coordinates (10) are holomorphic, hence their differentials are $\mathbb{C}$-linear. In matrix terms this is $S\,\mathscr{U}\;=\;\mathscr{U}\,S$, which yields

$$\big(\,\mathscr{U}\;\mathscr{P}\,\mathscr{V}\,|\,S\,\mathscr{U}\;\mathscr{P}\,\mathscr{V}\,\big)\;=\;\big(\,\mathscr{U}\;\mathscr{P}\,\mathscr{V}\,|\,\mathscr{U}\,S\,\mathscr{P}\,\mathscr{V}\,\big)\;=\;\mathscr{U}\,\big(\,\mathscr{P}\,|\,S\,\mathscr{P}\,\big)\,\begin{pmatrix}\mathscr{V} & 0\\ 0 & \mathscr{V}\end{pmatrix}.$$

Taking determinants we get

$$\det\big(\,\mathscr{Q}\,|\,S\,\mathscr{Q}\,\big)\;=\;\det(\mathscr{U})\det(\mathscr{M})\det(\mathscr{V})^2\,.$$

In terms of the functions D this relation becomes

$$D_{U'V'}\;=\;\lambda\,D_{UV}\,, \tag{18}$$

where

$$\lambda\;:=\;\det(\mathscr{U})\det(\mathscr{V})^2 : \mathscr{J}\,(U\cap U', V\cap V')\to\mathbb{R}_{>}\,. \tag{19}$$

In fact, $\det(\mathscr{U})>0$ because X is canonically oriented. It is a simple exercise to check that *the functions λ satisfy the cocycle condition.*

Therefore *$\mathscr{R}$ can be defined coherently by the vanishing of the functions D on the coordinate charts of $\mathscr{J}\,(\Gamma, X)$.*

Let us analyze more closely the functions D. Elementary operations on the matrix $\mathscr{M}$ in the block form (16) transform it into

$$\begin{pmatrix}\mathscr{A}+i\mathscr{B} & 0\\ \\ \frac{i}{2}(\mathscr{A}-i\mathscr{B}) & \mathscr{A}-i\mathscr{B}\end{pmatrix}\qquad\text{and finally into}\qquad\begin{pmatrix}\mathscr{A}+i\mathscr{B} & 0\\ \\ 0 & \mathscr{A}-i\mathscr{B}\end{pmatrix}.$$

Note that the rank of the first n columns in the above matrices changes only when the last group of elementary operations is performed. Note also that $\mathscr{A}+i\mathscr{B}$ and $\mathscr{A}-i\mathscr{B}$ have the same rank, hence

$$rk(\mathscr{M})\;=\;2\;rk(\mathscr{A}+i\mathscr{B})\,. \tag{20}$$

Moreover, $\det(\,\mathscr{A}+i\mathscr{B}\,)$ is a homogeneous polynomial in the indeterminates p_{ij}, with complex coefficients, of degree n. It is convenient to write it in the form

$$E\;:=\;\det(\,\mathscr{A}+i\mathscr{B}\,)\;=\;R+iI\,, \tag{21}$$

where *R and I are both homogeneous polynomials with real coefficients, of degree n*. Therefore

$$D_{UV}\;=\;\det\begin{pmatrix}\mathscr{A}+i\mathscr{B} & 0\\ \\ 0 & \mathscr{A}-i\mathscr{B}\end{pmatrix}\;=\;(\,R+iI\,)\,(\,R-iI\,)\;=\;R^2+I^2\,. \tag{22}$$

COROLLARY 3.1. *D_{UV} is a homogeneous polynomial with real coefficients, in the indeterminates p_{ij}, of degree $2n$. Moreover, as a function, $D_{UV} \geq 0$.*

For future use we have also to analyze the behaviour of the maps D outside $\mathscr{R}$. For this, set

$$\mathscr{F} \;:=\; \mathscr{J}(U,V) \;-\; \mathscr{R}\,.$$

The restriction of $D = D_{UV}$ to $\mathscr{F}$ is a *smooth* map (actually, an algebraic one, hence real-analytic) $\mathscr{F} \to \mathbb{R}_{>}$. It is elementary to check that *such a D is a surjective submersion.*

COROLLARY 3.2. *For any $a > 0$ the set $D^{-1}(a)$ is a smooth hypersurface of $\mathscr{F}$.*

Assume now that U, U' and V, V' are domains of coordinate charts for Γ and X respectively, such that $U \cap U' \neq \varnothing$ and $V \cap V' \neq \varnothing$. We have the restrictions of both D_{UV} and $D_{U'V'}$ on $\mathscr{J}(U \cap U', V \cap V')$. But the map $\lambda : \mathscr{J}(U \cap U', V \cap V') \to \mathbb{R}_{>}$ defined in (19) is not constant in general, hence the hypersurfaces $D_{UV}^{-1}(a)$ and $D_{U'V'}^{-1}(a')$ of $\mathscr{F}$ *do not glue*, however a, a' are choosen.

The throubles with λ disappear if we restrict to a fiber of (s,b). In fact, take any $c \in U \cap U'$ and $x \in V \cap V'$, and set

$$\Phi \;:=\; (s,b)^{-1}(c,x)\,.$$

Because of (18) we then have

$$\Phi \,\cap\, D_{U'V'}^{-1}\left(\,\lambda(c,x)\,a\,\right) \;=\; \Phi \,\cap\, D_{UV}^{-1}(a)\,. \tag{23}$$

Hence *these hypersurfaces of $\Phi \simeq \mathbb{R}^{2n^2}$ are independent from the system of local coordinates on $\mathscr{J}$ used to define them.* They will play an important role in the sequel, mainly bacause of the following proposition, quite similar to Corollary 3.2. From now on we will denote by D both the restriction to Φ of the map D_{UV}, and the homogeneous polynomial which is the determinant of the matrix (16).

PROPOSITION 3.3. *For any $a > 0$ the subsets*

$$\mathscr{D}_a \;:=\; D^{-1}(a)$$

of Φ are smooth hypersurfaces, and $\Phi - \mathscr{R}$ is foliated by them when a runs into $\mathbb{R}_{>}$.

Proof. Take any $P \in \Phi$, such that $D(P) > 0$. Since D is homogeneous, of degree $2n$, by Euler formula we have

$$\sum_{i,j} p_{ij}(P) \frac{\partial D}{\partial p_{ij}}(P) = 2n\, D(P) > 0\,.$$

Hence the various

$$\frac{\partial D}{\partial p_{ij}}(P)$$

cannot be all zero. □

4. Proof of Theorem 1.1

The outline of the construction of a retraction map $r : \mathscr{J} \to \mathscr{R}$ is rather simple. In fact, recall that $\mathscr{J}$ has a structure of real vector bundle over $\Gamma \times X$, given by the map (s, b), as was already remarked in § 2. Hence r can be constructed fiberwise. In any fiber Φ there are the level hypersurfaces of the maps D. Though the "levels" actually depend on the function D, hence on the local coordinates used to define it, the hypersurfaces themselves do not because of (18), and we can therefore consider the corresponding normal directions field, with respect to some metric on Φ. This metric will be supplyied by a Riemannian structure on $\mathscr{J}$, namely a smoothly varying positive definite symmetric bilinear form on each fiber. It is well known that any vector bundle over a smooth base can be endowed with such a structure.

The directions field mentioned above corresponds to several (nowhere vanishing) vector fields, e.g. the gradient of D. The integral curves of any of these vector fields foliate $\Phi - \mathscr{R}$, and the key point is that every integral curve "ends" on $\mathscr{R}$. Then, given any $P \in \Phi - \mathscr{R}$, there is exactly one integral curve containing it, and we can define $r(P)$ to be the limit point of this curve into $\mathscr{R}$.

Let us fix on $\mathscr{J}$ a Riemannian structure $\mathcal{M}$. On $\Phi = (s, b)^{-1}(c, x)$ we fix an ortonormal basis with respect to the metric $\mathcal{M}(c, x)$. On Φ we will use the coordinates q_{ij} given by the dual basis, instead of the p_{ij} introduced previously, to simplify somewhat the computations. In the new coordinates the function D has still the form (22), namely

$$D = \widetilde{R}^{\,2} + \widetilde{I}^{\,2}\,, \tag{24}$$

where $\widetilde{R}$ and $\widetilde{I}$ are both homogeneous polynomials of degree n in the variables q_{ij}. Therefore, D is homogeneous, of degree $2n$. Moreover, the set $\mathscr{C} = \Phi \cap \mathscr{R}$ is defined into Φ by the equation $D = 0$.

We are interested to the family of ortogonal curves to the level hypersurfaces of the function D. Hence, by definition, the more general system of differential equations with integral curves the family of curves we want is

$$\frac{\mathrm{d}\, q_{ij}}{\mathrm{d}\, t} \;=\; \nu\, \nabla D\,, \tag{25}$$

where ∇D denotes the gradient vector field of D, and ν is a nowhere vanishing real function defined in a suitable open set of Φ, to be determined in order that any solution of (25) satisfies some desired property.

Notice that ∇D vanishes exactly along $\mathscr{C}$. In fact, (24) implies that ∇D vanishes along $\mathscr{C}$, and at any point where ∇D vanishes, D vanishes as well by Euler formula. This allows us to consider the following specialization of (25) on $\Phi - \mathscr{C}$

$$\frac{\mathrm{d}\, q_{ij}}{\mathrm{d}\, t} \;=\; \frac{\nabla D}{\parallel \nabla D \parallel^2}\,. \tag{26}$$

The reason for (26) is that the relation of our integral curves with the level hypersurfaces of D makes reasonable to try to parametrize the integral curves, at least locally, by the "level" itself. More precisely, if $\varphi(t)$ is a function $\mathbb{R} \to \Phi$ whose image is an integral curve, then we want the following relation to be identically satisfied

$$D(\,\varphi(t)\,) \;\equiv\; t\,. \tag{27}$$

To determine the function ν in (25) such that (27) will be satisfied, we differentiate (27), where $\varphi(t)$ is assumed to be a solution of (25), thus getting

$$\nu \;\parallel \nabla D \parallel^2 \equiv\; 1\,.$$

Conversely, let $\varphi(t)$ be a solution of (26). Then,

$$\frac{\mathrm{d}}{\mathrm{d}\, t}\, D(\,\varphi(t)\,) \;=\; \sum_{i,j} \frac{\partial\, D}{\partial\, q_{ij}}\big(\,\varphi(t)\,\big)\, \varphi'_{ij}(t) \;\equiv\; 1$$

and there is a real constant C such that

$$D(\,\varphi(t)\,) \;\equiv\; t + C\,.$$

But the system (26) is autonomous, and we can safely assume that $C = 0$.

Lemma 4.1. *Every solution φ of (26) is maximally defined on $(0, +\infty)$. Moreover, the function $t \mapsto \parallel \varphi(t) \parallel^2$ is strictly increasing, and*

$$\lim_{t \to +\infty} \;\parallel \varphi(t) \parallel = \;\infty\,. \tag{28}$$

Proof. Take any $P \in \Phi$ not in $\mathscr{C}$, and set $t_0 = D(P)$. Moreover, let $\varphi(t)$ be the solution of (26) such that $\varphi(t_0) = P$. It is customary to consider ∇D as a column vector; if P is considered as a row vector, then by Euler formula we get

$$P \cdot \nabla D(P) \,=\, 2n\, D(P) \,=\, 2n\, t_0 \,>\, 0$$

and Schwarz inequality yields

$$2n\, t_0 \,=\, |\, P \cdot \nabla D(P)\,| \,\leq\, \|\, P \,\|\, \|\, \nabla D(P) \,\|\ .$$

Hence

$$\left\| \frac{\nabla D(P)}{\|\, \nabla D(P)\, \|^2} \right\| \,=\, \frac{1}{\|\, \nabla D(P)\, \|} \,\leq\, \frac{1}{2n\, t_0}\ \|\, P \,\|\ .$$

Therefore, if a is any real number such that $0 < a < t_0$, then for every $P' \in \Phi - \mathscr{C}$ such that $D(P') \geq a$, the following inequality is satisfied

$$\left\| \frac{\nabla D(P')}{\|\, \nabla D(P')\, \|^2} \right\| \,\leq\, \frac{1}{2n\, a}\ \|\, P' \,\|\ .$$

This shows that $\varphi(t)$ is defined on any $[t_1, t_2] \subseteq \mathbb{R}$, where $a < t_1 < t_0 < t_2$, hence on $[t_1, \infty)$. Since $a > 0$ is arbitrary, we conclude that every solution of (26) is defined on $(0, +\infty)$.

Moreover, we have by Euler formula and (26) (here ${}^{\mathbf{t}}\varphi(t)$ denotes the transposed of the column vector $\varphi(t)$)

$$\frac{\mathrm{d}}{\mathrm{d}\, t}\ \|\, \varphi(t) \,\|^2 = \, 2\, {}^{\mathbf{t}}\varphi(t) \cdot \frac{\nabla D(\varphi(t))}{\|\, \nabla D(\varphi(t))\, \|^2} \,=\, 4n\, \frac{D(\varphi(t))}{\|\, \nabla D(\varphi(t))\, \|^2} \,>\, 0\,,$$

for every t, hence $t \mapsto \|\, \varphi(t) \,\|^2$ is a strictly increasing function.

Finally, set $\mathscr{D}_a := D^{-1}(a)$ for every $a > 0$. Note that, if $b > 0$ is another real number, then the ubiquitous Euler formula yields also the diffeomorphism

$$\mathscr{D}_a \to \mathscr{D}_b \qquad \text{given by} \qquad P \mapsto \left(\frac{b}{a} \right)^{\frac{1}{2n}} P\,.$$

Therefore, if we set $\mu_a := \inf \{\, \|\, P \,\| \;|\, P \in \mathscr{D}_a \,\}$ (clearly $\mu_a > 0$), then μ_a and μ_b are related by

$$\mu_b \,=\, \left(\frac{b}{a} \right)^{\frac{1}{2n}} \mu_a$$

and (28) follows because $\varphi(t) \in \mathscr{D}_t$ for any $t > 0$ by (27), hence

$$\|\, \varphi(t) \,\| \geq\, \mu_t\,.$$

□

It remains to analyze the behaviour of the solutions of (26) when $t \to 0^+$. The key point is disposed by the following result (for the proof see [4]).

THEOREM 4.2. *Let $D : \Phi \to \mathbb{R}$ be a real-analytic function ≥ 0. Then, for every $P \in \mathscr{C}$ there is a neighborhood W_P of P inside Φ such that for every $Q \in W_P$ the solution q_Q of the Cauchy problem $q_Q(0) = Q$ for the system of first order ODE*

$$\left(q_{ij} \right)' = -\nabla D \tag{29}$$

is defined in $[0, \infty)$, has finite length, and converges uniformly to a point of $\mathscr{C}$ when $t \to \infty$. Moreover, if $Q \in W_P$ then $q_Q(t) \in W_P$ for every $t \geq 0$.

REMARK 4.3. *To keep close to [4] we stated the above theorem with the orientation of the integral curves reversed with respect to our conventions. Moreover, notice for future use that this result is local, namely it is sufficient to consider the restriction of D to any neighborhood L of a given $P \in \mathscr{C}$. In this case the solution q_Q will converge to a point of $\mathscr{C} \cap L$ when $t \to \infty$.*

Consider, now, an arbitrary solution ψ of (26). Since $t \mapsto \| \psi(t) \|$ is a strictly increasing function, for every fixed $b > 0$ we have $\| \psi(t) \| \leq \| \psi(b) \|$ whenever $t \leq b$. Let K denote the intersection of $\mathscr{C}$ with the closed ball B of vectors with norm $\leq \| \psi(b) \|$; then K is compact and there are finitely many points $P_1, \ldots, P_s \in K$ such that

$$K \subset W_{P_1} \cup \ldots \cup W_{P_s},$$

where any W_{P_i} is an open neighborhood of P_i like in Theorem 4.2.

The function D has a minimum on $B - (W_{P_1} \cup \ldots \cup W_{P_s})$, and this minimum is > 0, because this set is compact and disjoint from $\mathscr{C}$. Then, for $a > 0$ sufficiently small (and $a < b$), we get

$$\mathscr{D}_a \cap B \subset W_{P_1} \cup \ldots \cup W_{P_s}. \tag{30}$$

But every hypersurface $\mathscr{D}_a$ of Φ can be used to assign the initial condition for the solutions of (26), uniformly with respect to the time t. In fact, we have the straightforward consequence of (27).

COROLLARY 4.4. *For any fixed real number $a > 0$, every solution φ of (26) intersects $\mathscr{D}_a$ in exactly one point.*

Therefore (30) implies that $\psi(a) \in W_{P_i}$ for a suitable i. Then Theorem 4.2 applied to ψ (cum grano salis!) yields

$$\lim_{t \to 0^+} \psi(t) \in \mathscr{C}.$$

We are in position now to define a map $\rho : \Phi - \mathscr{C} \to \mathscr{C}$, the first step toward the retraction $r : \mathscr{J} \to \mathscr{R}$. In fact, if $Q \in \Phi - \mathscr{C}$ is arbitrary, let φ be the unique solution of (26) such that $\varphi(D(Q)) = Q$. We set

$$\rho(Q) := \lim_{t \to 0^+} \varphi(t) .$$

LEMMA 4.5. *The map ρ is continuous.*

Proof. For the proof we need another lemma. To state it, let us introduce a small piece of notation. If P is any point of $\Phi - \mathscr{C}$, and $D(P) = a$, we will denote by φ_P the unique solution of (26) such that $\varphi_P(a) = P$.

LEMMA 4.6. *For any fixed real number $c > 0$*

$$\chi : (0, +\infty) \times \mathscr{D}_c \to \Phi - \mathscr{C} \qquad \textit{given by} \qquad \chi(t, P) = \varphi_P(t)$$

is a homeomorphism. It follows, in particular, that for any two strictly positive real numbers a, b, the hypersurfaces $\mathscr{D}_a$ and $\mathscr{D}_b$ of Φ are homeomorphic via

$$P \mapsto \varphi_P(b) \qquad \textit{for every} \qquad P \in \mathscr{D}_a .$$

Proof. Corollary 4.4 implies that χ is bijective. Moreover, χ is the restriction to $(0, +\infty) \times \mathscr{D}_c$ of

$$(0, +\infty) \times (\Phi - \mathscr{C}) \to (\Phi - \mathscr{C}) , \qquad \text{defined by } (t, P) \mapsto \varphi_P(t) , \tag{31}$$

which gives the flow of the vector field at the R.H.S. of (26), and it is well known that this map is continuous. Finally, $\chi^{-1} : \Phi - \mathscr{C} \to (0, +\infty) \times \mathscr{D}_c$ is given by

$$P \mapsto (D(P) , \varphi_P(c))$$

and to show that it is continuous it is sufficient to check that $P \mapsto \varphi_P(c)$ is such. But this is a standard consequence of the theorem of the continuous dependence of solutions on initial data. □

To conclude the proof of Lemma 4.5, for an arbitrary $P \in \Phi - \mathscr{C}$, set $Q = \rho(P)$. Here we use the fact that Theorem 4.2 is of local nature. In fact, for any neighborhood L of Q, we can consider the neighborhood $W_Q \subset L$ as in the statement of Theorem 4.2, referred now to $D|_L$. Then, for $b > 0$ sufficiently small we have $\varphi_P(b) \in W_Q$. Fix one of such b.

Let M denote an open neighborhood of $\varphi_P(b)$ into $\mathscr{D}_b$, such that

$$M \subset W_Q . \tag{32}$$

If $D(P) = a$ and $0 < \eta < a$ is real, then Lemma 4.6 tells us that

$$\mathscr{L} = \{ R \in \Phi - \mathscr{C} \mid a - \eta < D(R) < a + \eta \quad \text{and} \quad \varphi_R(b) \in M \}$$

is an open neighborhood of P inside $\Phi - \mathscr{C}$. Then $\rho(\mathscr{L}) \subseteq L$ by Theorem 4.2 because of (32), and the proof of Lemma 4.5 is complete. □

REMARK 4.7. *I believe that* $\rho : \Phi - \mathscr{C} \to \mathscr{C}$ *is surjective, but I don't know how to prove this. Notice however that, as a straightforward consequence of Theorem 4.2, the set* $\rho(\Phi - \mathscr{C})$ *is dense inside* $\mathscr{C}$.

LEMMA 4.8. *The map* $\rho : \Phi - \mathscr{C} \to \mathscr{C}$ *can be extended to a continuous map* $\rho_0 : \Phi \to \mathscr{C}$ *by setting* $\rho_0(P) = P$ *when* $P \in \mathscr{C}$.

Proof. It remains to check the continuity at the points of $\mathscr{C}$. But this follows immediately from Theorem 4.2. □

The next step is the extension of ρ_0 to a coordinate neighborhood of $\mathscr{J}$. For this, let U and V be the usual coordinate neighborhoods for Γ and X respectively. Then we can define

$$\rho_1 : \mathscr{J}(U, V) \to \mathscr{J}(U, V) \cap \mathscr{R}$$

by assuming that it acts fiberwise (the fibres are those of (s, b)) like the map ρ_0 defined above. Since the restriction of $\mathscr{J}(\Gamma, X)$ to $U \times V$ is a trivial vector bundle, ρ_1 is continuous.

To extend ρ_1 to the desired map $r : \mathscr{J} \to \mathscr{R}$, the only delicate point is the following verification. Assume that U' and V' are other coordinate neighborhoods for Γ and X such that $U \cap U' \neq \varnothing$ and $V \cap V' \neq \varnothing$. Then we have also

$$\rho_1' : \mathscr{J}(U', V') \to \mathscr{J}(U', V') \cap \mathscr{R}$$

and we have to check that

$$\rho_1|_{\mathscr{J}(U \cap U', V \cap V')} = \rho_1'|_{\mathscr{J}(U \cap U', V \cap V')} . \tag{33}$$

Here we exploit the fact that both ρ_1 and ρ_1' are defined fiberwise. So, let $\Phi = (s, b)^{-1}(c, x)$ be an arbitrary fiber contained into $\mathscr{J}(U \cap U', V \cap V')$. The two coordinate neighborhoods of $\mathscr{J}$ containing Φ give us the two maps $D, D' : \Phi \to \mathbb{R}$ related by

$$D' = \lambda_0 D$$

because of (18), where $\lambda_0 = \lambda(c, x)$ (see (19)). Therefore

$$\nabla D' = \lambda_0 \nabla D \qquad \text{and} \qquad \frac{\nabla D'}{\| \nabla D' \|^2} = \frac{1}{\lambda_0} \frac{\nabla D}{\| \nabla D \|^2} . \tag{34}$$

The system of ODE (26) for the local coordinates corresponding to U' and V' is then

$$\frac{\mathrm{d}\, q_{ij}'}{\mathrm{d}\, t} = \frac{1}{\lambda_0} \frac{\nabla D}{\| \nabla D \|^2} . \tag{35}$$

Now, let $Q \in \Phi - \mathscr{C}$, ad assume that $D(Q) = a$, hence $D'(Q) = \lambda_0 a$. With the notation introduced in the proof of Lemma 4.5, let φ'_Q be the solution of (35) such that $\varphi'_Q(\lambda_0 a) = Q$. It is easily checked that the map

$$\varphi(t) := \varphi'_Q(\lambda_0 t) : (0, +\infty) \to \Phi - \mathscr{C}$$

satisfies identically (26) thanks to the (34). Moreover, since $\varphi(a) = Q$, we can conclude

$$\varphi_Q(t) = \varphi'_Q(\lambda_0 t) \qquad \text{for every } t > 0\,. \tag{36}$$

Hence,

$$\rho_1(Q) = \lim_{t \to 0^+} \varphi_Q(t) = \lim_{t \to 0^+} \varphi'_Q(t) = \rho'_1(Q)$$

and the equality (33) is completely proved.

Therefore, by (33) we can define a map $r : \mathscr{J} \to \mathscr{R}$ by just requiring that its restriction to any coordinate neighborhood $\mathscr{J}(U, V)$ of $\mathscr{J}$ is the corresponding ρ_1. It is clear that such an r is continuous, and that, if the inclusion $\mathscr{R} \subset \mathscr{J}$ is denoted by u, then $r \circ u = id_{\mathscr{R}}$.

To complete the proof of Theorem 1.1 it remains to show that $u \circ r$ is homotopic to $id_{\mathscr{J}}$. Since r was substantially defined fiberwise, it seems reasonable to try to construct in this way also an homotopy

$$H : [0, 1] \times \mathscr{J} \to \mathscr{J} \tag{37}$$

between $u \circ r$ and $id_{\mathscr{J}}$.

Then, let Φ, $\mathscr{C}$ and ρ_0 be as usual, and denote by i the inclusion $\mathscr{C} \subset \Phi$. For every $P \in \Phi - \mathscr{C}$ we have $\varphi_P : (0, +\infty) \to \Phi - \mathscr{C}$. This map can be extended to a continuous map

$$\widetilde{\varphi}_P : [0, +\infty) \to \Phi \qquad \text{by setting} \quad \widetilde{\varphi}_P(0) = \rho(P)\,.$$

Moreover, if $P \in \mathscr{C}$ we will define $\widetilde{\varphi}_P : [0, +\infty) \to \Phi$ to be the constant map with value P. After these preparations, we set

$$h : [0, 1] \times \Phi \to \Phi \qquad \text{where} \quad h(\tau, P) := \widetilde{\varphi}_P(\,\tau\, D(P)\,)\,. \tag{38}$$

The relations

$$h(1, \_\,) = id_{\mathbb{P}hi}\,, \qquad h(0, \_\,) = i \circ \rho_0\,,$$

follow from the definition. It remains to check that h is continuous. Only the continuity at a point (τ_0, P) where $\tau_0 > 0$ and $P \in \mathscr{C}$ deserves some comment. In this case $h(\tau_0, P) = P$, so let U be an arbitrary neighborhood of P. As usual, we will consider a neighborhood W_P of P like in the statement of Theorem 4.2,

and such that $W_{_P} \subset U$. Moreover, let $a > 0$ be such that $\tau_0 - a > 0$. Finally, let $b > 0$ such that $L := W_{_P} \cap \mathscr{D}_b \neq \varnothing$. We set

$$V := \left\{ Q \in \Phi \mid Q \in W_{_P} \ , \ D(Q) < \frac{b}{\tau_0 + a} \ , \ \text{ if } \ Q \notin \mathscr{C} \ \text{ then } \ \varphi_{_Q}(b) \in L \right\}.$$

Thanks to Lemma 4.6, V is an open neighborhood of P. Assume, now, that $\tau \in (\tau_0 - a, \tau_0 + a)$, and $Q \in V$. If $Q \in \mathscr{C}$, then

$$h(\tau, Q) \ = \ Q \ \in \ W_{_P} \subset U \, .$$

If $Q \notin \mathscr{C}$, then $h(\tau, Q) = \varphi_{_Q}(\tau \, D(Q))$. Therefore, the definition of V yields both the relations $\tau \, D(Q) < b$ and $\varphi_{_Q}(b) \in L \subset W_{_P}$. Hence $h(\tau, Q) \in W_{_P} \subset U$ by the last sentence of Theorem 4.2, and we conclude that the map h in (38) is continuous.

As with the definition of the retraction r, the key point to define the homotopy (37) is the verification that the map (38) actually does not depend on the choice of the local coordinate system $\mathscr{J}(U, V)$ of $\mathscr{J}$ containing the fiber $\mathbb{P}hi$. In fact, with the usual notations,

$$h'(\tau, P) \ = \ h(\tau, P)$$

holds trivially true if $\tau = 0$ or $P \in \mathscr{C}$. Otherwise, by (36),

$$h'(\tau, P) \ = \ \varphi'_{_Q}(\tau \, D'(P)) \ = \ \varphi'_{_Q}(\tau \, \lambda_0 \, D(P)) \ = \ \varphi_{_Q}(\tau \, D(P)) \ = \ h(\tau, P) \, .$$

Therefore we can define fiberwise the map (37), and it is continuous.

The proof of Theorem 1.1 is now complete.

5. Some geometric property of $\mathscr{R}$

To understand $\mathscr{R}$ it is useful to first focus on the geometry of

$$\mathscr{C} \ := \ \Phi \cap \mathscr{R}$$

where, as usual, $\mathbb{P}hi$ is any fibre of the map $(s, b) : \mathscr{J}^{\,1}(\Gamma, X) \to \Gamma \times X$. In particular, we are interested in the dimension of $\mathscr{C}$, and in the structure of its singular locus. To this aim, it is easier to first study the affine variety $\mathscr{C}_{\mathbb{C}}$ defined in $\mathbb{C}^{2\,n^2}$ by the same equations than $\mathscr{C}$, namely

$$R \ = \ 0 \qquad I \ = \ 0 \tag{39}$$

because of (21). Then one can investigate the set of real points of $\mathscr{C}_{\mathbb{C}}$, which is in fact $\mathscr{C}$.

The geometry of $\mathscr{C}_{\mathbb{C}}$ becomes perfectly clear if we replace the equations (39) used to define it, by those we get from the following change of variables in the ring of polynomials $B := \mathbb{C}[\, p_{ij} \,|\, 1 \le i \le 2n \ , \ 1 \le j \le n \,]$. For every pair of integers h, k such that $1 \le h, k \le n$, set

$$Z_{hk} := p_{hk} + i p_{h+n,k} \,, \quad W_{hk} := i p_{hk} + p_{h+n,k} = i(p_{hk} - i p_{h+n,k}) = i\,\overline{Z}_{hk} \,. \tag{40}$$

Under this change of variables B becomes $\mathbb{C}[\, Z_{11}, \dots, Z_{nn}, W_{11}, \dots, W_{nn}]$. By (40) the *generic* $n \times n$ matrices

$$\mathscr{Z} := (\, Z_{ij} \,) \quad \text{and} \quad \mathscr{W} := (\, W_{ij} \,)$$

are related to the matrices $\mathscr{A}$, $\mathscr{B}$ introduced in (16) by the obvious relations

$$\mathscr{Z} \ = \ \mathscr{A} + i\,\mathscr{B} \quad \text{and} \quad -\,i\,\mathscr{W} \ = \ \mathscr{A} - i\,\mathscr{B} \,.$$

Hence by (21) (possibly up to a constant factor $\neq 0$ for the second case)

$$\det(\,\mathscr{Z}\,) \ = \ \det(\,\mathscr{A} + i\mathscr{B}\,) \ = \ E \quad \text{and} \quad \det(\,\mathscr{W}\,) \ = \ \overline{E} \,.$$

The meaning of these relations is as follows. The change of variables (40) induces a change of coordinates

$$\omega : \mathbb{C}_{p_{ij}}^{\,2\,n^2} \longrightarrow \mathbb{C}_{zw}^{\,2\,n^2} \,. \tag{41}$$

Let $\omega(P) \ = \ \big((z),(w)\big)$. Then the coordinates (p_{ij}) of $P \in \mathbb{C}^{\,2\,n^2}$ satisfy the equation $E = 0$ if and only if

$$rk(\mathscr{Z}(z)) < n \,.$$

Therefore, if we set

$$\begin{aligned} Y \ &:= \ \big\{\, (z) \in \mathbb{C}_z^{\,n^2} \ \big|\ rk(\mathscr{Z}(z)) < n \,\big\} \\ Y' \ &:= \ \big\{\, (z) \in \mathbb{C}_w^{\,n^2} \ \big|\ rk(\mathscr{W}(w)) < n \,\big\} \end{aligned} \tag{42}$$

we can conclude that

$$\mathscr{C}_{\mathbb{C}} \ = \ Y \times Y' \,. \tag{43}$$

In fact, $D_{UV} = E \cdot \overline{E}$ because of (22). Moreover, if $P \in \mathscr{J}(U,V)$ annihilates E, i.e. if $E(P) = 0$, then we have also $\overline{E}(P) = 0$, and conversely.

Moreover, Y and Y' are generic determinantal varieties by (42), so that they are irreducible and reduced (see e.g. [1], Ch. II, § § 2 and 3). Hence $\mathscr{C}_{\mathbb{C}}$ is also *irreducible and reduced, of dimension* $2n^2 - 2$ because Y, Y' are both hypersurfaces of $\mathbb{C}^{\,n^2}$.

Finally, from (43) it is also easily seen that

$$Sing(\mathscr{C}_{\mathbb{C}}) \;=\; Sing(Y)\times Y' \;\cup\; Y\times Sing(Y')\,, \tag{44}$$

where (see e.g. [1])

$$Sing(Y) \;=\; \big\{\,(z)\in\mathbb{C}_z^{n^2} \;\big|\; rk(\mathscr{Z}(z))<n-1\,\big\} \tag{45}$$

and similarly for Y'. We can summarize all this as

THEOREM 5.1. *The variety $\mathscr{C}_{\mathbb{C}}$ is irreducible and reduced, of dimension $2n^2-2$. Its singular locus is given by (44), and has codimension 2 inside $\mathscr{C}_{\mathbb{C}}$.*

We are ready to start the study of the set $\mathscr{C}$ of real points of $\mathscr{C}_{\mathbb{C}}$. We will use (39) as equations for both $\mathscr{C}$ and $\mathscr{C}_{\mathbb{C}}$, inside $\mathbb{R}^{2n^2}$ and $\mathbb{C}^{2n^2}$ respectively. Then, the jacobian criterion yields

$$Sing(\mathscr{C}) \;=\; \mathscr{C}\cap Sing(\mathscr{C}_{\mathbb{C}}) \quad \text{or, equivalently} \quad \mathscr{C}_{\rm sm} \;=\; \mathscr{C}\cap\big(\mathscr{C}_{\mathbb{C}}\big)_{\rm sm}\,. \tag{46}$$

To get a better understanding of the above relations, and to exploit them, we have to be able to detect real points of $\mathscr{C}_{\mathbb{C}}$ when they are given in the coordinates z,w. For this, consider the following set-up, where γ is the conjugation map, and ω was defined in (41)

$$\begin{array}{ccccc}
\mathbb{R}^{n^2} & \subseteq & \mathbb{C}_{p_{ij}}^{n^2} & \xrightarrow{\;\omega\;} & \mathbb{C}_{zw}^{n^2} \;\supseteq\; \mathscr{C}_{\mathbb{C}} \\
 & & \gamma\big\downarrow & & \\
\mathbb{R}^{n^2} & \subseteq & \mathbb{C}_{p_{ij}}^{n^2} & \xrightarrow[\;\omega\;]{} & \mathbb{C}_{zw}^{n^2} \;\supseteq\; \mathscr{C}_{\mathbb{C}}\,.
\end{array}$$

Then set

$$\delta \;:=\; \omega\circ\gamma\circ\omega^{-1} : \mathbb{C}_{zw}^{n^2} \longrightarrow \mathbb{C}_{zw}^{n^2}\,.$$

It is clear that, for every $P\in\mathbb{C}_{p_{ij}}^{n^2}$, we have

$$P \;=\; \overline{P} \iff \delta\big(\,\omega(P)\,\big) \;=\; \omega(P)\,. \tag{47}$$

It is easily checked that the map δ is given in coordinates by

$$\delta : (z_{11},\ldots,z_{nn},w_{11},\ldots,w_{nn}) \mapsto (i\,\overline{w}_{11},\ldots,i\,\overline{w}_{nn},i\,\overline{z}_{11},\ldots,i\,\overline{z}_{nn})\,. \tag{48}$$

This allows us to write condition (47) explicitly, namely a point $Q=\omega(P)=(z_{11},\ldots,z_{nn},w_{11},\ldots,w_{nn})$ is such that $Q=\delta(Q)$ if and only if all the following conditions are satisfied

$$\begin{cases} z_{11} \;=\; i\,\overline{w}_{11} \\ \quad\vdots \\ z_{nn} \;=\; i\,\overline{w}_{nn}\,, \end{cases} \qquad\qquad \begin{cases} w_{11} \;=\; i\,\overline{z}_{11} \\ \quad\vdots \\ w_{nn} \;=\; i\,\overline{z}_{nn}\,. \end{cases} \tag{49}$$

Note that the conditions of one block are equivalent to those of the other block.

At this point we are able to describe explicitly the points of $\mathscr{C}$ by means of the map

$$u : Y \to \mathscr{C} \quad \text{given by} \quad (z) \mapsto ((z) \,|\, i\,(\overline{z})) \,. \tag{50}$$

In fact, by (49) the matrix $((z) \,|\, i\,(\overline{z}))$ represents a real point of $\mathscr{C}_{\mathbb{C}}$, hence a point of $\mathscr{C}$. Notice that the restriction p to $\mathscr{C}$ of the canonical projection $\mathscr{C}_{\mathbb{C}} = Y \times Y' \to Y$ is such that

$$p \circ u \,=\, id_Y \,. \tag{51}$$

Now, if $(z) \in Y_{\rm sm}$, i.e. by (42) and (45), if $rk(\mathscr{Z}(z)) = n - 1$, then $u((z)) \in (\mathscr{C}_{\mathbb{C}})_{\rm sm}$. Hence $u((z)) \in \mathscr{C}_{\rm sm}$ because of (46).

On the other hand, if $P = ((z) \,|\, i\,(\overline{z})) \in \mathscr{C}_{\rm sm}$ then it is also a point of $(\mathscr{C}_{\mathbb{C}})_{\rm sm}$, hence $rk(\mathscr{Z}(z)) = n - 1$ and $p(((z) \,|\, i\,(\overline{z}))) \in Y_{\rm sm}$. By (51) the point P of $\mathscr{C}_{\rm sm}$ then comes via u from a smooth point of Y.

To summarize, we have constructed a real-analytic, bijective map

$$u : Y_{\rm sm} \to \mathscr{C}_{\rm sm}$$

with real-analytic inverse. Since Y is an integral variety over $\mathbb{C}$, of dimension $n^{\,2} - 1$, we can conclude

PROPOSITION 5.2. *$\mathscr{C}_{\rm sm}$ is a real-analytic variety, of dimension $2\,(n^{\,2} - 1)$.*

Acknowledgements: I wish to thank Daniele Del Santo and Martino Prizzi for some very useful conversations. In particular, they pointed out to me the paper [4].

REFERENCES

[1] E. ARBARELLO, M. CORNALBA, P. A. GRIFFITHS, AND J. HARRIS, *Geometry of algebraic curves. Vol. I*, Grundlehren Math. Wiss., vol. 267, Springer, New York, 1985.

[2] N. BOURBAKI, *Éléments de mathématique. Fasc. XXXVI. Variétés différentielles et analytiques. Fascicule de résultats (Paragraphes 8 à 15)*, Actualités Scientifiques et Industrielles, No. 1347, Hermann, Paris, 1971.

[3] M. GROMOV, *Partial differential relations*, Ergeb. Math. Grenzgeb. (3), vol. 9, Springer, Berlin, 1986.

[4] S. ŁOJASIEWICZ, *Sur les trajectoires du gradient d'une fonction analytique*, Geometry seminars, 1982–1983 (Bologna, 1982/1983), Univ. Stud. Bologna, Bologna, 1984, pp. 115–117.

[5] R. THOM, *Quelques propriétés globales des variétés différentiables*, Comment. Math. Helv. **28** (1954), 17–86.

[6] R. Thom, *Remarques sur les problèmes comportant des inéquations différentielles globales*, Bull. Soc. Math. France **87** (1959), 455–461.
[7] C. Voisin, *Théorie de Hodge et géométrie algébrique complexe*, Cours Spécialisés, vol. 10, Société Mathématique de France, Paris, 2002.

Author's address:

Dario Portelli
Dipartimento di Matematica e Geoscienze
Università di Trieste
Via Valerio 12/1, 34127 Trieste, Italy
E-mail: porteda@units.it

Received March 28, 2012
Revised October 18, 2012

Rend. Istit. Mat. Univ. Trieste
Volume 44 (2012), 371–388

Semilinear evolution equations in abstract spaces and applications[1]

Irene Benedetti, Luisa Malaguti
and Valentina Taddei

Dedicated to professor Fabio Zanolin on the occasion of his 60th birthday

Abstract. *The existence of mild solutions is obtained, for a semilinear multivalued equation in a reflexive Banach space. Weakly compact valued nonlinear terms are considered, combined with strongly continuous evolution operators generated by the linear part. A continuation principle or a fixed point theorem are used, according to the various regularity and growth conditions assumed. Applications to the study of parabolic and hyperbolic partial differential equations are given.*

Keywords: semilinear multivalued evolution equation, mild solution, evolution system, compact operator, continuation principle
MS Classification 2010: 34G25, 34A60, 47H04, 28B20

1. Introduction

The paper deals with the initial value problem associated to a semilinear multivalued evolution equation

$$\begin{cases} x'(t) \in A(t)x(t) + F(t, x(t)), & \text{for a.a. } t \in [a, b], \\ x(0) = x_0 \in E \end{cases} \tag{1}$$

in a reflexive Banach space $(E, \|\cdot\|)$ where

(A) $\{A(t)\}_{t\in[a,b]}$ is a family of linear, not necessarily bounded, operators with $A(t) : D(A) \subset E \to E$, $D(A)$ dense in E, which generates a strongly continuous evolution operator $U : \Delta \to \mathcal{L}(E)$ (see Section 2 for details);

(F1) $F(\cdot, x) : [a, b] \multimap E$ has a measurable selection for any $x \in E$ and $F(t, x)$ is nonempty, convex and weakly compact for any $t \in [a, b]$ and $x \in E$.

[1]Supported by the national research project PRIN 2009 "Ordinary Differential Equations and Applications".

When E is a separable Banach space, the measurability of $F(\cdot, x)$ for any $x \in E$ implies the existence of a selection as in (F1) (see the Theorem of Kuratowski-Ryll-Nardzewski [6, Theorem A]). Sufficient conditions are given in [6] in order to obtain the existence of a strongly measurable selection for the multivalued map (multimap for short) $F(\cdot, x)$ in a not necessarily separable Banach space.

Two different sets of regularity and growth assumptions on F are assumed, which cause the use of different techniques for studying (1). In Section 3 we treat the case when the evolution operator $U(t, s)$ is compact for $t > s$ and we assume that

(F2) $F(t, \cdot) : E \multimap E_\sigma$ is upper semicontinuous (u.s.c. for short) for a.a. $t \in [a, b]$.

We denote with X_σ the topological space obtained when $X \subseteq E$ is equipped with the weak topology.
If we further impose the growth condition

(F3) $\sup_{x \in \Omega} \|F(t, x)\| \leq \eta_\Omega(t)$ for a.a. $t \in [a, b]$, with $\Omega \subset E$ bounded and $\eta_\Omega \in L^1([a, b]; \mathbb{R})$,

which allows the nonlinearity F to have a superlinear growth, we make use of a classical continuation principle for compact multivalued fields (see Theorem 2.3).

In Section 4 we allow $U(t, s)$ to be non-compact, but we replace (F2) with the stronger regularity condition

(F2′) $F(t, \cdot) : E_\sigma \to E_\sigma$ is u.s.c. for a.a. $t \in [a, b]$

and we use a recent continuation principle in Frechét spaces due to the same authors (see Theorem 2.4). To this aim we also need the following condition

(F2″) $F(t, \cdot)$ is locally compact for a.a. $t \in [a, b]$.

Moreover, in Sections 3 and 4 we also show that, if we restrict the growth condition on F to

(F3′) $\|F(t, x)\| \leq \alpha(t)(1 + \|x\|)$ for a.a. $t \in [a, b]$, every $x \in E$ and some $\alpha \in L^1([a, b]; \mathbb{R})$,

then Ky Fan fixed point Theorem (see Theorem 2.5) can be used in both regularity assets and the solution set is compact in the appropriate topology.
We always investigate the existence of mild solutions of problem (1).

Definition 1.1. *A continuous function $x : [a, b] \to E$ is said to be a mild solution of the problem (1) if there exists a function $f \in L^1([a, b]; E)$ such that $f(t) \in F(t, x(t))$ for a.a. $t \in [a, b]$ and*

$$x(t) = U(t, a)x_0 + \int_a^t U(t, s)f(s)\, ds, \quad \forall\, t \in [a, b].$$

We refer to [5, 10] for the study of problem (1) when $F(t,\cdot)\colon E \multimap E$ is u.s.c. for a.a. $t \in [a,b]$ and it has compact values. Instead, the case when the linear part $A(t)$ is defined and bounded on all the space E was treated in [2, 12] under different regularity conditions. Nonlocal boundary value problems associated to the evolution equation in (1) are investigated in [4, 13] respectively in the case when F satisfies (F2′) and (F2). Many differential operators satisfy condition (A) and frequently they generate a compact evolution operator (see e.g. [14, 16]; see also Example 2.1). The introduction of a multivalued equation is often motivated by the study of a control problem. In Sections 5 we propose an application of our theory to the study of a parabolic partial differential inclusion, hence generating a compact evolution operator. In Section 6 we investigate a feedback control problem associated to an hyperbolic partial differential equation, and thus with a non-compact associated evolution operator. Section 2 contains some preliminary results.

2. Preliminary results

This part contains some preliminary results, of different types, which are useful in the sequel.
Throughout the paper we denote with B the closed unit ball of E centered at 0. Given the measure space (S,Σ,μ) and the Banach space X, we denote with $\|\cdot\|_p$ the norm of the Lebesgue space $L^p(S;X)$.

Let $\Delta = \{(t,s) \in [a,b]\times[a,b] \,:\, a \le s \le t \le b\}$. A two parameter family $\{U(t,s)\}_{(t,s)\in\Delta}$, where $U(t,s) : E \to E$ is a bounded linear operator and $(t,s) \in \Delta$, is called an *evolution system* if the following conditions are satisfied:

1. $U(s,s) = I,\ a \le s \le b\ ;\quad U(t,r)U(r,s) = U(t,s),\ a \le s \le r \le t \le b$;
2. $(t,s) \longmapsto U(t,s)$ is strongly continuous on Δ, i.e. the map $(t,s) \to U(t,s)x$ is continuous on Δ for every $x \in E$.

For every evolution system, we can consider the respective *evolution operator* $U : \Delta \to \mathcal{L}(E)$, where $\mathcal{L}(E)$ is the space of all bounded linear operators in E. Since the evolution operator U is strongly continuous on the compact set Δ, by the uniform boundedness theorem there exists a constant $D = D_\Delta > 0$ such that

$$\|U(t,s)\|_{\mathcal{L}(E)} \le D\,, \quad (t,s) \in \Delta. \tag{2}$$

An evolution operator is said to be *compact* when $U(t,s)$ is a compact operator for all $t-s>0$, i.e. $U(t,s)$ sends bounded sets into relatively compact sets. We refer to [14] for details on this topic.

EXAMPLE 2.1. *Let $\Omega \subset \mathbb{R}^n$ be a bounded domain with a smooth boundary $\partial\Omega$ and consider the linear elliptic partial differential operator in divergence form $A\colon W^{2,2}(\Omega;\mathbb{R}) \cap W_0^{1,2}(\Omega;\mathbb{R}) \to L^2(\Omega;\mathbb{R})$ given by*

$$(A\ell)(x) = \sum_{i,j=1}^{n} \frac{\partial}{\partial x_i}\left(a_{ij}(x)\frac{\partial \ell(x)}{\partial x_i}\right),$$

under the following conditions

(i) $a_{ij} \in L^\infty(\Omega)$, $a_{ij} = a_{ji}$ *for* $i, j = 1, 2, .., n$;

(ii) $c\|\xi\|^2 \leq \sum_{i,j=1}^{n} a_{ij}(x)\xi_i\xi_j$ *a.e. for every* $\xi \in \mathbb{R}^n$ *with* $c > 0$.

It is known that A (see e.g. [16]) generates a strongly continuous semigroup of contractions $S(t)$ with $S(t)$ compact for $t > 0$. Notice that, whenever $a_{ij} = 0$ for $i \neq j$ and $a_{ii} = 1$ for $i = 1, 2, ...n$, then $A\ell = \Delta\ell$.

Given $q \in C([a,b];E)$, let us denote with

$$S_q = \{f \in L^1([a,b];E) \ : \ f(t) \in F(t,q(t)) \text{ a.a. } t \in [a,b]\}.$$

PROPOSITION 2.2. *For a multimap $F : [a,b] \times E \multimap E$ satisfying properties (F1), (F2) and (F3), the set S_q is nonempty for any $q \in C([a,b];E)$.*

Proof. Let $q \in C([a,b];E)$; by the uniform continuity of q there exists a sequence $\{q_n\}$ of step functions, $q_n : [a,b] \to E$ such that

$$\sup_{t\in[a,b]} \|q_n(t) - q(t)\| \to 0, \quad \text{for } n \to \infty. \tag{3}$$

Hence, by (F1), there exists a sequence of functions $\{w_n\}$ such that $w_n(t) \in F(t, q_n(t))$ for a.a. $t \in [a,b]$ and $w_n : [a,b] \to E$ is measurable for any $n \in \mathbb{N}$. From (3) there exists a bounded set $\Omega \subset E$ such that $q_n(t), q(t) \in \Omega$ for any $t \in [a,b]$ and $n \in \mathbb{N}$ and by (F3) there exists $\eta_\Omega \in L^1([a,b];\mathbb{R})$ such that

$$\|w_n(t)\| \leq \|F(t,q_n(t))\| \leq \eta_\Omega(t) \quad \forall\, n \in \mathbb{N}, \text{ and a.a. } t \in [a,b].$$

Hence $\{w_n\} \subset L^1([a,b];E)$, $\{w_n\}$ is bounded and uniformly integrable and $\{w_n(t)\}$ is bounded in E for a.a. $t \in [a,b]$. According to the reflexivity of the space E and by the Dunford-Pettis Theorem (see [7, p. 294]), we have the existence of a subsequence, denoted as the sequence, such that

$$w_n \rightharpoonup w \in L^1([a,b];E).$$

By Mazur's convexity Theorem we obtain a sequence

$$\tilde{w}_n = \sum_{i=0}^{k_n} \lambda_{n,i} w_{n+i}, \qquad \lambda_{n,i} \geq 0, \ \sum_{i=0}^{k_n} \lambda_{n,i} = 1$$

such that $\tilde{w}_n \to w$ in $L^1([a,b];E)$ and, up to a subsequence, $\tilde{w}_n(t) \to w(t)$ for a.a. $t \in [a,b]$.
To conclude we have only to prove that $w(t) \in F(t,q(t))$ for a.a. $t \in [a,b]$. Indeed, let N_0 with Lebesgue measure zero be such that $F(t,\cdot) : E \multimap E_\sigma$ is u.s.c., $w_n(t) \in F(t,q_n(t))$ and $\tilde{w}_n(t) \to w(t)$ for all $t \in [a,b] \setminus N_0$ and $n \in \mathbb{N}$. Fix $t_0 \notin N_0$ and assume by contradiction that $w(t_0) \notin F(t_0,q(t_0))$.
Since $F(t_0,q(t_0))$ is closed and convex, from the Hahn Banach Theorem there is a weakly open convex set $V \supset F(t_0,q(t_0))$ satisfying $w(t_0) \notin \overline{V}$. Since $F(t_0,\cdot) : E \multimap E_\sigma$ is u.s.c., we can find a neighborhood U of $q(t_0)$ such that $F(t_0,x) \subset V$ for all $x \in U$. The convergence $q_n(t_0) \to q(t_0)$ implies the existence of $n_0 \in \mathbb{N}$ such that $q_n(t_0) \in U$ for all $n > n_0$. Therefore $w_n(t_0) \in F(t_0,q_n(t_0)) \subset V$ for all $n > n_0$. Since V is convex we also have that $\tilde{w}_n(t_0) \in V$ for all $n > n_0$ and, by the convergence, we arrive to the contradictory conclusion that $w(t_0) \in \overline{V}$. We conclude that $w(t) \in F(t,q(t))$ for a.a. $t \in [a,b]$. □

We propose now the two continuation principles (see Theorems 2.3 and 2.4) that we use, respectively in Sections 3 and 4, and recall Ky Fan fixed point Theorem (see Theorem 2.5).

THEOREM 2.3 ([1]). *Let Q be a closed, convex subset of a Banach space Y with nonempty interior and $H\colon Q \times [0,1] \multimap Y$ be such that*

(a) H is nonempty convex valued and it has closed graph;

(b) H is compact;

(c) $H(Q,0) \subset Q$;

(d) $H(\cdot,\lambda)$ is fixed points free on the boundary of Q for all $\lambda \in [0,1)$.

Then there exists $y \in Q$ such that $y \in H(y,1)$.

A metric space X is *contractible* if the identity map on it, i.e. $\mathrm{id}_X : X \to X$ is homotopic to a constant map. A compact nonempty metric space X is called an *R_δ-set* if there exists a decreasing sequence $\{X_n\}$ of compact, contractible sets X_n such that $X = \cap\{X_n : n \in \mathbb{N}\}$. Every convex compact subset of a metric space is an R_δ-set (see e.g. [1] for details).

THEOREM 2.4 ([3, Theorem 2.1]). *Let F be a Hausdorff locally convex topological vector space, $X \subset F$ be a convex metrizable set, $Z \subset X$ be an open set in X and $H : Z \times [0,1] \multimap F$ be a compact u.s.c. multimap with R_δ values satisfying*

$$\begin{aligned} &\text{if } \{x_n\} \subset Z \text{ converges to } x \in H(x,\lambda), \text{ for some } \lambda \in [0,1)\,, \text{ there is } n_0 \\ &\text{such that } H(\{x_n\} \times [0,1]) \subset X, \text{ for all } n \geq n_0 \end{aligned} \tag{4}$$

and such that

(1) $H(\cdot,0)(Z) \subset X$;

(2) there exists a compact u.s.c. multimap with R_δ values $H' \colon X \multimap X$ such that $H'\lfloor_Z = H(\cdot,0)$ and $Fix(H') \cap X \setminus Z = \emptyset$.

Then there exists $x \in Z$ such that $x \in H(x,1)$.

When making use of a continuation principle it is often very delicate to show the so called transversality condition, i.e. condition *(d)* in Theorem 2.3 and condition (4) in Theorem 2.4. In both cases we assume here, to this aim, the existence of $R > \|x_0\|$ satisfying

$$D\left[\|x_0\| + \|\eta_{RB\setminus\|x_0\|B}\|_1\right] \leq R \tag{5}$$

with D given in (2) and η appearing in (F3).

THEOREM 2.5. *Let X be a Hausdorff locally convex topological vector space, V be a compact convex subset of X and $G : V \multimap V$ an u.s.c. multimap with closed, convex values. Then G has a fixed point.*

We finally propose a useful compactness result for semicompact sequences (see Theorem 2.7).

DEFINITION 2.6. *We say that a sequence $\{f_n\} \subset L^1([a,b];E)$ is* semicompact *if it is integrably bounded and the set $\{f_n(t)\}$ is relatively compact for a.a. $t \in [a,b]$.*

THEOREM 2.7 ([10, Theorem 5.1.1]). *Let $S \colon L^1([a,b];E) \to C([a,b];E)$ be an operator satisfying the following conditions*

(i) there is $L>0$ such that $\|Sf-Sg\|_C \leq L\|f-g\|_1$ for all $f,g \in L^1([a,b];E)$;

(ii) for any compact $K \subset E$ and sequence $\{f_n\} \subset L^1([a,b];E)$ such that $\{f_n(t)\} \subset K$ for a.a. $t \in [a,b]$ the weak convergence $f_n \rightharpoonup g$ implies $Sf_n \to Sg$.

Then for every semicompact sequence $\{f_n\} \subset L^1([a,b];E)$ the sequence $\{Sf_n\}$ is relatively compact in $C([a,b];E)$ and, moreover, if $f_n \rightharpoonup f_0$ then $Sf_n \to Sf_0$.

3. The case of a compact evolution operator

In this Section we assume that the family $\{A(t)\}$ generates a compact evolution operator and that the nonlinear term F satisfies the regularity condition (F2) and, when not explicitly mentioned, the growth condition (F3).

First we introduce the solution multioperator $T : C([a,b];E) \times [0,1] \multimap C([a,b];E)$ defined as

$$T(q,\lambda) = \left\{ \begin{array}{r} x \in C([a,b];E) \,:\, x(t) = U(t,a)x_0 + \lambda \displaystyle\int_a^t U(t,s)f(s)\,ds, \\ \text{for all } t \in [a,b] \text{ and } f \in S_q \end{array} \right\} \quad (6)$$

which is well-defined according to Proposition 2.2 and we investigate its regularity properties. Notice that the fixed points of $T(\cdot,1)$ are mild solutions of the problem (1).

PROPOSITION 3.1. *The multioperator T has a closed graph.*

Proof. Since $C([a,b];E)$ is a metric space, it is sufficient to prove the sequential closure of the graph. Let $\{q_n\}$, $\{x_n\} \subset C([a,b];E)$ and $\{\lambda_n\} \subset [0,1]$ satisfying $x_n \in T(q_n,\lambda_n)$ for all n and $q_n \to q$, $x_n \to x$ in $C([a,b];E)$, $\lambda_n \to \lambda$ in $[0,1]$. We prove that $x \in T(q,\lambda)$.
The fact that $x_n \in T(q_n,\lambda_n)$ means that there exists a sequence $\{f_n\}$, $f_n \in S_{q_n}$, such that

$$x_n(t) = U(t,a)x_0 + \lambda_n \int_a^t U(t,s)f_n(s)\,ds, \quad \forall\, t \in [a,b]. \quad (7)$$

Let $\Omega \subset E$ be such that $q_n(t)$, $q(t) \in \Omega$ for all $t \in [a,b]$ and $n \in \mathbb{N}$. Since $q_n \to q$ in $C([a,b];E)$, it follows that Ω is bounded and according to (F3) there is $\eta_\Omega \in L^1([a,b];\mathbb{R})$ satisfying $\|f_n(t)\| \le \eta_\Omega(t)$ for a.a. t and every n, implying that $\{f_n\}$ is bounded and uniformly integrable in $L^1([a,b];E)$ and $\{f_n(t)\}$ is bounded in E for a.a. $t \in [a,b]$. Hence, by the reflexivity of the space E and by the Dunford-Pettis Theorem (see [7, p. 294]), we have the existence of a subsequence, denoted as the sequence, and a function g such that $f_n \rightharpoonup g$ in $L^1([a,b];E)$. It is also easy to show that $U(t,\cdot)f_n \rightharpoonup U(t,\cdot)g$ in $L^1([a,t];E)$ for all $t \in [a,b]$. Since $\lambda_n \to \lambda$, we obtain that

$$x_n(t) \rightharpoonup x_0(t) := U(t,a)x_0 + \lambda \int_a^t U(t,s)g(s)\,ds \quad (8)$$

for all $t \in [a,b]$. By the uniqueness of the weak limit in E, we get that $x_0(t) = x(t)$ for all $t \in [a,b]$. Finally, reasoning as in the second part of the proof of Proposition 2.2 it is possible to show that $g(t) \in F(t,q(t))$ for a.a. $t \in [a,b]$. □

PROPOSITION 3.2. *$T(Q \times [0,1])$ is relatively compact, for every bounded $Q \subset C([a,b];E)$.*

Proof. Let $Q \subset C([a,b];E)$ be bounded. Since $C([a,b];E)$ is a metric space it is sufficient to prove the relative sequential compactness of $T(Q\times[0,1])$. Consider $\{q_n\} \subset Q$, $\{x_n\} \subset C([a,b];E)$ and $\{\lambda_n\} \subset [0,1]$ satisfying $x_n \in T(q_n,\lambda_n)$ for all n. By the definition of the multioperator T, there exist a sequence $\{f_n\}$, $f_n \in S_{q_n}$, such that x_n satisfies (7). Let $\Omega \subset E$ be such that $q_n(t) \in \Omega$ for all t and n. Since Q is bounded, we have that Ω is bounded too and according to (F3) there exists $\eta_\Omega \in L^1([a,b];\mathbb{R})$ such that $\|f_n(t)\| \le \eta_\Omega(t)$ for a.a. $t \in [a,b]$ and all n.
According to (2) and the compactness of the evolution operator U, the sequence $\{U(t,\cdot)f_n\}$ is semicompact in $[a,t]$ for every fixed $t \in (a,b]$ (see Definition 2.6). Since the operator $S\colon L^1([a,t];E) \to C([a,t];E)$ defined by $Sf(\tau) = \int_a^\tau f(s)\,ds$ for $\tau \in [a,t]$ satisfies conditions *(i)* and *(ii)* in Theorem 2.7 we obtain that the sequence

$$\tau \mapsto \int_a^\tau U(t,s)f_n(s)\,ds, \qquad \tau \in [0,t],\ n \in \mathbb{N}$$

is relatively compact in $C([a,t];E)$; in particular $\left\{\int_a^t U(t,s)f_n(s)\,ds\right\}$ is a relatively compact set in E for all $t \in [a,b]$.
Now consider $a < t_0 < t \le b$. For every $\sigma \in (0, t_0 - a)$ we have that

$$\begin{aligned}\left\|\int_a^t U(t,s)f_n(s)\,ds - \int_a^{t_0} U(t_0,s)f_n(s)\,ds\right\| &\le \left\|\int_a^{t_0-\sigma} [U(t,s) - U(t_0,s)]\,f_n(s)\,ds\right\| \\ &\quad + \left\|\int_{t_0-\sigma}^{t_0} [U(t,s) - U(t_0,s)]\,f_n(s)\,ds\right\| + \left\|\int_{t_0}^t U(t,s)f_n(s)\,ds\right\|. \end{aligned} \tag{9}$$

Since it is known that $t \to U(t,s)$ is continuous in the operator norm topology, uniformly with respect to s such that $t - s$ is bounded away from zero (see e.g. [13]), for each $\epsilon > 0$ there is $\delta \in (0, t_0 - a)$ satisfying

$$\left\|\int_a^{t_0-\delta} [U(t,s) - U(t_0,s)]\,f_n(s)\,ds\right\| \le \epsilon \int_a^{t_0-\delta} \eta_\Omega(s)\,ds;$$

whenever $t - t_0 < \delta$; hence, according to (9), we obtain that

$$\left\|\int_a^t U(t,s)f_n(s)\,ds - \int_a^{t_0} U(t_0,s)f_n(s)\,ds\right\| \le \epsilon\int_a^{t_0-\delta} \eta_\Omega(s)\,ds + 2D\int_{t_0-\delta}^t \eta_\Omega(s)\,ds.$$

Thanks to the absolute continuity of the integral function, it implies that the sequence $\{\int_a^t U(t,s)f_n(s)\,ds\}$ is equicontinuous in $[a,b]$. Consequently, passing

to a subsequence, denoted as the sequence, such that $\lambda_n \to \lambda \in [0,1]$ and using Arzelá-Ascoli theorem, we obtain that $\{x_n\}$ is relatively compact in $C([a,b];E)$ and the proof is complete. □

PROPOSITION 3.3. *The multioperator T has convex and compact values.*

Proof. Fix $q \in C([a,b];E)$ and $\lambda \in [0,1]$, since F is convex valued, the set $T(q,\lambda)$ is convex from the linearity of the integral and of the operator $U(t,s)$ for all $(t,s) \in \Delta$. The compactness of $T(q,\lambda)$ follows by Propositions 3.1 and 3.2. □

THEOREM 3.4. *Problem* (1) *under conditions (A) (F1), (F2), (F3),* (5) *and with $\{A(t)\}_{t\in[a,b]}$ generating a compact evolution operator has at least one solution.*

Proof. Consider the set $Q = C([a,b];RB)$ with R defined in (5). We show that the solution multioperator T defined in (6), when restricted to Q, satisfies the assumptions of Theorem 2.3. In fact Q is closed, convex, bounded and with a nonempty interior. According to Propositions 3.1, 3.2 and 3.3, T satisfies conditions *(a)* and *(b)* in Theorem 2.3.
Notice that $T(Q\times\{0\}) \subset D\|x_0\|B \subset \operatorname{int} Q$, hence condition *(c)* in Theorem 2.3 holds and $T(\cdot,0)$ is fixed point free on ∂Q. Let us now prove that T satisfies condition *(d)* also for $\lambda \in (0,1)$. Let $q \in Q$ and $\lambda \in (0,1)$ be such that $q \in T(q,\lambda)$ and assume, by contradiction, the existence of $t_0 \in (a,b]$ such that $q(t_0) \in \partial Q$ which is equivalent to $\|q(t_0)\| = R$. Since q is continuous and $q \in T(q,\lambda)$, from $\|x_0\| < R$ it follows that there exist $\hat{t}_0, \hat{t}_1 \in (a,t_0]$ with $\hat{t}_0 < \hat{t}_1$ such that $\|q(\hat{t}_0)\| = \|x_0\|$, $\|x_0\| < \|q(t)\| < R$ for $t \in (\hat{t}_0,\hat{t}_1)$ and $\|q(\hat{t}_1)\| = R$. Moreover there exists $f \in S_q$ such that $q(t) = U(t,\hat{t}_0)q(\hat{t}_0) + \lambda\int_{\hat{t}_0}^t U(t,s)f(s)\,ds$ for $t \in [\hat{t}_0,\hat{t}_1]$. According to (F3), $\|f(t)\| \le \eta_{RB\setminus\|x_0\|B}(t)$ for $t \in (\hat{t}_0,\hat{t}_1)$; so we arrive to the contradiction $R = \|q(\hat{t}_1)\| \le D[\|x_0\| + \lambda\|\eta_{RB\setminus\|x_0\|B}\|_1] < R$, and also condition *(d)* in Theorem 2.3 is satisfied.
Hence $T(\cdot,1)$ has a fixed point in Q which is a mild solution of problem (1). □

When the nonlinear term F has an at most linear growth, i.e. when it satisfies (F3′) instead of condition (F3), then the transversality condition (5) can be eliminated and the compactness of the solution set can be obtained too.

THEOREM 3.5. *Under conditions (A), (F1), (F2), (F3′) and with $\{A(t)\}_{t\in[a,b]}$ generating a compact evolution operator, the solution set of problem* (1) *is nonempty and compact.*

Proof. Consider the set Q defined as

$$Q = \left\{q \in C([a,b];E) \ : \ \|q(t)\| \le Re^{Lt} \text{ a.a. } t \in [a,b]\right\}$$

where L and R are such that

$$\max_{t\in[a,b]} D\int_a^t e^{L(s-t)}\alpha(s)ds := \overline{\beta} < 1,$$

$$R \geq e^{-La}D(\|x_0\| + \|\alpha\|_1)(1-\overline{\beta})^{-1}$$

and α was given in (F3′). Define the operator $\Gamma := T(\cdot, 1)$. According to Propositions 3.1, 3.2 and 3.3, it is easy to see that Γ is locally compact, with nonempty convex compact values and it has a closed graph. Hence it is also u.s.c. (see e.g. [10, Theorem 1.1.5]). We prove now that Γ maps the set Q into itself.
Indeed if $q \in Q$ and $x \in \Gamma(q)$ there exists a function $f \in S_q$ such that

$$x(t) = U(t,a)x_0 + \int_a^t U(t,s)f(s)\,ds.$$

By hypothesis $(F3')$ we have that

$$\begin{aligned}\|x(t)\| &= \left\|U(t,a)x_0 + \int_a^b U(t,s)f(s)\,ds\right\| \leq D\left(\|x_0\| + \int_a^t \alpha(s)(1+Re^{Ls})ds\right)\\ &\leq D\left(\|x_0\| + \|\alpha\|_1\right) + D\int_a^t \alpha(s)Re^{Ls}ds \leq D\left(\|x_0\| + \|\alpha\|_1\right) + Re^{Lt}\overline{\beta}\\ &\leq Re^{La}(1-\overline{\beta}) + Re^{Lt}\overline{\beta} \leq Re^{Lt}.\end{aligned}$$

Then $\Gamma(Q) \subseteq Q$. Let $V = \Gamma(Q)$ and $W = \overline{\text{co}}(V)$, where $\overline{\text{co}}(V)$ denotes the closed convex hull of V. Since $\overline{V}$ is a compact set, W is compact too. Moreover from the fact that $\Gamma(Q) \subset Q$ and that Q is a convex closed set we have that $W \subset Q$ and hence

$$\Gamma(W) = \Gamma(\overline{\text{co}}(\Gamma(Q))) \subseteq \Gamma(Q) = V \subset W.$$

Hence, according to Theorem 2.5, Γ has a fixed point, which is a solution of (1).
We prove now that the solution set is compact. Indeed a solution of the problem (1) is a fixed point of the operator Γ. If $x \in \Gamma(x)$, by the definition of Γ and $(F3')$ we have the existence of $f \in S_x$ and reasoning as above

$$\begin{aligned}\|x(t)\| &\leq \|U(t,s)x_0\| + \int_0^t \|U(t,s)f(s)\|\,ds\\ &\leq D\left(\|x_0\| + \|\alpha\|_1 + \int_0^t \alpha(s)\|x(s)\|\,ds\right).\end{aligned}$$

By the Gronwall's inequality it holds

$$\|x(t)\| \leq D(\|x_0\| + \|\alpha\|_1)e^{D\|\alpha\|_1} := \overline{n}.$$

Hence $\operatorname{Fix}\Gamma$ is a bounded set and so $\Gamma(\operatorname{Fix}\Gamma)$ is relatively compact. Since $\operatorname{Fix}\Gamma \subset \Gamma(\operatorname{Fix}\Gamma)$, then $\operatorname{Fix}\Gamma$ is relatively compact too. Finally, according to the closure of the graph of Γ, $\operatorname{Fix}\Gamma$ is also closed and hence compact. □

4. The case of a non-compact evolution operator

If we drop the assumption that the family $\{A(t)\}$ generates a compact evolution operator, we need stronger regularity hypotheses on F to consider the richer class of evolution operators which we discuss now. We take, precisely, F satisfying $(F2')$; moreover, when not explicitly mentioned, we always assume the growth restriction (F3).
Since an u.s.c. multimap from E_σ to E_σ is u.s.c. from E to E_σ, the Proposition 2.2 is still true under the condition $(F2')$. Hence the set $S_q \neq \emptyset$ for any $q \in C([a,b];E)$ and the solution operator $T : C([a,b];E) \times [0,1] \multimap C([a,b];E)$ can be defined as in (6) and it has nonempty convex values. With a similar reasoning as in Proposition 3.1 it is also possible to prove that T has a weakly sequentially closed graph. Now we show that T is locally weakly compact.

PROPOSITION 4.1. *$T(Q \times [0,1])$ is weakly relatively compact for every bounded $Q \subset C([a,b];E)$.*

Proof. Let $Q \subset C([a,b];E)$ be bounded. We first prove that $T(Q \times [0,1])$ is weakly relatively sequentially compact.
Consider $\{q_n\} \subset Q$, $\{x_n\} \subset C([a,b];E)$ and $\{\lambda_n\} \subset [0,1]$ satisfying $x_n \in T(q_n, \lambda_n)$ for all n. By the definition of T, there exist a sequence $\{f_n\}$, $f_n \in S_{q_n}$ such that x_n satisfies (7). Passing to a subsequence, denoted as the sequence, we have that $\lambda_n \to \lambda \in [0,1]$. Moreover, reasoning as in the proof of Proposition 3.1, we obtain that there exists a subsequence, denoted as the sequence, and a function g such that $f_n \rightharpoonup g$ in $L^1([a,b];E)$, implying that $x_n(t)$ satisfies (8) for all $t \in [a,b]$. Furthermore, by (2) and the weak convergence of $\{f_n\}$ we have

$$\|x_n(t)\| \leq D\|x_0\| + D\|f_n\|_1 \leq N$$

for all $n \in \mathbb{N}$, $t \in [a,b]$, and for some $N > 0$. Hence $x_n \rightharpoonup x_0$ in $C([a,b];E)$. Thus $T(Q \times [0,1])$ is weakly relatively sequentially compact, hence weakly relatively compact by the Eberlein-Smulian Theorem (see [11, Theorem 1, p. 219]). □

REMARK 4.2. *Notice that, since T has weakly sequentially closed graph and according to Proposition 4.1, T has also weakly compact values.*

THEOREM 4.3. *Assume conditions (A), (F1), $(F2')$, $(F2'')$, (F3) and* (5). *If E is separable, then problem* (1) *has at least one solution.*

Proof. Put $\hat{R} := D\|x_0\| + \|\eta_{RB}\|_1 + 1$ with R defined in (5) and η in (F3) and define $Q = C([a,b]; \hat{R}B)$. The set Q is closed, convex and bounded. Since E is separable, $C([a,b];E)$ is separable too and then Q is also metrizable. Consider the solution operator T defined in (6). Now we prove that it satisfies Theorem 2.4 with $F = (C([a,b];E))_\sigma$ and $X = Z = Q_\sigma$. According to Proposition 4.1, $T(Q \times [0,1])$ is weakly relatively compact so, in particular, $T(Q \times [0,1])$ is bounded and then $(T(Q \times [0,1]))_\sigma$ is metrizable. Since $T\colon Q \times [0,1] \multimap C([a,b];E)$ is weakly sequentially closed then it has weakly compact values and hence it is R_δ-valued. Moreover, according to Eberlein-Smulian Theorem and [10, Theorem 1.1.5], T is u.s.c. when both Q and $C([a,b];E)$ are endowed with the weak topology. Reasoning as in the proof of Theorem 3.4, it is also possible to show that condition *(1)* in Theorem 2.4 is satisfied; while condition *(2)* is trivially true. It remains to prove (4). So take $q_n \rightharpoonup q \in T(q, \lambda_0)$ for some $\lambda_0 \in [0,1)$. Let $x_n \in T(q_n, \lambda_n)$ for some $\lambda_n \in [0,1]$ and all n; then x_n satisfies (7) for some $f_n \in S_{q_n}$ and according to (F3) $\|f_n(t)\| \leq \eta_{\hat{R}B}(t)$ for a.a. $t \in [a,b]$. Reasoning as in the proof of Proposition 3.1 we obtain a subsequence, denoted as the sequence, such that $f_n \rightharpoonup g \in L^1([a,b];E)$. Up to a subsequence we also have that $\lambda_n \to \lambda \in [0,1]$. Moreover, since $\{f_n(t)\} \subset F(t, \hat{R}B)$, according to $(F2'')$ we have that $\{f_n(t)\}$ is relatively compact for a.a. t. Let $G\colon L^1([a,b];E) \to C([a,b];E)$ be the generalized Cauchy operator associated to U, i.e. let $Gf(t) = \int_a^t U(t,s)f(s)\,ds$ for $t \in [a,b]$. It satisfies condition *(i)* in Theorem 2.7 and according to [5, Theorem 2], it also satisfies condition *(ii)* in Theorem 2.7. Hence $x_n \to x$ in $C([a,b];E)$ where $x(t) := U(t,a)x_0 + \lambda \int_a^t U(t,s)g(s)\,ds$ for $t \in [a,b]$. Since T has sequentially weakly closed graph, we obtain that $x \in T(q,\lambda)$. According to (5) and with a similar reasoning as in the proof of Theorem 3.4, we can show that $\|q(t)\| < R$ for all $t \in [a,b]$. Condition (F3) then implies that $\|g(t)\| \leq \eta_{RB}(t)$ a.e. in $[a,b]$ and hence $\|x(t)\| \leq D\|x_0\| + D\|\eta_{RB}\|_1 < \hat{R}$ for all $t \in [a,b]$ and we can find n_0 such that $x_n \in Q$ for every $n \geq n_0$. All the assumptions of Theorem 2.4 are then satisfied and hence $T(\cdot,1)$ has a fixed point which is a solution of problem (1) thus the proof is complete. □

If we assume, as in the previous section, the stronger growth condition $(F3')$, instead of $(F3)$, we can remove conditions (5) and $(F2'')$ as well as the requirement of the separability of the space E. Indeed, recalling that by the Krein Smulian Theorem (see e.g. [7, p. 434]) the convex closure of a weakly compact set is weakly compact, it is possible to reason exactly as in the proof of Theorem 3.2 to obtain the following result.

THEOREM 4.4. *Under assumptions (A), (F1),$(F2')$ and $(F3')$ the solution set of problem* (1) *is nonempty and weakly compact.*

5. Application to a parabolic partial differential inclusion

Let $t \in [0,T]$ and $\Omega \subseteq \mathbb{R}^n$ be a bounded domain with a sufficiently regular boundary. Consider the initial value problem

$$\begin{cases} u_t \in \Delta u + \left[p_1\left(t,x,\int_\Omega k(x,y)u(t,y)dy\right), p_2\left(t,x,\int_\Omega k(x,y)u(t,y)dy\right)\right] f(t,u(t,x)), \\ \qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad t \in [0,T]\, x \in \Omega \\ u(t,x) = 0 \quad t \in [0,T],\, x \in \partial\Omega \\ u(0,x) = u_0(x), \quad x \in \Omega \end{cases} \tag{10}$$

under the following hypotheses:

(a) $k\colon \Omega \times \Omega \to \mathbb{R}$ is measurable with $k(x,\cdot) \in L^2(\Omega;\mathbb{R})$ and $\|k(x,\cdot)\|_2 \le 1$ for all $x \in \Omega$;

(b) $f\colon [0,T] \times \mathbb{R} \to \mathbb{R}$ is a Carathéodory function with $f(t,\cdot)$ L-Lipschitzian and $f(t,0) = 0$ for a.a. $t \in [0,T]$;

(c) $u_0 \in L^2(\Omega;\mathbb{R})$;

(d) $p_1, p_2\colon [0,T] \times \Omega \times \mathbb{R} \to \mathbb{R}$ satisfy the following conditions:

  (i) $p_i(\cdot,\cdot,r)$ is measurable for $i = 1,2$ and all $r \in \mathbb{R}$;

  (ii) $-p_1(t,x,\cdot)$ and $p_2(t,x,\cdot)$ are u.s.c. for a.a. $t \in [0,T]$ and all $x \in \Omega$;

  (iii) $p_1(t,x,r) \le p_2(t,x,r)$ in $[0,T] \times \Omega \times \mathbb{R}$;

  (iv) there exist $\psi \in L^1([0,T];\mathbb{R})$, $M\colon [0,\infty) \to \mathbb{R}$ increasing and $R > \|u_0\|_2$ such that $|p_i(t,x,r)| \le \psi(t)M(|r|)$ for $i = 1,2$ and all x and

$$\|u_0\|_2 + \|\psi\|_1 LRM(R) \le R. \tag{11}$$

We search for solutions $u \in C([a,b];L^2(\Omega;\mathbb{R}))$ of the initial value problem (10). Namely the following abstract formulation

$$\begin{cases} y'(t) \in Ay(t) + F(t,y(t)), & t \in [0,T] \\ y(0) = y_0, \end{cases} \tag{12}$$

should be satisfied, with $y(t) = u(t,\cdot) \in L^2(\Omega;\mathbb{R})$ for any $t \in [0,T]$. $A : W^{2,2}(\Omega;\mathbb{R}) \cap W_0^{1,2}(\Omega;\mathbb{R}) \to L^2(\Omega;\mathbb{R})$ is the linear operator defined as $Ay = \Delta y$ and $y_0 = u_0(\cdot)$. Given $\alpha \in L^2(\Omega;\mathbb{R})$, let $I_\alpha\colon \Omega \to \mathbb{R}$ be the function defined by $I_\alpha(x) = \int_\Omega k(x,y)\alpha(y)\,dy$. I_α is well-defined and measurable, according to (a), and it satisfies $|I_\alpha(x)| \le \|\alpha\|_2$ for all $x \in \Omega$. Given $(t,\alpha) \in [0,T] \times L^2(\Omega;\mathbb{R})$, we define the multimap $F : [0,T] \times L^2(\Omega;\mathbb{R}) \multimap L^2(\Omega;\mathbb{R})$ as $y \in F(t,\alpha)$ if and only

if there is a measurable function $\beta\colon \Omega \to \mathbb{R}$ satisfying $p_1\,(t,x,I_\alpha(x)) \le \beta(x) \le p_2\,(t,x,I_\alpha(x))$ for all $x \in \Omega$ such that $y(x) = \beta(x) f(t,\alpha(x))$ for all $x \in \Omega$.

Notice that, given $(t,\alpha) \in [0,T] \times L^2(\Omega;\mathbb{R})$ and according to *(d)(i)(ii)*, the maps $x \longmapsto p_i\,(t,x,I_\alpha(x))$, $i = 1,2$ are measurable in Ω; hence F has nonempty values and it is easy to see that they are also convex. Moreover $\|y\|_2 \le LM(\|\alpha\|_2)\|\alpha\|_2\psi(t)$, for all $y \in F(t,\alpha)$. Consequently, if $W \subset L^2(\Omega;\mathbb{R})$ is bounded, that is if $\|w\|_2 \le \mu$ for some $\mu > 0$ and all $w \in W$ we have that

$$\|F(t,W)\|_2 \le L\mu M(\mu)\psi(t) \tag{13}$$

implying (F3).

Now we investigate (F2) and hence we fix $t \in [a,b]$ and consider two sequences $\{\alpha_n\}, \{y_n\} \subset L^2(\Omega;\mathbb{R})$ satisfying $\alpha_n \to \alpha$, $y_n \rightharpoonup y$ in $L^2(\Omega;\mathbb{R})$ and $y_n \in F(t,\alpha_n)$ for all $n \in \mathbb{N}$. Notice that $I_{\alpha_n}(x) \to I_\alpha(x)$ for all x. Since $\{\alpha_n\}$ is bounded, there is $\sigma > 0$ such that $\|\alpha_n\|_2 \le \sigma$ for all n. According to *(b)* the sequence $f(t,\alpha_n(\cdot)) \to f(t,\alpha(\cdot))$ in $L^2(\Omega;\mathbb{R})$ and then, passing to a subsequence denoted as usual as the sequence, we obtain that $f(t,\alpha_n(x)) \to f(t,\alpha(x))$ for a.a. $x \in \Omega$. By Mazur's convexity Theorem we have the existence of a sequence

$$\tilde{y}_n = \sum_{i=0}^{k_n} \delta_{n,i} y_{n+i}, \qquad \delta_{n,i} \ge 0,\ \sum_{i=0}^{k_n} \delta_{n,i} = 1$$

such that $\tilde{y}_n \to y$ in $L^2(\Omega;\mathbb{R})$ and up to a subsequence, denoted as the sequence, $\tilde{y}_n(x) \to y(x)$ for a.a. $x \in \Omega$. We prove now that $y \in F(t,\alpha)$. In fact, if $f(t,\alpha(x)) > 0$ then also $f(t,\alpha_n(x)) > 0$ for n sufficiently large, and it implies that $p_1(t,x,I_{\alpha_n}(x))f(t,\alpha_n(x)) \le y_n(x) \le p_2(t,x,I_{\alpha_n}(x))f(t,\alpha_n(x))$ for a.a. x. Consequently

$$\sum_{i=0}^{k_n} \delta_{n,i} p_1(t,x,I_{\alpha_{n+i}}) f(t,\alpha_{n+i}(x)) \le \tilde{y}_n(x) \le \sum_{i=0}^{k_n} \delta_{n,i} p_2(t,x,I_{\alpha_{n+i}}) f(t,\alpha_{n+i}(x)).$$

Passing to the limit as $n \to \infty$ and according to *(d)(ii)*, we obtain that $p_1(t,x,I_\alpha(x))f(t,\alpha(x)) \le y(x) \le p_2(t,x,I_\alpha(x))f(t,\alpha(x))$. With a similar reasoning we arrive to the estimate

$$p_2(t,x,I_\alpha(x))f(t,\alpha(x)) \le y(x) \le p_1(t,x,I_\alpha(x))f(t,\alpha(x))$$

when $f(t,\alpha(x)) < 0$. So, it remains to consider $\Omega_0 = \{x \in \Omega : f(t,\alpha(x)) = 0\}$. Notice that $f(t,\alpha_n(x)) \to 0$ in Ω_0. Since $y_n(\cdot) = \beta_n(\cdot) f(t,\alpha_n(\cdot))$ for some bounded and measurable $\beta_n\colon \Omega \to \mathbb{R}$ satisfying $p_1(t,x,I_{\alpha_n}(x)) \le \beta_n(x) \le p_2(t,x,I_{\alpha_n}(x))$ a.e. in Ω, it follows that $y_n(x) \to 0$ and then also $\tilde{y}_n(x) \to 0$, implying $y(x) \equiv 0$ in Ω_0. Therefore, it is possible to define a measurable function $\beta\colon \Omega \to \mathbb{R}$ such that $p_1(t,x,I_\alpha(x)) \le \beta(x) \le p_2(t,x,I_\alpha(x))$ and $y(x) =$

$\beta(x)f(t,\alpha(x))$ a.e. in Ω. We have showed that F has closed graph. Then by (13) $F(t,\cdot)$ has weakly compact values and it is locally weakly compact, since $L^2(\Omega;\mathbb{R})$ is reflexive, thus it satisfies (F2) (see e.g. [10, Theorem 1.1.5]). Moreover, according to Pettis measurability Theorem (see [15, p. 278]) it is possible to see that, for all $\alpha \in L^2(\Omega;\mathbb{R})$, the map $t \mapsto p_1\left(t,\cdot,I_\alpha(\cdot)\right) f(t,\alpha(\cdot))$ is a measurable selection of $F(\cdot,\alpha)$, hence condition (F1) is satisfied. According to (13), for $\Theta = RB \setminus \|u_0\|B$ we can define η_Θ in (F3) as $\eta_\Theta(t) = LRM(R)\psi(t)$ and hence, according to *(d)(iv)* also condition (5) is satisfied.
All the assumptions of Theorem 3.4 are then satisfied and hence problem (12) is solvable, implying that (10) has at least one solution $u \in C([a,b];L^2(\Omega;\mathbb{R}))$.

6. Applications to an hyperbolic partial differential inclusion

Let Ω be a bounded domain in $\mathbb{R}^n$ with a sufficiently regular boundary. Consider the feedback control problem associated to a partial differential equation

$$\begin{cases} u_{tt} = \Delta u + p\left(t,x,\displaystyle\int_\Omega u(t,\xi)\,d\xi\right) u(t,x) + a(t,x)w(t,x) + b(t,x), \ \text{in}\,[0,d]\times\Omega \\ w(t,x) \in W(u(t,x)) \\ u(t,x) = 0 \ \ t \in [0,d], \ x \in \partial\Omega \\ u(0,x) = u_0(x); u_t(0,x) = u_1(x)\ , \ x \in \Omega \end{cases} \tag{14}$$

where $W(r) = \{s \in \mathbb{R} \ : \ \ell r + m_1 \le s \le \ell r + m_2\}$, with $\ell > 0$ and $m_1 < m_2$. Assume the following hypotheses:

(i) a and b are globally measurable in $[0,d]\times\Omega$ and there exist two functions $\varphi_1, \varphi_2 \in L^1([0,d];\mathbb{R})$ such that

$$|a(t,x)| \le \varphi^1(t) \ \text{ for a.a. } x \in \Omega \text{ and } \forall\, t \in [0,d];$$

$$|b(t,x)| \le \varphi^2(t) \ \text{ for a.a. } x \in \Omega \text{ and } \forall\, t \in [0,d];$$

the map $p : [0,d]\times\Omega\times\mathbb{R} \to \mathbb{R}$ satisfies the following conditions

(ii) $p(\cdot,\cdot,r) : [0,d]\times\Omega \to \mathbb{R}$ is measurable, for all $r \in \mathbb{R}$;

(iii) $p(t,x,\cdot) : \mathbb{R} \to \mathbb{R}$ is continuous, for a.a. $(t,x) \in [0,d]\times\Omega$;

(iv) there exists $\varphi^3 \in L^1([0,d];\mathbb{R})$ such that

$$|p(t,x,r)| \le \varphi^3(t) \ \text{ for a.e. } x \in \Omega, \ \forall\, t \in [0,d] \text{ and } \forall\, r \in \mathbb{R}.$$

Let $y : [0,d] \to L^2(\Omega;\mathbb{R})$, $v : [0,d] \to L^2(\Omega;\mathbb{R})$, $f : [0,d] \times L^2(\Omega;\mathbb{R}) \times L^2(\Omega;\mathbb{R}) \to L^2(\Omega;\mathbb{R})$, and $V : L^2(\Omega;\mathbb{R}) \multimap L^2(\Omega;\mathbb{R})$ be the maps defined by

$$
\begin{aligned}
&y(t) = u(t,\cdot);\\
&v(t) = w(t,\cdot);\\
&f(t,\alpha,\beta) : \Omega \to \mathbb{R},\ f(t,\alpha,\beta)(x) = p\left(t,x,\int_\Omega \alpha(\xi)\,d\xi\right)\alpha(x) + a(t,x)\beta(x) + b(t,x);\\
&V(z) = \{v \in L^2(\Omega;\mathbb{R}) \ :\ \ell z(x) + m_1 \le v(x) \le \ell z(x) + m_2,\ \text{a.a. } x \in \Omega\}.
\end{aligned}
$$

In the Hilbert space $L^2(\Omega;\mathbb{R})$ problem (14) can be rewritten as a second order inclusion of the following form

$$
\begin{cases} y''(t) \in Ay(t) + F(t,y(t)), \quad t \in [0,d],\ y(t) \in L^2(\Omega;\mathbb{R}) \\ y(0) = y_0;\ y'(0) = y_1 \end{cases} \tag{15}
$$

where $F(t,y(t)) = f(t,y(t),V(y(t)))$, $y_0 = u_0(\cdot)$, $y_1 = u_1(\cdot)$ and $A : D(A) = W^{2,2}(\Omega;\mathbb{R}) \cap W_0^{1,2}(\Omega;\mathbb{R}) \to L^2(\Omega;\mathbb{R})$ is the linear operator defined as $Ay = \Delta y$.

From the fact that $-A$ is a self-adjoint and positive definite operator on $L^2(\Omega;\mathbb{R})$ with a compact inverse, we have that there exists a unique positive definite square root $(-A)^{1/2}$ with domain $D((-A)^{1/2}) = W_0^{1,2}(\Omega;\mathbb{R})$. Introduce the Hilbert space $\mathcal{E} = W_0^{1,2}(\Omega;\mathbb{R}) \times L^2(\Omega;\mathbb{R})$ with the inner product

$$
\left\langle \begin{pmatrix} p_0 \\ p_1 \end{pmatrix} \cdot \begin{pmatrix} q_0 \\ q_1 \end{pmatrix} \right\rangle = \int_\Omega \nabla p_0\, \nabla q_0\, dx + \int_\Omega p_0\, q_0\, dx + \int_\Omega p_1\, q_1\, dx.
$$

Since the operator

$$
\mathcal{A} = \begin{pmatrix} 0 & I \\ A & 0 \end{pmatrix}, \quad D(\mathcal{A}) = D(A) \times W_0^{1,2}(\Omega;\mathbb{R})
$$

generates a strongly continuous semigroup (see e.g. [8]), we can treat (15) as a first order semilinear differential inclusion in $\mathcal{E}$

$$
\begin{cases} z'(t) \in \mathcal{A}z(t) + \mathcal{F}(t,z(t)),\ t \in [0,d] \\ z(0) = \begin{pmatrix} y_0 \\ y_1 \end{pmatrix} \end{cases} \tag{16}
$$

where $\mathcal{F} : [0,d] \times \mathcal{E} \multimap \mathcal{E}$ is defined as

$$
\mathcal{F}\left(t, \begin{pmatrix} c^0 \\ c^1 \end{pmatrix}\right) = \begin{pmatrix} 0 \\ F(t,c^0) \end{pmatrix}.
$$

Observe that the semigroup generated by $\mathcal{A}$ is noncompact.

Denoted $I_\alpha = \int_\Omega \alpha(y)dy$, by the separability of the space $L^2(\Omega;\mathbb{R})$ and the

Pettis measurability Theorem [15], we have that the map $t \to p(t,\cdot,I_\alpha)\alpha(\cdot) + a(t,\cdot)(\ell\alpha(\cdot)+m_1)+b(t,\cdot)$ is a measurable selection of $F(\cdot,\alpha)$. We prove, now, that the map F satisfies condition (F2′). Reasoning like in Section 5 it is possible to prove that the multimap V is weakly sequentially closed. Let, now, $t \in [0,d]$ be fixed, let $\{\alpha_n\} \subset L^2(\Omega;\mathbb{R})$, be weakly convergent to $\alpha \in L^2(\Omega;\mathbb{R})$ and let $\{w_n\} \subset L^2(\Omega;\mathbb{R})$ with $w_n \in F(t,\alpha_n)$ for any $n \in \mathbb{N}$, be weakly convergent to $w \in L^2(\Omega;\mathbb{R})$. By the definition of the multimap F we have

$$w_n = f_1(t,\alpha_n) + f_2(t,\beta_n), \quad \text{with} \quad \beta_n \in V(\alpha_n) \text{ for any } n \in \mathbb{N},$$

where $f_1(t,\alpha)(x) = p(t,x,I_\alpha)\alpha(x)$ and $f_2(t,\beta)(x) = a(t,x)\beta(x) + b(t,x)$. By the definition of the multimap V and the weak convergence of $\{\alpha_n\}$ we have that the sequence $\{\beta_n\}$ is norm bounded. Hence, by the reflexivity of the space $L^2(\Omega;\mathbb{R})$, up to subsequence, $\{\beta_n\}$ weakly converges to $\beta \in L^2(\Omega;\mathbb{R})$ and the weak closure of the multimap V implies $\beta \in V(\alpha)$. Moreover by the continuity of the map p we have that $\{f_1(t,\alpha_n)\}$ converges weakly to $f_1(t,\alpha)$ and it is easy to see that $\{f_2(t,\beta_n)\}$ converges weakly to $f_2(t,\beta)$. In conclusion we have obtained

$$w = f_1(t,\alpha) + f_2(t,\beta) \in f(t,\alpha,V(\alpha)) = F(t,\alpha).$$

Furthermore, easily, V has convex and closed values, thus, by the linearity of the map f_2 and following the same reasonings above, F is convex closed valued as well.
Finally (see e.g. [4])

$$\|F(t,\alpha)\|_2 \le \left(\varphi^3(t) + 2\ell\varphi^1(t)\right)\|\alpha\|_2 + |\Omega|^{1/2}\left[(m_1+m_2)\varphi^1(t) + \varphi^2(t)\right],$$

obtaining both that for any $t \in [0,d]$ and $\alpha \in L^2(\Omega;\mathbb{R})$ the set $F(t,\alpha)$ is bounded (hence relatively compact by the reflexivity of $L^2(\Omega;\mathbb{R})$), and that condition $(F3')$ is satisfied.
Let $z = (y_0, y_1)$ be a solution of (16). Applying the Implicit Function Theorem of Filippov's type (see [9, Theorem 7.2]) we have that there exists $v : [0,d] \to L^2(\Omega;\mathbb{R})$ such that $v(t) \in V(y_0(t))$ and $g(t) = f(t,y_0(t),v(t))$, $t \in [0,d]$. Hence the feedback control problem (14) admits a weakly compact set of solutions.

References

[1] J. Andres and L. Górniewicz, *Topological Fixed Point Principles for Boundary Value Problems*, Kluwer, Dordrecht, 2003.

[2] I. Benedetti, L. Malaguti and V. Taddei, *Semilinear differential inclusions via weak topologies*, J. Math. Anal. Appl. **368** (2010), 90–102.

[3] I. Benedetti, L. Malaguti and V. Taddei, *Erratum and addendum to: "Two-point b.v.p. for multivalued equations with weakly regular r.h.s."*, Nonlinear Anal. **75** (2012), 2376–2377.

[4] I. Benedetti, L. Malaguti and V. Taddei, *Nonlocal semilinear evolution equations without strong compactness: theory and applications*, preprint.
[5] T. Cardinali and P. Rubbioni, *On the existence of mild solutions of semilinear evolution differential inclusions*, J. Math. Anal. Appl. **308** (2005), 620–635.
[6] B. Cascales, V. Kadets and J. Rodriguez, *Measurability and selections of multi-functions in Banach spaces*, J. Convex Anal. **17** (2010), 229–240.
[7] N. Dunford and J. T. Schwartz, *Linear Operators*, John Wiley and Sons, Inc., New York, 1988.
[8] K. J. Engel and R. Nagel, *One-Parameter Semigroups for Linear Evolution Equations*, in: Graduate Texts in Mathematics, vol. 194, Springer, New York, 2000.
[9] C. Himmelberg, *Measurable relations*, Fund. Math. **87** (1975), 53–72.
[10] M. I. Kamenskii, V. Obukhovskii and P. Zecca, *Condensing Multivalued Maps and Semilinear Differential Inclusions in Banach Spaces*, de Gruyter, Berlin, 2001.
[11] L. V. Kantorovich and G. P. Akilov, *Functional Analysis*, Pergamon Press, Oxford, 1982.
[12] G. Marino, *Nonlinear boundary value problems for multivaued differential equations in Banach spaces*, Nonlinear Anal. **14** (1990), 545–558.
[13] N. S. Papageorgiu, *Existence of solutions of boundary value prpoblems of semilinear evolution inclusions*, Indian J. Pure Appl. Math. **23** (1992), 477–488.
[14] A. Pazy, *Semigroups of Linear Operators and Applications to Partial Differential Equations*, Springer, Berlin, 1983.
[15] B. J. Pettis, *On the integration in vector spaces*, Trans. Amer. Math. Soc. **44** (1938), 277–304.
[16] I. I. Vrabie, *Compactness Methods for Nonlinear Evolutions*, 2nd ed., Longman House, Harlow, 1990.

Authors' addresses:

Irene Benedetti
Dip. di Matematica e Informatica,
Università di Perugia
E-mail: irene.benedetti@dmi.unipg.it

Luisa Malaguti
Dip. di Scienze e Metodi dell'Ingegneria,
Università di Modena e Reggio Emilia
E-mail: luisa.malaguti@unimore.it

Valentina Taddei
Dip. di Scienze Fisiche, Informatiche e Matematiche,
Università di Modena e Reggio Emilia
E-mail: valentina.taddei@unimore.it

Received October 31, 2012
Revised November 12, 2012

Section 2

Rend. Istit. Mat. Univ. Trieste
Volume 44 (2012), 393–397

On repdigits as product of consecutive Fibonacci numbers[1]

Diego Marques and Alain Togbé

Abstract. *Let $(F_n)_{n\geq 0}$ be the Fibonacci sequence. In 2000, F. Luca proved that $F_{10} = 55$ is the largest repdigit (i.e. a number with only one distinct digit in its decimal expansion) in the Fibonacci sequence. In this note, we show that if $F_n \cdots F_{n+(k-1)}$ is a repdigit, with at least two digits, then $(k, n) = (1, 10)$.*

Keywords: Fibonacci, repdigits, sequences (mod m)
MS Classification 2010: 11A63, 11B39, 11B50

1. Introduction

Let $(F_n)_{n\geq 0}$ be the Fibonacci sequence given by $F_{n+2} = F_{n+1} + F_n$, for $n \geq 0$, where $F_0 = 0$ and $F_1 = 1$. These numbers are well-known for possessing amazing properties. In 1963, the Fibonacci Association was created to provide an opportunity to share ideas about these intriguing numbers and their applications. We remark that, in 2003, Bugeaud et al. [2] proved that the only perfect powers in the Fibonacci sequence are $0, 1, 8$ and 144 (see [6] for the Fibonomial version). In 2005, Luca and Shorey [5] showed, among other things, that a non-zero product of two or more consecutive Fibonacci numbers is never a perfect power except for the trivial case $F_1 \cdot F_2 = 1$.

Recall that a positive integer is called a *repdigit* if it has only one distinct digit in its decimal expansion. In particular, such a number has the form $a(10^m - 1)/9$, for some $m \geq 1$ and $1 \leq a \leq 9$. The problem of finding all perfect powers among repdigits was posed by Obláth [8] and completely solved, in 1999, by Bugeaud and Mignotte [1]. One can refer to [3] and its extensive annotated bibliography for additional references, history and related results.

In 2000, F. Luca [4], using elementary techniques, proved that $F_{10} = 55$ is the largest repdigit in the Fibonacci sequence. In a very recent paper, the authors [7] used bounds for linear forms in logarithms *à la Baker*, in order to

[1]The first author is grateful to FAP-DF, CNPq-Brazil and FEMAT-Brazil for the Financial support. The second author is supported in part by Purdue University North Central.

prove that there is no Fibonacci number of the form $B\cdots B$ (concatenation of B, m times), for $m>1$ and $B\in\mathbb{N}$ with at most 10 digits.

In this note, we follow the same ideas by using elementary tools for searching repdigits as product of consecutive Fibonacci numbers. More precisely, our main result is the following.

THEOREM 1.1. *The only solution of the Diophantine equation*

$$F_n\cdots F_{n+(k-1)}=a\left(\frac{10^m-1}{9}\right), \tag{1}$$

in positive integers n,k,m,a, with $1\leq a\leq 9$ and $m>1$ is $(n,k,m,a)=(10,1,2,5)$.

We need to point out that all relations which will appear in the proof of the above result can be easily proved by elementary ways (mathematical induction, the Fibonacci recurrence pattern, congruence properties etc). So, we will leave them as exercises to the reader.

2. The proof

First, we claim that $k\leq 4$. Indeed, we suppose the contrary, i.e. there exist at least 5 consecutive numbers among $n,...,n+(k-1)$. Thus, $3|(n+i)$ and $5|(n+j)$, for some $i,j\in\{0,...,k-1\}$. This implies that $2|F_{n+i}$ and $5|F_{n+j}$ leading to an absurdity as $10|F_n\cdots F_{n+(k-1)}=a(10^m-1)/9$ and hence $k\in\{1,2,3,4\}$. If $k=1$, Luca's result [4, Theorem 1] ensures that $(n,m,a)=(10,2,5)$. Hence, we must prove that Eq. (1) has no solution for $k\in\{2,3,4\}$.

Note that $a(10^2-1)/9=a\cdot 11$ and $a(10^3-1)/9=a\cdot 3\cdot 37$ are not products of at least two Fibonacci numbers, for $1\leq a\leq 9$. So, from now on, we can assume that $m\geq 4$.

a	1	2	3	4	5	6	7	8	9	
$a\cdot(\frac{10^m-1}{9})$	7	14	5	12	3	10	1	8	15	(mod 16)

Table 1: Residue classes modulo 16, for $m\geq 4$.

Case $k=4$. The sequence $(F_nF_{n+1}F_{n+2}F_{n+3})_{n\geq 1}$ has period 12 modulo 16. In fact,

$$F_nF_{n+1}F_{n+2}F_{n+3}\equiv 6,14,0,8,8,0,14,6,0,0,0,0 \pmod{16}.$$

So, by Table 1, it suffices to consider $a = 2$ and 8. Since 4 divides one of the numbers $n, n+1, n+2, n+3$, then

$$3 = F_4 | F_n F_{n+1} F_{n+2} F_{n+3} = a\left(\frac{10^m - 1}{9}\right)$$

and so $3|(10^m - 1)/9$. Thus we deduce that $3|m$ (in what follows, we will use this fact on several occasions).

For $a = 2$ and 8, one has $n \equiv 2, 7 \pmod{12}$ and $n \equiv 4, 5 \pmod{12}$, respectively. Therefore $F_n F_{n+1} F_{n+2} F_{n+3} \equiv 0, 1 \pmod 5$. Thus, Eq. (1) is not valid, since $2 \cdot \left(\frac{10^m-1}{9}\right) \equiv 2 \pmod 5$ and $8 \cdot \left(\frac{10^m-1}{9}\right) \equiv 3 \pmod 5$, for $m \geq 2$. We conclude that the assumption $k = 4$ is impossible.

Case $k = 3$. The period of $(F_n F_{n+1} F_{n+2})_{n\geq 1}$ modulo 16 is 12. Actually, we have

$$F_n F_{n+1} F_{n+2} \equiv 2, 6, 14, 8, 8, 8, 2, 6, 14, 0, 0, 0 \pmod{16}.$$

Again, by looking at Table 1, we deduce that $a = 2$ or 8.

First, we suppose that $a = 2$. Thus, one has $n \equiv 3, 9 \pmod{12}$. If $n \equiv 3 \pmod{12}$, then $F_n F_{n+1} F_{n+2} \equiv 25, 29, 22, 18, 30 \pmod{31}$. Since $3|m$ then $4|(n+1)$ and we get

$$2\left(\frac{10^m - 1}{9}\right) \equiv 5, 14, 24, 11, 0 \pmod{31}.$$

Thus Eq. (1) is not true in this case. In the case of $n \equiv 9 \pmod{12}$, we have $4 \nmid (n+j)$, for $j \in \{0, 1, 2\}$. Thus $3 \nmid m$ and we split the proof in two subcases:

- $m \equiv 1 \pmod 3$: In this case, $2(10^m - 1)/9 \equiv 14 \pmod{32}$, but on the other hand $F_n F_{n+1} F_{n+2} \equiv 30 \pmod{32}$;
- $m \equiv 2 \pmod 3$: Then $2(10^m - 1)/9 \equiv 4, 1 \pmod 7$, while $F_n F_{n+1} F_{n+2} \equiv 2, 5 \pmod 7$.

So, we have no solutions in the case $a = 2$.

Second, we take $a = 8$. One has $n \equiv 4, 5, 6 \pmod{12}$. In the case of $n \equiv 4 \pmod{12}$, we have $F_n F_{n+1} F_{n+2} \equiv 0, 1, 4 \pmod 5$. Since $4|n$, then $3|m$ yields $8(10^m - 1)/9 \equiv 3 \pmod 5$. When $n \equiv 6 \pmod{12}$, we obtain $F_n F_{n+1} F_{n+2} \equiv 0, 6, 9 \pmod{15}$. Again $3|m$, because $4|(n+2)$ and so $8(10^m - 1)/9 \equiv 3 \pmod{15}$. Therefore, a possible solution may appear for $n \equiv 5 \pmod{12}$. In this case, $3 \nmid m$, so we have the following two cases:

- $m \equiv 1 \pmod 3$ implies $8(10^m - 1)/9 \equiv 15, 4, 5, 17, 9, 8 \pmod{19}$. On the other hand, $F_n F_{n+1} F_{n+2} \equiv 0, 12, 7 \pmod{19}$;

- $m \equiv 2 \pmod 3$ yields $8(10^m - 1)/9 \equiv 7, 10 \pmod{13}$, while

$$F_n F_{n+1} F_{n+2} \equiv 9, 2, 0, 11, 4, 0, 0 \pmod{13}.$$

Thus, we also have no solution for $k = 3$.

Case $k = 2$. Since

$$F_n F_{n+1} \equiv 1, 2, 6, 15, 8, 8, 1, 10, 14, 15, 0, 0 \pmod{16},$$

we need to consider $a = 2, 6, 7, 8$, and 9. For $a = 6$, we have $n \equiv 8 \pmod{12}$ and then $F_n F_{n+1} \equiv 0, 2, 4 \pmod 5$, while $6(10^m - 1)/9 \equiv 1 \pmod 5$. When $a = 9$, one has $n \equiv 10 \pmod{12}$ and therefore Eq. (1) becomes $F_n F_{n+1} = 10^m - 1 \equiv 0 \pmod 9$. However, $F_n F_{n+1} \equiv 8 \pmod 9$, for $n \equiv 10 \pmod{12}$. In the case of $a = 7$, one gets $n \equiv 1, 7 \pmod{12}$ (and then $4 \nmid n$). On the other hand, Eq. (1) implies that $7|F_n$ or $7|F_{n+1}$ and thus $n \equiv 0 \pmod 8$ or $n \equiv -1 \pmod 8$. Therefore, $n \equiv 7 \pmod{12}$ and $n \equiv -1 \pmod 8$. We then get $n \equiv 7 \pmod{24}$ leading to $F_n F_{n+1} \equiv 0, 1, 3 \pmod 5$, but $7(10^m - 1)/9 \equiv 2 \pmod 5$. For $a = 2$, one has $n \equiv 9 \pmod{12}$ and so $4 \nmid (n + j)$, for $j \in \{0, 1\}$. Thus $3 \nmid m$ and then $2(10^m - 1)/9 \equiv 2 \pmod 5$, but $F_n F_{n+1} \equiv 0, 1, 3 \pmod 5$. For $a = 8$, we have $n \equiv 5, 6 \pmod{12}$. If $n \equiv 5 \pmod{12}$, similarly as in previous cases, we deduce that $3 \nmid m$.

- $m \equiv 1 \pmod 3$ implies $8(10^m - 1)/9 \equiv 5, 2, 8 \pmod 9$, however $F_n F_{n+1} \equiv 4 \pmod 9$;

- $m \equiv 2 \pmod 3$ yields $8(10^m - 1)/9 \equiv 2, 4 \pmod 7$, again Eq. (1) is not valid, since $F_n F_{n+1} \equiv 1, 5 \pmod 7$.

We finish by considering the case $n \equiv 6 \pmod{12}$. Again $3 \nmid m$ and so $8(10^m - 1)/9 \equiv 3 \pmod 5$, while $F_n F_{n+1} \equiv 0, 2, 4 \pmod 5$. In conclusion, Eq. (1) has no solution for $k > 1$. □

References

[1] Y. Bugeaud and M. Mignotte, *On integers with identical digits*, Mathematika **46** (1999), 411–417.

[2] Y. Bugeaud, M. Mignotte, and S. Siksek, *Classical and modular approaches to exponential diophantine equations I. Fibonacci and Lucas powers*, Ann. of Math. **163** (2006), 969–1018.

[3] Y. Bugeaud and P. Mihailescu, *On the Nagell–Ljunggren equation* $(x^n - 1)/(x - 1) = y^q$, Math. Scand. **101** (2007), 177–183.

[4] F. Luca, *Fibonacci and Lucas numbers with only one distinct digit*, Portugal. Math. **57** (2000), 243–254.

[5] F. Luca and T. N. Shorey, *Diophantine equations with products of consecutive terms in Lucas sequences*, J. Number Theory **114** (2005), 298–311.
[6] D. Marques and A. Togbé, *Perfect powers among C-nomial coefficients*, C. R. Math. Acad. Sci. Paris **348** (2010), 717–720.
[7] D. Marques and A. Togbé, *On terms of a linear recurrence sequence with only one distinct block of digits*, Colloq. Math. **124** (2011), 145–155.
[8] R. Obláth, *Une propriété des puissances parfaites*, Mathesis **65** (1956), 356–364.

Authors' addresses:

Diego Marques
Departamento de Matemática,
Universidade de Brasília,
Brasília, 70910-900, Brazil
E-mail: diego@mat.unb.br

Alain Togbé
Department of Mathematics,
Purdue University North Central,
1401 S, U.S. 421,
Westville, IN 46391, USA
E-mail: atogbe@pnc.edu

Received October 7, 2011
Revised January 9, 2012

Rend. Istit. Mat. Univ. Trieste
Volume 44 (2012), 399–411

On $\theta_{(\mathcal{I},\mathcal{J})}$-continuous functions

Ahmad Al-Omari and Takashi Noiri

Abstract. *In this paper we investigate some properties of $\theta_{(\mathcal{I},\mathcal{J})}$-continuous functions in ideal topological spaces. Moreover the relationships with other related functions are discussed.*

Keywords: ideal topological space, θ-continuous, weakly $\mathcal{J}$-continuous, strongly θ-continuous, $\theta_{(\mathcal{I},\mathcal{J})}$-continuous
MS Classification 2010: 54A05, 54C10

1. Introduction

The concept of ideals in topological spaces is treated in the classic text by Kuratowski [11] and Vaidyanathaswamy [17]. Janković and Hamlett [9] investigated further properties of ideal spaces. An ideal $\mathcal{I}$ on a topological space (X,τ) is a non-empty collection of subsets of X which satisfies the following properties: (1) $A \in \mathcal{I}$ and $B \subseteq A$ implies $B \in \mathcal{I}$; (2) $A \in \mathcal{I}$ and $B \in \mathcal{I}$ implies $A \cup B \in \mathcal{I}$. An ideal topological space (or an ideal space) is a topological space (X,τ) with an ideal $\mathcal{I}$ on X and is denoted by $(X,\tau,\mathcal{I})$. For a subset $A \subseteq X$, $A^*(\mathcal{I},\tau) = \{x \in X : A \cap U \notin \mathcal{I} \text{ for every } U \in \tau(X,x)\}$ is called the local function of A with respect to $\mathcal{I}$ and τ [11]. We simply write A^* in case there is no chance for confusion. A Kuratowski closure operator $Cl^*(.)$ for a topology $\tau^*(\mathcal{I},\tau)$ called the $*$-topology, finer than τ, is defined by $Cl^*(A) = A \cup A^*$ [17]. The notion of θ-continuity [6] in topological spaces is widely known and investigated. Recently, Yüksel et al. [19] have introduced the notion of $\theta_{(\mathcal{I},\mathcal{J})}$-continuous functions between ideal topological spaces. In the present paper, we obtain several characterizations and many properties of $\theta_{(\mathcal{I},\mathcal{J})}$-continuous functions.

2. Preliminaries

Let (X,τ) be a topological space with no separation axioms assumed. If $A \subseteq X$, $Cl(A)$ and $Int(A)$ will denote the closure and interior of A in (X,τ), respectively.

In 1968, Veličko [18] introduced the class of θ-open sets. A set A is said to be θ-open [18] if every point of A has an open neighborhood whose closure is contained in A. The θ-interior [18] of A in X is the union of all θ-open subsets

of A and is denoted by $Int_\theta(A)$. Naturally, the complement of a θ-open set is said to be θ-closed. Equivalently $Cl_\theta(A) = \{x \in X : Cl(U) \cap A \neq \phi, U \in \tau$ and $x \in U\}$ and a set A is θ-closed if and only if $A = Cl_\theta(A)$. Note that all θ-open sets form a topology on X, coarser than τ, denoted by τ_θ and that a space (X, τ) is regular if and only if $\tau = \tau_\theta$. Note also that the θ-closure of a given set need not be a θ-closed set.

Let $(X, \tau, \mathcal{I})$ be an ideal topological space and $A \subseteq X$. A point x of X is called a $\theta_\mathcal{I}$-cluster point of A if $Cl^*(U) \cap A \neq \phi$ for every open set U of X containing x. The set of all $\theta_\mathcal{I}$-cluster points of A is called the $\theta_\mathcal{I}$-closure of A and is denoted by $Cl_{\theta_\mathcal{I}}(A)$. A is said to be $\theta_\mathcal{I}$-closed if $Cl_{\theta_\mathcal{I}}(A) = A$. The complement of a $\theta_\mathcal{I}$-closed set is called a $\theta_\mathcal{I}$-open set.

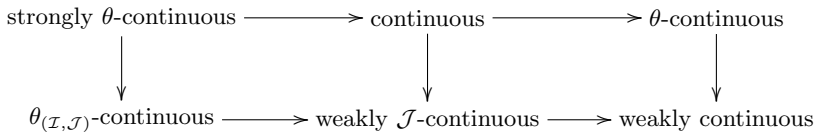
Definition 2.1. *Let $(X, \tau, \mathcal{I})$ be an ideal topological space. A point x of X is called a $\theta_\mathcal{I}$-interior point of A if there exists an open set U containing x such that $Cl^*(U) \subseteq A$. The set of all $\theta_\mathcal{I}$-interior points of A is called the $\theta_\mathcal{I}$-interior of A and is denoted by $Int_{\theta_\mathcal{I}}(A)$.*

Remark 2.2. *For a set A of X, $Int_{\theta_\mathcal{I}}(X - A) = X - Cl_{\theta_\mathcal{I}}(A)$ so that A is $\theta_\mathcal{I}$-open if and only if $A = Int_{\theta_\mathcal{I}}(A)$.*

Definition 2.3. *A function $f : (X, \tau) \to (Y, \sigma)$ is said to be θ-continuous [6] (resp. strongly θ-continuous [14], weakly continuous [13]) if for each $x \in X$ and each open set V in Y containing $f(x)$, there exists an open set U containing x such that $f(Cl(U)) \subseteq Cl(V)$ (resp. $f(Cl(U)) \subseteq V$, $f(U) \subseteq Cl(V)$).*

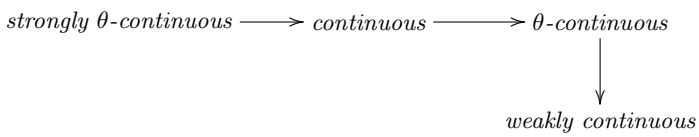
Definition 2.4. *A function $f : (X, \tau, \mathcal{I}) \to (Y, \sigma, \mathcal{J})$ is said to be weakly $\mathcal{J}$-continuous [1] (resp. $\theta_{(\mathcal{I},\mathcal{J})}$-continuous [19]) if for each $x \in X$ and each open set V in Y containing $f(x)$, there exists an open set U containing x such that $f(U) \subseteq Cl^*(V)$ (resp. $f(Cl^*(U)) \subseteq Cl^*(V)$).*

By the above definitions, we have the following diagram and none of these implications is reversible



Remark 2.5. *In [1, Example 2.1], it is shown that not every weakly continuous function is weakly $\mathcal{J}$-continuous.*

REMARK 2.6. *The following strict implications are well-known:*



EXAMPLE 2.7. *Let* $X = \{1, 2, 3, 4\}$, $\tau = \{X, \phi, \{1, 2, 3\}, \{3\}, \{3, 4\}\}$ *with* $\mathcal{I} = \{\phi, \{1\}, \{2\}, \{1, 2\}\}$ *and* $Y = \{a, b, c, d\}$, $\sigma = \{Y, \phi, \{a, b\}, \{b\}, \{d\}, \{b, d\}, \{a, b, d\}, \{b, c, d\}\}$ *with* $\mathcal{J} = \{\phi\}$. *We define a function* $f : (X, \tau, \mathcal{I}) \to (Y, \sigma, \mathcal{J})$ *as* $f = \{(1, a), (2, b), (3, c), (4, d)\}$. *Then* f *is weakly* $\mathcal{J}$*-continuous but not* $\theta_{(\mathcal{I},\mathcal{J})}$*-continuous. In [12, Example 10], it is shown that* f *is weakly* $\mathcal{J}$*-continuous. We show that* $f : (X, \tau, \mathcal{I}) \to (Y, \sigma, \mathcal{J})$ *is not* $\theta_{(\mathcal{I},\mathcal{J})}$*-continuous. Let* $1 \in X$ *and* $V = \{a, b\} \in \sigma$ *such that* $f(1) = a \in V \in \sigma$. *But, for every open set* $U \subseteq X$ *such that* $1 \in U$, *where* $U = \{1, 2, 3\}$ *or* $U = X$, $Cl^*(U) = X$. *Then* $f(Cl^*(U) = Y \nsubseteq Cl^*(V) = \{a, b, c\}$. *Therefore,* $f : (X, \tau, \mathcal{I}) \to (Y, \sigma, \mathcal{J})$ *is not* $\theta_{(\mathcal{I},\mathcal{J})}$*-continuous.*

EXAMPLE 2.8. *Let* $X = \{a, b, c\}$, $\tau = \{X, \phi, \{b, c\}\}$ *with* $\mathcal{I} = \{\phi, \{a\}\}$ *and* $Y = \{b, c\}$, $\sigma = \{Y, \phi, \{c\}\}$ *with* $\mathcal{J} = \{\phi, \{b\}\}$. *We define a function* $f : (X, \tau, \mathcal{I}) \to (Y, \sigma, \mathcal{J})$ *as* $f = \{(a, b), (b, c), (c, b)\}$. *Then* f *is* $\theta_{(\mathcal{I},\mathcal{J})}$*-continuous but not continuous.*

1. *Let* $a \in X$ *and* $V = Y \in \sigma$ *such that* $f(a) = b \in V$, *then there exists an open set* $U = X \in \tau$ *containing* a *such that* $f(Cl^*(U)) \subseteq Cl^*(V) = Y$.
2. *Let* $b \in X$ *and* $V = \{c\}$ *or* $V = Y$ *such that* $f(b) = c \in V$, *then there exists an open set* $U = \{b, c\}$ *or* $U = X$ *containing* b *such that* $f(Cl^*(U)) \subseteq Cl^*(V) = Y$.
3. *Let* $c \in X$ *and* $V = Y$ *such that* $f(c) = b \in V$, *then there exists an open set* $U = \{b, c\}$ *or* $U = X$ *containing* c *such that* $f(Cl^*(U)) \subseteq Cl^*(V) = Y$.

By (1), (2) and (3) f *is* $\theta_{(\mathcal{I},\mathcal{J})}$*-continuous. On the other hand, let* $b \in X$ *and* $V = \{c\} \in \sigma$ *such that* $f(b) = c \in V \in \sigma$. *But, for every open set* $U \subseteq X$ *such that* $b \in U$, *where* $U = \{b, c\}$ *or* $U = X$. *Then* $f(U) = Y \nsubseteq V = \{c\}$. *Therefore,* $f : (X, \tau, \mathcal{I}) \to (Y, \sigma, \mathcal{J})$ *is not continuous.*

The following lemma is useful in the sequel:

LEMMA 2.9 ([9]). *Let* $(X, \tau, \mathcal{I})$ *be an ideal topological space and* A, B *subsets of* X. *Then the following properties hold:*

1. *If* $A \subseteq B$, *then* $A^* \subseteq B^*$.

2. $A^* = Cl(A^*) \subseteq Cl(A)$.

3. $(A^*)^* \subseteq A^*$.

4. $(A \cup B)^* = A^* \cup B^*$.

3. Characterizations of $\theta_{(\mathcal{I},\mathcal{J})}$-continuous functions

In this section, we obtain several characterizations of $\theta_{(\mathcal{I},\mathcal{J})}$-continuous functions in ideal topological spaces.

THEOREM 3.1. *For a function $f : (X, \tau, \mathcal{I}) \to (Y, \sigma, \mathcal{J})$, the following properties are equivalent:*

1. *f is $\theta_{(\mathcal{I},\mathcal{J})}$-continuous;*

2. *$Cl_{\theta_{\mathcal{I}}}(f^{-1}(B)) \subseteq f^{-1}(Cl_{\theta_{\mathcal{J}}}(B))$ for every subset B of Y;*

3. *$f(Cl_{\theta_{\mathcal{I}}}(A)) \subseteq Cl_{\theta_{\mathcal{J}}}(f(A))$ for every subset A of X.*

Proof. (1) $\Rightarrow$ (2): Let B be any subset of Y. Suppose that $x \notin f^{-1}(Cl_{\theta_{\mathcal{J}}}(B))$. Then $f(x) \notin Cl_{\theta_{\mathcal{J}}}(B)$ and there exists an open set V containing $f(x)$ such that $Cl^*(V) \cap B = \phi$. Since f is $\theta_{(\mathcal{I},\mathcal{J})}$-continuous, there exists an open set U containing x such that $f(Cl^*(U)) \subseteq Cl^*(V)$. Therefore, we have $f(Cl^*(U)) \cap B = \phi$ and $Cl^*(U) \cap f^{-1}(B) = \phi$. This shows that $x \notin Cl_{\theta_{\mathcal{I}}}(f^{-1}(B))$. Thus, we obtain $Cl_{\theta_{\mathcal{I}}}(f^{-1}(B)) \subseteq f^{-1}(Cl_{\theta_{\mathcal{J}}}(B))$.
(2) $\Rightarrow$ (1): Let $x \in X$ and V be an open set of Y containing $f(x)$. Then we have $Cl^*(V) \cap (Y - Cl^*(V)) = \phi$ and $f(x) \notin Cl_{\theta_J}(Y - Cl^*(V))$. Therefore, $x \notin f^{-1}(Cl_{\theta_J}(Y - Cl^*(V)))$ and by (2) we have $x \notin Cl_{\theta_I}(f^{-1}(Y - Cl^*(V)))$. There exists an open set U containing x such that $Cl^*(U) \cap f^{-1}(Y - Cl^*(V)) = \phi$ and hence $f(Cl^*(U)) \subseteq Cl^*(V)$. Therefore, f is $\theta_{(\mathcal{I},\mathcal{J})}$-continuous.
(2) $\Rightarrow$ (3): Let A be any subset of X. Then we have $Cl_{\theta_{\mathcal{I}}}(A) \subseteq Cl_{\theta_{\mathcal{I}}}(f^{-1}(f(A)))$ $\subseteq f^{-1}(Cl_{\theta_{\mathcal{J}}}(f(A)))$ and hence $f(Cl_{\theta_{\mathcal{I}}}(A)) \subseteq Cl_{\theta_{\mathcal{J}}}(f(A))$.
(3) $\Rightarrow$ (2): Let B be a subset of Y. We have $f(Cl_{\theta_{\mathcal{I}}}(f^{-1}(B))) \subseteq Cl_{\theta_{\mathcal{J}}}(f(f^{-1}(B)))$ $\subseteq Cl_{\theta_{\mathcal{J}}}(B)$ and hence $Cl_{\theta_{\mathcal{I}}}(f^{-1}(B)) \subseteq f^{-1}(Cl_{\theta_{\mathcal{J}}}(B))$. □

DEFINITION 3.2 ([1]). *An ideal topological space $(X, \tau, \mathcal{I})$ is called an $F\mathcal{I}^*$-space if $Cl(U) \subseteq U^*$ for every open set U of X.*

DEFINITION 3.3 ([3]). *Let $(X, \tau, \mathcal{I})$ be an ideal topological space. $\mathcal{I}$ is said to be codense if $\tau \cap \mathcal{I} = \phi$.*

REMARK 3.4. *In [12], Kuyucu et al. showed the following properties:*

1. *an ideal topological space $(X, \tau, \mathcal{I})$ is an $F\mathcal{I}^*$-space if and only if $\mathcal{I}$ is codense,*

2. *if $(X,\tau,\mathcal{I})$ is an $F\mathcal{I}^*$-space, then $V^* = Cl^*(V) = Cl(V)$ for every open set V of X.*

THEOREM 3.5. *For a function $f : (X,\tau,\mathcal{I}) \to (Y,\sigma,\mathcal{J})$, the following implications: (1) $\Leftrightarrow$ (2) $\Rightarrow$ (3) $\Leftrightarrow$ (4) hold. Moreover, the implication (4) $\Rightarrow$ (1) holds if $(Y,\sigma,\mathcal{J})$ is an $F\mathcal{J}^*$-space.*

1. *f is $\theta_{(\mathcal{I},\mathcal{J})}$-continuous;*
2. *$f^{-1}(V) \subseteq Int_{\theta_\mathcal{I}}(f^{-1}(Cl^*(V)))$ for every open set V of Y;*
3. *$Cl_{\theta_\mathcal{I}}(f^{-1}(V)) \subseteq f^{-1}(Cl(V))$ for every open set V of Y;*
4. *For each $x \in X$ and each open set V of Y containing $f(x)$, there exists an open set U of X containing x such that $f(Cl^*(U)) \subseteq Cl(V)$.*

Proof. (1) $\Rightarrow$ (2): Suppose that V is any open set of Y and $x \in f^{-1}(V)$. Then $f(x) \in V$ and there exists an open set U containing x such that $f(Cl^*(U)) \subseteq Cl^*(V)$. Therefore, $x \in U \subseteq Cl^*(U) \subseteq f^{-1}(Cl^*(V))$. This shows that $x \in Int_{\theta_\mathcal{I}}(f^{-1}(Cl^*(V)))$. Therefore, we obtain $f^{-1}(V) \subseteq Int_{\theta_\mathcal{I}}(f^{-1}(Cl^*(V)))$.
(2) $\Rightarrow$ (1): Let $x \in X$ and $V \in \sigma$ containing $f(x)$. Then, by (2) $f^{-1}(V) \subseteq Int_{\theta_\mathcal{I}}(f^{-1}(Cl^*(V)))$. Since $x \in f^{-1}(V)$, there exists an open set U containing x such that $Cl^*(U) \subseteq f^{-1}(Cl^*(V))$. Therefore, $f(Cl^*(U)) \subseteq Cl^*(V)$ and hence f is $\theta_{(\mathcal{I},\mathcal{J})}$-continuous.
(2) $\Rightarrow$ (3): Suppose that V is any open set of Y and $x \notin f^{-1}(Cl(V))$. Then $f(x) \notin Cl(V)$ and there exists an open set W containing $f(x)$ such that $W \cap V = \phi$; hence $Cl^*(W) \cap V \subseteq Cl(W) \cap V = \phi$. Therefore, we have $f^{-1}(Cl^*(W)) \cap f^{-1}(V) = \phi$. Since $x \in f^{-1}(W)$, by (2) $x \in Int_{\theta_\mathcal{I}}(f^{-1}(Cl^*(W)))$. There exists an open set U containing x such that $Cl^*(U) \subseteq f^{-1}(Cl^*(W))$. Thus we have $Cl^*(U) \cap f^{-1}(V) = \phi$ and hence $x \notin Cl_{\theta_\mathcal{I}}(f^{-1}(V))$. This shows that $Cl_{\theta_\mathcal{I}}(f^{-1}(V)) \subseteq f^{-1}(Cl(V))$.
(3) $\Rightarrow$ (4): Suppose that $x \in X$ and V is any open set of Y containing $f(x)$. Then $V \cap (Y - Cl(V)) = \phi$ and $f(x) \notin Cl(Y - Cl(V))$. Therefore $x \notin f^{-1}(Cl(Y - Cl(V)))$ and by (3) $x \notin Cl_{\theta_\mathcal{I}}(f^{-1}(Y - Cl(V)))$. There exists an open set U containing x such that $Cl^*(U) \cap f^{-1}(Y - Cl(V)) = \phi$. Therefore, we obtain $f(Cl^*(U)) \subseteq Cl(V)$.
(4) $\Rightarrow$ (3): Let V be any open set of Y. Suppose that $x \notin f^{-1}(Cl(V))$. Then $f(x) \notin Cl(V)$ and there exists an open set W containing $f(x)$ such that $W \cap V = \phi$. By (4), there exists an open set U containing x such that $f(Cl^*(U)) \subseteq Cl(W)$. Since $V \in \sigma$, $Cl(W) \cap V = \phi$ and $f(Cl^*(U)) \cap V \subseteq Cl(W) \cap V = \phi$. Therefore, $Cl^*(U) \cap f^{-1}(V) = \phi$ and hence $x \notin Cl_{\theta_\mathcal{I}}(f^{-1}(V))$. This shows that $Cl_{\theta_\mathcal{I}}(f^{-1}(V)) \subseteq f^{-1}(Cl(V))$.
(4) $\Rightarrow$ (1): Since $(Y,\sigma,\mathcal{J})$ is an $F\mathcal{J}^*$-space, $Cl(V) \subseteq Cl^*(V)$ for every open set V of Y and hence f is $\theta_{(\mathcal{I},\mathcal{J})}$-continuous. □

PROPOSITION 3.6. *A function $f : (X, \tau, \mathcal{I}) \to (Y, \sigma, \mathcal{J})$ from an $F\mathcal{I}^*$-space to an $F\mathcal{J}^*$-space is $\theta_{(\mathcal{I},\mathcal{J})}$-continuous if and only if it is θ-continuous.*

Proof. This follows from the Remark 3.4. □

4. Some properties of $\theta_{(\mathcal{I},\mathcal{J})}$-continuous functions

DEFINITION 4.1. *An ideal topological space $(X, \tau, \mathcal{I})$ is said to be $\theta_{\mathcal{I}}$-T_2 (resp. $*$-Urysohn) if for each distinct points $x, y \in X$, there exist two $\theta_{\mathcal{I}}$-open (resp. open) sets $U, V \in X$ containing x and y, respectively, such that $U \cap V = \phi$ (resp. $Cl^*(U) \cap Cl^*(V) = \phi$).*

THEOREM 4.2. *If $f, g : (X, \tau, \mathcal{I}) \to (Y, \sigma, \mathcal{J})$ are $\theta_{(\mathcal{I},\mathcal{J})}$-continuous functions and $(Y, \sigma, \mathcal{J})$ is $*$-Urysohn , then $A = \{x \in X : f(x) = g(x)\}$ is a $\theta_{\mathcal{I}}$-closed set of $(X, \tau, \mathcal{I})$.*

Proof. We prove that $X - A$ is a $\theta_{\mathcal{I}}$-open set. Let $x \in X - A$. Then $f(x) \neq g(x)$. Since Y is $*$-Urysohn, there exist open sets V_1 and V_2 containing $f(x)$ and $g(x)$, respectively, such that $Cl^*(V_1) \cap Cl^*(V_2) = \phi$. Since f and g are $\theta_{(\mathcal{I},\mathcal{J})}$-continuous, there exists an open set U containing x such that $f(Cl^*(U)) \subseteq Cl^*(V_1))$ and $g(Cl^*(U)) \subseteq Cl^*(V_2))$. Hence we obtain that $Cl^*(U) \subseteq f^{-1}(Cl^*(V_1))$ and $Cl^*(U) \subseteq g^{-1}(Cl^*(V_2))$. From here we have $Cl^*(U) \subseteq f^{-1}(Cl^*(V_1)) \cap g^{-1}(Cl^*(V_2))$. Moreover $f^{-1}(Cl^*(V_1)) \cap g^{-1}(Cl^*(V_2)) \subseteq X - A$. This shows that $X - A$ is $\theta_{\mathcal{I}}$-open. □

DEFINITION 4.3. *An ideal topological space $(X, \tau, \mathcal{I})$ is said to be $*$-regular if for each closed set F and each point $x \in X - F$, there exist an open set V and an $*$-open set $U \in \tau^*$ such that $x \in V$, $F \subseteq U$ and $U \cap V = \phi$.*

EXAMPLE 4.4. *Let $X = \{a, b, c\}$, $\tau = \{\phi, X, \{a\}, \{a, b\}\}$ and $\mathcal{I} = \mathcal{P}(X)$, then $(X, \tau, \mathcal{I})$ is an $*$-regular space which is not regular.*

LEMMA 4.5 ([1]). *1. A function $f : (X, \tau) \to (Y, \sigma, \mathcal{J})$ is weakly $\mathcal{J}$-continuous if and only if for each open set V, $f^{-1}(V) \subseteq Int(f^{-1}(Cl^*(V)))$.*

2. If an ideal space $(Y, \sigma, \mathcal{I})$ is an $F\mathcal{J}^*$-space and a function $f : (X, \tau, \mathcal{I}) \to (Y, \sigma, \mathcal{I})$ is weakly $\mathcal{J}$-continuous, then $Cl^*(f^{-1}(G)) \subseteq f^{-1}(Cl^*(G))$ for every open set G in Y.

The equivalence of (1) and (2) in the following theorem is suggested by the referee.

THEOREM 4.6. *Let $(Y, \sigma, \mathcal{J})$ be an $F\mathcal{J}^*$-space. For a function $f : (X, \tau) \to (Y, \sigma, \mathcal{J})$, the following properties are equivalent:*

1. *f is weakly $\mathcal{J}$-continuous;*
2. *$Cl(f^{-1}(V)) \subseteq f^{-1}(Cl^*(V))$ for every open set V of Y;*
3. *f is weakly continuous.*

Proof. (1) $\Rightarrow$ (2): Let V be any open set of Y. Suppose that $x \notin f^{-1}(Cl^*(V))$. Then $f(x) \notin Cl^*(V)$. Since $(Y, \sigma, \mathcal{J})$ is an $F\mathcal{J}^*$-space, $f(x) \notin Cl(V)$ and there exists $W \in \sigma$ containing $f(x)$ such that $W \cap V = \phi$, hence $Cl^*(W) \cap V = Cl(W) \cap V = \phi$. Since f is weakly $\mathcal{J}$-continuous, there exists $U \in \tau$ containing x such that $f(U) \subseteq Cl^*(W)$. Therefore, we have $f(U) \cap V = \phi$ and $U \cap f^{-1}(V) = \phi$. Since $U \in \tau$, $U \cap Cl(f^{-1}(V)) = \phi$ and hence $x \notin Cl(f^{-1}(V))$. Therefore, we obtain $Cl(f^{-1}(V)) \subseteq f^{-1}(Cl^*(V))$.
(2) $\Rightarrow$ (3): Let V be any open set of Y. Since $(Y, \sigma, \mathcal{J})$ is an $F\mathcal{J}^*$-space, by (2) we have $Cl(f^{-1}(V)) \subseteq f^{-1}(Cl(V))$. It follows from [16, Theorem 7] that f is weakly continuous.
(3) $\Rightarrow$ (1): Let f be weakly continuous. By [13, Theorem 1]

$$f^{-1}(V) \subseteq Int(f^{-1}(Cl(V)))$$

for every open set V of Y. Since $(Y, \sigma, \mathcal{J})$ is an $F\mathcal{J}^*$-space, $Cl(V) = Cl^*(V)$ and we have $f^{-1}(V) \subseteq Int(f^{-1}(Cl^*(V)))$. Therefore, by Lemma 4.5 (1) f is weakly $\mathcal{J}$-continuous. □

Definition 4.7 ([5]). *An ideal space $(X, \tau, \mathcal{I})$ is said to be $*$-extremally disconnected if the $*$-closure of every open subset of X is open.*

Lemma 4.8. *An ideal topological space $(X, \tau, \mathcal{I})$ is $*$-regular if and only if for each open set U containing x there exists an open set V such that $x \in V \subseteq Cl^*(V) \subseteq U$.*

Proposition 4.9. *Let $(X, \tau, \mathcal{I})$ be an $*$-regular space. Then $f : (X, \tau, \mathcal{I}) \to (Y, \sigma, \mathcal{J})$ is $\theta_{(\mathcal{I},\mathcal{J})}$-continuous if and only if it is weakly $\mathcal{J}$-continuous.*

Proof. Every $\theta_{(\mathcal{I},\mathcal{J})}$-continuous function is weakly $\mathcal{J}$-continuous. Suppose that f is weakly $\mathcal{J}$-continuous. Let $x \in X$ and V be any open set of Y containing $f(x)$. Then, there exists an open set U containing x such that $f(U) \subseteq Cl^*(V)$. Since X is $*$-regular, by Lemma 4.8 there exists an open set W such that $x \in W \subseteq Cl^*(W) \subseteq U$. Therefore, we obtain $f(Cl^*(W)) \subseteq Cl^*(V)$. This shows that f is $\theta_{(\mathcal{I},\mathcal{J})}$-continuous. □

Theorem 4.10. *Let an ideal space $(Y, \sigma, \mathcal{J})$ be an $F\mathcal{J}^*$-space and $*$-extremally disconnected. Then $f : (X, \tau, \mathcal{I}) \to (Y, \sigma, \mathcal{J})$ is $\theta_{(\mathcal{I},\mathcal{J})}$-continuous if and only if it is weakly $\mathcal{J}$-continuous.*

Proof. It is clear that every $\theta_{(\mathcal{I},\mathcal{J})}$-continuous function is weakly $\mathcal{J}$-continuous. Conversely, suppose that f is weakly $\mathcal{J}$-continuous. Let $x \in X$ and V be an open set of Y containing $f(x)$. Then by Lemma 4.5 (1), $x \in f^{-1}(V) \subseteq Int(f^{-1}(Cl^*(V)))$. Let $U = Int(f^{-1}(Cl^*(V)))$. Since $(Y, \sigma, \mathcal{I})$ is an $F\mathcal{J}^*$-space and $*$-extremally disconnected, by using Lemma 4.5 (2) $f(Cl^*(U)) = f(Cl^*(Int(f^{-1}(Cl^*(V))))) \subseteq f(Cl^*(f^{-1}(Cl^*(V))) \subseteq f(f^{-1}(Cl^*(Cl^*(V))) \subseteq Cl^*(V)$. Hence f is $\theta_{(\mathcal{I},\mathcal{J})}$-continuous. □

Corollary 4.11. *Let an ideal space $(Y, \sigma, \mathcal{J})$ be an $F\mathcal{J}^*$-space and $*$-extremally disconnected. For a function $f : (X, \tau, \mathcal{I}) \to (Y, \sigma, \mathcal{J})$, the following properties are equivalent:*

1. *f is $\theta_{(\mathcal{I},\mathcal{J})}$-continuous;*
2. *f is weakly $\mathcal{J}$-continuous;*
3. *$f^{-1}(V) \subseteq Int(f^{-1}(V^*))$ for every open set V in Y;*
4. *$f^{-1}(V) \subseteq Int(f^{-1}(Cl(V)))$ for every open set V of Y;*
5. *f is weakly continuous.*

Proof. By Theorem 4.10, we have the equivalence of (1) and (2). The equivalences of (2), (3) and (4) follow from Lemma 4.5 (1) and Remark 3.4. The equivalence of (4) and (5) is shown in [13, Theorem 1]. □

A subset A of an ideal space $(X,\tau,\mathcal{I})$ is said to be pre-$\mathcal{I}$-open [4] if $A \subseteq Int(Cl^*(A))$. A function $f : (X, \tau, \mathcal{I}) \to (Y, \sigma, \mathcal{J})$ is said to be pre-$\mathcal{I}$-continuous [4] if the inverse image of every open set of Y is pre-$\mathcal{I}$-open in X.

Theorem 4.12. *If $f : (X, \tau, \mathcal{I}) \to (Y, \sigma, \mathcal{J})$ is a pre-$\mathcal{I}$-continuous function and $Cl^*(f^{-1}(U)) \subseteq f^{-1}(Cl^*(U))$ for every open set U in Y, then f is $\theta_{(\mathcal{I},\mathcal{J})}$-continuous.*

Proof. Let $x \in X$ and U be an open set in Y containing $f(x)$. By hypothesis, $Cl^*(f^{-1}(U)) \subseteq f^{-1}(Cl^*(U))$. Since f is pre-$\mathcal{I}$-continuous, $f^{-1}(U)$ is pre-$\mathcal{I}$-open in X and so $f^{-1}(U) \subseteq Int(Cl^*(f^{-1}(U)))$. Since $x \in f^{-1}(U) \subseteq Int(Cl^*(f^{-1}(U)))$, there exists an open set V containing x such that $x \in V \subseteq Cl^*(V) \subseteq Cl^*(f^{-1}(U)) \subseteq f^{-1}(Cl^*(U))$ and so $f(Cl^*(V)) \subseteq Cl^*(U)$ which implies that f is $\theta_{(\mathcal{I},\mathcal{J})}$-continuous. □

The following corollary follows from Lemma 4.5 and Theorems 4.6 and 4.12.

Corollary 4.13. *Let $f : (X, \tau, \mathcal{I}) \to (Y, \sigma, \mathcal{J})$ be pre-$\mathcal{I}$-continuous and $(Y, \sigma, \mathcal{J})$ is an $F\mathcal{J}^*$-space. The following properties are equivalent:*

1. *f is $\theta_{(\mathcal{I},\mathcal{J})}$-continuous;*

2. *$Cl^*(f^{-1}(V)) \subseteq f^{-1}(Cl^*(V))$ for every open set V in Y;*

3. *$Cl(f^{-1}(V)) \subseteq f^{-1}(Cl^*(V))$ for every open set V in Y;*

4. *f is weakly $\mathcal{J}$-continuous.*

5. Preservation theorems

A subset A of a space X is said to be quasi H^*-closed relative to X if for every cover $\{V_\alpha : \alpha \in \Lambda\}$ of A by open sets of X, there exists a finite subset Λ_0 of Λ such that $A \subseteq \cup\{Cl^*(V_\alpha) : \alpha \in \Lambda_0\}$. A space X is said to be quasi H^*-closed if X is quasi H^*-closed relative to X

THEOREM 5.1. *If $f : (X, \tau, \mathcal{I}) \to (Y, \sigma, \mathcal{J})$ is $\theta_{(\mathcal{I},\mathcal{J})}$-continuous and K is quasi H^*-closed relative to X, then $f(K)$ is quasi H^*-closed relative to Y.*

Proof. Suppose that $f : (X, \tau, \mathcal{I}) \to (Y, \sigma, \mathcal{J})$ is a $\theta_{(\mathcal{I},\mathcal{J})}$-continuous function and K is quasi H^*-closed relative to X. Let $\{V_\alpha : \alpha \in \Lambda\}$ be a cover of $f(K)$ by open sets of Y. For each point $x \in K$, there exists $\alpha(x) \in \Lambda$ such that $f(x) \in V_{\alpha(x)}$. Since f is $\theta_{(\mathcal{I},\mathcal{J})}$-continuous, there exists an open set U_x containing x such that $f(Cl^*(U_x)) \subseteq Cl^*(V_{\alpha(x)})$. The family $\{U_x : x \in K\}$ is a cover of K by open sets of X and hence there exists a finite subset K_* of K such that $K \subseteq \cup_{x \in K_*} Cl^*(U_x)$. Therefore, we obtain $f(K) \subseteq \cup_{x \in K_*} Cl^*(V_{\alpha(x)})$. This shows that $f(K)$ is quasi H^*-closed relative to Y. □

DEFINITION 5.2. *A function $f : (X, \tau, \mathcal{I}) \to (Y, \sigma, \mathcal{J})$ is said to be $\theta_{(\mathcal{I},\mathcal{J})}$-irresolute if for every $\theta_{\mathcal{J}}$-open set U in Y, $f^{-1}(U)$ is $\theta_{\mathcal{I}}$-open in X.*

THEOREM 5.3. *Every $\theta_{(\mathcal{I},\mathcal{J})}$-continuous function is $\theta_{(\mathcal{I},\mathcal{J})}$-irresolute.*

Proof. Let $f : X \to Y$ be a $\theta_{(\mathcal{I},\mathcal{J})}$-continuous function and U be a $\theta_{\mathcal{J}}$-open set in Y. Let $x \in f^{-1}(U)$. Then, $f(x) \in U$. Since U is $\theta_{\mathcal{J}}$-open, there exists an open set V in Y such that $f(x) \in V \subseteq Cl^*(V) \subseteq U$. By $\theta_{(\mathcal{I},\mathcal{J})}$-continuity of f, there exists an open set W in X containing x such that $f(Cl^*(W)) \subseteq Cl^*(V) \subseteq U$. Thus $x \in W \subseteq Cl^*(W) \subseteq f^{-1}(U)$. Hence $f^{-1}(U)$ is $\theta_{\mathcal{I}}$-open and hence f is $\theta_{(\mathcal{I},\mathcal{J})}$-irresolute. □

DEFINITION 5.4. *(1) An ideal space $(X, \tau, \mathcal{I})$ is said to be $\theta_{\mathcal{I}}$-compact if every cover of X by $\theta_{\mathcal{I}}$-open sets admits a finite subcover.*
(2) A subset A of an ideal space $(X, \tau, \mathcal{I})$ is said to be $\theta_{\mathcal{I}}$-compact relative to X if every cover of A by $\theta_{\mathcal{I}}$-open sets of X admits a finite subcover.

PROPOSITION 5.5. *Every quasi H^*-closed space $(X, \tau, \mathcal{I})$ is $\theta_{\mathcal{I}}$-compact.*

Proof. More generally, we show that if A is quasi H^*-closed relative to a space X, then A is $\theta_{\mathcal{I}}$-compact relative to X. Let $A \subseteq \cup\{V_\alpha : \alpha \in \Lambda\}$, where each V_α is $\theta_{\mathcal{I}}$-open, and A be quasi H^*-closed relative to X, then for each $x \in A$ there exists an $\alpha(x) \in \Lambda$ with $x \in V_{\alpha(x)}$. Then there exists an open set $U_{\alpha(x)}$ with $x \in U_{\alpha(x)}$ such that $Cl^*(U_{\alpha(x)}) \subseteq V_{\alpha(x)}$. Since $\{U_{\alpha(x)} : x \in A\}$ is a cover of A by open set in X, then there is a finite subset $\{x_1, x_2, ..., x_n\} \subseteq A$ such that $A \subseteq \cup\{Cl^*(U_{\alpha(x_i)}) : i = 1, 2, ..., n\} \subseteq \cup\{V_{\alpha(x_i)} : i = 1, 2, ..., n\}$. Hence A is $\theta_{\mathcal{I}}$-compact relative to X. □

THEOREM 5.6. *If $f : (X, \tau, \mathcal{I}) \to (Y, \sigma, \mathcal{J})$ is a $\theta_{(\mathcal{I},\mathcal{J})}$-irresolute surjection and $(X, \tau, \mathcal{I})$ is $\theta_{\mathcal{I}}$-compact, then Y is $\theta_{\mathcal{J}}$-compact.*

Proof. Let $\mathcal{V}$ be a $\theta_{\mathcal{J}}$-open covering of Y. Then, since f is $\theta_{(\mathcal{I},\mathcal{J})}$-irresolute, the collection $\mathcal{U} = \{f^{-1}(U) : U \in \mathcal{V}\}$ is a $\theta_{\mathcal{I}}$-open covering of X. Since X is $\theta_{\mathcal{I}}$-compact, there exists a finite subcollection $\{f^{-1}(U_i) : i = 1, ..., n\}$ of $\mathcal{U}$ which covers X. Now since f is onto, $\{U_i : i = 1, ..., n\}$ is a finite subcollection of $\mathcal{V}$ which covers Y. Hence Y is a $\theta_{\mathcal{J}}$-compact space. □

COROLLARY 5.7. *The $\theta_{(\mathcal{I},\mathcal{J})}$-continuous surjective image of a $\theta_{\mathcal{I}}$-compact space is $\theta_{\mathcal{J}}$-compact.*

DEFINITION 5.8. *An ideal topological space $(X, \tau, \mathcal{I})$ is said to be $*$-Lindelöf if for every open cover $\{U_\alpha : \alpha \in \Lambda\}$ of X there exists a countable subset $\{\alpha_n : n \in \mathbb{N}\} \subseteq \Lambda$ such that $X = \cup_{n\in\mathbb{N}} Cl^*(U_{\alpha_n})$.*

THEOREM 5.9. *Let $f : (X, \tau, \mathcal{I}) \to (Y, \sigma, \mathcal{J})$ be a $\theta_{(\mathcal{I},\mathcal{J})}$-continuous (resp. weakly $\mathcal{J}$-continuous) surjection. If X is $*$-Lindelöf (resp. Lindelöf), then Y is $*$-Lindelöf.*

Proof. Suppose that f is $\theta_{(\mathcal{I},\mathcal{J})}$-continuous and X is $*$-Lindelöf. Let $\{V_\alpha : \alpha \in \Lambda\}$ be an open cover of Y. For each $x \in X$, there exists $\alpha(x) \in \Lambda$ such that $f(x) \in V_{\alpha(x)}$. Since f is $\theta_{(\mathcal{I},\mathcal{J})}$-continuous, there exists an open set $U_{\alpha(x)}$ of X containing x such that $f(Cl^*(U_{\alpha(x)})) \subseteq Cl^*(V_{\alpha(x)})$. Now $\{U_{\alpha(x)} : x \in X\}$ is an open cover of the $*$-Lindelöf space X. So there exists a countable subset $\{U_{\alpha(x_n)} : n \in \mathbb{N}\}$ such that $X = \cup_{n\in\mathbb{N}}(Cl^*(U_{\alpha(x_n)}))$. Thus $Y = f(\cup_{n\in\mathbb{N}}(Cl^*(U_{\alpha(x_n)}))) \subseteq \cup_{n\in\mathbb{N}} f(Cl^*(U_{\alpha(x_n)})) \subseteq \cup_{n\in\mathbb{N}} Cl^*(V_{\alpha(x_n)})$. This shows that Y is $*$-Lindelöf. In case X is Lindelöf the proof is similar. □

A function $f : (X, \tau, \mathcal{I}) \to (Y, \sigma, \mathcal{J})$ is said to be $\theta_{(\mathcal{I},\mathcal{J})}$-closed if for each $\theta_{\mathcal{I}}$-closed set F in X, $f(F)$ is $\theta_{\mathcal{J}}$-closed in Y.

The following characterization of $\theta_{(\mathcal{I},\mathcal{J})}$-closed functions will be used in the sequel.

THEOREM 5.10. *A surjective function $f : (X, \tau, \mathcal{I}) \to (Y, \sigma, \mathcal{J})$ is $\theta_{(\mathcal{I},\mathcal{J})}$-closed if and only if for each set $B \subseteq Y$ and for each $\theta_{\mathcal{I}}$-open set U containing $f^{-1}(B)$, there exists a $\theta_{\mathcal{J}}$-open set V containing B such that $f^{-1}(V) \subseteq U$.*

Proof. Necessity. Suppose that f is $\theta_{(\mathcal{I},\mathcal{J})}$-closed. Since U is $\theta_{\mathcal{I}}$-open in X, $X - U$ is $\theta_{\mathcal{I}}$-closed and so $f(X - U)$ is $\theta_{\mathcal{J}}$-closed in Y. Now, $V = Y - f(X - U)$ is $\theta_{\mathcal{J}}$-open, $B \subseteq V$ and $f^{-1}(V) = f^{-1}(Y - f(X - U)) = X - f^{-1}(f(X - U)) \subseteq X - (X - U) = U$.

Sufficiency. Let A be a $\theta_{\mathcal{I}}$-closed set in X. To prove that $f(A)$ is $\theta_{\mathcal{J}}$-closed, we shall show that $Y - f(A)$ is $\theta_{\mathcal{J}}$-open. Let $y \in Y - f(A)$. Then $f^{-1}(y) \cap f^{-1}(f(A)) = \phi$ and so $f^{-1}(y) \subseteq X - f^{-1}(f(A)) \subseteq X - A$. By hypothesis there exists a $\theta_{\mathcal{J}}$-open set V containing y such that $f^{-1}(V) \subseteq X - A$. So $A \subseteq X - f^{-1}(V)$ and hence $f(A) \subseteq f(X - f^{-1}(V)) = Y - V$. Thus $V \subseteq Y - f(A)$ and so the set $Y - f(A)$ being the union of $\theta_{\mathcal{J}}$-open sets is $\theta_{\mathcal{J}}$-open. □

THEOREM 5.11. *Let $f : (X, \tau, \mathcal{I}) \to (Y, \sigma, \mathcal{J})$ be a $\theta_{(\mathcal{I},\mathcal{J})}$-closed surjection such that for each $y \in Y$, $f^{-1}(y)$ is $\theta_{\mathcal{I}}$-compact relative to X. If Y is $\theta_{\mathcal{J}}$-compact, then X is $\theta_{\mathcal{I}}$-compact.*

Proof. Let $\mathcal{U} = \{U_\alpha : \alpha \in \Lambda\}$ be a $\theta_{\mathcal{I}}$-open covering of X. Since for each $y \in Y$, $f^{-1}(y)$ is $\theta_{\mathcal{I}}$-compact relative to X, we can choose a finite subset Λ_y of Λ such that $\{U_\beta : \beta \in \Lambda_y\}$ is a covering of $f^{-1}(y)$. Now, by Theorem 5.10, there exists a $\theta_{\mathcal{J}}$-open set V_y containing y such that $f^{-1}(V_y) \subseteq \cup\{U_\beta : \beta \in \Lambda_y\}$. The collection $\mathcal{V} = \{V_y : y \in Y\}$ is a $\theta_{\mathcal{J}}$-open covering of Y. In view of $\theta_{\mathcal{J}}$-compactness of Y there exists a finite subcollection $\{V_{y_1}, ..., V_{y_n}\}$ of $\mathcal{V}$ which covers Y. Then the finite subcollection $\{U_\beta : \beta \in \Lambda_{y_i}, i = 1, ..., n\}$ of $\mathcal{U}$ covers X. Hence X is a $\theta_{\mathcal{I}}$-compact space. □

Let (X, τ) be a space with an ideal $\mathcal{I}$ on X and $D \subseteq X$. Then $\mathcal{I}_D = \{D \cap A : A \in \mathcal{I}\}$ is obviously an ideal on D.

THEOREM 5.12. *Let $f : (X, \tau, \mathcal{I}) \to (Y, \sigma, \mathcal{J})$ be a function, D be a dense subset in the topological space (Y, σ^*) and $f(X) \subseteq D$. Then the following properties are equivalent:*

1. *$f : (X, \tau, \mathcal{I}) \to (Y, \sigma, \mathcal{J})$ is $\theta_{(\mathcal{I},\mathcal{J})}$-continuous;*
2. *$f : (X, \tau, \mathcal{I}) \to (D, \sigma_D, \mathcal{J}_D)$ is $\theta_{(\mathcal{I},\mathcal{J}_D)}$-continuous.*

Proof. (1) ⇒ (2): Let $x \in X$ and W be any open set of D containing $f(x)$, that is $f(x) \in W \in \sigma_D$. Then exists a $V \in \sigma$ such that $W = D \cap V$. Since $f : (X, \tau, \mathcal{I}) \to (Y, \sigma, \mathcal{J})$ is $\theta_{(\mathcal{I},\mathcal{J})}$-continuous and $f(x) \in V \in \sigma$, there exists $U \in \tau$ such that $x \in U$ and $f(Cl^*(U)) \subseteq Cl^*(V)$. If D is a dense subset in the topological space (Y, σ^*), then D is a dense subset in the topological space

(Y, σ) since $Cl^*(D) \subseteq Cl(D)$. Since $\sigma \subseteq \sigma^*$, $V \in \sigma^*$. So, $Cl^*(D \cap V) = Cl^*(V)$ since D is dense. Thus $f(Cl^*(U)) \subseteq Cl^*(V) \cap f(X) \subseteq Cl^*(D \cap V) \cap D \subseteq Cl^*(V) \cap D$. Since $W = D \cap V$, $Cl^*_D(W) = Cl^*(V) \cap D$ by [7, Lemma 4] $f(Cl^*(U)) \subseteq Cl^*_D(W)$. Hence we obtain that $f : (X, \tau, \mathcal{I}) \to (D, \sigma_D, \mathcal{J}_D)$ is $\theta_{(\mathcal{I},\mathcal{J}_D)}$-continuous.
(2) $\Rightarrow$ (1): Let $x \in X$ and V be any open set Y containing $f(x)$. Since $f(x) \in D \cap V$ and $D \cap V \in \sigma_D$, by (2) there exists $U \in \tau$ containing x such that $f(Cl^*(V)) \subseteq Cl^*_D(D \cap V) = Cl^*(D \cap V) \cap D \subseteq Cl^*(V)$. This shows that f is $\theta_{(\mathcal{I},\mathcal{J})}$-continuous. □

Acknowledgements. The authors wish to thank the referee for useful comments and suggestions. Especially, Theorems 3.5, 4.6 and 5.12, Corollary 4.13 and Proposition 5.5 have been improved by the suggestions of the referee.

References

[1] A. Açikgöz, T. Noiri and S. Yüksel, *A decomposition of continuity in ideal topological spaces*, Acta Math. Hungar. **105** (2004), 285–289.
[2] A. Açikgöz, T. Noiri and S. Yüksel, *On $*$-operfect sets and α-$*$-closed sets*, Acta Math. Hungar. **127** (2010), 146–153.
[3] J. Dontchev, M. Ganster and D.Rose, *Ideal resolvability*, Topology Appl. **93** (1999), 1–16.
[4] J. Dontchev, *On pre-$\mathcal{I}$-open sets and a decomposition of $\mathcal{I}$-continuity*, Banyan Math. J. **2** (1996).
[5] E. Ekici and T. Noiri, *$*$-extremally disconnected ideal topological spaces*, Acta Math. Hungar. **122** (2009), 81–90.
[6] S. Fomin, *Extensions of topological spaces*, Ann. of Math. (2) **44** (1943), 471–480.
[7] E. Hatir, A. Keskin and T. , *A note on strong β-$\mathcal{I}$-sets and strongly β-$\mathcal{I}$ continuous functions*, Acta Math. Hungar. **108** (2005) 87–94.
[8] E. Hayashi, *Topologies defined by local properties*, Math. Ann. **156** (1964), 205–215.
[9] D. Janković and T. R. Hamlett, *New topologies from old via ideals*, Amer. Math. Monthly **97** (1990), 295–310.
[10] V. Jeyanthi, V. Renuka Devi and D. Sivaraj, *Weakly $\mathcal{I}$-continuous functions*, Acta Math. Hungar. **113** (2006), 319–324.
[11] K. Kuratowski, *Topology I*, Academic Press, New York (1966).
[12] F. Kuyucu, T. Noiri and A. A. Özkurt, *A note in W-$\mathcal{I}$-continuous functions*, Acta Math. Hungar. **119** (2008), 393–400.
[13] N. Levine, *A decomposition of continuity in topological spaces*, Amer. Math. Monthly **68** (1961), 44–46.
[14] T. Noiri, *On δ-continuous functions*, J. Korean Math. Soc. **16** (1980), 161–166.
[15] V. Renuka Devi, D. Sivaraj and T. Tamizh Chelvam, *Codense and completely codense ideals*, Acta Math. Hungar. **108** (2005), 197–205.

[16] D. A. Rose, *Weak continuity and almost continuity*, Internat. J. Math. Math. Sci. **7** (1984), 311–318.
[17] R. Vaidyanathaswamy, *Set Topology*, Chelsea Publishing Co., New York (1960).
[18] N. V. Veličko, *H-closed topological spaces*, Amer. Math. Soc. Transl. **78** (1968), 103–118.
[19] S. Yüksel, A. Açikgöz and T. Noiri, *On δ-$\mathcal{I}$-continuous functions*, Turkish J. Math. **29** (2005), 39–51.

Authors' addresses:

Ahmad Al-Omari
Department of Mathematics, Faculty of Science
Al al-Bayt University,
P.O. Box 130095,
Mafraq 25113, Jordan
E-mail: omarimutah1@yahoo.com

Takashi Noiri
2949-1 Shiokita-cho
Hinagu, Yatsushiro-shi,
Kumamoto-ken, 869-5142, Japan
E-mail: t.noiri@nifty.com

Received October 10, 2011
Revised January 30, 2012

Rend. Istit. Mat. Univ. Trieste
Volume 44 (2012), 413–422

Rank two globally generated vector bundles with $c_1 \leq 5$

Ludovica Chiodera and Philippe Ellia

Abstract. *We classify globally generated rank two vector bundles on $\mathbb{P}^n$, $n \geq 3$, with $c_1 \leq 5$. The classification is complete but for one case ($n = 3$, $c_1 = 5$, $c_2 = 12$).*

Keywords: Vector bundles, rank two, globally generated, projective space
MS Classification 2010: 14F05, 14M15

1. Introduction.

Vector bundles generated by global sections are basic objects in projective algebraic geometry. Globally generated line bundles correspond to morphisms to a projective space, more generally higher rank bundles correspond to morphism to (higher) Grassmann varieties. For this last point of view (that won't be touched in this paper) see [10, 12, 13]. Also globally generated vector bundles appear in a variety of problems ([7] just to make a single, recent example).

In this paper we classify globally generated rank two vector bundles on $\mathbb{P}^n$ (projective space over k, $\overline{k} = k$, $ch(k) = 0$), $n \geq 3$, with $c_1 \leq 5$. The result is:

Theorem 1.1. *Let E be a rank two vector bundle on $\mathbb{P}^n$, $n \geq 3$, generated by global sections with Chern classes c_1, c_2, $c_1 \leq 5$.*

1. *If $n \geq 4$, then E is the direct sum of two line bundles*
2. *If $n = 3$ and E is indecomposable, then*

$$(c_1, c_2) \in S = \{((2,2),(4,5),(4,6),(4,7),(4,8),(5,8),(5,10),(5,12)\}.$$

If E exists there is an exact sequence:

$$0 \to \mathcal{O} \to E \to \mathcal{I}_C(c_1) \to 0 \ (*)$$

where $C \subset \mathbb{P}^3$ is a smooth curve of degree c_2 with $\omega_C(4 - c_1) \simeq \mathcal{O}_C$. The curve C is irreducible*, except maybe if $(c_1, c_2) = (4, 8)$: in this case C can be either irreducible or the disjoint union of two smooth conics.*

3. *For every* $(c_1, c_2) \in S$, $(c_1, c_2) \neq (5, 12)$, *there exists a rank two vector bundle on* $\mathbb{P}^3$ *with Chern classes* (c_1, c_2) *which is globally generated (and with an exact sequence as in 2.).*

The classification is complete, but for one case: we are unable to say if there exist or not globally generated rank two vector bundles with Chern classes $c_1 = 5, c_2 = 12$ on $\mathbb{P}^3$.

2. Rank two vector bundles on $\mathbb{P}^3$.

2.1. General facts.

For completeness let's recall the following well known results:

LEMMA 2.1. *Let* E *be a rank* r *vector bundle on* $\mathbb{P}^n$, $n \geq 3$. *Assume* E *is generated by global sections.*

1. *If* $c_1(E) = 0$, *then* $E \simeq r.\mathcal{O}$

2. *If* $c_1(E) = 1$, *then* $E \simeq \mathcal{O}(1) \oplus (r-1).\mathcal{O}$ *or* $E \simeq T(-1) \oplus (r-n).\mathcal{O}$.

Proof. If $L \subset \mathbb{P}^n$ is a line then $E|L \simeq \bigoplus_{i=1}^{r} \mathcal{O}_L(a_i)$ by a well known theorem and $a_i \geq 0, \forall i$ since E is globally generated. It turns out that in both cases: $E|L \simeq \mathcal{O}_L(c_1) \oplus (r-1).\mathcal{O}_L$ for every line L, i.e. E is uniform. Then 1. follows from a result of Van de Ven ([14]), while 2. follows from IV. Prop. 2.2 of [4]. □

LEMMA 2.2. *Let* E *be a rank two vector bundle on* $\mathbb{P}^n$, $n \geq 3$. *If* E *has a nowhere vanishing section then* E *splits. If* E *is generated by global sections and doesn't split then* $h^0(E) \geq 3$ *and a general section of* E *vanishes along a smooth curve,* C, *of degree* $c_2(E)$ *such that* $\omega_C(4 - c_1) \simeq \mathcal{O}_C$. *Moreover* $\mathcal{I}_C(c_1)$ *is generated by global sections.*

LEMMA 2.3. *Let* E *be a non split rank two vector bundle on* $\mathbb{P}^3$ *with* $c_1 = 2$. *If* E *is generated by global sections then* E *is a null-correlation bundle.*

Proof. We have an exact sequence: $0 \to \mathcal{O} \to E \to \mathcal{I}_C(2) \to 0$, where C is a smooth curve with $\omega_C(2) \simeq \mathcal{O}_C$. It follows that C is a disjoint union of lines. Since $h^0(\mathcal{I}_C(2)) \geq 2$, $d(C) \leq 2$. Finally $d(C) = 2$ because E doesn't split. □

This settles the classification of rank two globally generated vector bundles with $c_1(E) \leq 2$ on $\mathbb{P}^3$.

2.2. Globally generated rank two vector bundles with $c_1 = 3$.

The following result has been proved in [10] (with a different and longer proof).

PROPOSITION 2.4. *Let E be a rank two globally generated vector bundle on $\mathbb{P}^3$. If $c_1(E) = 3$ then E splits.*

Proof. Assume a general section vanishes in codimension two, then it vanishes along a smooth curve C such that $\omega_C \simeq \mathcal{O}_C(-1)$. Moreover $\mathcal{I}_C(3)$ is generated by global sections. We have $C = \cup_{i=1}^r C_i$ (disjoint union) where each C_i is smooth irreducible with $\omega_{C_i} \simeq \mathcal{O}_{C_i}(-1)$. It follows that each C_i is a smooth conic. If $r \geq 2$ let $L = \langle C_1 \rangle \cap \langle C_2 \rangle$ ($\langle C_i \rangle$ is the plane spanned by C_i). Every cubic containing C contains L (because it contains the four points $C_1 \cap L$, $C_2 \cap L$). This contradicts the fact that $\mathcal{I}_C(3)$ is globally generated. Hence $r = 1$ and $E = \mathcal{O}(1) \oplus \mathcal{O}(2)$. □

2.3. Globally generated rank two vector bundles with $c_1 = 4$.

Let's start with a general result:

LEMMA 2.5. *Let E be a non split rank two vector bundle on $\mathbb{P}^3$ with Chern classes c_1, c_2. If E is globally generated and if $c_1 \geq 4$ then:*

$$c_2 \leq \frac{2c_1^3 - 4c_1^2 + 2}{3c_1 - 4}.$$

Proof. By our assumptions a general section of E vanishes along a smooth curve, C, such that $\mathcal{I}_C(c_1)$ is generated by global sections. Let U be the complete intersections of two general surfaces containing C. Then U links C to a smooth curve, Y. We have $Y \neq \varnothing$ since E doesn't split. The exact sequence of liaison: $0 \to \mathcal{I}_U(c_1) \to \mathcal{I}_C(c_1) \to \omega_Y(4 - c_1) \to 0$ shows that $\omega_Y(4 - c_1)$ is generated by global sections. Hence $\deg(\omega_Y(4 - c_1)) \geq 0$. We have $\deg(\omega_Y(4 - c_1)) = 2g' - 2 + d'(4 - c_1)$ ($g' = p_a(Y)$, $d' = \deg(Y)$). So $g' \geq \frac{d'(c_1-4)+2}{2} \geq 0$ (because $c_1 \geq 4$). On the other hand, always by liaison, we have: $g' - g = \frac{1}{2}(d' - d)(2c_1 - 4)$ ($g = p_a(C)$, $d = \deg(C)$). Since $d' = c_1^2 - d$ and $g = \frac{d(c_1-4)}{2} + 1$ (because $\omega_C(4 - c_1) \simeq \mathcal{O}_C$), we get: $g' = 1 + \frac{d(c_1-4)}{2} + \frac{1}{2}(c_1^2 - 2d)(2c_1 - 4) \geq 0$ and the result follows. □

Now we have:

PROPOSITION 2.6. *Let E be a rank two globally generated vector bundle on $\mathbb{P}^3$. If $c_1(E) = 4$ and if E doesn't split, then $5 \leq c_2 \leq 8$ and there is an exact sequence: $0 \to \mathcal{O} \to E \to \mathcal{I}_C(4) \to 0$, where C is a smooth irreducible elliptic curve of degree c_2 or, if $c_2 = 8$, C is the disjoint union of two smooth elliptic quartic curves.*

Proof. A general section of E vanishes along C where C is a smooth curve with $\omega_C = \mathcal{O}_C$ and where $\mathcal{I}_C(4)$ is generated by global sections. Let $C = C_1 \cup ... \cup C_r$ be the decomposition into irreducible components: the union is disjoint, each C_i is a smooth elliptic curve hence has degree at least three.
By Lemma 2.5 $d = \deg(C) \leq 8$. If $d \leq 4$ then C is irreducible and is a complete intersection which is impossible since E doesn't split. If $d = 5$, C is smooth irreducible.
Claim: If $8 \geq d \geq 6$, C cannot contain a plane cubic curve.
Assume $C = P \cup X$ where P is a plane cubic and where X is a smooth elliptic curve of degree $d - 3$. If $d = 6$, X is also a plane cubic and every quartic containing C contains the line $\langle P \rangle \cap \langle X \rangle$. If $deg(X) \geq 4$ then every quartic, F, containing C contains the plane $\langle P \rangle$. Indeed $F|H$ vanishes on P and on the $deg(X) \geq 4$ points of $X \cap \langle P \rangle$, but these points are not on a line so $F|H = 0$. In both cases we get a contradiction with the fact that $\mathcal{I}_C(4)$ is generated by global sections. The claim is proved.
It follows that, if $8 \geq d \geq 6$, then C is irreducible except if $C = X \cup Y$ is the disjoint union of two elliptic quartic curves. □

Now let's show that all possibilities of Proposition 2.6 do actually occur. For this we have to show the existence of a smooth irreducible elliptic curve of degree d, $5 \leq d \leq 8$ with $\mathcal{I}_C(4)$ generated by global sections (and also that the disjoint union of two elliptic quartic curves is cut off by quartics).

LEMMA 2.7. *There exist rank two vector bundles with $c_1 = 4, c_2 = 5$ which are globally generated. More precisely any such bundle is of the form $\mathcal{N}(2)$, where $\mathcal{N}$ is a null-correlation bundle (a stable bundle with $c_1 = 0, c_2 = 1$).*

Proof. The existence is clear (if $\mathcal{N}$ is a null-correlation bundle then it is well known that $\mathcal{N}(k)$ is globally generated if $k \geq 1$). Conversely if E has $c_1 = 4, c_2 = 5$ and is globally generated, then E has a section vanishing along a smooth, irreducible quintic elliptic curve (cf 2.6). Since $h^0(\mathcal{I}_C(2)) = 0$, E is stable, hence $E = \mathcal{N}(2)$. □

LEMMA 2.8. *There exist smooth, irreducible elliptic curves, C, of degree 6 with $\mathcal{I}_C(4)$ generated by global sections.*

Proof. Let X be the union of three skew lines. The curve X lies on a smooth quadric surface, Q, and has $\mathcal{I}_X(3)$ globally generated (indeed the exact sequence $0 \to \mathcal{I}_Q \to \mathcal{I}_X \to \mathcal{I}_{X,Q} \to 0$ twisted by $\mathcal{O}(3)$ reads like: $0 \to \mathcal{O}(1) \to \mathcal{I}_C(3) \to \mathcal{O}_Q(3,0) \to 0$). The complete intersection, U, of two general cubics containing X links X to a smooth curve, C, of degree 6 and arithmetic genus 1. Since, by liaison, $h^1(\mathcal{I}_C) = h^1(\mathcal{I}_X(-2)) = 0$, C is irreducible. The exact sequence of liaison: $0 \to \mathcal{I}_U(4) \to \mathcal{I}_C(4) \to \omega_X(2) \to 0$ shows that $\mathcal{I}_C(4)$ is globally generated. □

In order to prove the existence of smooth, irreducible elliptic curves, C, of degree $d = 7, 8$, with $\mathcal{I}_C(4)$ globally generated, we have to recall some results due to Mori ([11]).

According to [11] Remark 4, Prop. 6, there exists a smooth quartic surface $S \subset \mathbb{P}^3$ such that $Pic(S) = \mathbb{Z}H \oplus \mathbb{Z}X$ where X is a smooth elliptic curve of degree d ($7 \leq d \leq 8$). The intersection pairing is given by: $H^2 = 4$, $X^2 = 0$, $H.X = d$. Such a surface doesn't contain any smooth rational curve ([11, p. 130]). In particular: $(*)$ every integral curve, Z, on S has degree ≥ 4 with equality if and only if Z is a planar quartic curve or an elliptic quartic curve.

LEMMA 2.9. *With notations as above, $h^0(\mathcal{I}_X(3)) = 0$.*

Proof. A curve $Z \in |3H - X|$ has invariants $(d_Z, g_Z) = (5, -2)$ (if $d = 7$) or $(4, -5)$ (if $d = 8$), so Z is not integral. It follows that Z must contain an integral curve of degree < 4, but this is impossible. □

LEMMA 2.10. *With notations as above $|4H - X|$ is base point free, hence there exist smooth, irreducible elliptic curves, X, of degree d, $7 \leq d \leq 8$, such that $\mathcal{I}_X(4)$ is globally generated.*

Proof. Let's first prove the following: *Claim:* Every curve in $|4H - X|$ is integral.

If $Y \in |4H - X|$ is not integral then $Y = Y_1 + Y_2$ where Y_1 is integral with $\deg(Y_1) = 4$ (observe that $\deg(Y) = 9$ or 8).

If Y_1 is planar then $Y_1 \sim H$, so $4H - X \sim H + Y_2$ and it follows that $3H \sim X + Y_2$, in contradiction with $h^0(\mathcal{I}_X(3)) = 0$ (cf 2.9).

So we may assume that Y_1 is a quartic elliptic curve, i.e. (i) $Y_1^2 = 0$ and (ii) $Y_1.H = 4$. Setting $Y_1 = aH + bX$, we get from (i): $2a(2a + bd) = 0$. Hence (α) $a = 0$, or (β) $2a + bd = 0$.

(α) In this case $Y_1 = bX$, hence (for degree reasons and since S doesn't contain curves of degree < 4), $Y_2 = \varnothing$ and $Y = X$, which is integral.

(β) Since $Y_1.H = 4$, we get $2a + (2a + bd) = 2a = 4$, hence $a = 2$ and $bd = -4$ which is impossible ($d = 7$ or 8 and $b \in \mathbb{Z}$).

This concludes the proof of the claim.

Since $(4H - X)^2 \geq 0$, the claim implies that $4H - X$ is numerically effective. Now we conclude by a result of Saint-Donat (cf. [11, Theorem 5]) that $|4H - X|$

is base point free, i.e. $\mathcal{I}_{X,S}(4)$ is globally generated. By the exact sequence: $0 \to \mathcal{O} \to \mathcal{I}_X(4) \to \mathcal{I}_{X,S}(4) \to 0$ we get that $\mathcal{I}_X(4)$ is globally generated. □

REMARK 2.11. *If $d = 8$, a general element $Y \in |4H - X|$ is a smooth elliptic curve of degree 8. By the way $Y \neq X$ (see [1]). The exact sequence of liaison: $0 \to \mathcal{I}_U(4) \to \mathcal{I}_X(4) \to \omega_Y \to 0$ shows that $h^0(\mathcal{I}_X(4)) = 3$ (i.e. X is of maximal rank). In case $d = 8$ Lemma 2.10 is stated in [2], however the proof there is incomplete, indeed in order to apply the enumerative formula of [8] one has to know that X is a* connected *component of $\bigcap_{i=1}^{3} F_i$; this amounts to say that the base locus of $|4H - X|$ on F_1 has dimension ≤ 0.*

To conclude we have:

LEMMA 2.12. *Let X be the disjoint union of two smooth, irreducible quartic elliptic cuvres, then $\mathcal{I}_X(4)$ is generated by global sections.*

Proof. Let $X = C_1 \sqcup C_2$. We have: $0 \to \mathcal{O}(-4) \to 2.\mathcal{O}(-2) \to \mathcal{I}_{C_1} \to 0$, twisting by $\mathcal{I}_{C_2}$, since $C_1 \cap C_2 = \varnothing$, we get:
$0 \to \mathcal{I}_{C_2}(-4) \to 2.\mathcal{I}_{C_2}(-2) \to \mathcal{I}_X \to 0$ and the result follows. □

Summarizing:

PROPOSITION 2.13. *There exists an indecomposable rank two vector bundle, E, on $\mathbb{P}^3$, generated by global sections and with $c_1(E) = 4$ if and only if $5 \leq c_2(E) \leq 8$ and in these cases there is an exact sequence:*

$$0 \to \mathcal{O} \to E \to \mathcal{I}_C(4) \to 0$$

where C is a smooth irreducible elliptic curve of degree $c_2(E)$ or, if $c_2(E) = 8$, the disjoint union of two smooth elliptic quartic curves.

2.4. Globally generated rank two vector bundles with $c_1 = 5$.

We start by listing the possible cases:

PROPOSITION 2.14. *If E is an indecomposable, globally generated, rank two vector bundle on $\mathbb{P}^3$ with $c_1(E) = 5$, then $c_2(E) \in \{8, 10, 12\}$ and there is an exact sequence:*

$$0 \to \mathcal{O} \to E \to \mathcal{I}_C(5) \to 0$$

where C is a smooth, irreducible curve of degree $d = c_2(E)$, with $\omega_C \simeq \mathcal{O}_C(1)$.

In any case E is stable.

Proof. A general section of E vanishes along a smooth curve, C, of degree $d = c_2(E)$ with $\omega_C \simeq \mathcal{O}_C(1)$. Hence every irreducible component, Y, of C is a smooth, irreducible curve with $\omega_Y \simeq \mathcal{O}_Y(1)$. In particular $\deg(Y) = 2g(Y) - 2$ is even and $\deg(Y) \geq 4$.

1. If $d = 4$, then C is a planar curve and E splits.
2. If $d = 6$, C is necessarily irreducible (of genus 4). It is well known that any such curve is a complete intersection $(2, 3)$, hence E splits.
3. If $d = 8$ and C is not irreducible, then $C = P_1 \sqcup P_2$, the disjoint union of two planar quartic curves. If $L = \langle P_1 \rangle \cap \langle P_2 \rangle$, then every quintic containing C contains L in contradiction with the fact that $\mathcal{I}_C(5)$ is generated by global sections. Hence C is irreducible.
4. If $d = 10$ and C is not irreducible, then $C = P \sqcup X$, where P is a planar curve of degree 4 and where X is a degree 6 curve (X is a complete intersection $(2, 3)$). Every quintic containing C vanishes on P and on the 8 points of $X \cap \langle P \rangle$, since these 8 points are not on a line, the quintic vanishes on the plane $\langle P \rangle$. This contradicts the fact that $\mathcal{I}_C(5)$ is globally generated.
5. If $d = 12$ and C is not irreducible we have three possibilities:
   (a) $C = P_1 \sqcup P_2 \sqcup P_3$, P_i planar quartic curves
   (b) $C = X_1 \sqcup X_2$, X_i complete intersection curves of types $(2, 3)$
   (c) $C = Y \sqcup P$, Y a canonical curve of degree 8, P a planar curve of degree 4.

   (a) This case is impossible (consider the line $\langle P_1 \rangle \cap \langle P_2 \rangle$).
   (b) We have $X_i = Q_i \cap F_i$. Let Z be the quartic curve $Q_1 \cap Q_2$. Then $X_i \cap Z = F_i \cap Z$, i.e. X_i meets Z in 12 points. It follows that every quintic containing C meets Z in 24 points, hence such a quintic contains Z. Again this contradicts the fact that $\mathcal{I}_C(5)$ is globally generated.
   (c) This case too is impossible: every quintic containing C vanishes on P and on the points $\langle P \rangle \cap Y$, hence on $\langle P \rangle$.

   We conclude that if $d = 12$, C is irreducible.

The normalized bundle is $E(-3)$, since in any case $h^0(\mathcal{I}_C(2)) = 0$ (every smooth irreducible subcanonical curve on a quadric surface is a complete intersection), E is stable. □

Now we turn to the existence part.

LEMMA 2.15. *There exist indecomposable rank two vector bundles on $\mathbb{P}^3$ with Chern classes $c_1 = 5$ and $c_2 \in \{8, 10\}$ which are globally generated.*

Proof. Let $R = \sqcup_{i=1}^{s} L_i$ be the union of s disjoint lines, $2 \leq s \leq 3$. We may perform a liaison $(s,3)$ and link R to $K = \sqcup_{i=1}^{s} K_i$, the union of s disjoint conics. The exact sequence of liaison: $0 \to \mathcal{I}_U(4) \to \mathcal{I}_K(4) \to \omega_R(5-s) \to 0$ shows that $\mathcal{I}_K(4)$ is globally generated (n.b. $5-s \geq 2$).
Since $\omega_K(1) \simeq \mathcal{O}_K$ we have an exact sequence: $0 \to \mathcal{O} \to \mathcal{E}(2) \to \mathcal{I}_K(3) \to 0$, where $\mathcal{E}$ is a rank two vector bundle with Chern classes $c_1 = -1, c_2 = 2s-2$. Twisting by $\mathcal{O}(1)$ we get: $0 \to \mathcal{O}(1) \to \mathcal{E}(3) \to \mathcal{I}_K(4) \to 0$ $(*)$. The Chern classes of $\mathcal{E}(3)$ are $c_1 = 5$, $c_2 = 2s+4$ (i.e. $c_2 = 8, 10$). Since $\mathcal{I}_K(4)$ is globally generated, it follows from $(*)$ that $\mathcal{E}(3)$ too, is generated by global sections. □

Remark 2.16.

1. *If $\mathcal{E}$ is as in the proof of Lemma 2.15 a general section of $\mathcal{E}(3)$ vanishes along a smooth, irreducible (because $h^1(\mathcal{E}(-2)) = 0$) canonical curve, C, of genus $1 + c_2/2$ ($g = 5, 6$) such that $\mathcal{I}_C(5)$ is globally generated. By construction these curves are not of maximal rank ($h^0(\mathcal{I}_C(3)) = 1$ if $g = 5$, $h^0(\mathcal{I}_C(4)) = 2$ if $g = 6$). As explained in [9] 4 this is a general fact: no canonical curve of genus $g, 5 \leq g \leq 6$ in $\mathbb{P}^3$ is of maximal rank. We don't know if this is still true for $g = 7$.*

2. *According to [9] the general canonical curve of genus* 6 *lies on a unique quartic surface.*

3. *The proof of 2.15 breaks down with four conics: $\mathcal{I}_K(4)$ is no longer globally generated, every quartic containing K vanishes along the lines L_i ($5 - s = 1$). Observe also that four disjoint lines always have a quadrisecant and hence are an exception to the* normal generation conjecture *(the omogeneous ideal is not generated in degree three as it should be).*

Remark 2.17. *The case $(c_1, c_2) = (5, 12)$ remains open. It can be shown that if E exists, a general section of E is linked, by a complete intersections of two quintics, to a smooth, irreducible curve, X, of degree* 13, *genus* 10 *having $\omega_X(-1)$ as a base point free g^1_5. One can prove the existence of curves $X \subset \mathbb{P}^3$, smooth, irreducible, of degree* 13, *genus* 10, *with $\omega_X(-1)$ a base point free pencil and lying on* one *quintic surface. But we are unable to show the existence of such a curve with $h^0(\mathcal{I}_X(5)) \geq 3$ (or even with $h^0(\mathcal{I}_X(5)) \geq 2$). We believe that such bundles do not exist.*

3. Globally generated rank two vector bundles on $\mathbb{P}^n$, $n \geq 4$.

For $n \geq 4$ and $c_1 \leq 5$ there is no surprise:

Proposition 3.1. *Let E be a globally generated rank two vector bundle on $\mathbb{P}^n$, $n \geq 4$. If $c_1(E) \leq 5$, then E splits.*

Proof. It is enough to treat the case $n = 4$. A general section of E vanishes along a smooth (irreducible) subcanonical surface, S: $0 \to \mathcal{O} \to E \to \mathcal{I}_S(c_1) \to 0$. By [5], if $c_1 \leq 4$, then S is a complete intersection and E splits. Assume now $c_1 = 5$. Consider the restriction of E to a general hyperplane H. If E doesn't split, by 2.14 we get that the normalized Chern classes of E are: $c_1 = -1$, $c_2 \in \{2, 4, 6\}$. By Schwarzenberger condition: $c_2(c_2 + 2) \equiv 0$ (*mod* 12). The only possibilities are $c_2 = 4$ or $c_2 = 6$. If $c_2 = 4$, since E is stable (because $E|H$ is, see 2.14), we have ([3]) that E is a Horrocks-Mumford bundle. But the Horrocks-Mumford bundle (with $c_1 = 5$) is not globally generated.

The case $c_2 = 6$ is impossible: such a bundle would yield a smooth surface $S \subset \mathbb{P}^4$, of degree 12 with $\omega_S \simeq \mathcal{O}_S$, but the only smooth surface with $\omega_S \simeq \mathcal{O}_S$ in $\mathbb{P}^4$ is the abelian surface of degree 10 of Horrocks-Mumford. □

REMARK 3.2. *For $n > 4$ the results in [6] give stronger and stronger (as n increases) conditions for the existence of indecomposable rank two vector bundles generated by global sections.*

Putting everything together, the proof of Theorem 1.1 is complete.

REFERENCES

[1] V. BEORCHIA AND PH. ELLIA, *Normal bundle and complete intersections*, Rend. Sem. Mat. Univ. Politec. Torino **48** (1990), 553–562.

[2] J. D'ALMEIDA, *Une involution sur un espace de modules de fibrés instantons*, Bull. Soc. Math. France **128** (2000), 577–584.

[3] W. DECKER, *Stable rank 2 vector bundles with Chern classes $c_1 = -1, c_2 = 4$*, Math. Ann. **275** (1986), 481–500.

[4] PH. ELLIA, *Sur les fibrés uniformes de rang $n + 1$ sur $\mathbb{P}^n$*, Mém. Soc. Math. France **7** (1982).

[5] PH. ELLIA, D. FRANCO, AND L. GRUSON, *On subcanonical surfaces of $\mathbb{P}^4$*, Math. Z. **251** (2005), 257–265.

[6] PH. ELLIA, D. FRANCO, AND L. GRUSON, *Smooth divisors of projective hypersurfaces*, Comment. Math. Helv. **83** (2008), 371–385.

[7] M.L. FANIA AND E. MEZZETTI, *Vector spaces of skew-symmetric matrices of constant rank*, Linear Algebra Appl. **434** (2011), 2383–2403.

[8] W. FULTON, *Intersection theory*, Ergeb. Math. Grenzgeb., no. 2, Springer, Berlin, 1984.

[9] L. GRUSON AND CH. PESKINE, *Genre des courbes de l'espace projectif*, Lecture Notes in Math. **687** (1978), 31–59.

[10] S. HUH, *On triple Veronese embeddings of $\mathbb{P}^n$ in the Grassmannians*, Math. Nachr. **284** (2011), 1453–1461.

[11] S. MORI, *On degrees and genera of curves on smooth quartic surfaces in $\mathbb{P}^3$*, Nagoya Math. J. **96** (1984), 127–132.

[12] J.C. SIERRA AND L. UGAGLIA, *On double Veronese embeddings in the Grassmannian $G(1, N)$*, Math. Nachr. **279** (2006), 798–804.

[13] J.C. Sierra and L. Ugaglia, *On globally generated vector bundles on projective spaces*, J. Pure Appl. Algebra **213** (2009), 2141–2146.
[14] A. Van de Ven, *On uniform vector bundles*, Math. Ann. **195** (1972), 245–248.

Authors' addresses:

Ludovica Chiodera
Dipartimento di Matematica
Università di Ferrara
via Machiavelli 35, 44100 Ferrara, Italy
E-mail: ludovica.chiodera@unife.it

Philippe Ellia
Dipartimento di Matematica
Università di Ferrara
via Machiavelli 35, 44100 Ferrara, Italy
E-mail: phe@unife.it

Received November 30, 2011
Revised April 2, 2012

Rend. Istit. Mat. Univ. Trieste
Volume 44 (2012), 423–437

Contra continuity on weak structure spaces

Ahmad Al-Omari

Abstract. *We introduce some contra continuous functions in weak structure spaces such as contra $(\mathcal{M}, w)$-continuous functions, contra $(\alpha(m), w)$-continuous functions, contra $(\sigma(m), w)$-continuous functions, contra $(\pi(m), w)$-continuous functions and contra $(\beta(m), w)$-continuous functions. We investigate their characterization and relationships among such functions.*

Keywords: weak structure, contra continuity, contra $(\mathcal{M}, w)$-continuity
MS Classification 2010: 54A05, 54C10

1. Introduction and Preliminaries

Császár [4] introduced a generalized structure called generalized topology. Recently, Császár [5] has introduced a new notion of structures called a weak structure which is weaker than both a generalized topology [4] and a minimal structure [8, 9]. Let X be a nonempty set and $w \subseteq \mathcal{P}(X)$, where $\mathcal{P}(X)$ is the power set of X. Then w is called a weak structure (briefly WS) on X if $\emptyset \in w$. Each member of w is said to be w-open and the complement of a w-open set is said to be w-closed. Let w be a weak structure on X and $A \subseteq X$. Császár [5] defined (as in the general case) $i_w(A)$ as the union of all w-open subsets of A (e.g. $\emptyset$) and $c_w(A)$ as the intersection of all w-closed sets containing A (e.g. X). Quite recently, Al-Omari and Noiri [1, 2, 3, 7] has obtained several fundamental properties of weak structure spaces.

Let X be a nonempty set and $\mathcal{M} \subseteq \mathcal{P}(X)$. Then $\mathcal{M}$ is called a minimal structure on X if $\emptyset$, $X \in \mathcal{M}$ [8], in this case $(X, \mathcal{M})$ is called a minimal space. Each member of $\mathcal{M}$ is said to be m-open and the complement of an m-open set is said to be m-closed. Let $\mathcal{M}$, be a minimal structure on X and $A \subseteq X$. Maki, Umehara and Noiri [8] defined (as in the general case) $i_m(A)$ as the union of all m-open subsets of A and $c_m(A)$ as the intersection of all m-closed sets containing A.

We call a class $\mu \subseteq \mathcal{P}(X)$ a generalized topology [4] (briefly GT) if $\phi \in \mu$

and the arbitrary union of elements of μ belongs to μ. A set X with a GT μ on it is called a generalized topological space (briefly GTS) and is denoted by (X, μ). In this paper, We introduce some contra continuous functions in weak structure spaces such as contra $(\mathcal{M}, w)$-continuous functions, contra $(\alpha(m), w)$-continuous functions, contra $(\sigma(m), w)$-continuous functions, contra $(\pi(m), w)$-continuous functions and contra $(\beta(m), w)$-continuous functions. We investigate their characterization and relationships among such functions.

The following lemmas are useful in the sequel:

LEMMA 1.1 ([5]). *Let w be a WS on X and A, B subsets of X, then the following properties hold:*

1. $i_w(A) \subseteq A \subseteq c_w(A)$.
2. *If $A \subseteq B$ implies that $i_w(A) \subseteq i_w(B)$ and $c_w(A) \subseteq c_w(B)$.*
3. $i_w(i_w(A)) = i_w(A)$ *and* $c_w(c_w(A)) = c_w(A)$.
4. $i_w(X - A) = X - c_w(A)$ *and* $c_w(X - A) = X - i_w(A)$.

LEMMA 1.2 ([5]). *Let w be a WS on X and A a subset of X, then the following properties hold:*

1. $x \in i_w(A)$ *if and only if there is* $W \in w$ *such that* $x \in W \subseteq A$.
2. $x \in c_w(A)$ *if and only if* $W \cap A \neq \emptyset$ *whenever* $x \in W \in w$.
3. *If* $A \in w$, *then* $A = i_w(A)$ *and if A is w-closed, then* $A = c_w(A)$.

REMARK 1.3. *If* w *is a WS on X, then*

1. $i_w(\emptyset) = \emptyset$ *and* $c_w(X) = X$.
2. $i_w(X)$ *is the union of all* w*-open sets in X.*
3. $c_w(\emptyset)$ *is the intersection of all* w*-closed sets in X.*

THEOREM 1.4 ([1]). *For a WS space (X, w), the following properties are equivalent:*

1. $w = \mu$ *i.e. w is a generalized topology in the sense of Császár;*
2. $i_w(A)$ *is w-open for every subset A of X;*
3. $c_w(A)$ *is w-closed for every subset A of X.*

THEOREM 1.5 ([1]). *Let w be a WS on X and* $w^* = \{A \subset X : A = i_w(A)\}$. *Then, the following properties hold:*

1. w^* *is a GT containing w;*
2. w *is a GT if and only if* $w = w^*$.

2. Contra $(\mathcal{M}, w)$-continuity on weak structure spaces

DEFINITION 2.1. *Let $\mathcal{M}$ be minimal structure on X and w be weak structure on Y. A function $f : (X, \mathcal{M}) \to (Y, w)$ is said to be*

1. *contra $(\mathcal{M}, w)$-continuous if for each w-open set U in Y, $f^{-1}(U)$ is m-closed in X.*
2. *contra $(\mathcal{M}, w)$-continuous at some $x \in X$ if for each w-closed set V containing $f(x)$, there exists $U \in \mathcal{M}$ containing x such that $f(U) \subseteq V$.*

THEOREM 2.2. *Let $\mathcal{M}$ be minimal structure on X and w be weak structure on Y. For a function $f : (X, \mathcal{M}) \to (Y, w)$. The implications (1) $\Rightarrow$ (2) $\Rightarrow$ (3) $\Rightarrow$ (4) hold. If $\mathcal{M} = \mathcal{M}^*$, then the following statements are equivalent:*

1. *f is contra $(\mathcal{M}, w)$-continuous.*
2. *f is contra $(\mathcal{M}, w)$-continuous at any $x \in X$.*
3. *$f^{-1}(F) \subseteq i_m(f^{-1}(F))$ for any w-closed F of Y.*
4. *$c_m(f^{-1}(V)) \subseteq f^{-1}(V)$ for any w-open V of Y.*

Proof. (1)$\Rightarrow$ (2). Let $x \in X$ and V be w-closed set containing $f(x)$. By (1), $f^{-1}(V) \in \mathcal{M}$. Put $U = f^{-1}(V)$. We have U is m-open containing x and $f(U) \subseteq V$.
(2)$\Rightarrow$ (3). Let F be w-closed F of Y. For each $x \in f^{-1}(F)$, $f(x) \in F$. By (2), there exists $U \in \mathcal{M}$ containing x such that $f(U) \subseteq F$. Since $x \in U \subseteq f^{-1}(F)$, we have $x \in i_m(f^{-1}(F))$. This implies $f^{-1}(F) \subseteq i_m(f^{-1}(F))$.
(3)$\Rightarrow$ (4). Let $V \in w$. Then $Y - V$ is w-closed. By (3) and Lemma 1.1, $f^{-1}(Y - V) \subseteq i_m(f^{-1}(Y - V)) = i_m(X - f^{-1}(V)) = X - c_m(f^{-1}(V))$. Thus $c_m(f^{-1}(V)) \subseteq f^{-1}(V)$.
(4)$\Rightarrow$ (1). Let $V \in w$. By (4), we have $c_m(f^{-1}(V)) \subseteq f^{-1}(V)$ and hence $c_m(f^{-1}(V)) = f^{-1}(V)$. Since $\mathcal{M} = \mathcal{M}^*$, then $f^{-1}(V)$ is m-closed. Hence f is contra $(\mathcal{M}, w)$-continuous. □

The implication (2)$\Rightarrow$ (1) of Theorem 2.2 need not be true in general.

EXAMPLE 2.3. *Let $X = \{a, b, c\}$ and $\mathcal{M} = \{\phi, \{a\}, \{b\}, \{c\}, X\}$ be a minimal structure on X. Let $f : (X, \mathcal{M}) \to (X, \mathcal{M})$ be the identity function. Then f is contra $(\mathcal{M}, \mathcal{M})$-continuous at any $x \in X$ but not contra-$(\mathcal{M}, \mathcal{M})$-continuous.*

THEOREM 2.4. *Let $\mathcal{M}$ be minimal structure on X and w be weak structure on Y. For a function $f : (X, \mathcal{M}) \to (Y, w)$. The implications (1) $\Rightarrow$ (2) $\Rightarrow$ (3) hold. If $w = w^*$, then the following statements are equivalent:*

1. *$c_m(f^{-1}(i_w(B))) \subseteq f^{-1}(i_w(B))$ for any $B \subseteq Y$.*

2. $f^{-1}(c_w(B)) \subseteq i_m(f^{-1}(c_w(B)))$ *for any* $B \subseteq Y$.

3. $c_m(f^{-1}(V)) \subseteq f^{-1}(V)$ *for any* w*-open* V *of* Y.

Proof. (1)⇒ (2). Let $B \subseteq Y$. By (1), $c_m(f^{-1}(i_w(Y - B))) \subseteq f^{-1}(i_w(Y - B))$. By Lemma 1.1, $c_m(f^{-1}(i_w(Y - B))) = c_m(f^{-1}(Y - c_w(B))) = c_m(X - f^{-1}(c_w(B))) = X - i_m(f^{-1}(c_w(B)))$ and $X - i_m(f^{-1}(c_w(B))) \subseteq X - f^{-1}(c_w(B))$. Thus $f^{-1}(c_w(B)) \subseteq i_m(f^{-1}(c_w(B)))$.
(2)⇒ (3). Let $V \in w$. Then $Y - V$ is w-closed and hence $c_w(Y - V) = Y - V$. Now by (2), we have $f^{-1}(c_w(Y-V)) \subseteq i_m(f^{-1}(c_w(Y-V)))$ and hence $f^{-1}(Y - V) \subseteq i_m(f^{-1}(Y - V)) = X - c_m(f^{-1}(V))$. Then $c_m(f^{-1}(V)) \subseteq f^{-1}(V)$.
(3)⇒ (1). Let $B \subseteq Y$. Since $w = w^*$, then $i_w(B)$ is w-open set, by (3) $c_m(f^{-1}(i_w(B))) \subseteq f^{-1}(i_w(B))$. □

Definition 2.5 ([1]). *Let* (X, w) *be a* WS *space. Then the weak kernel of* $A \subseteq X$ *is denoted by* w-*ker*(A) *and defined as* w-*ker*$(A) = \cap\{G \in w : A \subseteq G\}$.

Lemma 2.6 ([1]). *Let* A *and* B *be two subsets of a* WS *space* (X, w). *Then the following properties hold:*

1. $x \in w$-*ker*(A) *if and only if* $A \cap F \neq \phi$ *for any* w*-closed* F *containing* x.

2. $A \subseteq w$-*ker*(A) *and* $A = w$-*ker*(A) *if* $A \in w$.

3. *If* $A \subseteq B$, *then* w-*ker*$(A) \subseteq w$-*ker*(B).

Lemma 2.7. *Let* A *be a subset of a* WS *space* (X, w). *Then* w-*ker*$(A) = w$-*ker*$(w$-*ker*$(A))$

Proof. By Lemma 2.6, we have w-ker$(A) \subseteq w$-ker$(w$-ker $(A))$. Conversely, if $x \notin w$-ker(A) there exists F which is w-closed such that $x \in F$ and $F \cap A = \phi$. Since $X - F \in w$ and $A \subseteq X - F$, and since w-ker(A) is the intersection of all w-open sets containing A, we have w-ker$(A) \subseteq X - F$ so that $F \cap w$-ker$(A) = \phi$. Since $x \in F$, we have that $x \notin w$-ker$(w$-ker $(A))$. Thus w-ker$(w$-ker $(A)) \subseteq w$-ker(A). □

Theorem 2.8. *Let* $\mathcal{M}$ *be minimal structure on* X *and* w *be weak structure on* Y. *For a function* $f : (X, \mathcal{M}) \to (Y, w)$. *The implications (1)* ⇒ *(2)* ⇒ *(3) hold. If* $\mathcal{M} = \mathcal{M}^*$, *then the following statements are equivalent:*

1. f *is contra* $(\mathcal{M}, w)$*-continuous;*

2. $f(c_m(A)) \subseteq w$-*ker*$(f(A))$ *for any* $A \subseteq X$;

3. $c_m(f^{-1}(B)) \subseteq f^{-1}(w$-*ker*$(B))$ *for any* $B \subseteq Y$.

Proof. (1) $\Rightarrow$ (2). Let $A \subseteq X$. Suppose that $f(c_m(A)) - w\text{-ker}(f(A)) \neq \phi$. Pick $y \in f(c_m(A)) - w\text{-ker}(f(A))$. By $y \notin w\text{-ker}(f(A))$, there exists w-closed set F containing y such that $f(A) \cap F = \phi$. Then $A \cap f^{-1}(F) = \phi$ and $c_m(A) \cap f^{-1}(F) = \phi$, since $f^{-1}(F) \in m$. This implies that $f(c_m(A)) \cap F = \phi$ and $y \notin f(c_m(A))$. Thus $f(c_m(A)) \subseteq w\text{-ker}(f(A))$.
(2) $\Rightarrow$ (3). Let $B \subseteq Y$. By (2), $f(c_m(f^{-1}(B))) \subseteq w\text{-ker}(f(f^{-1}(B))) \subseteq w\text{-}ker(B)$. Thus $c_m(f^{-1}(B)) \subseteq f^{-1}(w\text{-ker}(B))$.
(3) $\Rightarrow$ (1). Let $B \in w$. By (3) $c_m(f^{-1}(B)) \subseteq f^{-1}(w\text{-ker}(B))$. By Lemma 2.6, $B = w\text{-ker}(B))$. Thus $c_m(f^{-1}(B)) \subseteq f^{-1}(B)$. Since $\mathcal{M} = \mathcal{M}^*$ implies that $f^{-1}(B)$ is m-closed. Hence f is contra $(\mathcal{M}, w)$-continuous. □

Definition 2.9. *Let (X, w) be a WS space. X is called w-connected, if there are no nonempty disjoint w-open subsets U, V of X such that $U \cup V = X$.*

Lemma 2.10. *Let (X, w) be a WS space. If U, V are nonempty disjoint w-open subsets of X and $U \cup V = X$, then U and V are w-closed.*

Theorem 2.11. *Let $f : (X, \mathcal{M}) \to (Y, w)$ be a contra $(\mathcal{M}, w)$-continuous surjection. If X is m-connected, then Y is w-connected.*

Proof. Let $f : (X, \mathcal{M}) \to (Y, w)$ be a contra $(\mathcal{M}, w)$-continuous surjection and let X be m-connected. Suppose Y is not w-connected. Then there exists nonempty disjoint w-open subsets V_1 and V_2 of Y such that $V_1 \cup V_2 = Y$. By Lemma 2.10, V_1 and V_2 are w-closed. Since f is contra $(\mathcal{M}, w)$-continuous, then $f^{-1}(V_1)$, $f^{-1}(V_2) \in \mathcal{M}$. Note that $f^{-1}(V_1) \cap f^{-1}(V_2) \neq \phi$ and $f^{-1}(V_1) \cup f^{-1}(V_2) = X$. Then X is not m-connected, contradiction. Thus Y is w-connected. □

Definition 2.12. *A WS space (X, w) is said to be* strongly w-closed *if every cover of X by w-closed sets of (X, w) has a finite subcover.*

Definition 2.13. *A minimal space $(X, \mathcal{M})$ is said to be m-compact if every m-open cover of X has a finite subcover.*

Theorem 2.14. *Let $f : (X, \mathcal{M}) \to (Y, w)$ be a contra-$(\mathcal{M}, w)$-continuous surjection. If $(X, \mathcal{M})$ is m-compact, then (Y, w) is strongly w-closed.*

Proof. Let $(X, \mathcal{M})$ be m-compact and $\{V_\alpha : \alpha \in \Delta\}$ any cover of Y by w-closed sets of (Y, w). Since f is contra-$(\mathcal{M}, w)$-continuous, the family $\{f^{-1}(V_\alpha) : \alpha \in \Delta\}$ is a m-open cover of X. Since $(X, \mathcal{M})$ is m-compact, there exists a finite subset Δ_0 of Δ such that $X = \cup\{f^{-1}(V_\alpha) : \alpha \in \Delta_0\}$. Therefore, $Y = f(X) = \cup\{V_\alpha : \alpha \in \Delta_0\}$. This shows that (Y, w) is strongly w-clsoed. □

3. Contra continuity on weak structure spaces

Definition 3.1 ([10]). *Let $(X, \mathcal{M})$ be a minimal structure space and $A \subseteq X$. Then A is said to be*

1. *m-semi-open if $A \subseteq c_m(i_m(A))$,*
2. *m-preopen if $A \subseteq i_m(c_m(A))$,*
3. *m-α-open if $A \subseteq i_m(c_m(i_m(A)))$,*
4. *m-β-open if $A \subseteq c_m(i_m(c_m(A)))$,*
5. *mr-open if $A = i_m(c_m(A))$.*

The complement of m-semi-open (resp. m-preopen, m-α-open, m-β-open, mr-open) is said to be m-semi-closed (resp. m-preclosed, m-α-closed, m-β-closed, wr-closed). Let us denote by $\sigma(m)$ (resp. $\pi(m)$, $\alpha(m)$, $\beta(m)$) the class of all m-semi-open (resp. m-preopen, m-α-open, m-β-open) sets of $(X, \mathcal{M})$.

Definition 3.2. *Let $\mathcal{M}$ be minimal structure on X and w be weak structure on Y. A function $f : (X, \mathcal{M}) \to (Y, w)$ is said to be*

1. *contra $(\alpha(m), w)$-continuous if for each w-open set U in Y, $f^{-1}(U)$ is m-α-closed in X.*
2. *contra $(\sigma(m), w)$-continuous if for each w-open set U in Y, $f^{-1}(U)$ is m-σ-closed in X.*
3. *contra $(\pi(m), w)$-continuous if for each w-open set U in Y, $f^{-1}(U)$ is m-π-closed in X.*
4. *contra $(\beta(m), w)$-continuous if for each w-open set U in Y, $f^{-1}(U)$ is m-β-closed in X.*
5. *contra $(\sigma(m), w^*)$-continuous if for each w^*-open set U in Y, $f^{-1}(U)$ is m-σ-closed in X.*
6. *contra $(\pi(m), w^*)$-continuous if for each w^*-open set U in Y, $f^{-1}(U)$ is m-π-closed in X.*

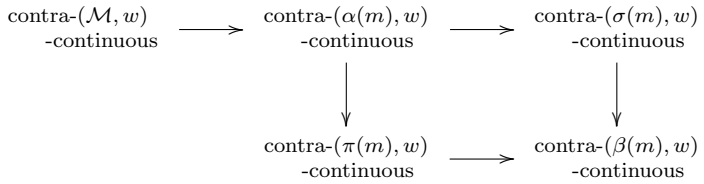
Lemma 3.3 ([5]). *For a WS w on X, the following relations hold:*

1. *$w \subseteq \alpha(w) \subseteq \sigma(w) \subseteq \beta(w)$.*
2. *$w \subseteq \alpha(w) \subseteq \pi(w) \subseteq \beta(w)$.*

Theorem 3.4 ([5]). *If w is a WS, each of the structures $\alpha(w)$, $\sigma(w)$, $\pi(w)$ and $\beta(w)$ is a generalized topology.*

For several functions defined above, we have the following implications.

DIAGRAM



The reverse implication may be not true in general and this can be clearly seen from the following examples.

EXAMPLE 3.5. *Let $X = \{a, b, c, d\}$ and $\mathcal{M} = \{\phi, \{a\}, \{b\}, \{a, b, c\}, X\}$ be a minimal structure on X. Define $f : (X, \mathcal{M}) \to (X, \mathcal{M})$ as follows: $f(a) = f(b) = d$ and $f(c) = f(d) = a$. Then $f^{-1}(\{a\}) = \{c, d\}$, $f^{-1}(\{b\}) = \phi$ and $f^{-1}(\{a, b, c\}) = \{c, d\}$. We have f is contra-$(\alpha(m), \mathcal{M})$-continuous but not contra-$(\mathcal{M}, \mathcal{M})$-continuous.*

EXAMPLE 3.6. *Let $X = Y = \{a, b, c\}$, $\mathcal{M} = \{\phi, \{a\}, \{b\}, X\}$ be a minimal structure on X and $w = \{\phi, \{a\}, \{b\}\}$ a WS on Y. Define $f : (X, \mathcal{M}) \to (Y, w)$ be the identity function. We have f is contra-$(\sigma(m), w)$-continuous but not contra-$(\alpha(m), w)$-continuous.*

EXAMPLE 3.7. *Let $X = Y = \{a, b, c\}$, $\mathcal{M} = \{\phi, \{a\}, \{b\}, X\}$ be a minimal structure on X and $w = \{\phi, \{a, c\}, \{b\}\}$ a WS on Y. Define $f : (X, \mathcal{M}) \to (Y, w)$ as follows: $f(a) = a$, $f(b) = c$ and $f(c) = c$. Then $f^{-1}(\{a, b\}) = \{a\}$ and $f^{-1}(\{b\}) = \phi$. We have f is contra-$(\beta(m), w)$-continuous but not contra-$(\pi(m), w)$-continuous.*

EXAMPLE 3.8. *Let $X = Y = \{a, b, c\}$, $\mathcal{M} = \{\phi, \{a, c\}, \{b, c\}, X\}$ be a minimal structure on X and $w = \{\phi, \{a, c\}\}$ a WS on Y. Define $f : (X, \mathcal{M}) \to (Y, w)$ as follows: $f(a) = f(b) = a$ and $f(c) = b$. Then $f^{-1}(\{a, c\}) = \{a, b\}$. We have f is contra-$(\pi(m), w)$-continuous but not contra-$(\sigma(m), w)$-continuous.*

THEOREM 3.9. *Let $\mathcal{M}$ be a minimal structure on X and w be weak structures on Y. A function $f : (X, \mathcal{M}) \to (Y, w)$ is contra-$(\alpha(m), w)$-continuous if and only if it is both contra-$(\pi(m), w)$-continuous and contra-$(\sigma(m), w)$-continuous.*

Proof. Necessity. It is clear from the above diagram.
Sufficiency. Follows from the fact that $\alpha(w) = \pi(w) \cap \sigma(w)$. □

DEFINITION 3.10. *Let $\mathcal{M}$ be a minimal structure on X and w be weak structures on Y. A function $f : (X, \mathcal{M}) \to (Y, w)$ is said to be*

1. *$(\sigma(m), w)$-continuous if $f^{-1}(V)$ is m-semi-open in X for each w-open set V of Y,*
2. *$(\pi(m), w)$-continuous if $f^{-1}(V)$ is m-preopen in X for each w-open set V of Y.*

LEMMA 3.11. *For a subset A of a WS space (X, w), the following properties are equivalent:*

1. *A is wr-closed;*
2. *A is w-preclosed and w-semi-open;*
3. *A is w-α-closed and w-β-open.*

Proof. (1)⇒ (2). Let A be wr-closed. Then $A = c_w(i_w(A))$ and A is w-preclosed and w-semi-open.
(2)⇒ (3). Let A be w-preclosed and w-semi-open. Then $A \subseteq c_w(i_w(A))$ and $c_w(i_w(A)) \subseteq A$. Therefore, we have $c_w(A) = c_w(i_w(A))$ and hence $c_w(i_w(c_w(A)))$ $= c_w(i_w(c_w(i_w(A)))) = c_w(i_w(A)) \subseteq A$. This shows that A is w-α-closed. Since $\sigma(w) \subseteq \beta(w)$, it is obvious that A is w-β-open.
(3)⇒ (1). Let A be w-α-closed and w-β-open. Then $A = c_w(i_w(c_w(A)))$ and hence $c_w(i_w(A)) = c_w(i_w(c_w(i_w(c_w(A))))) = c_w(i_w(c_w(A))) = A$. Therefore, A is wr-closed. □

DEFINITION 3.12. *Let $\mathcal{M}$ be a minimal structure on X and w be weak structures on Y. A function $f : (X, \mathcal{M}) \to (Y, w)$ is said to be RC-$(\mathcal{M}, w)$-continuous if $f^{-1}(V)$ is mr-closed in X for each w-open set of Y.*

As a consequence of Lemma 3.11, we have the following result:

THEOREM 3.13. *Let $\mathcal{M}$ be a minimal structure on X and w be weak structures on Y. For a function $f : (X, \mathcal{M}) \to (Y, w)$, the following statements are equivalent:*

1. *f is RC-$(\mathcal{M}, w)$-continuous;*
2. *f is contra-$(\pi(m), w)$-continuous and $(\sigma(m), w)$-continuous;*
3. *f is contra-$(\alpha(m), w)$-continuous and $(\beta(m), w)$-continuous.*

Let $\mathcal{M}$ be a minimal structure on X or w be a weak structures on X and $A \subseteq X$. The m-α-closure (resp. m-semi-closure, m-preclosure, m-β-closure, w^*-closure) of a subset A of X, denoted by $c_\alpha(A)$ (resp. $c_\sigma(A)$, $c_\pi(A)$, $c_\beta(A)$, $c_{w^*}(A)$), is the intersection of m-α-closed (resp. m-semi-closed, m-preclosed, m-β-closed, w^*-closed) sets including A. The m-α-interior (resp. m-semi-interior, m-preinterior, m-β-interior, w^*-interior) of a subset A of X, denoted by $i_\alpha(A)$ (resp. $i_\sigma(A)$, $i_\pi(A)$, $i_\beta(A)$, $i_{w^*}(A)$), is the union of m-α-open (resp. m-semi-open, m-preopen, m-β-open, w^*-open) sets contained in A.

THEOREM 3.14. *Let $\mathcal{M}$ be a minimal structure on X and w be a weak structures on Y. For a function $f : (X, \mathcal{M}) \to (Y, w)$, the following properties are equivalent:*

1. *f is contra $(\pi(m), w^*)$-continuous;*
2. *$f^{-1}(A)$ is m-preopen set in X for every w^*-closed set A in Y;*
3. *$f^{-1}(A) \subseteq i_m(c_m(f^{-1}(c_{w^*}(A))))$ for every subset A in Y;*
4. *$c_m(i_m(f^{-1}(i_{w^*}(A)))) \subseteq f^{-1}(A)$ for every subset A in Y;*
5. *$A \subseteq i_m(c_m(f^{-1}(c_{w^*}(f(A)))))$ for every subset A in X.*

Proof. (1) $\Leftrightarrow$ (2). It is obvious.
(2) $\Rightarrow$ (3). Let $A \subseteq Y$. Then $c_{w^*}(A)$ is w^*-closed set in Y. By (2) implies that $f^{-1}(c_{w^*}(A))$ is m-preopen set in X. Therefore, $f^{-1}(c_{w^*}(A)) \subseteq i_m(c_m(f^{-1}(c_{w^*}(A))))$. Hence $f^{-1}(A) \subseteq i_m(c_m(f^{-1}(c_{w^*}(A))))$.
(3) $\Leftrightarrow$ (4). It is obvious.
(3) $\Rightarrow$ (5). Let $A \subseteq X$. Then $f(A) \subseteq Y$. By (3) implies that $f^{-1}(f(A)) \subseteq i_m(c_m(f^{-1}(c_{w^*}(f(A))))).$ Therefore, $A \subseteq f^{-1}(f(A)) \subseteq i_m(c_m(f^{-1}(c_{w^*}(f(A))))).$
(5) $\Rightarrow$ (2). Let A be w^*-closed in Y. Then $f^{-1}(A) \subseteq X$. By hypothesis

$$\begin{aligned} f^{-1}(A) &\subseteq i_m(c_m(f^{-1}(c_{w^*}(f(f^{-1}(A)))))) \\ &\subseteq i_m(c_m(f^{-1}(c_{w^*}(A)))) \\ &= i_m(c_m(f^{-1}(A))) . \end{aligned}$$

Hence $f^{-1}(A)$ is m-preopen set in X.

□

REMARK 3.15. *Since every w-open set is w^*-open set in Y. Then every contra $(\pi(m), w^*)$-continuous is contra $(\pi(m), w)$-continuous.*

THEOREM 3.16. *Let $\mathcal{M}$ be a minimal structure on X and w be weak structures on Y. For a function $f : (X, \mathcal{M}) \to (Y, w)$, the following properties are equivalent:*

1. *f is contra $(\sigma(m), w^*)$-continuous;*
2. *$f^{-1}(A)$ is m-semi-open set in X for every w^*-closed set A in Y;*
3. *$f^{-1}(A) \subseteq c_m(i_m(f^{-1}(c_{w^*}(A))))$ for every subset A in Y;*
4. *$i_m(c_m(f^{-1}(i_{w^*}(A)))) \subseteq f^{-1}(A)$ for every subset A in Y;*
5. *$A \subseteq c_m(i_m(f^{-1}(c_{w^*}(f(A)))))$ for every subset A in X.*

Proof. (1) $\Leftrightarrow$ (2). It is obvious.
(2) $\Rightarrow$ (3). Let $A \subseteq Y$. Then $c_{w^*}(A)$ is w^*-closed set in Y. By (2) implies that $f^{-1}(c_{w^*}(A))$ is m-semi-open set in X. Therefore, $f^{-1}(c_{w^*}(A)) \subseteq c_m(i_m(f^{-1}(c_{w^*}(A))))$. Hence $f^{-1}(A) \subseteq f^{-1}(c_{w^*}(A)) \subseteq c_m(i_m(f^{-1}(c_{w^*}(A))))$.
(3) $\Leftrightarrow$ (4). It is obvious by taking complement.
(3) $\Rightarrow$ (5). Let $A \subseteq X$. Then $f(A) \subseteq Y$. By (3) implies that $f^{-1}(f(A)) \subseteq c_m(i_m(f^{-1}(c_{w^*}(f(A)))))$. Therefore, $A \subseteq f^{-1}(f(A)) \subseteq c_m(i_m(f^{-1}(c_{w^*}(f(A)))))$.
(5) $\Rightarrow$ (2). Let A be w^*-closed in Y. Then $f^{-1}(A) \subseteq X$. By hypothesis

$$\begin{aligned} f^{-1}(A) \subseteq & c_m(i_m(f^{-1}(c_{w^*}(f(f^{-1}(A)))))) \\ \subseteq & c_m(i_m(f^{-1}(c_{w^*}(A)))) \\ = & c_m(i_m(f^{-1}(A))) . \end{aligned}$$

Hence $f^{-1}(A)$ is m-semi-open set in X.

□

REMARK 3.17. *Since every w-open set is w^*-open set in Y. Then every contra (σ, w^*)-continuous is contra (σ, w)-continuous.*

THEOREM 3.18. *Let $\mathcal{M}$ be a minimal structure on X and w be weak structures on Y. A function $f : (X, \mathcal{M}) \to (Y, w)$ is contra $(\beta(m), w^*)$-continuous if and only if $f^{-1}(c_\beta(B)) \subseteq i_\beta(f^{-1}(c_{w^*}(B))$ for each subset B in Y.*

Proof. Necessity. Let $B \subseteq Y$. Then $c_{w^*}(B)$ is w^*-closed in Y. By hypothesis, $f^{-1}(c_{w^*}(B)) \in \beta(m)$ and since $w^* \subseteq \beta(w)$. Therefore, $f^{-1}(c_\beta(B)) \subseteq f^{-1}(c_{w^*}(B)) = i_\beta(f^{-1}(c_{w^*}(B))$. Hence $f^{-1}(c_\beta(B)) \subseteq i_\beta(f^{-1}(c_{w^*}(B))$.
Sufficiency. Let $B \subseteq Y$ be w^*-closed. Then $c_{w^*}(B) = B$. By hypothesis, $f^{-1}(c_\beta(B)) \subseteq i_\beta(f^{-1}(c_{w^*}(B))) = i_\beta(f^{-1}(B)$. Now $f^{-1}(B) \subseteq f^{-1}(c_\beta(B)) \subseteq i_\beta(f^{-1}(B)) \subseteq f^{-1}(B)$. This implies that $i_\beta(f^{-1}(B)) = f^{-1}(B)$ and by Theorem 3.4. Hence $f^{-1}(B) \in \beta(m)$ and hence f is contra $(\beta(m), w^*)$-continuous.

□

REMARK 3.19. *Since every w-open set is w^*-open set in Y. Then every contra $(\beta(m), w^*)$-continuous is contra $(\beta(m), w)$-continuous.*

THEOREM 3.20. *Let $\mathcal{M}$ be a minimal structure on X and w be weak structures on Y. A function $f : (X, \mathcal{M}) \to (Y, w)$ is contra $(\alpha(m), w^*)$-continuous if and only if $f^{-1}(c_\alpha(B)) \subseteq i_\alpha(f^{-1}(c_{w^*}(B))$ for each subset B in Y.*

Proof. Similar as in Theorem 3.18. □

THEOREM 3.21. *Let $\mathcal{M}$ be a minimal structure on X and w be weak structures on Y. For a function $f : (X, \mathcal{M}) \to (Y, w)$. Suppose that one of the following conditions holds:*

1. $f^{-1}(c_w(B)) \subseteq i_m(c_\beta(f^{-1}(B)))$ *for each subset* B *in* Y;
2. $c_m(i_\beta(f^{-1}(B))) \subseteq f^{-1}(i_w(B))$ *for each subset* B *in* Y;
3. $f(c_m(i_\beta(A))) \subseteq i_w(f(A))$ *for each subset* A *in* X;
4. $f(c_m(A)) \subseteq i_w(f(A))$ *for each* m-β-*open set* A *in* X.

Then f *is contra* $(\beta(m), w)$-*continuous.*

Proof. (1) $\Rightarrow$ (2). It is obvious by taking complement.
(2) $\Rightarrow$ (3). Let $A \subseteq X$, then $f(A) \subseteq Y$. By (2) implies that $c_m(i_\beta(f^{-1}(f(A)))) \subseteq f^{-1}(i_w(f(A)))$. That is $c_m(i_\beta(A)) \subseteq c_m(i_\beta(f^{-1}(f(A)))) \subseteq f^{-1}(i_w(f(A)))$. Hence $f(c_m(i_\beta(A))) \subseteq f(f^{-1}(i_w(f(A)))) \subseteq i_w(f(A))$.
(3) $\Rightarrow$ (4). Let $A \subseteq X$ be m-β-open. Then $f(c_m(i_\beta(A))) \subseteq i_w(f(A))$. That is $f(c_w(A)) = f(c_m(i_\beta(A))) \subseteq i_w(f(A))$, since $i_\beta(A) = A$. Hence $f(c_m(A)) \subseteq i_w(f(A))$.
Suppose (4) holds: Let $A \subseteq Y$ be w-open. Then $f^{-1}(A) \subseteq X$ and $i_\beta(f^{-1}(A))$ is m-β-open in X, by Theorem 3.4. By (4) implies that $f(c_m(i_\beta(f^{-1}(A)))) \subseteq i_w(f(i_\beta(f^{-1}(A)))) \subseteq i_w(f(f^{-1}(A))) \subseteq i_w(A) = A$. Now $c_m(i_\beta(f^{-1}(A))) \subseteq f^{-1}(f(c_m(i_\beta(f^{-1}(A))))) \subseteq f^{-1}(A)$. We have $c_m(i_m(f^{-1}(A))) \subseteq f^{-1}(A)$. Therefore, $f^{-1}(A)$ is a m-preclosed set and hence a m-β-closed set. Thus f is contra $(\beta(m), w)$-continuous. □

THEOREM 3.22. *Let* $\mathcal{M}$ *be a minimal structure on* X *and* w *be weak structures on* Y. *For a function* $f : (X, \mathcal{M}) \to (Y, w)$. *Suppose that one of the following conditions holds:*

1. $f^{-1}(c_w(B)) \subseteq i_m(c_\alpha(f^{-1}(B)))$ *for each subset* B *in* Y;
2. $c_m(i_\alpha(f^{-1}(B))) \subseteq f^{-1}(i_w(B))$ *for each subset* B *in* Y;
3. $f(c_m(i_\alpha(A))) \subseteq i_w(f(A))$ *for each subset* A *in* X;
4. $f(c_m(A)) \subseteq i_w(f(A))$ *for each* m-α-*open set* A *in* X.

Then f *is contra* $(\alpha(m), w)$-*continuous.*

Proof. Similar as in Theorem 3.21. □

THEOREM 3.23. *Let* $\mathcal{M}$ *be a minimal structure on* X *and* w *be weak structures on* Y. *For a function* $f : (X, \mathcal{M}) \to (Y, w)$. *Suppose that one of the following conditions holds:*

1. $f(c_\beta(A)) \subseteq i_w(f(A))$ *for each subset* A *in* X;
2. $c_\beta(f^{-1}(B))) \subseteq f^{-1}(i_w(B))$ *for each subset* B *in* Y;

3. $f^{-1}(c_w(B)) \subseteq i_\beta(f^{-1}(B)))$ *for each subset* B *in* Y.

Then f *is contra* $(\beta(m), w)$*-continuous*

Proof. (1) $\Rightarrow$ (2). Let $B \subseteq Y$. Then $f^{-1}(B) \subseteq X$. By (1) implies that $f(c_\beta(f^{-1}(B))) \subseteq i_w(f(f^{-1}(B))) \subseteq i_w(B)$. Therefore$f^{-1}(f(c_\beta(f^{-1}(B)))) \subseteq f^{-1}(i_w(B))$. So that $c_\beta(f^{-1}(B)) \subseteq f^{-1}(f(c_\beta(f^{-1}(B)))) \subseteq f^{-1}(i_w(B))$. Hence $c_\beta(f^{-1}(B)) \subseteq f^{-1}(i_w(B))$.
(2) $\Rightarrow$ (3). It is obvious by taking complement in (2).
Suppose (3) holds: Let $B \subseteq Y$ be w-closed. Then, by hypothesis, $f^{-1}(c_w(B)) \subseteq i_\beta(f^{-1}(B)))$. That is $f^{-1}(B) = f^{-1}(c_w(B)) \subseteq i_\beta(f^{-1}(B)) \subseteq f^{-1}(B)$ and by Theorem 3.4. Therefore, $f^{-1}(B)$ is m-β-open in X. Hence f is contra $(\beta(m), w)$-continuous.

□

THEOREM 3.24. *Let* $\mathcal{M}$ *be a minimal structure on* X *and* w *be weak structures on* Y. *For a function* $f : (X, \mathcal{M}) \to (Y, w)$. *Suppose that one of the following conditions holds:*

1. $f(c_\alpha(A)) \subseteq i_w(f(A))$ *for each subset* A *in* X;
2. $c_\alpha(f^{-1}(B))) \subseteq f^{-1}(i_w(B))$ *for each subset* B *in* Y;
3. $f^{-1}(c_w(B)) \subseteq i_\alpha(f^{-1}(B)))$ *for each subset* B *in* Y.

Then f *is contra* $(\alpha(m), w)$*-continuous.*

Proof. Similar as in Theorem 3.23. □

THEOREM 3.25. *Let* $\mathcal{M}$ *be a minimal structure on* X *and* w *be weak structures on* Y. *A function* $f : (X, \mathcal{M}) \to (Y, w)$ *is contra* $(\beta(m), w)$*-continuous if* $c_w(f(A)) \subseteq f(i_\beta(A))$ *for each subset* A *of* X *and* f *is bijective.*

Proof. Let $B \subseteq Y$ be w-closed. Then $f^{-1}(B) \subseteq X$. By hypothesis $c_w(f(f^{-1}(B))) \subseteq f(i_\beta(f^{-1}(B)))$. Now $B = c_w(B) = c_w(f(f^{-1}(B))) \subseteq f(i_\beta(f^{-1}(B)))$. Therefore, $f^{-1}(B) \subseteq f^{-1}(f(i_\beta(f^{-1}(B)))) = i_\beta(f^{-1}(B)) \subseteq f^{-1}(B)$ and by Theorem 3.4. Hence $f^{-1}(B) \in \beta(m)$ and hence f is contra $(\beta(m), w)$-continuous.

□

THEOREM 3.26. *Let* $\mathcal{M}$ *be a minimal structure on* X *and* w *be weak structures on* Y. *Let* $f : (X, \mathcal{M}) \to (Y, w)$ *be a contra* $(\beta(m), w)$*-continuous. Then the following properties hold:*

1. $c_\beta(f^{-1}(B)) \subseteq f^{-1}(i_w(c_\beta(B)))$ *for each* w*-open set* B *in* Y.
2. $f^{-1}(c_w(i_\beta(B))) \subseteq i_\beta(f^{-1}(B))$ *for each* w*-closed set* B *in* Y.

Proof. (1). Let $B \subseteq Y$ be w-open. By hypothesis, $f^{-1}(B)$ is m-β-closed in X. Then $c_\beta(f^{-1}(B)) = f^{-1}(B) = f^{-1}(i_w(B)) \subseteq f^{-1}(i_w(c_\beta(B)))$. Hence
$c_\beta(f^{-1}(B)) \subseteq f^{-1}(i_w(c_\beta(B)))$.
(2). It is obvious by taking complement in (1). □

THEOREM 3.27. *Let $\mathcal{M}$ be a minimal structure on X and w be weak structures on Y. For a function $f : (X, \mathcal{M}) \to (Y, w)$. The following conditions are equivalent:*

1. *f is contra $(\beta(m), w)$-continuous;*
2. *for each $x \in X$ and each w-closed set B containing $f(x)$, there exists $A \in \beta(m)$ and $x \in A$ such that $A \subseteq f^{-1}(B)$;*
3. *for each $x \in X$ and each w-closed set B containing $f(x)$, there exists $A \in \beta(m)$ and $x \in A$ such that $f(A) \subseteq B$.*

Proof. (1) ⇒ (2). Let $B \subseteq Y$ be w-closed and $f(x) \in B$. By hypothesis $f^{-1}(B) \in \beta(m)$. Therefore, $i_\beta(f^{-1}(B)) = f^{-1}(B)$. Put $A = i_\beta(f^{-1}(B))$. Then $A \in \beta(w)$ and $A \subseteq f^{-1}(B)$.
(2) ⇒ (3). Let $B \subseteq Y$ be w-closed and $f(x) \in B$. By hypothesis there exists $A \in \beta(m)$ and $x \in A$ such that $A \subseteq f^{-1}(B)$. Therefore, $f(A) \subseteq f(f^{-1}(B)) \subseteq B$. Thus $f(A) \subseteq B$.
(3) ⇒ (1). Let B be w-closed in Y. Let $x \in X$ and $f(x) \in B$. By hypothesis there exists $A \in \beta(m)$ and $x \in A$ such that $f(A) \subseteq B$. This implies that $x \in A \subseteq f^{-1}(f(A)) \subseteq f^{-1}(B)$. That is $x \in f^{-1}(B)$. Since $A \in \beta(m)$, $A = i_\beta(A) \subseteq i_\beta(f^{-1}(B))$. Hence $x \in i_\beta(f^{-1}(B))$. Therefore, $f^{-1}(B) = \cup\{x : x \in f^{-1}(B)\} \subseteq i_\beta(f^{-1}(B)) \subseteq f^{-1}(B)$. Thus $i_\beta(f^{-1}(B)) = f^{-1}(B)$ and by Theorem 3.4 we have $f^{-1}(B) \in \beta(m)$. Hence f is contra $(\beta(m), w)$-continuous. □

THEOREM 3.28. *Let $\mathcal{M}$ be a minimal structure on X and w be weak structures on Y. For a function $f : (X, \mathcal{M}) \to (Y, w)$. The following conditions are equivalent:*

1. *f is contra $(\pi(m), w)$-continuous;*
2. *$f^{-1}(A) \in \pi(m)$ for every w-closed set A in Y;*
3. *for each $x \in X$ and each w-closed set A containing $f(x)$, there exists $B \in \pi(m)$ containing x such that $f(B) \subseteq A$;*
4. *$f(c_\pi(A)) \subseteq w\text{-}ker(f(A))$ for every subset A of X;*
5. *$c_\pi(f^{-1}(B)) \subseteq f^{-1}(w\text{-}ker(B))$ for every subset B of Y.*

Proof. (1) $\Leftrightarrow$ (2). It is obvious.
(2) $\Rightarrow$ (3). Let $x \in X$ and A be w-closed set containing $f(x)$. By hypothesis, $f^{-1}(A) \in \pi(m)$. Now put $B = f^{-1}(A)$, then $f(B) = f(f^{-1}(A)) \subseteq A$. Thus $f(B) \subseteq A$.
(3) $\Rightarrow$ (2). Let A be a w-closed set in Y and $x \in f^{-1}(A)$. Then $f(x) \in A$. By (3) there exists $B_x \in \pi(m)$ containing x such that $f(B_x) \subseteq A$. This implies that $B_x \subseteq f^{-1}(f(B_x)) \subseteq f^{-1}(A)$. Now $f^{-1}(A) = \cup\{B_x : x \in f^{-1}(A)\}$ and since $\pi(m)$ is a generalized topology, $f^{-1}(A) \in \pi(m)$.
(2) $\Rightarrow$ (4). Let A be any subset of X. Suppose $y \notin w\text{-ker}(f(A))$, then by Lemma 2.6 there exists w-closed set B containing y such that $f(A) \cap B = \phi$. thus we have $A \cap f^{-1}(B) = \phi$ and $c_\pi(A) \cap f^{-1}(B) = \phi$. Therefore, $f(c_\pi(A)) \cap B = \phi$ and $y \notin f(c_\pi(A))$. This implies $f(c_\pi(A)) \subseteq w\text{-ker}(f(A))$.
(4) $\Rightarrow$ (5). Let B be any subset of Y. By (4) and Lemma 2.6, we have $f(c_\pi(f^{-1}(B))) \subseteq w\text{-ker}(f(f^{-1}(B))) \subseteq w\text{-ker}(B)$ and $c_\pi(f^{-1}(B)) \subseteq f^{-1}(w\text{-ker}(B))$.
(5) $\Rightarrow$ (1). Let B be any w-open set in Y. By Lemma 2.6, we have $c_\pi(f^{-1}(B)) \subseteq f^{-1}(w\text{-ker}(B)) = f^{-1}(B)$ and $c_\pi(f^{-1}(B)) = f^{-1}(B)$ and by Theorem 3.4. Hence $f^{-1}(B) \in \pi(m)$. □

Acknowledgements

The author wishes to thank the referees for their useful comments and suggestions.

References

[1] A. Al-Omari and T. Noiri, *(w,k)-continuity and weak (w,k)-continuity in weak structure spaces due to Császár*, submitted.
[2] A. Al-Omari and T. Noiri, *Some weak separation axioms in a weak structure space due to Császár*, Analele Universităţii Oradea **20** (2013), 105–111.
[3] A. Al-Omari and T. Noiri, *Λ_w-sets and $\vee_w$-sets in weak structures*, Annales Univ. Sci. Budapest. Sect. Math., in press.
[4] Á. Császár, *Generalized topology, generalized continuity*, Acta Math. Hungar. **96** (2002), 351–357.
[5] Á. Császár, *Weak structures*, Acta Math. Hungar. **131** (2011), 193–195.
[6] J. Dontchev, *Contra-continuous functions and strongly S-closed spaces*, Internat. J. Math. Math. Sci. **19** (1996), 303–310.
[7] T. Noiri and A. Al-Omari, *Characterizations of* w-T_0 *and* w-R_0 *via the topology generated by* Λ_w , Questions and Answers in General Topology, in press.
[8] H. Maki, J. Umehara and T. Noiri, *Every topological space is pre-$T_{\frac{1}{2}}$*, Mem. Fac. Sci. Kochi. Univ. Ser. A Math. **17** (1996), 33–42.
[9] V. Popa and T. Noiri, *On M-continuous functions*, Anal. Univ. "Dunarea de Jos" Galati, Ser. Mat. Fiz. Mec. Teor., Fasc. II, **18** (23) (2000), 31–41.

[10] L. Vásquez, M. S. Brown and E. Rosas, *Functions almost contra-super-continuity in m-spaces*, Bol. Soc. Parana. Mat. **29** (2011), 15-36.

Author's address:

Ahmad Al-Omari
Department of Mathematics
Faculty of Science
Al al-Bayt University
P.O. Box 130095
Mafraq 25113, Jordan
E-mail: omarimutah1@yahoo.com

Received July 10, 2012
Revised August 25, 2012

Rend. Istit. Mat. Univ. Trieste
Volume 44 (2012), 439–472

SBV-like regularity for general hyperbolic systems of conservation laws in one space dimension

Stefano Bianchini and Lei Yu

Abstract. *We prove the SBV regularity of the characteristic speed of the scalar hyperbolic conservation law and SBV-like regularity of the eigenvalue functions of the Jacobian matrix of flux function for general hyperbolic systems of conservation laws. More precisely, for the equation*

$$u_t + f(u)_x = 0, \quad u : \mathbb{R}^+ \times \mathbb{R} \to \Omega \subset \mathbb{R}^N,$$

we only assume that the flux f is a C^2 function in the scalar case ($N = 1$) and Jacobian matrix Df has distinct real eigenvalues in the system case ($N \geq 2$). Using a modification of the main decay estimate in [8] and the localization method applied in [17], we show that for the scalar equation $f'(u)$ belongs to the SBV space, and for system of conservation laws the i-th component of $D_x\lambda_i(u)$ has no Cantor part, where λ_i is the i-th eigenvalue of the matrix Df.

Keywords: hyperbolic conservation laws, SBV-like regular, wave-front tracking
MS Classification 2010: 35L65, 35D30

1. Introduction

The study of the regularity of solutions to a general hyperbolic system of conservation laws

$$u_t + f(u)_x = 0, \quad u : \mathbb{R}^+ \times \mathbb{R} \to \Omega \subset \mathbb{R}^N \tag{1}$$

with initial data

$$u(t = 0) = u_0 \in \mathrm{BV}(\mathbb{R}, \Omega) \tag{2}$$

is an important topic in the study of hyperbolic equations. In particular, recently there have been interesting advances in the analysis of the structure of the measure derivative $D_x u(t)$ of BV solution to genuinely nonlinear scalar equations and hyperbolic systems. The results obtained are that, in addition

to the BV bounds, the solution enjoys the strong regularity property that no Cantor part in the space derivative of $u(t)$ appears out of a countable set of times [1, 8, 17]: the fact that the measure $D_x u(t)$ has only absolutely continuous and jump part yields by definition that $u(t) \in \mathrm{SBV}(\mathbb{R})$.

The main idea of the proof is to find a positive bounded functional, which is monotonically decreasing in time: then one shows that at each time a Cantor part appears the functional has a jump downward, and hence one concludes that the SBV regularity of u holds outside a countable set of times.

This paper concerns the extension of the results of [8] to the case where the system is only strictly hyperbolic, i.e. no assumption on the nonlinear structure of the eigenvalues λ_i of Df is done. Clearly, by just considering a linearly degenerate eigenvalue, it is fairly easy to see that the solution u itself cannot be in the SBV function space, so the regularity concerns some nonlinear function of u.

We state the main theorems of this paper: in the following a BV function on $\mathbb{R}$ will be considered defined everywhere by taking the right continuous representative.

In the scalar case, one has

THEOREM 1.1. *Suppose that $u \in \mathrm{BV}(\mathbb{R}^+ \times \mathbb{R})$ is an entropy solution of the scalar conservation law* (3). *Then there exists a countable set $S \subset \mathbb{R}^+$ such that for every $t \in \mathbb{R}^+ \setminus S$ the following holds:*

$$f'(u(t,\cdot)) \in \mathrm{SBV}_{\mathrm{loc}}(\mathbb{R}).$$

After introducing the definition of i-th component of $D_x\lambda_i(u)$ (see (16)), we have the SBV-like regularity for the system case.

THEOREM 1.2. *Let u be a vanishing viscosity solution of the Cauchy problem for the strictly hyperbolic system* (6) *with small* BV *norm. Then there exists an at most countable set $S \subset \mathbb{R}^+$ such that i-th component of $D_x\lambda_i(u(t,\cdot))$ has no Cantor part for every $t \in \mathbb{R}^+ \setminus S$ and $i \in \{1,\ 2,\ \ldots,\ N\}$.*

Since in the genuinely nonlinear case $u \mapsto \lambda_i(u)$ is invertible along the i-th admissible curves $T_s^i[u]$ (see Theorem 3.2 for the definition), it follows that Theorem 4.1 is an extension of the results contained in [8] (and Theorem 1.1 is an extension of the results contained in [17] when the source is 0). The example contained in Remark 7.2 shows that the results are sharp.

The main point of the paper is the fact that the wave-front tracking approximation for the waves of a genuinely nonlinear family does not essentially differ from the wave-front approximations of genuinely nonlinear systems: in other words, the wave pattern of a genuinely nonlinear characteristic family for a (approximate) solution in a general hyperbolic system has the same structure as if all characteristic families are genuinely nonlinear. Thus the analysis carried out in [8] holds also in this case.

The proof of the above two theorems is done as follows. To introduce the argument in the easiest setting, in Section 2, we give a proof for the SBV regularity of the characteristic speed for the general scalar conservation laws. The proof is just a slight modification of the proof of [17, Theorem 1.1].

As one sees in the proof of Theorem 1.1, the main tool is to obtain the SBV regularity when only one characteristic field is genuinely nonlinear (Corollary 4.2). By inspection, the analysis of [8] relies on the wave-front tracking approximation of [9], which assumes that all characteristic fields are genuinely nonlinear or linearly degenerate. Thus we devote Sections 3.2, Section 5.1 to introduce the wave-front tracking approximation for general systems [3].

The focus of Section 5.2 is the observation that the convergence and regularity estimates of [Theorem 10.4][9] still holds for the i-th component of u_x, under the only assumption that the i-th characteristic field is genuinely nonlinear: these estimates are needed in order to define the i-th (ϵ_1, ϵ_0)-shocks and to pass to the limit the estimates concerning the interaction, cancellation and jump measures. The latter is responsible for the functional controlling the SBV regularity, Theorem 4.1.

After these estimates, for completeness we repeat the proof of the decay of negative waves in Section 6.2. Finally we show how to adapt the strategy of the scalar case in Section 7.

2. The scalar case

In this section, we restrict our attention to the scalar conservation laws and motivate our general strategy with this comparatively simpler situation. Let us consider the entropy solution to the hyperbolic conservation law in one space dimension

$$\begin{cases} u_t + f(u)_x = 0 & u : \mathbb{R}^+ \times \mathbb{R} \to \Omega \subset \mathbb{R},\ f \in C^2(\Omega, \mathbb{R}), \\ u_{|t=0} = u_0 & u_0 \in \mathrm{BV}(\mathbb{R}, \Omega). \end{cases} \tag{3}$$

In [17], it is proved the SBV regularity result for the convex or concave flux case.

LEMMA 2.1. *[17] Suppose $f \in C^2(\mathbb{R})$ and $|f''(u)| > 0$. Let $u \in L^\infty(\mathbb{R})$ be an entropy solution of the scalar conservation law* (3). *Then there exists a countable set $S \subset \mathbb{R}$ such that for every $\tau \in \mathbb{R}^+ \setminus S$ the following holds:*

$$u(\tau, \cdot) \in \mathrm{SBV}_{\mathrm{loc}}(\mathbb{R}).$$

Further, by Volpert's Chain Rule (see [2, Theorem 3.99]), it follows that $f'(u(\tau, \cdot)) \in \mathrm{SBV}_{\mathrm{loc}}(\mathbb{R})$ for $\tau \in \mathbb{R}^+ \setminus S$: actually, since $f'' \neq 0$, the two conditions $f'(u(\tau, \cdot)) \in \mathrm{SBV}_{\mathrm{loc}}(\mathbb{R})$ and $u(\tau, \cdot) \in \mathrm{SBV}_{\mathrm{loc}}(\mathbb{R})$ are equivalent.

Following the same argument together with the analysis in [17], we can get the SBV regularity of the slope of characteristics for the scalar conservation law with general flux as stated in Theorem 1.1.

Proof of Theorem 1.1. Recall that if $u \in \mathrm{BV}(\mathbb{R}^+ \times \mathbb{R})$ is an entropy solution, then by the theory of entropy solutions, it follows that $u(\tau, \cdot) \in \mathrm{BV}(\mathbb{R})$ is well defined for every $\tau \in \mathbb{R}^+$.

Define the sets

$$\begin{aligned} J_\tau &:= \big\{x \in \mathbb{R} :\ u(\tau, x-) \neq u(\tau, x+)\big\}, \\ F_\tau &:= \big\{x \in \mathbb{R} :\ f''(u(\tau, x)) = 0\big\}, \\ C &:= \big\{(\tau, \xi) \in \mathbb{R}^+ \times \mathbb{R} :\ \xi \in J_\tau \cup F_\tau\big\}. \end{aligned}$$

Set also $C_\tau := J_\tau \cup F_\tau$ as the τ-section of C.

Since the Cantor part $D^{\mathrm{c}}u(\tau, \cdot)$ of $Du(\tau, \cdot)$ and the jump part $D^{\mathrm{j}}u(\tau, \cdot)$ of $Du(\tau, \cdot)$ are mutually singular, then $|D^{\mathrm{c}}u(\tau, \cdot)|(J_\tau) = 0$. Using the fact that $f''(u(\tau, \cdot)) = 0$ on F_τ, by Volpert's Chain Rule one obtains

$$\begin{aligned} |D^{\mathrm{c}} f'(u(\tau, \cdot))|(C_\tau) &\leq |D^{\mathrm{c}} f'(u(\tau, \cdot))|(J_\tau) + |D^{\mathrm{c}} f'(u(\tau, \cdot))|(F_\tau) \\ &= |f''(u(\tau, \cdot)) D^{\mathrm{c}} u(\tau, \cdot)|(J_\tau) + |f''(u(\tau, \cdot)) D^{\mathrm{c}} u(\tau, \cdot)|(F_\tau) = 0. \end{aligned}$$

Let $(t_0, x_0) \in \mathbb{R}^+ \times \mathbb{R} \setminus C$. Using the finite speed of propagation and the maximum principle for entropy solutions and the fact that $u(t_0, x)$ is continuous at x_0 by the definition of C, it is possible to find a triangle of the form

$$T(t_0, x_0) := \Big\{(t, x) :\ |x - x_0| < b_0 - \bar\lambda (t - t_0),\ 0 < t - t_0 < b_0/\bar\lambda\Big\} \tag{4}$$

such that $|f''(u(t, x))| \geq c_0 > 0$, for any $(t, x) \in T(t_0, x_0)$. Here c_0 depends on (t_0, x_0) and $\bar\lambda$ is the maximal speed of propagation, which depends only on the L^∞-bound of u_{t_0} (and hence only depends on the L^∞-bound of u by maximal principle).

In particular, in $T(t_0, x_0)$ the solution u of (3) coincides with the solution of the following problem

$$\begin{cases} w_t + f(w)_x = 0, \\ w(t_0, x) = \begin{cases} u(t_0, x) & |x - x_0| < b_0, \\ \frac{1}{2b_0} \int_{x_0 - b_0}^{x_0 + b_0} u(t_0, y)\, dy & |x - x_0| \geq b_0. \end{cases} \end{cases}$$

By Lemma 2.1, $w(t, \cdot)$ is SBV regular for any $t > t_0$ out of a countable set of times $S(t_0, x_0)$. Write $T_\tau(t_0, x_0) := T(t_0, x_0) \cap \{t = \tau\}$, thus $u(\tau, \cdot)\llcorner_{T_\tau(t_0, x_0)}$ and $f'(u(\tau, \cdot))\llcorner_{T_\tau(t_0, x_0)}$ are SBV for $\tau \in]t_0,\ t_0 + b/\bar\lambda[\setminus S(t_0, x_0)$.

Let B be the set of all points of $\mathbb{R}^+ \times \mathbb{R} \setminus C$ which are contained in at least one of these triangles. (Notice that $T(t_0, x_0)$ is a open set and does not contain

the point (t_0, x_0).) Let $\{T(t_i, x_i)\}_{i\in\mathbb{N}}$ be a countable subfamily of the triangles covering B. From the previous observation on the function $u_{\llcorner T(t_i,x_i)}$, the set

$$S_i := \left\{\tau : u(\tau,\cdot)_{\llcorner T_\tau(t_i,x_i)} \notin \mathrm{SBV}(T_\tau(t_i, x_i))\right\}$$

is at most countable.

Let $C' := \mathbb{R}^+ \times \mathbb{R} \setminus (B \cup C)$ and $S_{C'} := \{\tau \in \mathbb{R}^+ : \{t = \tau\} \cap C' \neq \emptyset\}$. It is obvious that for every $t' \in \mathbb{R}^+ \setminus S_{C'}$, $x' \in \mathbb{R}$, either there is a triangle $T \in \{T(t_i, x_i)\}_{i\in\mathbb{N}}$ such that $(t', x') \in T$ and $u(t,\cdot)_{\llcorner T}$ is SBV function out of countable many times or $(t', x') \in C$.

We claim that the set $S_{C'}$ is at most countable. Indeed, it is enough to prove that the set $S_K := \{\tau \in \mathbb{R}^+ : \{t = \tau\} \cap C' \cap K \neq \emptyset\}$ is at most countable for every compact set $K \subset \mathbb{R}^+ \times \mathbb{R}$ when the triangles $T(t', x')$ have a base of fixed length for every $(t', x') \in C'$: it is fairly simple to see that in this case the set S_K is finite since (t', x') can not be contained in any other $T(t'', x'')$ for $t' \neq t''$ and $(t'', x'') \in C'$.

For any τ not in the countable set

$$S_{C'} \cup \bigcup_{i\in\mathbb{N}} S_i,$$

one obtains the following inequality:

$$\begin{aligned} |D^{\mathrm{c}} f'(u(\tau,\cdot))(\mathbb{R})| \quad &\leq |D^{\mathrm{c}} f'(u(\tau,\cdot))|\left(\bigcup_{i\in\mathbb{N}} T_\tau(t_i, x_i)\right) \\ &\quad + |D^{\mathrm{c}} f'(u(\tau,\cdot))|(C_\tau) = 0. \end{aligned} \tag{5}$$

This concludes the proof. □

By a standard argument in the theory of BV functions, we have the following result.

Corollary 2.2. *Let $u \in L^\infty(\mathbb{R}^+ \times \mathbb{R})$ be an entropy solution of the scalar conservation law* (3). *Then $f'(u) \in \mathrm{SBV}_{\mathrm{loc}}(\mathbb{R}^+ \times \mathbb{R})$.*

The difference is that now the function $f'(u)$ is considered as a function of two variable.

Proof. The starting point is that up to a countable set of times, $Df'(u(t,\cdot))$ has no Cantor part (Theorem 1.1). From the slicing theory of BV function ([2, Theorem 3.107-108]), we know that the Cantor part of the 2-dimensional measure $D_x f'(u)$ is the integral with respect of t of the Cantor part of $Df'(u(t,\cdot))$. This concludes that $D_x f'(u)$ has no Cantor part, i.e. $D^{\mathrm{c}}_x f'(u) = 0$.

By combining Volpert's Chain Rule and the conservation law (3), one has

$$D_t^c u = -f'(u) D_x^c u.$$

Using Volpert's rule once again, one obtains

$$D_t^c f'(u) = -f''(u) D_t^c u = -f''(u) f'(u) D_x^c u = -f'(u) D_x^c f'(u) = 0,$$

which concludes that also $D_t f(u)$ has no Cantor part. □

REMARK 2.3. *In [17], it is proved that if f in* (3) *has only countable many inflection points. i.e. the set*

$$\{u \in \Omega : f''(u) \neq 0\}$$

is at most countable, then the entropy solution of (3) *is SBV regular. It is easy to see that for general hyperbolic scalar conservation laws $f \in C^2$ is not enough to obtain the SBV regularity. In fact, we can consider $f' \equiv$ constant, which means* (3) *degenerates into a linear equation. Then the entropy solution u is not SBV regular unless the initial data u_0 is a SBV function.*

3. Notations and settings for general systems

Throughout the rest of the paper, the symbol $\mathcal{O}(1)$ always denotes a quantity uniformly bounded by a constant depending only on the system (1).

3.1. Preliminary notation

Consider the Cauchy problem

$$\begin{cases} u_t + f(u)_x = 0 & u : \mathbb{R}^+ \times \mathbb{R} \to \Omega \subset \mathbb{R}^N,\ f \in C^2(\Omega, \mathbb{R}), \\ u_{|t=0} = u_0 & u_0 \in \mathrm{BV}(\mathbb{R}, \Omega). \end{cases} \tag{6}$$

The only assumption is strict hyperbolicity in Ω: the eigenvalues $\{\lambda_i(u)\}_{i=1}^N$ of the Jacobi matrix $A(u) = Df(u)$ satisfy

$$\lambda_1(u) < \cdots < \lambda_N(u), \qquad u \in \Omega.$$

Furthermore, as we only consider the solutions with small total variation, it is not restrictive to assume that Ω is bounded and there exist constants $\{\check{\lambda}_j\}_{j=0}^N$, such that

$$\check{\lambda}_{k-1} < \lambda_k(u) < \check{\lambda}_k, \qquad \forall u \in \Omega,\ k = 1, \ldots, N. \tag{7}$$

Let $\{r_i(u)\}_{i=1}^N$ and $\{l_j(u)\}_{j=1}^N$ be a basis of right and left eigenvectors, depending smoothly on u, such that

$$l_j(u) \cdot r_i(u) = \delta_{ij} \text{ and } |r_i(u)| \equiv 1, \quad i = 1, \ldots, N. \tag{8}$$

DEFINITION 3.1. *For $i = 1, \dots, N$, we say that the i-th characteristic field (or i-th family) is* genuinely nonlinear *if*

$$\nabla\lambda_i(u) \cdot r_i(u) \neq 0 \quad \textit{for all } u \in \Omega,$$

and we say that the i-th characteristic field (or i-th family) is linearly degenerate *if instead*

$$\nabla\lambda_i(u) \cdot r_i(u) = 0 \quad \textit{for all } u \in \Omega.$$

In the following, if the i-th characteristic field is genuinely nonlinear, instead of (8) we normalize $r_i(u)$ such that

$$\nabla\lambda_i(u) \cdot r_i(u) \equiv 1. \tag{9}$$

In [7], it is proved that if the total variation of u_0 is sufficiently small, the solutions of the viscous parabolic approximation equations

$$\begin{cases} u_t + f(u)_x = \epsilon u_{xx}, \\ u(0,x) = u_0(x), \end{cases}$$

are uniformly bounded, and the limit of u^ϵ as $\epsilon \to 0$ is called *vanishing viscosity solution* of (6) and it is a BV function.

3.2. Construction of solutions to the Riemann problem

The Riemann problem is the Cauchy problem (6) with piecewise constant initial data of the form

$$u_0 = \begin{cases} u^{\mathrm{L}} & x < 0, \\ u^{\mathrm{R}} & x > 0. \end{cases} \tag{10}$$

The solution to this problem is the key ingredient for building the front-tracking approximate solution: the basic step is the construction of the admissible *elementary curve* of the k-th family for any given left state u^{L}.

A working definition of admissible elementary curves can be given by means of the following theorem.

THEOREM 3.2 ([4, 7]). *For every $u \in \Omega$ there exist*

1. *N Lipschitz continuous curves $s \mapsto T_s^k[u] \in \Omega$, $k = 1, \dots, N$, satisfying $\lim_{s\to 0} \dfrac{d}{ds} T_s^k[u] = r_k(u)$,*

2. *N Lipschitz functions $(s,\tau) \mapsto \sigma_s^k[u](\tau)$, with $0 \le \tau \le s$, $k = 1, \dots, N$, satisfying $\tau \mapsto \sigma_s^k[u](\tau)$ increasing and $\sigma_0^k[u](0) = \lambda_k(u)$*

with the following properties.
When $u^{\mathrm{L}} \in \Omega$, $u^{\mathrm{R}} = T_s^k[u^{\mathrm{L}}]$, *for some* s *sufficiently small, the unique vanishing viscosity solution of the Riemann problem* (6)-(10) *is defined a.e. by*

$$u(t,x) := \begin{cases} u^{\mathrm{L}} & x/t < \sigma_s^k[u^{\mathrm{L}}](0), \\ T_\tau^k[u^{\mathrm{L}}] & x/t = \sigma_s^k[u^{\mathrm{L}}](\tau), \tau \in \mathcal{I}, \\ u^{\mathrm{R}} & x/t > \sigma_s^k[u^{\mathrm{L}}](s). \end{cases}$$

where $\mathcal{I} := \big\{\tau \in [0,s] : \sigma_s^k[u^{\mathrm{L}}](\tau) \neq \sigma_s^k[u^{\mathrm{L}}](\tau') \text{ for all } \tau' \neq \tau\big\}$.

Remark 3.3. *If* i*-th family is genuinely nonlinear, then the Lipschitz curve* $T_s^i[\bar u]$ *can be written as*

$$T_s^i[\bar u] = \begin{cases} R_i[\bar u](s) & s \geq 0, \\ S_i[\bar u](s) & s < 0, \end{cases}$$

where $R_i[\bar u]$, $S_i[\bar u]$ *are respectively the rarefaction curve and the Rankine-Hugoniot curve of the* i*-th family with any given point* $\bar u$ *in* Ω. *Some certain elementary weak solution, called rarefaction waves and shock waves can be defined along the rarefaction curve and Rankine-Hugoniot curve, for example see [9]. The elementary curve* $T_s^i[\bar u]$ *is parametrized by*

$$s = l_i(\bar u) \cdot (T_s^i[\bar u] - \bar u). \tag{11}$$

The vanishing viscosity solution [7] of a Riemann problem for (6) is obtained by constructing a Lipschitz continuous map

$$(s_1, \dots, s_N) \mapsto T_{s_N}^N \Big[T_{s_{N-1}}^{N-1} \big[\cdots \big[T_{s_1}^1 \big[\, u^{\mathrm{L}} \,\big]\,\big]\,\big]\Big] = u^{\mathrm{R}},$$

which is one to one from a neighborhood of the origin onto a neighborhood of u^{L}. Then we can uniquely determine intermediate states $u^{\mathrm{L}} = \omega_0, \omega_1, \dots, \omega_N = u^{\mathrm{R}}$, and the *wave sizes* $s_1, s_2, \dots, s_N$ such that

$$\omega_k = T_{s_k}^k[\omega_{k-1}], \quad k = 1, \dots, N,$$

provided that $|u^{\mathrm{L}} - u^{\mathrm{R}}|$ is sufficiently small.

By Theorem 3.2, each Riemann problem with initial datum

$$u_0 = \begin{cases} \omega_{k-1} & x < 0, \\ \omega_k & x > 0, \end{cases} \tag{12}$$

admits a vanishing viscosity solution u_k, containing a sequence of rarefactions, shocks and discontinuities of the k-th family: we call u_k the k-th *elementary*

composite wave. Therefore, under the strict hyperbolicity assumption, the general solution of the Riemann problem with the initial data (10) is obtained by piecing together the vanishing viscosity solutions of the elementary Riemann problems given by (6)-(12).

Indeed, from the strict hyperbolicity assumption (7), the speed of each elementary k-th wave in the solution u_k is inside the interval $[\check{\lambda}_{k-1}, \check{\lambda}_k]$ if $s \ll 1$, so that the solution of the general Riemann problem (6)-(10) is then given by

$$u(t,x) = \begin{cases} u^{\mathrm{L}} & x/t < \check{\lambda}_0 \\ u_k(t,x) & \check{\lambda}_{k-1} < x/t < \check{\lambda}_k, k = 1, \dots, N, \\ u^{\mathrm{R}} & x/t > \check{\lambda}_N. \end{cases} \tag{13}$$

REMARK 3.4. *If the characteristic fields are either genuinely nonlinear or linearly degenerate, the admissible solution of Riemann problem* (6)-(10) *consists of N family of waves. Each family contains either only one shock, one rarefaction wave or one contact discontinuity. However, the general solution of a Riemann problem provided above may contain a countable number of rarefaction waves, shock waves and contact discontinuities.*

3.3. Cantor part of the derivative of characteristic for i-th waves

Recalling the solution (13) to the Riemann problem (6)-(10), let $\tilde{\lambda}_i(u^{\mathrm{L}}, u^{\mathrm{R}})$ denote the i-th eigenvalue of the average matrix

$$A(u^{\mathrm{L}}, u^{\mathrm{R}}) = \int_0^1 A(\theta u^{\mathrm{L}} + (1-\theta)u^{\mathrm{R}}) d\theta, \tag{14}$$

and $\tilde{l}_i(u^{\mathrm{L}}, u^{\mathrm{R}})$, $\tilde{r}_i(u^{\mathrm{L}}, u^{\mathrm{R}})$ are the corresponding left and right eigenvector satisfying $\tilde{l}_i \cdot \tilde{r}_i = \delta_{ij}$ and $|\tilde{r}_j| \equiv 1$, for every $i, j \in \{1, \dots, N\}$. Define thus

$$\tilde{\lambda}_i(t,x) = \tilde{\lambda}_i(u(t,x-), u(t,x+)), \tag{15a}$$

$$\tilde{r}_i(t,x) = \tilde{r}_i(u(t,x-), u(t,x+)), \tag{15b}$$

$$\tilde{l}_i(t,x) = \tilde{l}_i(u(t,x-), u(t,x+)). \tag{15c}$$

Since the $\tilde{r}_i$, $\tilde{l}_i$ have directions close to r_i, l_i, one can decompose $D_x u$ into the sum of N measures:

$$D_x u = \sum_{k=1}^{N} v_k \tilde{r}_k.$$

where $v_i = \tilde{l}_i \cdot D_x u$ is a scalar valued measure which we call as *i-th wave measure* [9].

In the same way we can decompose the a.c. part $D_x^{\rm ac}u$, the Cantor part $D_x^{\rm c}u$ and the jump part $D_x^{\rm jump}u$ of D_xu as

$$D_x^{\rm ac}u=\sum_{k=1}^N v_k^{ac}\tilde r_k,\quad D_x^{\rm c}u=\sum_{k=1}^N v_k^{\rm c}\tilde r_k,\quad D_x^{\rm jump}u=\sum_{k=1}^N v_k^{\rm jump}\tilde r_k.$$

We call $v_i^{\rm c}$ the Cantor part of v_i and denote by

$$v_i^{\rm cont}:=v_i^{\rm c}+v_i^{ac}=\tilde l_i\cdot(D_x^{\rm c}u+D_x^{\rm ac}u)$$

the continuous part of v_i. According to Volpert's Chain Rule

$$D_x\lambda_i(u)=\nabla\lambda_i(u)(D_x^{\rm ac}u+D_x^{\rm c}u)+[\lambda_i(u^+)-\lambda_i(u^-)]\delta_x,$$

and then

$$D_x^{\rm c}\lambda_i(u)=\nabla\lambda_i\cdot D_x^{\rm c}u=\sum_k\big(\nabla\lambda_i\cdot\tilde r_k\big)v_k^{\rm c}.$$

We define the *i-th component of* $D_x\lambda_i(u)$ as

$$[D_x\lambda_i(u)]_i:=\big(\nabla\lambda_i\cdot\tilde r_i\big)v_i^{\rm cont}+\big[\lambda_i(u^+)-\lambda_i(u^-)\big]\,\frac{|v_i^{\rm jump}(x)|}{\sum_k|v_k^{\rm jump}(x)|},\tag{16}$$

and the *Cantor part of i-th component of* $D_x\lambda_i(u)$ to be

$$[D_x^{\rm c}\lambda_i(u)]_i:=\big(\nabla\lambda_i\cdot\tilde r_i\big)v_i^{\rm c}.\tag{17}$$

4. Main SBV regularity argument

Following [8], the key idea to obtain SBV-like regularity for v_i is to prove a decay estimate for the continuous part of v_i. We state here the main estimate of our paper.

THEOREM 4.1. *Consider the general strictly hyperbolic system* (6)*, and suppose that the i-th characteristic field is genuinely nonlinear. Then there exists a finite, non-negative Radon measure* $\mu_i^{\rm ICJ}$ *on* $\mathbb{R}^+\times\mathbb{R}$ *such that for* $t>\tau>0$

$$\big|v_i^{\rm cont}(t)\big|(B)\le\mathcal{O}(1)\bigg\{\frac{\mathcal{L}(B)}{\tau}+\mu_i^{\rm ICJ}([t-\tau,t+\tau]\times\mathbb{R})\bigg\}\tag{18}$$

for all Borel subset B *of* $\mathbb{R}$.

Different from [8], we assume only one characteristic field to be genuinely nonlinear and no other requirement on the other characteristic fields. Once Theorem 4.1 is proved, then the SBV argument develops as follows [8].

Suppose at time $t = s$, $v_i(s)$ has a Cantor part. Then there exists a $\mathcal{L}^1$-negligible Borel set K with $v_i^{\rm cont}(s)(K) > 0$ and $D^{\rm jump}v_i(s)(K) = 0$. Then for all $s > \tau > 0$,

$$0 < |v_i(s)|(K) = |v_i^{\rm cont}(s)|(K) \leq \mathcal{O}(1)\Big\{ \frac{\mathcal{L}^1(K)}{\tau} + \mu_i^{\rm ICJ}([s-\tau, s+\tau] \times \mathbb{R}) \Big\}.$$

Since $\mathcal{L}^1(K) = 0$, we can let $\tau \to 0$, and deduce that $\mu_i^{\rm ICJ}(\{s\} \times \mathbb{R}) > 0$. This shows that the Cantor part appears at most countably many times because $\mu_i^{\rm ICJ}$ is finite.

Then, we can have the following result which generalizes [8, Corollary 3.2] to the case when only one characteristic field is genuinely nonlinear and no assumption is made on the others.

COROLLARY 4.2. *Let u be a vanishing viscosity solution of the Cauchy problem for the strictly hyperbolic system* (6)*, and assume that the i-th characteristic field is genuinely nonlinear. Then $v_i(t)$ has no Cantor part out of a countable set of times.*

As we see in the scalar case, by proving the SBV regularity of the solution under the genuinely nonlinearity assumption of one characteristic field, we can deduce a kind of SBV regularity of the characteristic speed for general systems.

Unlike the scalar case, we do not have the maximum principle to guarantee the small variation of u in the triangle $T(t_0, x_0)$ defined in (4). However, in the system case, we have the following estimates for the vanishing viscosity solutions.

For $a < b$ and $\tau \geq 0$, we denote by Tot.Var.$\{u(\tau);\ ]a, b[\}$ the total variation of $u(\tau)$ over the open interval $]a, b[$. Moreover, consider the triangle

$$\Delta_{a,b}^{\tau,\eta} := \Big\{ (t, x) :\ \tau < t < (b-a)/2\eta,\ a + \eta t < x < b - \eta t \Big\}.$$

The oscillation of u over $\Delta_{a,b}^{\tau,\eta}$ will be denoted by

$$\text{Osc.}\{u;\ \Delta_{a,b}^{\tau,\eta}\} := \sup \Big\{ |u(t,x) - u(t',x')| :\quad (t,x),\ (t',x') \in \Delta_{a,b}^{\tau,\eta} \Big\}.$$

We have the following results.

THEOREM 4.3 (Tame Oscillation, [7]). *There exists $C' > 0$ and $\bar\eta > 0$ such that for every $a < b$ and $\tau \geq 0$, one has*

$$\text{Osc.}\{u;\ \Delta_{a,b}^{\tau,\bar\eta}\} \leq C' \cdot \text{Tot.Var.}\{u(\tau);\ ]a, b[\}.$$

Adapting the proof of the scalar case, we can prove the main Theorem 1.2 of this paper: the proof of this theorem will be done in Section 7.

5. Review of wave-front tracking approximation for general system

To prove Theorem 4.1, we use the front tracking approximation in [3] which extends the one in [9] to the general systems. Since the construction is now standard, we only give a short overview about existence, compactness and convergence of the approximation, pointing to the properties needed in our argument: more precisely, we will only consider how one constructs the approximate wave pattern of the k-th genuinely nonlinear family (Section 5.1.2).

The main point is that, for general systems, the accurate/simplified/crude Riemann solvers for the k-th wave coincides with the approximate/simplified/crude Riemann solvers when all families are genuinely nonlinear (see below for the definition of accurate/simplified/crude Riemann solvers). This means that the wave pattern of the k-th genuinely nonlinear family will have the same structure as if all other families are genuinely nonlinear: by this, we mean that shock-shock interaction generates shocks, the jump in characteristic speed across k-th waves is proportional to their size, and one can thus use the k-component of the derivative of λ_k (16) to measure the total variation of v_k.

5.1. Description of the wave-front tracking approximation

The wave-front tracking approximation is an algorithm which produces piecewise constant approximate solutions to the Cauchy problem (6). Roughly speaking, we first choose a piecewise constant function u_0^ϵ which is a good approximation to the initial data u_0 such that

$$\text{Tot.Var.}\{u_0^\epsilon\} \leq \text{Tot.Var.}\{u_0\}, \quad ||u_0^\epsilon - u_0||_{L^1} < \epsilon, \tag{19}$$

and u_0^ϵ only has finite jumps. Let $x_1 < \cdots < x_m$ be the jump points of u_0^ϵ. For each $\alpha = 1, \ldots, m$, we approximately solve the Riemann problem (see Section 3.2, just shifting the center from $(0,0)$ to $(0, x_\alpha)$) with the initial data given by the jump $[u_0^\epsilon(x_\alpha-), [u_0^\epsilon(x_\alpha+)]$ by a function $w(t,x) = \phi(\frac{x-x_0}{t-t_0})$ where ϕ is a piecewise constant function. The straight lines where the discontinuities are located are called *wave-fronts* (or just *fronts* for shortness). The wave-fronts can be prolonged until they interact with other fronts, then at the interaction point, the corresponding Riemann problem is approximately solved and several new fronts are generated forward. Then one tracks the wave-fronts until they interact with other wave-fronts, etc... In order to avoid the algorithm to produce infinite many wave-fronts in finite time, different kinds of approximate Riemann solvers should be introduced.

5.1.1. Approximate Riemann solver

Suppose at the point (t_1, x_1) a wave-front of size s' belonging to k'-th family interacts from the left with a wave-front of size s'' belonging to k''-th family for some k', $k'' \in \{1, \cdots, N\}$ such that $k' < k''$ and (see Section 3.2 for the definition of T_s^k)

$$u^M = T_{s'}^{k'}[u^{\mathrm{L}}], \qquad u^{\mathrm{R}} = T_{s''}^{k''}[u^M].$$

Assuming that $|u^{\mathrm{L}} - u^{\mathrm{R}}|$ sufficiently small, at the interaction point, the Riemann problem with the initial data given by the jump $[u^{\mathrm{L}}, u^{\mathrm{R}}]$ will be solved by approximate Riemann solver. There are two kinds of approximate Riemann solvers defined for interactions between two physical wave-fronts.

- *Accurate Riemann Solver*: It replaces each elementary composite wave of the exact Riemann solution (refers to u_k in (13)) by an approximate elementary wave which is a finite collection of jumps traveling with a speed given by the average speed $\tilde{\lambda}_k$ (see (15a)), and the wave opening (i.e. the difference in speeds between any two consecutive fronts) is less than some small parameter ϵ controlling the accuracy of the approximation.

- *Simplified Riemann Solver*: It only generates approximate elementary waves belonging to k'-th and k''-th families with corresponding size s' and s'' as the incoming ones if $k' \neq k''$, and approximate elementary waves of size $s' + s''$ belonging to k'-th family if $k' = k''$. The simplified Riemann solver collects the remaining new waves into a single *nonphysical front*, traveling with a constant speed $\hat{\lambda}$, strictly larger than all characteristic speed $\hat{\lambda}$. Therefore, usually the simplified Riemann solver generate less outgoing fronts after interaction than the accurate Riemann solver.

Since the simplified Riemann solver produces nonphysical wave-fronts and they can not interact with each other, one only needs an approximate Riemann solver defined for the interaction between, for example, a physical front of the k-th family with size s, connecting u^M, u^{R} and a nonphysical front (coming from the left) connecting the left value u^{L} and u^M traveling with speed $\hat{\lambda}$.

- *Crude Riemann Solver* generates a k-th front connecting u^{L} and $\tilde{u}^M = T_s^k[u^{\mathrm{L}}]$ traveling with speed $\tilde{\lambda}_i$ and a nonphysical wave-front joining $\tilde{u}^M$ and u^{R}, traveling with speed $\hat{\lambda}$. In the following, for simplicity, we just say that the non-physical fronts belong to the $(N+1)$-th characteristic field.

REMARK 5.1. *We can assume that at each time $t > 0$, at most one interaction takes place, involving exactly two incoming fronts, because we can slightly change the speed of one of the incoming fronts if more than two fronts meet at*

the same point. It is sufficient to require that the error vanishes when $\epsilon \to 0$. To simplify the analysis, we assume that the fronts satisfy the Rankine-Hugoniot conditions exactly.

5.1.2. The approximate Riemann solvers for genuinely nonlinear waves

If the k-th characteristic family is genuinely nonlinear, the elementary wave u_k is either a shock wave or a rarefaction wave. The key example of the accurate Riemann solver is thus to consider how these two solutions are approximated.

If k-th elementary wave u_k in (13) is just a single shock, for example

$$u_k = \begin{cases} u^{\mathrm{L}} & x/t < \sigma, \\ u^{\mathrm{R}} & x/t > \sigma, \end{cases}$$

where σ is the speed of shock wave, then the approximated k-th wave coincides the exact one (apart from the speed in case, see the above remark).

If u_k is a rarefaction wave of the k-th family connecting the left value u^{L} and the right value u^{R}, for example, if $u^{\mathrm{R}} := T^k_s[u^{\mathrm{L}}]$ and

$$u_k = \begin{cases} u^{\mathrm{L}} & x/t < \lambda_k(u^{\mathrm{L}}), \\ T^k_{s^*}[u^{\mathrm{L}}] & x/t \in [\lambda_k(u^{\mathrm{L}}), \lambda_k(u^{\mathrm{R}})],\ x/t = \lambda_k(T^k_{s^*}[u^{\mathrm{L}}]), \\ u^{\mathrm{R}} & x/t > \lambda_k(u^{\mathrm{R}}), \end{cases}$$

where $s^* \in [0, s]$. Then the approximation $\tilde{u}_k$ is a rarefaction fan containing several rarefaction fronts. More precisely, we can choose real numbers $0 = s_0 < s_1 < \cdots < s_n = s$, and define the points $w_i := T^k_{s_i}[u^{\mathrm{L}}]$, $i = 0, \ldots, n$, with the following properties,

$$w_{i+1} = T^k_{(s_{i+1}-s_i)}[w_i],$$

and the wave opening of consecutive wave-fronts are sufficiently small, i.e.

$$\sigma^k_s[u^{\mathrm{L}}](s_{i+1}) - \sigma^k_s[u^{\mathrm{L}}](s_i) \le \epsilon, \quad \forall i = 0, \ldots, n-1.$$

where the function σ^k_s is defined in Theorem 3.2. We let the jump $[\omega_i, \omega_{i+1}]$ travel with the speed $\tilde{\sigma}_i := \tilde{\lambda}_k(\omega_i, \omega_{i+1})$ (15a), so that the rarefaction fan $\tilde{u}_k$ becomes

$$\tilde{u}_k = \begin{cases} u^{\mathrm{L}} & x/t < \tilde{\sigma}_1, \\ \omega_i & \tilde{\sigma}_i \le x/t < \tilde{\sigma}_{i+1},\ i = 1, \ldots, n-1, \\ u^{\mathrm{R}} & x/t \ge \tilde{\sigma}_n. \end{cases}$$

5.1.3. Interaction potential and BV estimates

Suppose two wave-fronts with size s' and s'' interact. In order to get the estimate on the difference between the size of the incoming waves and the size of the outgoing waves produced by the interaction, we need to define the *amount of interaction* $\mathcal{I}(s', s'')$ between s' and s''.

When s' and s'' belong to different characteristic families (including $N+1$-th family), set

$$\mathcal{I}(s', s'') = |s's''|. \tag{20}$$

If s', s'' belong to the same characteristic family, the definition of $\mathcal{I}(s', s'')$ is more complicated (see [3, Definition 3]). We just mention that if s', s'' are the sizes of two shocks which have the same sign, traveling with the speed σ' and σ'' respectively, then the amount of interaction takes the form

$$\mathcal{I}(s', s'') = |s's''|\big|\sigma' - \sigma''\big|, \tag{21}$$

i.e. the product of the size of the waves times the difference of their speeds (of the order of the angle between the two shocks).

To control the amount of interaction, the following potential is introduced. At each time $t > 0$ when no interaction occurs, and $u(t, \cdot)$ has jumps at $x_1, \dots, x_m$, we denote by

$$\omega_1, \dots, \omega_m, \quad s_1, \dots, s_m, \quad i_1, \dots, i_m,$$

their left states, signed sizes and characteristic families, respectively: the sign of s_α is given by the respective orientation of $dT_s^k[u]/ds$ and r_k, if the jump at x_α belongs to the k-th family. The Total Variation of u will be computed as

$$V(t) = V(u(t)) := \sum_\alpha \big|s_\alpha\big|.$$

Following [4], we define the *Glimm wave interaction potential* as follows:

$$\begin{aligned} \mathcal{Q}(t) = \mathcal{Q}(u(t)) &:= \sum_{\substack{i_\alpha > i_\beta \\ x_\alpha < x_\beta}} \big|s_\alpha s_\beta\big| \\ &+ \frac{1}{4} \sum_{i_\alpha = i_\beta < N+1} \int_0^{|s_\alpha|} \int_0^{|s_\beta|} \big|\sigma^{i_\beta}_{s_\beta}[\omega_\beta](\tau'') - \sigma^{i_\alpha}_{s_\alpha}[\omega_\alpha](\tau')\big| d\tau' d\tau''. \end{aligned} \tag{22}$$

Denoting the time jumps of the total variation and the Glimm potential as

$$\Delta V(\tau) = V(\tau+) - V(\tau-), \quad \Delta \mathcal{Q}(\tau) = \mathcal{Q}(\tau+) - \mathcal{Q}(\tau-),$$

the fundamental estimates are the following ([3, Lemma 5]): in fact, when two wave-fronts with size s', s'' interact,

$$\Delta\mathcal{Q}(\tau) = -\mathcal{O}(1)\mathcal{I}(s', s''), \tag{23a}$$

$$\Delta V(\tau) = \mathcal{O}(1)\mathcal{I}(s', s''). \tag{23b}$$

Thus one defines the *Glimm functional*

$$\Upsilon(t) := V(t) + C_0\mathcal{Q}(t) \tag{24}$$

with C_0 suitable constant, so that Υ decreases at any interaction. Using this functional, one can prove that ϵ-approximate solutions exist and their total variations are uniformly bounded (see [3, Section 6.1]).

5.1.4. Construction of the approximate solutions and their convergence to exact solution

The construction starts at initial time $t = 0$ with a given $\epsilon > 0$, by taking u_0^ϵ as a suitable piecewise constant approximation of initial data u_0, satisfying (19). At the jump points of u_0^ϵ, we locally solve the Riemann problem by accurate Riemann solver. The approximate solution u^ϵ then can be prolonged until a first time t_1 when two wave-fronts interact. Again we solve the Riemann problem at the interaction point by an approximate Riemann solver. Whenever the amount of interaction (see Section 5.1.3 for the definition) of the incoming waves is larger than some threshold parameter $\rho = \rho(\epsilon) > 0$, we shall adopt the accurate Riemann solver. Instead, in the case where the amount of interaction of the incoming waves is less than ρ, we shall adopt simplified Riemann solvers. And we will apply the crude Riemann solver if one of the incoming wave-front is non-physical front. One can show that the number of wave-fronts in approximate solution constructed by such algorithm remains finite for all times (see [3, Section 6.2]).

We call such approximate solutions ϵ-*approximate front tracking solutions*. At each time t when there is no interaction, the restriction $u^\epsilon(t)$ is a step function whose jumps are located along straight lines in the (t, x)-plane.

Let $\{\epsilon_\nu\}_{\nu=1}^\infty$ be a sequence of positive real numbers converging to zero. Consider a corresponding sequence of ϵ_ν-approximate front tracking solutions $u^\nu := u^{\epsilon_\nu}$ of (6): it is standard to show that the functions $t \mapsto u^\nu(t, \cdot)$ are uniformly Lipschitz continuous in L^1 norm, and the decay of the Glimm Functional yields that the solutions $u^\nu(t, \cdot)$ have uniformly bounded total variation. Then by Helly's theorem, u^ν converges up to a subsequence in $\mathbb{L}^1_{\mathrm{loc}}(\mathbb{R}^+ \times \mathbb{R})$ to some function u, which is a weak solution of (6).

It can be shown that by the choice of the Riemann Solver in Theorem 3.2, the solution obtained by the front tracking approximation coincides with the unique vanishing viscosity solution [7]. Furthermore, there exists a closed domain $\mathcal{D} \subset L^1(\mathbb{R}, \Omega)$ and a unique distributional solution u, which is a Lipschitz semigroup $\mathcal{D} \times [0, +\infty[\to \mathcal{D}$ and which for piecewise constant initial data coincides, for a small time, with the solution of the Cauchy problem obtained piecing together the standard entropy solutions of the Riemann problems. Moreover, it lives in the space of BV functions.

For simplicity, the pointwise value of u is its L^1 representative such that the restriction map $t \mapsto u(t)$ is continuous form the right in L^1 and $x \mapsto u(t, x)$ is right continuous from the right.

5.1.5. Further estimates

To each u^ν, we define the *measure* μ^{I}_ν *of interaction* and the *measure* μ^{IC}_ν *of interaction and cancellation* concentrated on the set of interaction points as follows. If two wave-fronts belonging to the families $i, i' \in \{1, \dots, N+1\}$ with size s', s'' interact at a point P, we define by

$$\mu^{\mathrm{I}}_\nu(\{P\}) := \mathcal{I}(s', s''),$$

$$\mu^{\mathrm{IC}}_\nu(\{P\}) := \mathcal{I}(s', s'') + \begin{cases} |s'| + |s''| - |s' + s''| & i = i', \\ 0 & i \neq i'. \end{cases} \tag{25}$$

the measure of interaction and the measure of interaction-cancellation.

The wave size estimates ([3, Lemma 1]) yields balance principles for the wave size of approximate solution. More precisely, given a polygonal region Γ with edges transversal to the waves it encounters, denote by $W^{i\pm}_{\nu,\mathrm{in}}$, $W^{i\pm}_{\nu,\mathrm{out}}$ the positive $(+)$ or negative $(-)$ i-th waves in u^ν entering or exiting Γ, and let $W^i_{\nu,\mathrm{in}} = W^{i+}_{\nu,\mathrm{in}} - W^{i-}_{\nu,\mathrm{in}}$, $W^i_{\nu,\mathrm{out}} = W^{i+}_{\nu,\mathrm{out}} - W^{i-}_{\nu,\mathrm{out}}$. Then the measure of interaction and the measure of interaction-cancellation control the difference between the amount of exiting i-th waves and the amount of entering i-th waves w.r.t. the region as follows:

$$|W^i_{\nu,\mathrm{out}} - W^i_{\nu,\mathrm{in}}| \leq \mathcal{O}(1)\mu^{\mathrm{I}}_\nu(\Gamma), \tag{26a}$$

$$|W^{i\pm}_{\nu,\mathrm{out}} - W^{i\pm}_{\nu,\mathrm{in}}| \leq \mathcal{O}(1)\mu^{\mathrm{IC}}_\nu(\Gamma). \tag{26b}$$

The above estimates are fairly easy consequence of the interaction estimates (23) and the definition of μ^{I}_ν, μ^{IC}_ν.

By taking a subsequence and using the weak compactness of bounded measures, there exist measures μ^I and μ^{IC} on $\mathbb{R}^+ \times \mathbb{R}$ such that the following weak convergence holds:

$$\mu^{\mathrm{I}}_\nu \rightharpoonup \mu^{\mathrm{I}}, \quad \mu^{\mathrm{IC}}_\nu \rightharpoonup \mu^{\mathrm{IC}}. \tag{27}$$

5.2. Jump part of i-th waves

The derivative of u^ν is clearly concentrated on polygonal lines, being a piecewise constant function with discontinuities along lines. Suppose the i-th family is genuinely nonlinear. To select the wave fronts belonging to i-th family of u^ν converging to the jump part of u, we use the following definition.

DEFINITION 5.2 (Maximal (ϵ^0, ϵ^1)-shock front). *[9] A maximal (ϵ^0, ϵ^1)-shock front for the i-th family of an ϵ_ν-approximate front-tracking solution u^ν is any maximal (w.r.t. inclusion) polygonal line $(t, \gamma^\nu(t))$ in the (t, x)-plane, $t_0 \le t \le t_1$, satisfying:*

(i) the segments of γ^ν are i-shocks of u^ν with size $|s^\nu| \ge \epsilon^0$, and at least once $|s^\nu| \ge \epsilon^1$;

(ii) the nodes are interaction points of u^ν;

(iii) it is on the left of any other polygonal line which it intersects and which have the above two properties.

Let $M^{\nu,i}_{(\epsilon^0,\epsilon^1)}$ be the number of maximal (ϵ^0, ϵ^1)-shock front for the i-th family. Denote

$$\gamma^{\nu,i}_{(\epsilon^0,\epsilon^1),m} : \left[t^{\nu,i,-}_{(\epsilon^0,\epsilon^1),m}, t^{\nu,i,+}_{(\epsilon^0,\epsilon^1),m}\right] \to \mathbb{R}, \quad m = 1, \dots, M^{\nu,i}_{(\epsilon^0,\epsilon^1)},$$

as the maximal (ϵ^0, ϵ^1)-shock fronts for the i-th family in u^ν. Up to a subsequence, we can assume that $M^{\nu,i}_{(\epsilon^0,\epsilon^1)} = \bar{M}^i_{(\epsilon^0,\epsilon^1)}$ is a constant independent of ν because the total variations of u^ν are bounded.

Consider the collection of all maximal (ϵ^0, ϵ^1)-shocks for the i-th family and define

$$\mathscr{T}^{\nu,i}_{(\epsilon^0,\epsilon^1)} = \bigcup_{m=1}^{\bar{M}^i_{(\epsilon^0,\epsilon^1)}} \mathrm{Graph}\left(\gamma^{\nu,i}_{(\epsilon^0,\epsilon^1),m}\right),$$

and let $\{\epsilon^0_k\}_{k\in\mathbb{N}}$, $\{\epsilon^1_k\}_{k\in\mathbb{N}}$ be two sequences satisfying $0 < 2^k\epsilon^0_k \le \epsilon^1_k \searrow 0$.

Up to a diagonal argument and by a suitable labeling of the curves, one can assume that for each fixed k, m the Lipschitz curves $\gamma^{\nu,i}_{(\epsilon^0_k,\epsilon^1_k),m}$ converge uniformly to a Lipschitz curve $\gamma^i_{(\epsilon^0_k,\epsilon^1_k),m}$. Let

$$\mathscr{T}^i := \bigcup_{m,k} \mathrm{Graph}\left(\gamma^i_{(\epsilon^0_k,\epsilon^1_k),m}\right).$$

denote the collection of all these limiting curves in u.

For fixed (ϵ^0, ϵ^1), we write for shortness

$$\tilde{l}^\nu_i(t, x) := \tilde{l}_i(u^\nu(t, x-), u^\nu(t, x+)) \tag{28}$$

and define

$$v^{\nu,\mathrm{jump}}_{i,(\epsilon^0,\epsilon^1)} := \tilde{l}^{\nu}_i \cdot u^{\nu}_x \llcorner_{\mathscr{T}^{\nu,i}_{(\epsilon^0,\epsilon^1)}}. \tag{29}$$

Following the same idea of the proof of [9, Theorem 10.4], the next lemma holds if only the i-th characteristic field is genuinely nonlinear.

LEMMA 5.3. *The jump part of v_i is concentrated on $\mathscr{T}^i$.*

Moreover there exists a countable set $\Theta \subset \mathbb{R}^+ \times \mathbb{R}$, such that for each point

$$P = (\tau, \xi) = (\tau, \gamma^i_m(\tau)) \notin \Theta$$

where i-th shock curve γ^i_m is approximated by the sequence of (ϵ^0, ϵ^1)-shock fronts $\gamma^{\nu,i}_{(\epsilon^0,\epsilon^1),m}$ of the approximate solutions u^ν, the following holds

$$\lim_{r\to 0+} \limsup_{\nu\to\infty} \left(\sup_{\substack{x<\gamma^{\nu,i}_{(\epsilon^0,\epsilon^1),m}(t) \\ (t,x)\in B(P,r)}} \left|u^\nu(t,x) - u^-\right| \right) = 0, \tag{30a}$$

$$\lim_{r\to 0+} \limsup_{\nu\to\infty} \left(\sup_{\substack{x>\gamma^{\nu,i}_{(\epsilon^0,\epsilon^1),m}(t) \\ (t,x)\in B(P,r)}} \left|u^\nu(t,x) - u^+\right| \right) = 0. \tag{30b}$$

Moreover, we can choose a sequence $\{\nu_k\}_{k=1}^{\infty}$ such that

$$v^{\mathrm{jump}}_i = \mathrm{weak}^* - \lim_k \sum_{i=1}^{N} v^{\nu_k,\mathrm{jump}}_{i,(\epsilon^0_k,\epsilon^1_k)}. \tag{31}$$

The key argument of the proof is that we can use the tools of the proof of [9, Theorem 10.4] because the wave structure of the i-th genuinely nonlinear family has the following properties:

1. the interaction among two shocks of the i-th family generates only one shock of the i-th family,
2. the strength of i-th waves can be measured by the jump of the i-th characteristic speed λ_i,
3. the speed of i-th waves is very close to the average of the jump of λ_i across the discontinuity.

These properties are a direct consequence of the behavior of the approximate Riemann solvers on the i-th waves if the i-th family is genuinely nonlinear (Section 5.1.2).

Before proving the lemma, we recall some definitions which will be used in the proof.

DEFINITION 5.4 ([9], Definition 7.2). *Let $\hat\lambda$ be a constant larger than the absolute value of all characteristic speed. We say a curve $x = y(t),\ t \in [a,b]$ is* space-like *if*

$$|y(t_2) - y(t_1)| > \hat\lambda(t_2 - t_1) \quad \text{for all } a < t_1 < t_2 < b.$$

We recall that a *minimal generalized i-characteristic* is an absolutely continuous curve starting from (t_0, x_0) satisfying the differential inclusion

$$x^\nu(t; t_0, x_0) := \min \left\{ \begin{array}{c} x^\nu(t) : x^\nu(t_0) = x_0\,, \\ \dot x^\nu(t) \in \big[\lambda_i\big(u^\nu(t, x(t)+)\big), \lambda_i\big(u^\nu(t, x(t)-)\big)\big] \end{array} \right\}$$

for a.e. $t \geq t_0$.

For any given $(T, \bar x) \in \mathbb{R}$, we consider the minimal (maximal) generalized i-characteristic through $(T, \bar x)$, defined as

$$\chi^{-(+)}(t) = \min(\max)\{\chi(t) : \chi \text{ is a generalized } i\text{-characteristic, } \chi(T) = \bar x\}.$$

From the properties of approximate solutions, we conclude that there is no wave-front of i-th family crossing χ^+ from the left or crossing χ^- from the right.

Sketch of the proof. Let Θ be the set defined by all jump points of the initial datum, the atoms of $\mu^{\rm IC}$ (see (27)). For any point $P \in \mathscr{T}^i \setminus \Theta$, if (30a) or (30b) does not hold, then this means that the approximate solutions u^ν have some uniform oscillation: Indeed, if (30a) not true , there exist $P_\nu,\ Q_\nu \to P$ and $P_\nu,\ Q_\nu$ on the left of $\gamma^{\nu,i}_{(\epsilon^0,\epsilon^1),m}$, $\overline{P_\nu Q_\nu}$ is space-like such that

$$u(P_\nu) \to u^-$$

and

$$|u_\nu(P_\nu) - u_\nu(Q_\nu)| \geq \epsilon,$$

for some constant $\epsilon > 0$. It is not restrictive to assume that the direction $\overrightarrow{P_\nu Q_\nu}$ towards $\gamma^{\nu,i}_{(\epsilon^0,\epsilon^1),m}$. Let $\Lambda_k(\overline{P_\nu Q_\nu})$ be the total wave strength of fronts of k-th family which across the segment $\overline{P_\nu Q_\nu}$. Then, one has $\Lambda_j(\overline{P_\nu Q_\nu}) \geq c\epsilon$ for some $j \in \{1, \cdots, d\}$ and some constant $c > 0$. We consider three cases.

1 $j > i$, we take the minimal forward generalized j-characteristic χ^+ through P_ν and maximal generalized j-characteristic χ^- through Q_ν.

If χ^+ and χ^- interact with each other at O_ν before hitting $\gamma^{\nu,i}_{(\epsilon^0,\epsilon^1),m}$, we consider the region Γ_ν bounded by $\overline{P_\nu Q_\nu}$, χ^+ and χ^-. Since no fronts can leave Γ_ν through χ^+ and χ^-. By (20) and (25), one obtains that there exists a constant $c_1 > 0$ such that $\mu^{IC}_\nu(\Gamma_\nu) \geq c_1\epsilon^2$.

If χ^+ interacts with $\gamma^{\nu,i}_{(\epsilon^0,\epsilon^1),m}$ at A_ν and χ^- interact $\gamma^{\nu,i}_{(\epsilon^0,\epsilon^1),m}$ at B_ν, we consider the region Γ_ν bounded by $\overline{P_\nu Q_\nu}$, χ^+, χ^- and $\gamma^{\nu,i}_{(\epsilon^0,\epsilon^1),m}$. Then

either there exists a constant $0 < c_0' < 1$ such that $\mu_\nu^{IC}(\Gamma_\nu) > c_0'\epsilon$ or there exists a constant $0 < c_0'' < 1$ such that fronts with total strength lager than $c_0''\epsilon_0$ hit $\overline{A_\nu B_\nu}$. By (20) and the fact that each front on $\gamma^{\nu,i}_{(\epsilon^0,\epsilon^1),m}$ has strength less than $-\epsilon_0$, we determine that $\mu_\nu^{IC}(\bar{\Gamma}_\nu) \geq c_0\epsilon\epsilon_0$ on the closure of Γ_ν.

Thus, let $B(P, r_\nu)$ be a ball with center at P containing Γ_ν with radius $r_\nu \to 0$ as $\nu \to 0$. This implies that $\mu^{IC}(\{P\}) > 0$ against the assumption $P \notin \Theta$.

2 $j < i$, we consider the minimal backward generalized j-characteristic through the point P_ν and the maximal backward generalized j-characteristic through the point Q_ν. Then by the similar argument for the case $j > i$, we get $\mu^{IC}(\{P\}) > 0$ against the assumptions.

3 $j = i$ and for any $j' \neq i,\ 1 \leq j' \leq N,\ \Lambda_{j'}(\overline{P_\nu Q_\nu}) \to 0$ as $\nu \to \infty$. In this case, suppose that $\overline{P_\nu Q_\nu}$ intersects the curve $\gamma^{\nu,i}_{(\epsilon^0,\epsilon^1),m}$ at B_ν. Because of genuine nonlinearity, the minimal generalized i-characteristic χ through P_ν will hit $\gamma^{\nu,i}_{(\epsilon^0,\epsilon^1),m}$ if no previous large interactions or cancellations occur on $\gamma^{\nu,i}_{(\epsilon^0,\epsilon^1),m}$. We consider the triangle region Γ_ν bounded by the segment $P_\nu B_\nu$, the curve $\gamma^{\nu,i}_{(\epsilon^0,\epsilon^1),m}$ and χ. Since no fronts of ith-family can exit from Γ_ν through χ, one obtains $\mu_\nu^{IC}(\Gamma_\nu)$ uniformly positive which contradicts the assumption $\mu^{IC}(P) = 0$.

Therefore, we conclude that (30a) is true. And (30b) is similar to prove.

For $P \notin \mathscr{T}^i \cup \Theta$, if $v_i^{\text{jump}}(P) > 0$, i.e. P is a jump point of u, by the similar argument of Step 8 in the proof of [9, Theorem 10.4], the waves present in the approximate solutions are canceled, and thus $\mu^{\text{IC}}(P) > 0$. It is impossible since $P \notin \Theta$. This concludes that v_i^{jump} is concentrated on $\mathscr{T}^i$, because by (30) the jumps in the approximate solutions are vanishing in a neighborhood of every $P \notin \mathscr{T}^i \cup \Theta$.

We are left with the proof of (31). At jump point $(t, \gamma^i_{(\epsilon^0,\epsilon^1),m}(t)) \in \mathscr{T}^i \setminus \Theta$, according to (30a), (30b), there exist a sequence $(t^\nu, \gamma^{\nu,i}_{(\epsilon^0,\epsilon^1),m}(t)(t^\nu))$ such that

$$\big(t, \gamma^i_{(\epsilon^0,\epsilon^1),m}(t)\big) = \lim_{\nu\to\infty} \big(t^\nu, \gamma^{\nu,i}_{(\epsilon^0,\epsilon^1),m}(t)(t^\nu)\big) \tag{32}$$

and its left and right values converges to the left and right values of the jump in $(t, \gamma^i_{(\epsilon^0,\epsilon^1),m}(t))$.

Since $f \in C^2$, by the definition (14) the matrix $A(u^{\text{L}}, u^{\text{R}})$ depends continuously on the value $(u^{\text{L}}, u^{\text{R}})$, and since its eigenvalues are uniformly separated the same continuity holds for its eigenvalues $\tilde{\lambda}_k(u^{\text{L}}, u^{\text{R}})$, left eigenvectors $\tilde{l}_k(u^{\text{L}}, u^{\text{R}})$ and right eigenvectors $\tilde{r}_k(u^{\text{L}}, u^{\text{R}})$. Using the notation (15a) and (28),

one obtains

$$\tilde{l}_i\big(t,\gamma^i_{(\epsilon^0,\epsilon^1),m}(t)\big)=\lim_\nu \tilde{l}^\nu_i\big(t^\nu,\gamma^{\nu,i}_{(\epsilon^0,\epsilon^1),m}(t^\nu)\big), \tag{33}$$

and similar limits hold for $\tilde{r}_i$, $\tilde{\lambda}_i$.

Up to a subsequence $\{\nu_k\}$, from the convergence of the graphs of $\mathscr{T}^{\nu_k,i}_{(\epsilon^0_k,\epsilon^1_k)}$ to $\mathscr{T}^i$ and (30a), (30b), it is fairly easy to prove that

$$Du\llcorner_{\mathscr{T}^i}=\text{weak}^*-\lim_{k\to\infty} Du^{\nu_k}\llcorner_{\mathscr{T}^{\nu_k,i}_{(\epsilon^0_k,\epsilon^1_k)}}. \tag{34}$$

According to (29), (33) and (34), one concludes the weak convergence of $v^{\nu_k,\text{jump}}_{i,(\epsilon^0_k,\epsilon^1_k)}$ to v^{jump}_i. □

6. Proof of Theorem 4.1

6.1. Decay estimate for positive waves

The Glimm Functional for BV functions to general systems has been obtained in [4], and when u is piecewise constant, it reduced to (22): and we will write it as $\mathcal{Q}$ also the formulation of the functional given in [4]. Moreover, for the same constant $C_0>0$ of the Glimm Functional $\Upsilon(t)$ (24), the sum Tot.Var.$(u)+C_0\mathcal{Q}(u)$ is lower semi-continuous w.r.t the L^1 norm (see [9, Theorem 10.1]).

For any Radon measure μ, we denote $[\mu]^+$ and $[\mu]^-$ as the positive and negative part of μ according to Hahn-Jordan decomposition. The same proof of the decay of the Glimm Functional $\Upsilon(t)$ yields that for every finite union of the open intervals $J=I_1\cup\cdots\cup I_m$

$$[v_i]^\pm(J)+C_0\mathcal{Q}(u)\le\liminf_{\nu\to\infty}\left\{[v^\nu_i]^\pm(J)+C_0\mathcal{Q}(u^\nu)\right\},\quad i=1,\dots,\ n, \tag{35}$$

as $u^\nu\to u$ in L^1.

In [9, 10] the authors prove a decay estimate for positive part of the i-th wave measure under the assumption that i-th characteristic field is genuinely nonlinear and the other characteristic fields are either genuinely nonlinear or linearly degenerate. In [12], a sharp decay estimate for positive waves is also given under the same assumptions as those in [9, 10]. By inspection, one can verify that the proof also works (with a little modification) under no assumptions on the nonlinearity on the other characteristic fields, since the essential requirements of strict hyperbolicity and of the controllability of interaction amounts by Glimm Potential still hold: the main variation is that one should replace the original Glimm Potential in [9] with the generalized one given in [4].

We thus state the following theorem, which is the analog of [9, Theorem 10.3].

THEOREM 6.1. *Let the system* (1) *be strictly hyperbolic and the i-th characteristic field be genuinely non-linear. Then there exists a constant C'' such that, for every $0 \le s < t$ and every solution u with small total variation obtained as the limit of wave-front tracking approximation, the measure $[v_i(t)]^+$ satisfies*

$$[v_i(t)]^+(B) \le C'' \left\{ \frac{\mathcal{L}^1}{t-s}(B) + [\mathcal{Q}(s) - \mathcal{Q}(t)] \right\} \tag{36}$$

for every B Borel set in $\mathbb{R}$.

The estimate (36) gives half of the bound (18). In this section, we always assume that the i-th family is genuinely nonlinear.

6.2. Decay estimate for negative waves

To simplify the notation, we omit the index (ϵ^0, ϵ^1) in $v^{\nu,\mathrm{jump}}_{i,(\epsilon^0,\epsilon^1)}$ in the rest of the proof. In order to get the uniform estimate for the *continuous part* $v_i^{\nu,\mathrm{cont}} := v_i^\nu - v_i^{\nu,\mathrm{jump}}$, we need to consider the distributions

$$\mu_i^\nu := \partial_t v_i^\nu + \partial_x(\tilde\lambda_i^\nu v_i^\nu), \qquad \mu_i^{\nu,\mathrm{jump}} := \partial_t v_i^{\nu,\mathrm{jump}} + \partial_x(\tilde\lambda_i^\nu v_i^{\nu,\mathrm{jump}}).$$

6.2.1. Estimate for the source

Let $y_m : [\tau_m^-, \tau_m^+] \to \mathbb{R}$, $m = 1, \dots, L^\nu$, be time-parameterized segments whose graphs are the i-th wave-fronts of u^ν and define

$$u_m^{\mathrm{L}} := u(t, y_m(t)-), \quad u_m^{\mathrm{R}} = u(t, y_m(t)+), \qquad t \in]\tau_m^-, \tau_m^+[.$$

For any test function $\phi \in C_c^\infty(\mathbb{R}^+ \times \mathbb{R})$ one obtains

$$-\int_{\mathbb{R}^+\times\mathbb{R}} \phi \, d\mu_i^\nu = \sum_{m=1}^{L^\nu} \big[\phi(\tau_m^+, y_m(\tau_m^+)) - \phi(\tau_m^-, y_m(\tau_m^-)\big] \tilde l_i \cdot (u_m^{\mathrm{R}} - u_m^{\mathrm{L}}). \tag{37}$$

For any m, since the i-th characteristic field is genuinely nonlinear, one has

$$|\tilde l_i(u^{\mathrm{L}}, u^{\mathrm{R}}) - l_i(u^{\mathrm{L}})| = \mathcal{O}(1)|u_m^{\mathrm{R}} - u_m^{\mathrm{L}}|,$$

where $u_m^{\mathrm{R}} = T^i_{s_i}[u_m^{\mathrm{L}}]$ for some size s_i. Then it follows from (11) that

$$s_i \cong \tilde l_i \cdot (u_m^{\mathrm{R}} - u_m^{\mathrm{L}}). \tag{38}$$

Let $\{(t_k, x_k)\}_k$ be the collection of points where the i-th fronts interact. The computation (37) yields that μ_i^ν concentrates on the interaction points, i.e.

$$\mu_i^\nu = \sum_k p_k \delta_{(t_k, x_k)},$$

where p_k is the difference between the strength of the i-th waves leaving at (t_k, x_k) and the i-th waves arriving at (t_k, x_k). We estimate the quantity p_k depending on the type of interaction:

Since in [8], it is proved that the total size of nonphysical wave-fronts are of the same order of ϵ_ν, when decomposing u^ν_x, we only consider the physical fronts. If at (t_k, x_k), two physical fronts with i-th component size s'_i, s''_i interact and generate an i-th wave or a rarefaction fan with total size $s_i = \sum_m s_i^m$, from (37) and (38), one has

$$p_k \cong s_i - s'_i - s''. \tag{39}$$

Notice that s' or s'' or both may vanish in (39) if one of incoming physical fronts does not belong to the i-th family.

According to the estimate in [3, Lemma 1], the difference of sizes between the incoming and outgoing waves of the same family is controlled by the amount of interaction (see Section 5.1.3), so that one concludes

$$|\mu^\nu_i|(\{(t_k, x_k)\}) \le \mathcal{O}(1)\mathcal{I}(s_i, s'_i)$$

and thus

$$|\mu^\nu_i|(\{t_k\} \times \mathbb{R}) \le \mathcal{O}(1)\{\Upsilon^\nu(t_k^-) - \Upsilon^\nu(t_k^+)\}.$$

This yields

$$|\mu^\nu_i|(\mathbb{R}^+ \times \mathbb{R}) \le \mathcal{O}(1)\Upsilon^\nu(0),$$

i.e. $|\mu^\nu_i|$ is a finite Radon measure.

6.2.2. Estimate for the jump part

Let $\gamma^i_m : [\tau^-_m, \tau^+_m] \to \mathbb{R}$, $m = 1, \dots, \bar{M}^i_{(\epsilon^0, \epsilon^1)}$, be the curves whose graphs are the segments supporting the fronts of u^ν belonging to $\mathscr{T}^{\nu,i}_{(\epsilon^0,\epsilon^1)}$, and write

$$u^{\mathrm{L}}_m := u\big(t, \gamma^i_m(t) - \big), \quad u^{\mathrm{R}}_m := u\big(t, \gamma^i_m(t) + \big), \qquad t \in]\tau^-_m, \tau^+_m[.$$

For any test function $\phi \in C^\infty_c(\mathbb{R}^+ \times \mathbb{R})$ by direct computation one has as in (37)

$$-\int_{\mathbb{R}^+\times\mathbb{R}} \phi \, d\mu^{\nu,\mathrm{jump}}_i = \sum_{m=1}^{\bar{M}^i_{(\epsilon^0,\epsilon^1)}} \big[\phi(\tau^+_m, y_m(\tau^+_m)) - \phi(\tau^-_m, y_m(\tau^-_m))\big]\tilde{l}_i \cdot (u^{\mathrm{R}} - u^{\mathrm{L}}),$$

which yields

$$\mu^{\nu,\mathrm{jump}}_i = \sum_k q_k \delta_{(\tau_k, x_k)},$$

where (τ_k, x_k) are the nodes of the jumps in $\mathscr{T}^{\nu,i}_{(\epsilon^0,\epsilon^1)}$ and the quantities q_k can be computed as follows: if the i-th incoming waves have sizes s' and s'', and the outgoing i-th shock has size s, then (see [8])

$$q_k \cong \begin{cases} -s' & (t_k, x_k) \text{ terminal point of a front not merging} \\ & \qquad \text{into another front,} \\ s & (t_k, x_k) \text{ initial point of a maximal front,} \\ s - s' - s'' & (t_k, x_k) \text{ is a triple point of } \mathscr{T}^{\nu,i}_{(\epsilon^0,\epsilon^1)}\ , \\ s - s' & (t_k, x_k) \text{ interaction point of a front with waves} \\ & \qquad \text{not belonging to } \mathscr{T}^{\nu,i}_{(\epsilon^0,\epsilon^1)}. \end{cases} \tag{40}$$

In fact, since $s \leq 0$ on shocks the second case of (40) implies $q_k \leq 0$. For the triple point, one has that

$$q_k \leq \mu^{\mathrm{IC}}_\nu(\tau_k, x_k).$$

When a shock front in $\mathscr{T}^{\nu,i}_{(\epsilon^0,\epsilon^1)}$ interacts with a front not belonging to $\mathscr{T}^{\nu,i}_{(\epsilon^0,\epsilon^1)}$, there are three situations:

- It interacts with a rarefaction front of i-th family, then one has by the interaction estimates
$$q_k \leq \mu^{\mathrm{IC}}_\nu(\tau_k, x_k).$$
- It interacts with a front of different family, then also one gets
$$q_k \leq \mu^{\mathrm{I}}_\nu(\tau_k, x_k).$$
- It interacts with a shock of i-th family which does not belong to $\mathscr{T}^{\nu,i}_{(\epsilon^0,\epsilon^1)}$, then
$$q_k \leq 0.$$

Suppose now that (τ_k, x_k) is a terminal point of an (ϵ^0, ϵ^1)-shock front γ_m. By the definition of (ϵ^0, ϵ^1)-shock, for some $t \leq \tau_k$ the shock front γ_m has size $s_0 \leq -\epsilon^1$, and at (τ_k, x_k) the size s_1 of the outgoing i-th front must be not less than $-\epsilon^0$ as a result of interaction between two wave-fronts belonging to different family or cancellation between two wave-fronts belonging to the same family along γ_k. Hence we obtain

$$\epsilon^1 - \epsilon^0 \leq |s_0| - |s_1| \leq \mathcal{O}(1)\mu^{\mathrm{IC}}_\nu(\gamma_k).$$

This yields

$$\begin{aligned} q_k \cong & - s_1 + (s_1 + q_k) \\ \leq & \frac{\epsilon^0}{\epsilon^1 - \epsilon^0}(\epsilon^1 - \epsilon^0) + \mathcal{O}(1)\mu^{\mathrm{I}}_\nu(t_k, x_k) \leq \frac{\mathcal{O}(1)\epsilon^0}{\epsilon^1 - \epsilon^0}\mu^{\mathrm{IC}}_\nu(\gamma_k) + \mathcal{O}(1)\mu^{\mathrm{I}}_\nu(t_k, x_k). \end{aligned}$$

Since the end points correspond to disjoint maximal i-th fronts, due to genuinely nonlinearity, it follows that

$$\sum_{(t_k,x_k)\text{ end point}} q_k \leq \mathcal{O}(1)\mu_\nu^{\mathrm{IC}}(\mathbb{R}^+\times\mathbb{R}),$$

so that it is a uniformly bounded measure. We thus conclude that the distribution

$$\bar\mu^\nu := -\mu_i^{\nu,\mathrm{jump}} + \mathcal{O}(1)\mu_\nu^{\mathrm{IC}} + \sum_{(t_k,x_k)\text{ end point}} q_k\delta_{(t_k,x_k)}$$

is non-negative, so it is a Radon measure and thus also $\mu_i^{\nu,\mathrm{jump}}$ is a Radon measure.

In order to obtain a lower bound, one considers the Lipschitz continuous test function

$$\phi_\alpha(t) := \chi_{[0,T+\alpha]}(t) - \frac{t-T}{\alpha}\chi_{[T,T+\alpha]}(t), \quad \alpha>0,$$

which is allowed because v_i^ν is a bounded measure. Since $\bar\mu$ is non-negative, one obtains

$$\begin{aligned}
&\bar\mu^\nu\big([0,T]\times\mathbb{R}\big) \leq \int_{\mathbb{R}^+\times\mathbb{R}} \phi_\alpha d\bar\mu\\
=\ & -\int_{\mathbb{R}^+\times\mathbb{R}} \phi_\alpha d\mu_i^{\nu,\mathrm{jump}} + \mathcal{O}(1)\int_{\mathbb{R}^+\times\mathbb{R}} \phi_\alpha d\mu_\nu^{\mathrm{IC}} + \sum_{(t_k,x_k)\text{ end point}} q_k\phi_\alpha(t_k)\\
\leq\ & \int_{\mathbb{R}^+\times\mathbb{R}} \big[(\phi_\alpha)_t + \tilde\lambda_i^\nu(\phi_\alpha)_x\big] d\big[v_i^{\nu,\mathrm{jump}}(t)\big]dt + \big[v_i^{\nu,\mathrm{jump}}(0)\big](\mathbb{R})\\
&\qquad + \mathcal{O}(1)\mu_\nu^{\mathrm{IC}}\big([0,T+\alpha]\times\mathbb{R}\big)\\
\leq\ & -\frac{1}{\alpha}\int_T^{T+\alpha} \big[v_i^{\nu,\mathrm{jump}}(t)\big](\mathbb{R})dt + \big[v_i^{\nu,\mathrm{jump}}(0)\big](\mathbb{R}) + \mathcal{O}(1)\mu_\nu^{\mathrm{IC}}\big([0,T+\alpha]\times\mathbb{R}\big).
\end{aligned}$$

Letting $\alpha\searrow 0$ and since $[v_i^{\nu,\mathrm{jump}}(\mathbb{R})](0)$ is negative, one concludes

$$\bar\mu^\nu\big([0,T]\times\mathbb{R}\big) \leq -\big[v_i^{\nu,\mathrm{jump}}(T)\big](\mathbb{R}) + \mathcal{O}(1)\mu_\nu^{\mathrm{IC}}\big([0,T+\alpha]\times\mathbb{R}\big) \leq \mathcal{O}(1)\Upsilon^\nu(0).$$

We conclude this section by writing the uniform estimate

$$-\mathcal{O}(1)\Upsilon^\nu(0) \leq \mu_i^{\nu,\mathrm{jump}} \leq \mathcal{O}(1)\mu_\nu^{\mathrm{IC}}.$$

In particular, the definitions of the measures μ_i^ν, $\mu_i^{\nu,\mathrm{jump}}$ give the following balances for the i-th waves across the horizontal lines:

$$\big[v_i^\nu(t+)\big](\mathbb{R}) - \big[v_i^\nu(t-)\big](\mathbb{R}) = \mu_i^\nu\big(\{t\}\times\mathbb{R}\big), \tag{41a}$$

$$\big[v_i^{\nu,\mathrm{jump}}(t+)\big](\mathbb{R}) - \big[v_i^{\nu,\mathrm{jump}}(t-)\big](\mathbb{R}) = \mu_i^{\nu,\mathrm{jump}}\big(\{t\}\times\mathbb{R}\big). \tag{41b}$$

The limits are taken in the weak topology. Notice that we can always take that $t\mapsto v_i^\nu(t), v_i^{\nu,\mathrm{jump}}(t)$ is right continuous in the weak topology.

6.2.3. Balances of waves in the region bounded by generalized characteristics

Given an interval $I = [a, b]$, we define the region $A^{\nu,(t_0,\tau)}_{[a,b]}$ bounded by the minimal i-th characteristics $a(t)$, $b(t)$ of u^ν starting at (t_0, a) and (t_0, b) by

$$A^{\nu,(t_0,\tau)}_{[a,b]} := \Big\{(t,x) : t_0 < t \le t_0 + \tau,\ a(t) \le x \le b(t)\Big\},$$

and its time-section by $I(t) := [a(t), b(t)]$. Let $J := I_1 \cup I_2 \cup \cdots \cup I_M$ be the union of the disjoint closed intervals $\{I_i\}_{i=1}^M$, and set

$$J(t) := I_1(t) \cup \cdots \cup I_M(t), \quad A^{\nu,(t_0,\tau)}_J := \bigcup_{m=1}^M A^{\nu,(t_0,\tau)}_{I_m}.$$

We will now obtain wave balances in regions of the form $A^{\nu,(t_0,\tau)}_J$. Due to the genuinely non-linearity of the i-th family, the corresponding proof in [8] works, we will repeat it for completeness.

The balance on the region $A^{\nu,(t_0,\tau)}_J$ has to take into account also the contribution of the flux Φ^ν_i across boundaries of the segments $I_m(t)$: due to the definition of generalized characteristic and the wave-front approximation, it follows that Φ^ν_i is an atomic measure on the characteristics forming the border of $A^{\nu,(t_0,\tau)}_J$, and moreover a positive wave may enter the domain $A^{\nu,(t_0,\tau)}_J$ only if an interaction occurs at the boundary point $(\hat t, \hat x)$, which gives the estimate

$$\Phi^\nu_i\big(\{(\hat t, \hat x)\}\big) \le \mathcal{O}(1)\mu^{\mathrm{IC}}_i\big(\{(\hat t, \hat x)\}\big). \tag{42}$$

One thus obtains that

$$\big[v^\nu_i(\tau)\big](J(\tau)) - \big[v^\nu_i(t_0)\big](J) = \mu^\nu_i\big(A^{\nu,(t_0,\tau)}_J\big) + \Phi^\nu_i\big(A^{\nu,(t_0,\tau)}_J\big) + \mathcal{O}(1)\epsilon_\nu, \tag{43}$$

where the last term depends on the errors due to the wave-front approximation (a single rarefaction front may exit the interval I_m at t_0).

The same computation can be done for the jump part $v^{\nu,\mathrm{jump}}_i$, obtaining

$$\begin{aligned}&\big[v^{\nu,\mathrm{jump}}_i(\tau)\big](J(t)) - \big[v^{\nu,\mathrm{jump}}_i(t_0)\big](J)\\ &\quad = \mu^{\nu,\mathrm{jump}}_i\big(A^{\nu,(t_0,\tau)}_J\big) + \Phi^{\nu,\mathrm{jump}}_i\big(A^{\nu,(t_0,\tau)}_J\big).\end{aligned} \tag{44}$$

Since the flux $\Phi^{\nu,\mathrm{jump}}_i$ only involves the contribution of (ϵ^0, ϵ^1)-shocks, it is clearly non-positive.

Subtracting (44) to (43), one finds the following equation for $v^{\nu,\mathrm{cont}}_i$:

$$\begin{aligned}&\big[v^{\nu,\mathrm{cont}}_i(\tau)\big](J(\tau)) - \big[v^{\nu,\mathrm{cont}}_i(t_0)\big](J)\\ &\quad = \big(\mu^\nu_i - \mu^{\nu,\mathrm{jump}}_i\big)\big(A^{\nu,\tau}_J\big) + \big(\Phi^\nu_i - \Phi^{\nu,\mathrm{jump}}_i\big)\big(A^{\nu,(t_0,\tau)}_J\big) + \mathcal{O}(1)\epsilon_\nu.\end{aligned}$$

Denote the difference between the two fluxes by

$$\Phi_i^{\nu,\mathrm{cont}} := \Phi_i^\nu - \Phi_i^{\nu,\mathrm{jump}}.$$

Since $\Phi_i^{\nu,\mathrm{jump}}$ removes only some terms in the negative part of Φ_i^ν, one concludes that

$$\Phi_i^\nu - \Phi_i^{\nu,\mathrm{jump}} \le \big[\Phi_i^\nu\big]^+ \le \mu_\nu^{\mathrm{IC}}. \tag{45}$$

Setting

$$\mu_{i,\nu}^{\mathrm{ICJ}} := \mu_\nu^{\mathrm{IC}} + \big|\mu_i^{\nu,\mathrm{jump}}\big|,$$

and using the estimate $|\mu_i^\nu| \le \mathcal{O}(1)\mu_\nu^{\mathrm{IC}}$, one has

$$\mu_i^\nu - \mu_i^{\nu,\mathrm{jump}} \le \mathcal{O}(1)\mu_{i,\nu}^{\mathrm{ICJ}}. \tag{46}$$

6.2.4. Decay estimate

Due to the semigroup property of solutions, it is sufficient to prove the estimate for the measure $[v_i^{\nu,\mathrm{cont}}(t=0)]^-$. Consider thus a closed interval $I=[a,b]$ and let $z(t) := b(t) - a(t)$ where

$$a(t) := x^\nu(t;0,a), \quad b(t) := x^\nu(t;0,b)$$

and the minimal forward characteristics stating at $t=0$ from a and b. For $\mathcal{L}^1$-a.e. t one has

$$\dot z(t) = \tilde\lambda_i(t,b(t)) - \tilde\lambda_i(t,a(t)).$$

By introducing a piecewise Lipschitz continuous non-decreasing potential Φ to control the waves on the other families [9], with $\Phi(0)=1$, one obtains

$$\Big|\dot z(t) + \xi(t) - \big[v_i^\nu(t)\big](I(t))\Big| \le \mathcal{O}(1)\epsilon_\nu + \dot\Phi(t)z(t), \tag{47}$$

where

$$\xi(t) := \big(\tilde\lambda_i(t,a(t)+) - \tilde\lambda_i(t,a(t)-)\big) + \big(\tilde\lambda_i(t,b(t)+) - \tilde\lambda_i(t,b(t)-)\big).$$

We consider two cases.

Case 1. If

$$\dot z(t) - \dot\Phi(t)z(t) < \frac{1}{4}\big[v_i^{\nu,cont}(0)\big](I)$$

for all $t>0$, then

$$\begin{aligned}\frac{d}{dt}\Big[e^{-\int_0^t\dot\Phi(s)ds}z(t)\Big] &= e^{-\int_0^t\dot\Phi(s)ds}\big\{\dot z(t) - \dot\Phi(t)z(t)\big\}\\ &< \frac{e^{-\int_0^t\dot\Phi(s)ds}}{4}\big[v_i^{\nu,cont}(0)\big](I).\end{aligned}$$

Integrating the above inequality from 0 to τ and remembering that $\Phi(0) = 1$ and $v_i^{\nu,\text{jump}}(0)$ is non-positive, one has

$$\begin{aligned} -\mathcal{L}^1(I) = -z(0) &\leq e^{-\int_0^\tau \dot{\Phi}(s)ds} z(\tau) - z(0) \\ &\leq \frac{1}{4}\int_0^\tau e^{-\int_0^t \dot{\Phi}(s)ds} d\tau \big[v_i^{\nu,cont}(0)\big](I) \\ &\leq \frac{1}{4}\tau \big[v_i^{\nu,\text{cont}}(0)\big](I), \end{aligned}$$

which reads as

$$-\big[v_i^{\nu,\text{cont}}(0)\big](I) \leq 4\frac{\mathcal{L}^1(I)}{\tau}.$$

Case 2. Assume instead that

$$\dot{z}(\bar{t}) - \dot{\Phi}(\bar{t})z(\bar{t}) \geq \frac{1}{4}\big[v_i^{\nu,cont}(0)\big](I) \tag{48}$$

at some time $\bar{t} > 0$. From (29) and the fact that the i-th family is genuinely nonlinear and the fronts in $\mathscr{T}^{\nu,i}_{(\epsilon^0,\epsilon^1)}$ satisfy Rankine-Hugoniot conditions (up to a negligible error), we have

$$v_i^{\nu,\text{jump}}(t, a(t)) = \lambda_i(t, a(t)+) - \lambda_i(t, a(t)-),$$

Then by the assumption of genuine nonlinearity, we conclude that

$$\begin{aligned} \xi(t) &\geq \frac{3}{4}\Big[\big[v_i^{\nu,\text{jump}}(t)\big](a(t)) + \big[v_i^{\nu,\text{jump}}(t)\big](b(t)) - 2\epsilon^1\Big] \\ &\geq \frac{3}{4}\Big[\big[v_i^{\nu,\text{jump}}(t)\big](I(t)) - 2\epsilon^1\Big]. \end{aligned} \tag{49}$$

As $v^{\nu,\text{jump}}$ is non-positive, (47) and (49) yield that

$$\begin{aligned} \dot{z}(t) - \dot{\Phi}z(t) &\leq \big[v_i^{\nu,\text{cont}}(t)\big](I(t)) + \big[v_i^{\nu,\text{jump}}(t)\big](I(t)) - \xi(t) + \mathcal{O}(1)\epsilon_\nu \\ &\leq \big[v_i^{\nu,\text{cont}}(t)\big](I(t)) + \mathcal{O}(1)\epsilon_\nu + 2\epsilon^1. \end{aligned}$$

Recall the assumption (48), at time $\bar{t}$, we get

$$\big[v_i^{\nu,\text{cont}}(0)\big](I)/4 \leq \big[v_i^{\nu,\text{cont}}(\bar{t})\big](I(\bar{t})) + \mathcal{O}(1)\epsilon_\nu + 2\epsilon^1.$$

By the balance for $v^{\nu,\text{cont}}$ we get in Section 6.2.3, one obtains

$$\big[v_i^{\nu,\text{cont}}(0)\big](I)/4 \leq \big[v_i^{\nu,\text{cont}}(0)\big](I) + \mu_\nu^{\text{ICJ}}\big(A_I^{\nu,(0,\bar{t})}\big) + \mathcal{O}(1)\epsilon_\nu + 2\epsilon^1.$$

Combining the conclusion for the two cases one gets the uniform bound r.w.t ν

$$-\big[v_i^{\nu,\text{cont}}(0)\big](I) \leq \mathcal{O}(1)\bigg\{\frac{\mathcal{L}^1(I)}{t} + \mu_\nu^{\text{ICJ}}\big(A_I^{\nu,(0,t)}\big) + \epsilon^1 + \epsilon_\nu\bigg\}.$$

This gives the estimate (18) for the case of a single interval for the approximate solution.

By analogous computation for the region which is a finite union of intervals, as we have done in Section 6.2.3, one obtains the same bound as above, and since $v_i^{\nu,\mathrm{cont}}$ is a Radon measure, the same result holds for any Borel sets, i.e.

$$-\big[v_i^{\nu,\mathrm{cont}}(0)\big](B) \leq \mathcal{O}(1)\bigg\{\frac{\mathcal{L}^1(B)}{t} + \mu_\nu^{\mathrm{ICJ}}\big(\overline{A_B^{\nu,(0,t)}}\big) + \epsilon^1 + \epsilon_\nu\bigg\},$$

where B is any Borel set in $\mathbb{R}$ and

$$A_B^{\nu,(0,t)} := \Big\{\big(\tau, x^\nu(\tau;0,x_0)\big) :\ x\in B,\ 0<\tau\leq t\Big\}.$$

As the solution is independent on the choice of the approximation, we can consider a particular converging sequence $\{u^\nu\}_{\nu\geq 1}$ of ϵ_ν-approximate solutions with the following additional properties:

$$\mathcal{Q}(u^\nu(0,\cdot)) \to \mathcal{Q}(u_0).$$

By lower semi-continuity of $[v_i(0)]^- + C_0\mathcal{Q}(u(0))$ (35), one gets

$$[v_i(0)]^- + C_0\mathcal{Q}(u(0)) \leq \text{weak}^* - \liminf_{\nu\to\infty}\left\{[v_i^\nu(0)]^- + C_0\mathcal{Q}(u^\nu(0))\right\}. \tag{50}$$

Since $v_i^{\mathrm{jump}}(0)$ has only negative part, from (50) and (31), up to a subsequence, one obtains for any open set $U\subset\mathbb{R}$,

$$\begin{aligned}
&\big[v_i^{\mathrm{cont}}(0)\big]^-(U)\\
&= [v_i(0)]^-(U) + \big[v_i^{\mathrm{jump}}(0)\big](U)\\
&\leq \liminf_{\nu\to\infty}\left\{\big[v_i^\nu(0)\big]^-(U) + C_0\mathcal{Q}(u^\nu(0))\right\} - C_0\mathcal{Q}(u(0)) + \lim_{\nu\to\infty}\big[v_i^{\nu,\mathrm{jump}}(0)\big](U)\\
&= \liminf_{\nu\to\infty}\left\{\big[v_i^{\nu,\mathrm{cont}}(0)\big]^-(U) + C_0\mathcal{Q}(u^\nu(0))\right\} - C_0\mathcal{Q}(u(0))\\
&\leq \liminf_{\nu\to\infty}\mathcal{O}(1)\bigg\{\frac{\mathcal{L}^1(U)}{t} + \mu_i^{\nu,\mathrm{ICJ}}\big(\overline{A_U^{\nu,(0,t)}}\big) + \epsilon^1 + \epsilon_\nu + \mathcal{Q}(u^\nu(0)) - \mathcal{Q}(u(0))\bigg\}\\
&\leq \mathcal{O}(1)\bigg\{\frac{\mathcal{L}^1(U)}{t} + \mu_i^{\mathrm{ICJ}}\big([0,t]\times\mathbb{R}\big)\bigg\},
\end{aligned}$$

where μ_i^{ICJ} is defined as weak*-limit of measure $\mu_i^{\nu,\mathrm{ICJ}}$ (up to a subsequence). Then the outer regularity of Radon measure yields the inequality for any Borel set.

The above estimate together with Theorem 6.1 gives (18).

7. SBV regularity for the i-th component of the i-th eigenvalue

This last section concerns the proof of Theorem 1.2, adapting the strategy of Section 2.

Proof of Theorem 1.2. As in the scalar case, we define the sets

$$\begin{aligned} J_\tau &:= \big\{x \in \mathbb{R} :\ u^{\mathrm{L}}(\tau,x) \neq u^{\mathrm{R}}(\tau,x)\big\},\\ F_\tau &:= \big\{x \in \mathbb{R} :\ \nabla\lambda_i(u(\tau,x)) \cdot r_i(u(\tau,x)) = 0\big\},\\ C &:= \big\{(\tau,\xi) \in \mathbb{R}^+ \times \mathbb{R} :\ \xi \in J_\tau \cup F_\tau\big\}, \quad C_\tau := J_\tau \cup F_\tau. \end{aligned}$$

By definition of continuous part

$$\big|v_i^{\mathrm{cont}}(\tau)\big|(J_\tau) = 0,$$

and since

$$\nabla\lambda_i\big(u(\tau, F_\tau \setminus J_\tau)\big) \cdot r_i\big(u(\tau, F_\tau \setminus J_\tau)\big) = 0,$$

we conclude that

$$\begin{aligned} &\big|\nabla\lambda_i(u) \cdot r_i(u) v_i^{\mathrm{cont}}(\tau)|(C_\tau)\\ &\quad = \big|\nabla\lambda_i(u) \cdot r_i(u) v_i^{\mathrm{cont}}(\tau)\big|(J_\tau) + \big|\nabla\lambda_i(u) \cdot r_i(u) v_i^{\mathrm{cont}}(\tau)\big|(F_\tau \setminus J_\tau)\\ &\quad\quad = 0. \end{aligned}$$

For any $(t_0,x_0) \in \mathbb{R}^+ \times \mathbb{R} \setminus C$, there exist strictly positive $b_0 = b_0(x_0,t_0)$, $c_0 = c_0(x_0,t_0)$ such that

$$\big|\nabla\lambda_i \cdot r_i(u(t_0,x))\big| \geq c_0 > 0$$

for every x in the open interval $I_0 :=]-b_0+x_0, x_0+b_0[$, because $u(t_0,x)$ is continuous at $x_0 \notin C_{t_0}$. Hence by Theorem 4.3, we know that there is a triangle

$$T_0 := \Big\{(t,x) :\ |x-x_0| < b_0' - \bar\eta(t-t_0),\ 0 < t-t_0 < b_0'/\bar\eta\Big\}$$

with the basis $I_0' :=]-b_0'+x_0, x_0+b_0'[\subset I_0$, such that

$$\big|\nabla\lambda_i \cdot r_i(u(t_0,x))\big| \geq \frac{c_0}{2} > 0, \tag{51}$$

by taking $b_0' \ll 1$ in order to have that the total variation remains sufficiently small.

Since $u_{\llcorner T_0}$ coincides with the solution to

$$\begin{cases} \partial_t w + f(w)_x = 0, \\ w(x,t_0) = \begin{cases} u_{t_0}(x) & |x-x_0| < b_0', \\ \frac{1}{2b_0'} \int_{x_0-b_0'}^{x_0+b_0'} u_{t_0}(y)dy & |x-x_0| \geq b_0, \end{cases} \end{cases} \tag{52}$$

and by taking b_0' sufficiently small, we still have that (51) holds for the range of w. In particular w is SBV outside a countable number of times, and the same happens for u in T_0.

As in the scalar case, one thus verifies that there is a countable family of triangles $\{T_i\}_{i=1}^{\infty}$ covering the complement of C outside a set whose projection on the t-axis is countable. The same computation of the scalar case concludes the proof: for any τ chosen as in (5)

$$\begin{aligned} \big|(\nabla\lambda_i \cdot r_i)v_i^{\mathrm{c}}\big|(\mathbb{R}) \leq & \big|(\nabla\lambda_i \cdot r_i)v_i^{\mathrm{c}}\big|(C_\tau) \\ & + \big|(\nabla\lambda_i \cdot r_i)v_i^{\mathrm{c}}\big|\left(\bigcup_i T_i \cap \{t=\tau\}\right) = 0. \end{aligned}$$

Recall the definition (17), we can finally conclude that the i-th component of $D_x\lambda_i(u(t,\cdot))$ has no Cantor part for every $t \in \mathbb{R}^+ \setminus S$ and $i \in \{1,\ 2,\ \ldots,\ N\}$. □

Similar to the scalar case, it is easy to get the following corollary from the Theorem 1.2 and (16).

Corollary 7.1. *Suppose u be a vanishing viscosity solution of the Cauchy problem for the strictly hyperbolic system* (1)-(2). *Let u be the vanishing viscosity solution of the problem* (1), (2). *Then the scalar measure $[D_x\lambda_i(u)]_i$ has no Cantor part in $\mathbb{R}^+ \times \mathbb{R}$.*

Remark 7.2. *As we mentioned in the introduction, it no longer holds the SBV regularity of admissible solution to the general strictly hyperbolic system of conservation laws.*

Consider the following equations

$$\begin{cases} u_t = 0, \\ v_t + ((1+v+u)v)_x = 0. \end{cases}$$

Since $D_x\lambda_2((u,v)) = D_xu + 2D_xv$, then it is clear that $D_x\lambda_2$ can have a Cantor part since the first equation is just trivial which means that the component u is not SBV regular if the initial data is not.

While from Theorem 4.1 we know that the Cantor part of the second component of $D_x\lambda_2(u)$*,*

$$\begin{aligned}[D_x^c\lambda_2(u)]_2 &= \big(D_u\lambda_2\cdot r_2\big)\big(l_2\cdot D_x^c(u,v)\big)\\ &= \frac{2}{1+u+2v}\big(vD_x^cu_x+(1+u+2v)D_x^cv\big)\end{aligned}$$

vanishes. (Notice that since the Cantor part of (D_xu, D_xv) *concentrates on the set of continuous points of* (u,v)*, we do not need to specify the coefficients at the jump points of* (u,v)*.)*

References

[1] L. Ambrosio and C. De Lellis, *A note on admissible solutions of 1d scalar conservation laws and 2d Hamilton-Jacobi equations*, J. Hyperbolic Differ. Equ. **1** (2004), 813–826.

[2] L. Ambrosio, N. Fusco, and D. Pallara, *Functions of bounded variation and free discontinuity problems*, Oxford Clarendon Press, 2000.

[3] F. Ancona and A. Marson, *Existence theory by front tracking for general nonlinear hyperbolic systems*, Arch. Ration. Mech. Anal. **185** (2007), 287–340.

[4] S. Bianchini, *Interaction estimates and Glimm functional for general hyperbolic systems*, Discrete Contin. Dyn. Syst. **9** (2003), 133–166.

[5] S. Bianchini, *On the Riemann problem for non-conservative hyperbolic systems*, Arch. Ration. Mech. Anal. **166** (2003), 1–26.

[6] S. Bianchini, *SBV regularity of genuinely nonlinear hyperbolic systems of conservation laws in one space dimension*, Acta Math. Sci. **32** (2012), 380–388.

[7] S. Bianchini and A. Bressan, *Vanishing viscosity solutions of nonlinear hyperbolic systems*, Ann. of Math. **161** (2005), 223–342.

[8] S. Bianchini and L. Caravenna, *SBV regularity for genuinely nonlinear, strictly hyperbolic systems of conservation laws in one space dimension*, arXiv:1111.6246v1, November 2011.

[9] A. Bressan, *Hyperbolic systems of conservation laws: the one-dimensional Cauchy problem*, Oxford Lecture Series in Mathematics and its Applications, Oxford University Press, USA, 2000.

[10] A. Bressan and R.M. Colombo, *Decay of positive waves in nonlinear systems of conservation laws*, Ann. Scuola Norm. Sup. Pisa Cl. Sci. (4) **26** (1998), 133–160.

[11] A. Bressan and P.G. LeFloch, *Structural stability and regularity of entropy solutions to hyperbolic systems of conservation laws*, Indiana Univ. Math. J. **48** (1999), 43–84.

[12] A. Bressan and T. Yang, *A sharp decay estimate for positive nonlinear waves*, SIAM J. Math. Anal. **36** (2004), 659–677.

[13] C. M. Dafermos, *Generalized characteristics and the structure of solutions of hyperbolic conservation laws*, Indiana Univ. Math. J. **26** (1977), 1097–1119.

[14] C. M. Dafermos, *Hyperbolic conservation laws in continuum physics*, Springer-Verlag, Berlin, 2009.

[15] C. De Lellis, *Hyperbolic equations and SBV functions*, Journées équations aux dérivées partielles **6** (2010), 1–10.
[16] J. Glimm, *Solutions in the large for nonlinear hyperbolic systems of equations*, Comm. Pure Appl. Math. **18** (1965), 697–715.
[17] R. Robyr, *SBV regularity of entropy solutions for a class of genuinely nonlinear scalar balance laws with non-convex flux function*, J. Hyperbolic Differ. Equ. **5** (2008), 449–475.

Authors' addresses:

Stefano Bianchini
SISSA - International School for Advanced Studies
via Bonomea 265, 34136 Trieste, ITALY
E-mail: bianchin@sissa.it

Lei Yu
SISSA - International School for Advanced Studies
via Bonomea 265, 34136 Trieste, ITALY
E-mail: yulei@sissa.it

Received March 1, 2012
Revised September 7, 2012

Rend. Istit. Mat. Univ. Trieste
Volume 44 (2012), 473–502

Strongly inessential elements of a perfect height 2 ideal

Giannina Beccari and Carla Massaza

Abstract. *In this paper we expand on some results exposed in a previous one, in which we introduced the concept of inessential and strongly inessential generators in a standard basis of a saturated homogeneous ideal. The appearance of strongly inessential elements seemed to be a non generic situation; in this paper we analyze their presence in a perfect height 2 ideal with the greatest number of generators, according to Dubreil's inequality.*

Keywords: perfect height 2 ideals, invariants of a standard basis, Hilbert-Burch matrix, Dubreil's inequality
MS Classification 2010: 13C14

1. Introduction

In a previous paper [4] we introduced the concept of *strongly inessential element* (briefly s.i.) in a homogeneous ideal $\mathfrak{I} \subset K[X_1, \ldots, X_n]$. Our first idea, when we started to think about essential and inessential elements of a standard basis (see [4], n.3), was that every homogeneous ideal should have a standard basis consisting of essential forms, but we very soon found many counterexamples. Therefore, our next conjecture was that the assertion might be true for a sufficiently general ideal. In this paper we thus investigate the structure of e-maximal bases ([4], Definition 5.1) and, as a consequence ([4], Theorem 5.1), the presence of s.i. elements, in what seemed to be the easiest situation, that is when $\mathfrak{I}$ is a perfect height 2 ideal. In this case, it is possible to associate to every $\mathcal{B}(\mathfrak{I})$ a Hilbert-Burch matrix ([13, 14]) and to decide the nature of the forms of $\mathcal{B}(\mathfrak{I})$, with respect to essentiality, just looking at the ideals generated by the entries of its columns ([4]).

We observe that, if the multiplicity $e(\mathfrak{I})$ ([10, 11, 15]) is low, our first idea was correct; more precisely, if $e(\mathfrak{I}) < 6$, then every standard basis consists of essential elements, while, if $6 < e(\mathfrak{I}) < 9$, $\mathfrak{I}$ has at least a standard basis whose elements are all essential.

To deal with the problem when the multiplicity is ≥ 9, we observe that strong inessentiality is preserved modulo a regular sequence (while essentiality

is not). So, the first case to be considered seems to be the one of zero depth. As the general case still appears hard to be analyzed, we replace the family of all perfect height 2 ideals with its subfamily $F = \bigcup_{n\geq 2} F[n]$, where $F[n]$ is the set of all perfect height two ideals in $S = K[X_1, \ldots, X_n], n \geq 2$, whose standard bases are of maximal cardinality with respect to Dubreil's inequality ([9]). In a previous paper [3], in fact, we found a description of F that is of help in dealing with the problem considered here. So, as we restrict our attention to the ideals of zero depth, we study $F[2]$. For every ideal $\mathfrak{I} \in F[2]$, we produce a *canonical* Hilbert matrix, with the property that its corresponding basis is e-maximal, which means that its inessential elements are s.i.. Using such a matrix, we prove that the number of s.i. elements appearing in an e-maximal basis is linked to the greatest common divisor Φ of its generators of minimal degree $\alpha(\mathfrak{I})$; in fact, it depends on the decomposition of Φ into linear factors (see Theorem 4.1). More precisely, we prove that $\mathfrak{I}$ has a basis of essential elements iff all the linear factors of Φ are distinct; therefore, the generic $\mathfrak{I} \in F[2]$ has this a property.

The description of the e-maximal bases is much more complicated when we pass from $F[2]$ to $F[3]$. The Hilbert-Burch matrix of any element $\mathfrak{I} \in F[3]$ can be obtained by lifting the Hilbert-Burch matrix of its image $\bar{\mathfrak{I}} \in F[2]$ modulo any linear form, regular for $S/\mathfrak{I}$ ([3]); however, it may happen that there exists some $\widetilde{\mathfrak{I}}$ with the same number of s.i. elements of $\bar{\mathfrak{I}}$ in any e-maximal basis, among the ideals of $F[3]$ lifting $\bar{\mathfrak{I}} \in F[2]$, but there are also cases in which no lifting of $\bar{\mathfrak{I}}$ preserves a s.i. element. We prove that the greatest expected number of s.i. generators in a standard basis of $\mathfrak{I} \in F[3]$ is $\alpha(\mathfrak{I})-2$ and that this number is attained. So, we focus on the set $\mathcal{S} \subset F[3]$ of the ideals with $\alpha(\mathfrak{I})-2$ s.i. generators in their e-maximal bases, finding some of their properties and giving examples. In particular, we completely describe the ideals $\mathfrak{I}$ generated in two different degrees, with $\alpha(\mathfrak{I}) = 3$ and a s.i. element in any e-maximal basis.

2. Background and Notation

Let $S = K[X_1, \ldots, X_n]$, where K is an algebraically closed field, be the coordinate ring of $\mathbb{P}^{n-1}$, $\mathfrak{I} = \bigoplus \mathfrak{I}_d, d \in \mathbf{N}$, a homogeneous ideal of S, and $\mathfrak{M} = (X_1, \ldots, X_n)$ be the irrelevant maximal ideal. We recall some basic definitions.

The Hilbert function of $S/\mathfrak{I}$ ([12]), denoted $H(S/\mathfrak{I}, -)$, is the function defined by

$$H(S/\mathfrak{I}, t) = dim_K(S/\mathfrak{I})_t.$$

It is well known that for $t \gg 0$ the function $H(S/\mathfrak{I}, t)$ is a polynomial, with rational coefficients, of degree $r(S/\mathfrak{I}) - 1$, where $r(S/\mathfrak{I})$ is the Krull dimension of $S/\mathfrak{I}$.

If Δ denotes the difference operator on maps from $\mathbb{Z}$ to $\mathbb{Z}$, defined by $\Delta\phi(t) = \phi(t) - \phi(t-1)$, the function

$$\Gamma(\mathfrak{I}, t) = \Delta^{r(S/\mathfrak{I})} H(S/\mathfrak{I}, t)$$

is called the Castelnuovo function of $\mathfrak{I}$, while $\Delta^{r(S/\mathfrak{I})} H(S/\mathfrak{I}, t)$ is, for large t, a natural number $e(\mathfrak{I})$, independent on t, which is called the *multiplicity* of $S/\mathfrak{I}$, or also of $\mathfrak{I}$.

DEFINITION 2.1 ([8]). *A* standard basis $\mathcal{B}(\mathfrak{I})$ *of* $\mathfrak{I}$ *is an ordered set of forms of* S*, generating* $\mathfrak{I}$*, such that its elements of degree* d *define a* K*-basis of* $\mathfrak{I}_d/(\mathfrak{I}_{d-1}S_1)$ *([5, 7, 8]).*

It is well known ([8]) that the degree vector of $\mathcal{B}(\mathfrak{I})$, with non decreasing entries, does not depend on the basis; $\alpha(\mathfrak{I})$ denotes its first entry, $\nu(\mathfrak{I})$ the number of entries, $\nu(\mathfrak{I}, t)$ the number of entries equal to t. Moreover, if $ht(\mathfrak{I}) > 1, \beta(\mathfrak{I})$ is the minimal degree t such that $GCD(\mathfrak{I}_t) = 1$.

The following theorem links $\alpha(\mathfrak{I})$ to $\nu(\mathfrak{I})$.

THEOREM 2.2 (Dubreil, [7, 8, 9]). *Let* $\mathfrak{I}$ *be a homogeneous perfect height 2 ideal. Then* $\nu(\mathfrak{I}) \leq \alpha(\mathfrak{I}) + 1$.

According to [3], $\mathcal{F}[n]$ denotes the set of all the homogeneous perfect height 2 ideals of $S = K[X_1, \ldots, X_n]$ such that $\nu(\mathfrak{I}) = \alpha(\mathfrak{I}) + 1$; in this paper they are called Dubreil' s ideals. In the special case $n = 2$, Theorem 1.7 ii) of [3] gives a description of every ideal of $\mathcal{F}[2]$ involving the greatest common divisor Φ of its elements of degree $\alpha(\mathfrak{I})$ and a decomposition of Φ as a product of forms.

A refinement of Theorem 2.2 ([5]) says, in particular, that, for every perfect height 2 ideal $\mathfrak{I}$ in S

$$t \leq \beta(\mathfrak{I}) \Rightarrow \nu(\mathfrak{I}, t) \leq -\Delta\Gamma(\mathfrak{I}, t). \tag{1}$$

We say that $\nu(\mathfrak{I}, t)$ is maximal when equality holds in (1).

If $\mathfrak{I}$ is a perfect height 2 ideal, then a minimal resolution of $S/\mathfrak{I}$ is defined by a Hilbert-Burch (shortly H.B.) matrix $M(\mathfrak{I})$ which, in turn, is uniquely determined by a standard basis $\mathcal{B}(\mathfrak{I})$ and by a minimal basis of its module of syzygies $Syz\,\mathcal{B}(\mathfrak{I})$. Its corresponding degree matrix $\partial M(\mathfrak{I})$ is uniquely determined by $\mathfrak{I}$.

We need some results, widely explained in [1, 2], that we summarize as follows.

THEOREM 2.3. *Let $\mathfrak{I}$ be a perfect height 2 ideal, $p+1$ a degree in which the number $\nu(\mathfrak{I}, p+1)$ of generators in degree $p+1$ satisfies the following relation of maximality with respect to Dubreil-Campanella inequality*

$$\nu(\mathfrak{I}, p+1) = \Gamma(\mathfrak{I}, p) - \Gamma(\mathfrak{I}, p+1), \tag{2}$$

D the greatest common divisor of $\mathfrak{I}_p$. Then $\mathfrak{I}$ admits a basis

$$\mathcal{B} = (DF_1, \ldots, DF_m,\ G_1, \ldots, G_n),$$

where $(DF_1, \ldots, DF_m)S = \mathfrak{I}_p$, so that $\mathfrak{I}$ splits into two ideals $\mathfrak{I}' = (F_1, \ldots, F_m)$ and $\mathfrak{I}'' = (D, G_1, \ldots, G_n)$, which are still perfect of height 2. Moreover, there is a H.B. matrix $M(\mathfrak{I})$ with respect to $\mathcal{B}$, with the following shape

$$M(\mathfrak{I}) = \begin{pmatrix} A & 0 \\ B & C \end{pmatrix},$$

where

i) $A \in K^{(m-1)\times m}$ is a H.B. matrix of $\mathfrak{I}'$,

ii) A H.B. matrix of $\mathfrak{I}''$ is $(B''\ C)$, where $B'' = B\ {}^t(F_1 \ldots F_m)$,

iii) $\det C = D$

3. Strongly inessential elements of an ideal: recalls and complements

Let $\mathfrak{I} = \oplus\mathfrak{I}_d, d \in \mathbf{N}, \mathfrak{I}_d \subset S_d$ be a homogeneous ideal of $S = K[X_1, \ldots, X_n]$. We recall some definitions and results appearing in [4].

DEFINITION 3.1 ([4], Definition 3.1). *An element f of a standard basis $\mathcal{B}(\mathfrak{I})$ is called an* inessential *generator of $\mathfrak{I}$ with respect to $\mathcal{B}(\mathfrak{I})$ iff*

$$\exists t \in \mathbf{N}, fM^t \subseteq (\mathcal{B}(\mathfrak{I}) - \{f\})S.$$

Otherwise we say that f is an essential *generator of $\mathfrak{I}$ with respect to $\mathcal{B}(\mathfrak{I})$.*

In the special case of perfect height 2 ideals, the essentiality of the r-th element f_r of $\mathcal{B}(\mathfrak{I})$ can be read on the ideal $\mathfrak{I}_{C_r}$ generated by the entries of the r-th column of any matrix of $Syz\,\mathcal{B}(\mathfrak{I})$. In fact, in [4], Proposition 4.1 says what follows.

PROPOSITION 3.2. *Let $\mathfrak{I}$ be a perfect codimension 2 ideal of S. Then $f_r \in \mathcal{B}(\mathfrak{I})$ is inessential for $\mathcal{B}(\mathfrak{I})$ iff the condition*

$$(\exists t \in \mathbf{N})\ M^t \subseteq \mathfrak{I}_{C_r}$$

is satisfied.

DEFINITION 3.3 ([4], Definition 3.2). *An element $f \in \mathfrak{I}_d$ is strongly inessential (s.i.) iff $f \notin (\mathfrak{I}_{d-1})S$ and it is inessential with respect to any standard basis containing it.*

DEFINITION 3.4 ([4], Definition 5.1). *A standard basis is called e-maximal iff it has, in every degree d, exactly $\nu_e(d)$ essential generators, where $\nu_e(d)$ is the greatest number of essential generators of degree d appearing in a standard basis of $\mathfrak{I}$.*

THEOREM 3.5 ([4], Theorem 5.1). *A standard basis is e-maximal iff its inessential elements are strongly inessential.*

Starting from Theorem 3.5 we can prove the following statement.

PROPOSITION 3.6. *The ideal $\mathfrak{I} \subset S$ admits a basis of essential elements iff none of its elements is s.i..*

Proof. Proposition 5.2 of [4] says that two different e-maximal bases contain the same number of inessential elements. So, $\mathfrak{I}$ has a basis of essential elements iff all its e-maximal bases do not contain inessential elements, and we know that they should be s.i., thanks to Theorem 3.5. Now, every s.i. element can be considered as an entry of a standard basis $\mathcal{B}(\mathfrak{I})$ and from any standard basis $\mathcal{B}(\mathfrak{I})$ it is possible to produce an e-maximal basis $\mathcal{B}_M(\mathfrak{I})$, containing as a subset all the s.i. elements appearing in $\mathcal{B}(\mathfrak{I})$ (see Proposition 5.4 in [4]). So, the e-maximal bases do not contain inessential elements iff s.i. elements do not exist in $\mathfrak{I}$. □

In other words, $\mathfrak{I}$ admits a basis of essential elements iff one of its e-maximal basis (and, as a consequence, all of them) consists of essential elements and this is equivalent to say that $\mathfrak{I}$ does not contain s.i. forms.

Next proposition says that a s.i. element of $\mathfrak{I}$ preserves its property modulo a linear form, regular for $S/\mathfrak{I}$. We will use the following notation.

Notation If z is any element of $S = K[X_1, \ldots, X_n]$ and $\phi : S \longrightarrow S/zS$ is the canonical morphism, then we set : $\phi(s) = \bar{s}, \forall s \in S$ and $\phi(\mathcal{A}) = \bar{\mathcal{A}}$ for any subset $\mathcal{A} \subseteq S$, if the element z can be understood.

We need the following lemma.

LEMMA 3.7 ([7, 8]). *If $\mathcal{B}$ is a standard basis of $\mathfrak{I}$ and $z \in S$ is a linear form, regular for $S/\mathfrak{I}$, then $\bar{\mathcal{B}}$ is a standard basis of $\bar{\mathfrak{I}}$.*

PROPOSITION 3.8. *Let $s \in \mathfrak{I}$ be a s.i. element and z a linear form regular for $S/\mathfrak{I}$. Then $\bar{s} \in \bar{\mathfrak{I}}$ is s.i..*

Proof. Without any loss of generality we can suppose $z = X_1$. At first we notice that if s is inessential for $\mathcal{B}(\mathfrak{I}) = \mathcal{B}$, then $\bar{s}$ is inessential for the standard basis $\bar{\mathcal{B}}(\bar{\mathfrak{I}}) = \bar{\mathcal{B}}$ of $\bar{\mathfrak{I}}$. In fact we have:

$$s\ \mathfrak{M}^t \subseteq (\mathcal{B} - \{s\})S \Rightarrow \bar{s}\ \bar{\mathfrak{M}}^t \subseteq (\bar{\mathcal{B}} - \{\bar{s}\}).$$

Now, let us suppose s to be s.i. and consider a standard basis $\mathcal{B}$ containing it, say $\mathcal{B} = (b_1, b_2, \ldots, b_h)$, where $b_i = s$. Then $\bar{\mathcal{B}}$ is a standard basis of $\bar{\mathfrak{I}}$, containing $\bar{b}_i = \bar{s}$ and any other standard basis $\mathcal{C}$ of $\bar{\mathfrak{I}}$ is of the form $\mathcal{C} = \bar{\mathcal{B}}P$, where $P = (p_{ji})$ is an invertible matrix, whose entries are forms in $K[X_2, \ldots, X_n]$. Let us observe that $\bar{s} = \bar{b}_i \in \mathcal{C}$ iff $p_{ii} \neq 0$ and $p_{ij} = 0$ when $j \neq i$. As a consequence, $\mathcal{B}' = \mathcal{B}P$ is a standard basis containing $s = b_i$; in $\mathcal{B}'$ the element s is inessential, as it is so in every basis in which it appears. The first part of the proof allows to conclude that $\bar{s}$ is inessential for $\mathcal{C}$. □

In Section 5 we will see that the lifting of a s.i. element of $\bar{\mathfrak{I}}$ is not necessarily s.i. in $\mathfrak{I}$. (see Remark 5.4).

A consequence of Proposition 3.8 is that if the image $\bar{\mathfrak{I}}$ of $\mathfrak{I} \subset K[X_1, \ldots, X_n]$ modulo a maximal regular sequence does not contain any s.i. element, the same property holds for $\mathfrak{I}$. So, it seems convenient to start considering the problem of the presence of s.i. elements when *depth* $(S/\mathfrak{I}) = 0$ (see Section 3).

In the sequel we use the following statement (see Theorem 2.3 for notation).

THEOREM 3.9. *Let $\mathfrak{I} \subset S$ be a perfect height 2 ideal and $p+1$ a degree in which the maximality condition (2) is verified. The following statements hold.*

i) If a form $F \in \mathfrak{I}'$ is s.i. in $\mathfrak{I}'$, then also $DF \in \mathfrak{I}$ is s.i. in $\mathfrak{I}$.

ii) $G \in \mathfrak{I}''$ is s.i. iff $G \in \mathfrak{I}_t, t > p$ and G is s.i. as an element of $\mathfrak{I}$.

Proof. i) Let $F \in \mathfrak{I}'_u, u \leq p - d$, where d is the degree of D, be s.i.. Then $F \notin \mathfrak{I}'_{u-1}S_1$, because it is an element of a standard basis of $\mathfrak{I}'$. As a consequence $FD \in \mathfrak{I}_{d+u}$, $FD \notin \mathfrak{I}_{d+u-1}S_1$, so that FD can be an element of some standard basis of $\mathfrak{I}$. Let $\mathcal{B} = (DF_1, \ldots, DF_m, G_1, \ldots, G_n)$ be any basis of $\mathfrak{I}$ such that $F = F_i$. As $(F_1, \ldots, F_m)$ is a standard basis of $\mathfrak{I}'$, we have

$$(\exists t)\ F\mathfrak{M}^t \subset (F_1, \ldots, F_{i-1}, F_{i+1}, \ldots, F_m).$$

So, for some t, the relation

$$(DF)\mathfrak{M}^t \subset (DF_1, \ldots, DF_{i-1}, DF_{i+1}, \ldots, DF_m, G_1, \ldots, G_n)$$

holds.

ii) Let G be a s.i. element of $\mathfrak{I}''$. Thanks to Proposition 3.4 in [4], stating that no element of degree $\alpha(\mathfrak{I})$ can be s.i., G cannot be of the form kD, $k \in K$, so that $t = \deg G \geq p + 1$. First we observe that, as an element of $\mathfrak{I}$, G can belong to a standard basis. In fact, as it is a form of a standard basis of $\mathfrak{I}''$, we have $G \notin (\mathfrak{I}''_{t-1})S_1 \supseteq (\mathfrak{I}_{t-1})S_1$, so that $G \notin \mathfrak{I}_{t-1}S_1$. Now, let $\mathcal{B} = (DF_1, \ldots, DF_m, G_1, \ldots, G_{i-1}, G, G_{i+1}, \ldots, G_n)$ be any standard basis of $\mathfrak{I}$ containing G. Then $(D, G_1, \ldots, G_{i-1}, G, G_{i+1}, \ldots, G_n)$ is a standard basis of $\mathfrak{I}''$. The hypothesis of inessentiality of G as an element of $\mathfrak{I}''$ implies that

$$(\exists t)\ G\mathfrak{M}^t \subset (D, G_1, \ldots, G_{i-1}, G_{i+1}, \ldots, G_n).$$

In other words, for every form $P \in \mathfrak{M}^t$, we have

$$GP = DV + \sum_{j \neq i} V_j G_j,$$

so that $(V, V_1, \ldots, V_{i-1}, -P, V_{i+1}, \ldots, V_n) \in Syz\ \mathfrak{I}''$. From (c) of Theorem 3.1 ([2]) it follows $V \in \mathfrak{I}'$, so that $GP \in (DF_1, ..., DF_m, G_1, ..., G_{i-1}, G_{i+1}, ..., G_n)$. This means that G is s.i. also as an element of $\mathfrak{I}$.

Viceversa, let $G \in \mathfrak{I}_t, t > p$ be a s.i. element in $\mathfrak{I}$. If $\mathcal{B} = (DF_1, \ldots, DF_m, G_1, \ldots, G_{i-1}, G, G_{i+1}, \ldots, G_n)$ is a basis of $\mathfrak{I}$ containing G, then $\mathcal{B}'' = (D, G_1, \ldots, G_{i-1}, G\ , G_{i+1}, \ldots, G_n)$ is a basis of $\mathfrak{I}''$. Thanks to Proposition 5.1 in [4], it is enough to prove that G is inessential with respect to any basis

$$\tilde{\mathcal{B}''} = (D, G_1 + A_1G, \ldots, G_{i-1} + A_{i-1}G, G, G_{i+1} + A_{i+1}G, \ldots, G_n + A_nG)$$

for every (degree allowed) choice of $A_1, \ldots, A_{i-1}, A_{i+1}, \ldots, A_n$.
As $\tilde{\mathcal{B}} = (DF_1, ..., DF_m, G_1 + A_1G, ..., G_{i-1} + A_{i-1}G, G, G_{i+1} + A_{i+1}G, ..., G_n + A_nG)$ is still a standard basis of $\mathfrak{I}$, G is inessential with respect to it. This means that

$$\begin{aligned}(\exists t \in \mathbf{N})\ G\mathfrak{M}^t &\subset (DF_1, \ldots, DF_m, G_1 + A_1G, \ldots, G_{i-1} + A_{i-1}G,\\ &\qquad G_{i+1} + A_{i+1}G, \ldots, G_n + A_nG)\\ &\subset (D, G_1 + A_1G, \ldots, G_{i-1} + A_{i-1}G,\\ &\qquad G_{i+1} + A_{i+1}G, \ldots, G_n + A_nG).\end{aligned}$$

As a consequence, G is inessential also with respect to the basis $\tilde{\mathcal{B}''}$. □

REMARK 3.10. *Theorem 3.9 can also be proved by working on a suitable H.B. matrix of $\mathfrak{I}$, taking into account Proposition 1.2 in [1] and Corollary 4.1 in [4].*

REMARK 3.11. *It may happen that in $\mathfrak{I}$ there exist s.i. elements that do not produce s.i. elements in $\mathfrak{I}'$ (see Remark 4.18)*

REMARK 3.12. *For every $\mathfrak{I} \in F[n]$, the maximality condition required in Theorem 3.9 is verified at any degree.*

Proposition 3.2 suggests a situation in which all the elements of every basis of $\mathfrak{I}$ are essential because the columns of its H.B. matrix are "short", so that they cannot generate a power of $\mathfrak{M}$.

COROLLARY 3.13. *Let $\mathfrak{I}$ be a perfect height 2 ideal of $S = K[X_1, \ldots, X_n]$. Each of the following conditions is enough to guaranty that in any standard basis of $\mathfrak{I}$ all the elements are essential:*

i) $\nu(\mathfrak{I}) < n + 1$

ii) $\alpha(\mathfrak{I}) < n$

iii) $e(\mathfrak{I}) < \dfrac{n(n+1)}{2}$.

Proof. i) and ii) are the statement of Corollary 5.1 and Remark in [4]; iii) comes from the inequality $\dfrac{\alpha(\alpha+1)}{2} \leq e(\mathfrak{I})$, where $\alpha = \alpha(\mathfrak{I})$, just observing that $e(\mathfrak{I}) < \dfrac{n(n+1)}{2}$ implies $\alpha < n$. □

REMARK 3.14. *It is easy to find examples of ideals with $e(\mathfrak{I}) = \dfrac{n(n+1)}{2}$ containing inessential elements in some standard basis; see, for instance, Example 3.1 in [4], where $n = 3, e = 6$.*

Proposition 3.4 of [4] says that in degree $\alpha(\mathfrak{I})$ no element is s.i.. So, the existence of a basis of essential elements is assured if the generators of degree $> \alpha$ are essential. Such a condition is verified when in the degree matrix $\partial M(\mathfrak{I}) = (d_{ij}), i = 1, \ldots, \nu(\mathfrak{I}) - 1, j = 1, \ldots, \nu(\mathfrak{I})$ the inequality $d_{h,\nu(\mathfrak{I},\alpha)} \leq 0$ is verified for $h = \nu(\mathfrak{I}) - n$ (and, as a consequence, for $h < \nu(\mathfrak{I}) - n$), because it assures that the columns $C_j, j \geq \nu(\mathfrak{I}, \alpha)$, have at most $n-1$ elements different from zero. This justifies the following statement.

PROPOSITION 3.15. *Let $\mathfrak{I}$ be a perfect height 2 ideal of $S = K[X_1, \ldots, X_n]$, with degree matrix $\partial M(\mathfrak{I}) = (d_{ij}), i = 1, \ldots, \nu(\mathfrak{I}) - 1, j = 1, \ldots, \nu(\mathfrak{I})$. If $d_{\nu(\mathfrak{I})-n,\ \nu(\alpha,\mathfrak{I})} \leq 0$, then $\mathfrak{I}$ has a basis of essential elements.*

A consequence of Proposition 3.15 is the following statement.

COROLLARY 3.16. *Let $\mathfrak{I}$ be a perfect height 2 ideal of $S = K[X_1, \ldots, X_n]$. If*

$$e(\mathfrak{I}) < \frac{n(n+3)}{2},$$

then $\mathfrak{I}$ has a standard basis whose elements are all essential.

Proof. Taking into account the inequality $\frac{\alpha(\alpha+1)}{2} \leq e(\mathfrak{I})$, we see that the hypothesis implies $\alpha \leq n$. In case $\alpha < n$ we apply Corollary 3.13 ii). In case $\alpha = n$ and $\nu = \nu(\mathfrak{I}) < \alpha + 1$ we apply Corollary 3.13 i). So, the only case to be considered is $\alpha = n, \nu = n + 1$. In this situation the degree matrix $\partial M(I) = (d_{ij})$ satisfies the conditions $d_{i,i+1} = 1,\ i = 1, \ldots, n$. Taking into account the rule of computation of $e(\mathfrak{I})$ starting from $\partial M(\mathfrak{I})$ (see [6]), it is easy to verify that the only values of d_{ii} compatible with the hypothesis are the following ones:

a) $d_{ii} = 1,\ i = 1, \ldots, n$

b) $d_{ii} = 1,\ i \neq i_0, d_{i_0 i_0} = 2$ for some $i_0 \neq 1$.

In case a) the ideal is generated in degree α, so that we apply Proposition 3.4 of [4].
In case b) we have necessarily $d_{i_0(i_0+1)} = 0$, so that Proposition 3.15 can be used. □

REMARK 3.17. *If the inequality of Corollary 3.16 is not satisfied, there exist examples of ideals with s.i. elements. For instance, let us consider in $S = K[X_1, \ldots, X_n]$ the ideal $\mathfrak{I}$, with H.B. matrix*

$$M(\mathfrak{I}) = \begin{pmatrix} X_2^2 & -X_1 & & & & \\ & X_2 & -X_1 & & & \\ & X_3 & X_2 & -X_1 & & \\ \ldots. & \ldots & \ldots & \ldots & \ldots & \ldots \\ & X_n & & & X_2 & -X_1 \end{pmatrix},$$

where the unwritten entries are zero forms.
$\mathfrak{I}$ satisfies the condition $e(\mathfrak{I}) = \frac{n^2+3n}{2}$ and its second generator is s.i..

We observe that the ideals, with multiplicity $e(\mathfrak{I}) = \frac{n^2+3n}{2}$, that do not admit a basis of essential elements must necessarily have as a degree matrix the one defined by

$$d_{11} = 2;\ d_{ii} = 1,\ i \neq 1;\ d_{i(i+1)} = 1,\ i = 1, \ldots, n,$$

so that they have only one generator in degree α.

On the other side, it is possible to produce ideals with a basis of essential elements and with no upper limit on $e(\mathfrak{I})$. For instance, every ideal $\mathfrak{I}$ whose $\partial M(\mathfrak{I})$ is defined by

$$d_{i(i+1)} = 1, i = 1, \dots, n;\ d_{ii} = 1,\ i = 1, \dots, n-1;\ d_{nn} = h \geq 2$$

satisfies the condition of Proposition 3.15 and has multiplicity $e(\mathfrak{I}) = \dfrac{n(n+1)}{2} + h - 1$, which is arbitrarily large if $h \gg 0$.

We see that if $e(\mathfrak{I}) \geq \dfrac{n^2+3n}{2}$ the situation is hard to be examined , also if $\mathfrak{I}$ is a perfect height 2 ideal. That is a reason why we restrict our attention to the subfamily $\mathcal{F}[n]$ (see Section 2), starting with $n = 2$.

4. An e-maximal basis of $\mathfrak{I} \in \mathcal{F}[2]$

Relation (1.10) in Remark 1 to Theorem 17 in [3] gives a good description of every $\mathfrak{I} \in \mathcal{F}[2]$. With some change of notation, we rewrite it as follows:

$$\mathfrak{I} = \sum_{i=0}^{r} \Phi_{i+1} \dots \Phi_{r+1} S_{\beta_i} S, \tag{3}$$

where Φ_i is a form of degree δ_i, $\Phi_{r+1} = 1$ and S_t is the subset of $S = K[X, Y]$ consisting of the forms of degree t.

Let us denote $\Delta_0 = 0, \Delta_i = \delta_1 + \dots + \delta_i, i = 1, \dots, r$ and $\Delta_r = \delta$ the degree of $\Phi = \Phi_1 \dots \Phi_r$, so that we have

$$\delta = \sum_{i=1}^{r} \delta_i, \quad \beta_i = \beta_{i-1} + \delta_i + t_i,$$

where $t_i = \alpha_i - \alpha_{i-1} > 0,\ i = 1, \dots, r$ is the difference between two successive different degrees of the generators appearing in a standard basis and $\alpha_0 = \alpha(\mathfrak{I}) = \alpha$.

In (3), $r + 1$ is the number of distinct elements appearing in any degree vector $\mathbf{a}$ of a standard basis of $\mathfrak{I}$; moreover, we have

$$\begin{aligned} \mathbf{a} &= ((\beta_0 + \delta)^{[\beta_0+1]}, \dots, (\beta_i + \delta - \Delta_i)^{[\delta_i]}, \dots, \beta_r^{[\delta_r]}) \\ &= (\alpha_0^{[\beta_0+1]}, \alpha_1^{[\delta_1]}, \dots, \alpha_i^{[\delta_i]}, \dots, \alpha_r^{[\delta_r]}), \end{aligned}$$

where $c^{[n]}$ is the sequence $(c, \dots, c)$, with c repeated n times.

The degree matrix $\partial M(\mathfrak{I}) = (d_{ij})$ is completely determined by its elements in position $(i, i+1)$ (which are necessarily 1, as $M(\mathfrak{I})$ is a $\alpha \times (\alpha+1)$ matrix) and by $\mathbf{a}$, or, equivalently, by its elements in position (i,i), which are

$$d_{ii} = 1 \text{ if } i \neq \beta_0 + 1 + \Delta_j,\ j = 0, \dots, r-1\,,$$
$$d_{ii} = t_{j+1} + 1 \text{ if } i = \beta_0 + 1 + \Delta_j\,.$$

Our aim is to produce an e-maximal basis of $\mathfrak{I}$ (see Definition 3.4), that allows to prove the following theorem.

THEOREM 4.1. *Let $\mathfrak{I}$ be as in (3) and let*

$$\Phi = \Phi_1 \dots \Phi_r = H_1^{\mu_1} \dots H_v^{\mu_v}, \quad \sum_{i=1}^{v} \mu_i = \delta \geq 1 \tag{4}$$

be a factorization of Φ as a product of linear forms pairwise linearly independent. The number of s.i. elements appearing in every e-maximal basis of $\mathfrak{I}$ is $\delta - v$. If $\delta = 0$, then $\mathfrak{I} = S_\alpha S$ does not contain s.i. elements.

In order to prove Theorem 4.1 it is convenient (and possible) to produce an e-maximal basis $\mathcal{B}(\mathfrak{I})$ satisfying the following condition.

(*) *There is a basis of its module of syzygies linking only couples of adjacent elements .*

The Hilbert matrix corresponding to such a basis of $Syz(\mathcal{B}(\mathfrak{I}))$ will be called *the canonical matrix* of $\mathcal{B}(\mathfrak{I})$ or *a canonical matrix* of $\mathfrak{I}$.
Condition (*) will be of help in checking that $\mathcal{B}(\mathfrak{I})$ is an e-maximal basis.

Let us consider first two special cases, useful to face the general situation.

Case 1. $\mathfrak{I} = S_\alpha S$.

Thanks to Proposition 3.4 of [4], we know that an ideal generated in minimal degree cannot have s.i. elements. However, in the sequel we need an e-maximal basis, satisfying condition (*), constructed according to the following Proposition. The notation $\hat{L}$ will always mean that the element L is omitted.

PROPOSITION 4.2. *Let $\mathfrak{I} = S_\alpha S$. If $\{L_0, \dots, L_\alpha\}$ is a set of linear forms, pairwise linearly independent, then $\mathcal{B}(\mathfrak{I}) = (F_i), i = 0, \dots, \alpha,\ F_i = L_0 \dots \hat{L_i} \dots L_\alpha$, is a standard basis, consisting entirely of essential elements, whose canonical matrix M looks as follows:*

$$M = (m_{ij}),\ \ i = 1, \dots, \alpha,\ \ j = 1, \dots, (\alpha+1),$$
$$\text{where: } m_{ii} = L_{i-1},\ m_{i(i+1)} = -L_i,\ m_{ij} = 0 \text{ otherwise.}$$

Proof. It is immediate to verify that the F_i's are a set of $\alpha+1$ linearly independent elements of S_α, so that they are a basis of it as a K-space. The rows of M are syzygies linking adjacent elements; as they are linearly independent, they are a basis of $Syz\ (\mathcal{B}(\mathfrak{I}))$ (see Hilbert-Burch Theorem, [13]), so that M is a matrix of syzygies of $\mathfrak{I}$. The entries of every column C_i generate a principal ideal $\mathfrak{I}_{C_i}$; so, Proposition 3.2 says that all the elements of $\mathcal{B}(\mathfrak{I})$ are essential. □

Case 2. $\mathfrak{I}$ is generated in two different degrees and in the lower one there is just one generator, so that

$$\mathfrak{I}=\Phi S+S_bS,\ \deg\Phi=\delta=\alpha(\mathfrak{I}),\ b=\beta(\mathfrak{I})=\delta+t,\ t>0. \tag{5}$$

Let us consider the decomposition of Φ as in (4), with $r=1$. We prove first the following lemma.

LEMMA 4.3. *Let $\Phi=H_1^{\mu_1}\dots H_v^{\mu_v}$ be any form of degree δ in $S=K[X,Y]$. The K-space S_b, $b=\delta+t, t\geq 0$, admits a decomposition*

$$S_b=\Phi S_t\bigoplus T,\quad T=\bigoplus_{i=1}^{v}T_i, \tag{6}$$

where a K-basis of T_i is the ordered set $\mathcal{B}_i=(F_{ij}), j=1,\dots,\mu_i$, described as follows:

$$F_{ij}=A_{ij}C_i, \tag{7}$$

with

$$A_{ij}=H_i^{\mu_i-j}H_{i+1}^{\mu_{i+1}}\dots H_v^{\mu_v}U^{j-1},\ GCD(U,H_h)=1,\ h=1,\dots,v, \tag{8}$$

and C_i any form of degree $t+\mu_1+\dots+\mu_{i-1}+1, (\mu_0=0)$, satisfying the relation

$$GCD(C_i,H_i)=1. \tag{9}$$

Proof. We use induction on v.

For $v=1$ we have $\Phi=H^\mu, \delta=\mu$ and the statement becomes $T=T_1$, with basis $\mathcal{B}_1=(F_{1j}),\ j=1,\dots,\mu$, where

$$F_{1j}=F_j=A_jC=H^{\mu-j}U^{j-1}C,\ \deg C=t+1,\ GCD(C,H)=1. \tag{10}$$

It is immediate to prove that $F_1,\dots,F_\mu$ are linearly independent, so we only have to show that $\Phi S_t\bigcap T=(0)$.

For $\Lambda\in S_t$, let us suppose $\Lambda\Phi=\sum_{j=1}^{\mu}a_jF_j=(\sum_{j=1}^{\mu}a_jH^{\mu-j}U^{j-1})C$. This implies that H^μ must divide $A=\sum_{j=1}^{\mu}a_jH^{\mu-j}U^{j-1}$. For degree reason, A must be zero, so that $a_j=0,\ j=1,\dots,\mu$.

Let us suppose the statement true until $v-1$ and prove it for v. We set $\Phi = \Psi H_v^{\mu_v}$ and use the decomposition of case $v=1$ with H^μ replaced by $H_v^{\mu_v}$, so obtaining

$$S_b = H_v^{\mu_v} S_{b-\mu_v} \bigoplus T_v,$$

where $T_v = (F_{v1}, \dots, F_{v\mu_v})$, with $F_{vj} = A_{vj}C_v,\ A_{vj} = H_v^{\mu_v - j}U^{j-1}$, $GCD(C_v, H_v) = 1,\ \deg C_v = b - \mu_v + 1$, according to (10).

Using induction, we have $S_{b-\mu_v} = \Psi S_t \bigoplus T',\ T' = \bigoplus_{i=1}^{v-1} T_i'$, where T_i' has the basis (F_{ij}') described in the statement of Lemma 4.3, that is $F_{ij}' = H_i^{\mu_i - j} H_{i+1}^{\mu_{i+1}} \dots H_{v-1}^{\mu_{v-1}} U^{j-1} C_i$. So, we finally obtain

$$S_b = H_v^{\mu_v}(\Psi S_t \bigoplus T') \bigoplus T_v = \Phi S_t \bigoplus T,$$

where $T = \bigoplus H_v^{\mu_v} T' \bigoplus T_v = \bigoplus_{i=1}^{v-1} H_v^{\mu_v} T_i' \bigoplus T_v$. It is immediate to check that $(F_{ij}) = (H_v^{\mu_v} F_{ij}'),\ j = 1, \dots, \mu_i,$ is the required basis of $T_i = H_v^{\mu_v} T_i',\ i = 1, \dots, (v-1)$. □

REMARK 4.4. *Each space T_i depends on the choice of the form C_i, with the link (9). So, there are infinitely many decompositions of the type described in (6). Later on, we will use some of them, properly chosen accordingly to the situation.*

REMARK 4.5. *The basis of $S_b,\ b = \alpha$, used in Proposition 4.2 is obtained accordingly to Lemma 4.3, with the choice $\Phi = L_0 \dots L_b,\ t = -1,\ C_i = H_1 \dots H_{i-1}$. In this situation, the first summand of (6) is empty, so that $S_b = T$.*

PROPOSITION 4.6. *Let us consider the ideal*

$$\mathfrak{I} = \Phi S + S_b S,\ b = \delta + t,\ t > 0, \quad \Phi = H_1^{\mu_1} \dots H_v^{\mu_v}, \deg \Phi = \delta.$$

i) $\mathfrak{I}$ has as a standard basis the set

$$\mathcal{B}(\mathfrak{I}) = (\Phi, F_{ij}),\ i = 1, \dots, v,\ j = j(i) = 1, \dots, \mu_i,$$

where:

$$F_{ij} = A_{ij} C_i, \tag{11}$$

$$A_{ij} = H_i^{\mu_i - j} H_{i+1}^{\mu_{i+1}} \dots H_v^{\mu_v} U^{j-1}, \quad GCD(H_i, U) = 1 \tag{12}$$

$$C_1 = U^{t+1}, C_i = H_1 H_2 \dots H_{i-1} U^{\nu_i}, \quad \nu_i = t + \mu_1 + \dots + \mu_{i-1} - i + 2,\ i > 1, \tag{13}$$

ii) The basis $\mathcal{B}(\mathfrak{I})$ satisfies condition (*). Its canonical matrix $M(\mathfrak{I})$ has as rows the basis of syzygies $\{s_{ij}\}$, $i = 1, \ldots, v$, $j = j(i) = 1, \ldots, \mu_i$, with the lexicographic order, where:

$$\begin{aligned} s_{11} &= (U^{t+1}, -H_1, 0, \ldots, 0), \\ s_{i1} &= (0, \ldots, 0, H_{i-1}, -H_i, 0, \ldots 0),\ i = 2, \ldots, v\,, \\ &\qquad -H_i \text{ in position } \mu_1 + \mu_2 + \ldots + \mu_{i+2}, \\ s_{ij} &= (0, \ldots, 0, U, -H_i, 0, \ldots, 0),\ i = 1, \ldots, v,\ j = 2, \ldots, \mu_i\,, \\ &\qquad -H_i \text{ in position } \mu_1 + \ldots + \mu_{i-1} + j + 1,\ \mu_0 = 0. \end{aligned}$$

So, $M(\mathfrak{I})$ looks as follows:

$$M(\mathfrak{I}) = \begin{pmatrix} U^{t+1} & & \\ & \mathcal{A} & \\ \mathcal{O} & & \end{pmatrix},$$

where $\mathcal{A} = (a_{ij})$ is a square $\delta \times \delta$ matrix, whose elements different from $a_{ii}, a_{(i+1)i}$ are zero, and $(a_{11}, \ldots, a_{\delta\delta}) = ([-H_1]^{\mu_1}, [-H_2]^{\mu_2}, \ldots, [-H_v]^{\mu_v})$, $a_{(i+1),i} = -a_{ii}$ if $a_{ii} \neq a_{jj}, j > i$ and $a_{(i+1),i} = U$ otherwise.

iii) The essential elements of $\mathcal{B}(\mathfrak{I})$ are : $\Phi, F_{(i,\mu_i)}, i = 1, \ldots, v$. All the other $\delta - v$ elements of $\mathcal{B}(\mathfrak{I})$ are s.i..

Proof. i) This assertion is an immediate consequence of Lemma 4.3. In fact, thanks to the inequality $\mu_1 + \mu_2 + \ldots + \mu_{i-1} \geq i - 1$, we can choose $C_i = H_1 \ldots H_{i-1} U^{\nu_i}$, so obtaining F_{ij} as a basis of the K-space T complementary to ΦS_t in S_b.

ii) The fact that the $\{s_{ij}\}$'s are syzygies can be verified with an easy direct computation. Moreover, they are clearly linearly independent, of the expected degree and their number δ is the rank of the module of syzygies, according to Hilbert theorem. It is easy to verify that the first maximal minor of $M(\mathfrak{I})$ is Φ and (apart from a sign) the other maximal minors are the F_{ij}'s. Using Proposition 3.2, we see immediately that the essential columns of $M(\mathfrak{I})$ (that is the columns corresponding to essential elements, see [4]) are the first one and the $(\mu_1 + \mu_2 + \ldots + \mu_i + 1) - th, i = 1 \ldots v$; so, the essential elements of $\mathcal{B}$ are $\{\Phi, F_{i,\mu_i}, i = 1 \ldots v\}$.

iii) The proof that all the inessential elements are s.i. is a consequence of the following Lemma 4.8, stated in a form sufficiently general to be used later in a more general situation. In fact, the submatrix $\mathcal{A}$ appearing in $M(\mathfrak{I})$ satisfies the hypothesis of Lemma 4.8 .

□

It is convenient to generalize the notion of inessential and strongly inessential columns of a matrix, as considered in [4].

DEFINITION 4.7. *Let A be a matrix whose entries a_{ij} are forms of $K[X_1, ..., X_n]$ such that $\deg a_{ij} - \deg a_{i(j+1)}$ is independent from i. A column C_j is inessential when the ideal $\mathfrak{I}_{C_j}$ generated by its entries contains a power of the irrelevant ideal. C_j is strongly inessential when every column $C'_j = \sum_i \lambda_i C_i = (a'_{ij})$, $\lambda_i \in K[X_1, \ldots, X_n]$, $\lambda_j = 1$, $\deg a'_{ij} = \deg a_{ij}$, replacing C_j, is still inessential.*

LEMMA 4.8. *Let $A = (a_{ij}), a_{ij} \in K[X, Y]$, be a square $m \times m$ matrix such that $\deg a_{ij} - \deg a_{i(j+1)} \leq 0$ is independent from i and satisfying the conditions:*

i) $a_{ij} = 0$ if $i \neq j, j+1$

ii) a_{ii} is a linear form L_i,

iii) $a_{(i+1)i}$, is any form G_i, such that L_j is not a factor of G_i if $L_j \neq L_i$ and G_i is a multiple of L_i iff every $L_j, j > i$ is different from L_i.

Then the inessential columns of A are s.i.

Proof. The inessential column we are considering is of the form

$$C_j =^t (0, \ldots, 0, L_j, G_j, 0, \ldots, 0),$$

where L_j does not divide G_j, so that no L_i divides G_j. Let us replace such a C_j with $C'_j = \sum_i C_i$, $\lambda_j = 1$ and prove that C'_j is still inessential. If h is the first index for which $\lambda_h \neq 0$, we point our attention on the column C_h (clearly, $h \geq j$). Let us distinguish two possible situations.

i) C_h *is essential*, so that $C_h =^t (0, \ldots, 0, -L_h, aL_h, 0, \ldots, 0)$. In this case we have $C_h \neq C_j$, so that $h < j$ and the entries of C_h must have the same degree of the corresponding entries of C_j (in particular, $a \in K^*$.)

- If $\lambda_{h+1} \neq 0$, let us consider $C_{h+1} =^t (0, \ldots 0, L_{h+1}, G_{h+1}, 0, \ldots, 0)$. The entries of C'_j in position (h, j) and $(h+1, j)$ are respectively $c'_{hj} = \lambda_h L_h$ and $c'_{(h+1)j} = a\lambda_h L_h + \lambda_{h+1} L_{h+1}$, so that $\mathfrak{I}_{C'_j} = (L_h, L_{h+1}) = \mathfrak{M}$, as L_h is independent from L_{h+1}.

- If $\lambda_{h+1} = \lambda_{h+2} = \ldots \lambda_{h+u-1} = 0, \lambda_{h+u} \neq 0, u > 1$, then necessarily $h + u \geq j$, so that C'_j has as elements $c'_{hj} = \lambda_h L_h$ and $c'_{h+u} = \lambda_{h+u} L_{h+u}$, $\lambda_{h+u} \in K^*$; as a consequence, also in this case $\mathfrak{I}_{C'_j} = \mathfrak{M}$.

ii) C_h *is inessential*, so that $C_h =^t (0, \ldots, 0, -L_h, aL_h, 0, \ldots, 0)$, where G_h is not divisible for $L_q, q = 1, \ldots, m$. (As a special case, C_h might coincide with C_j.) Let us denote $h + u$ the least integer v for which $L_v = L_h$.

1- If $u = 1$, then $c'_{h,j} = \lambda_h L_h$, $c'_{(h+1)j} = \lambda_h G_h + \lambda_{h+1} L_h$, so that $\mathfrak{I}_{C'_j} \supseteq (L_h, G_h) \supseteq \mathfrak{M}^t$, for some $t \in \mathbf{N}$.

2- If $u \neq 1$ but $\lambda_{h+1} = 0$, then $c'_{hj} = L_h$, $c'(h+1)j = G_h$, so that $\mathfrak{I}_{C'_j} \supseteq (L_h, G_h)$, as in the previous case.

3- If $u \neq 1$, $\lambda_{h+1} \neq 0$, then $c'_{hj} = \lambda_h L_h$, $c'_{(h+1)j} = \lambda_h G_h + \lambda_{h+1} L_{h+1}$, where $\lambda_h \in K^*$ (as $h \leq j$).

If λ_{h+1} is such that $c'_{(h+1)j}$ is not a multiple of L_h, we get $C'_j \supseteq \mathfrak{M}^t$, for some $t \in \mathbf{N}$. However, for some choice of λ_{h+1} it may happen $c'_{(h+1)j} = L_h P$. In fact, if $G_h = M_1 \dots M_s$ is a decomposition of G_h into linear factors, there exists $a \in K^*$ such that $\lambda_{h+1} = aM_1 \dots M_{s-1}$ gives $c'_{(h+1)j} = M_1 \dots M_{s-1}(\lambda_h M_s + aL_{h+1})$, where $\lambda_h M_s + aL_{h+1} = bL_h$, as M_s, L_{h+1} are linearly independent linear forms. Let us observe that such a λ_{h+1} cannot be a multiple of L_h, as G_h is not so. If we replace C_h with $C^*_h = \lambda C_h + \lambda_{h+1} C_{\lambda+1} C_{h+1}$ and consider $C'_j = C^*_h + \sum_{i=h+2}^{m} \lambda_i C_i$, we have a situation very similar to the previous one. In fact $c*_{hj} = \lambda_h L_h$, $c^*_{(h+1)j} = PL_h$, $c^*_{h+2,j} = \lambda_{h+1} G_{h+1}$, so that G_h is replaced with $\lambda_{h+1} G_{h+1}$, which is not a multiple of L_h. Now, we can repeat the same reasoning until when we find either case 2, if $\lambda_i = 0$ for some i with $h+1 < i \leq h+u$, or case 1, for $i = h+u$.

□

REMARK 4.9. *The essential generators of $\mathcal{B}(\mathfrak{I}) - \{\Phi\}$ are exactly the ones that do not contain as their factors all the linear factors of Φ; more precisely, $F_{i\mu_i}$ does not contain H_i, while it contains as factors $H_j, j \neq i$.*

In the sequel we will need also bases slightly different from the one produced in Proposition 4.6. We introduce them in the following Remarks.

REMARK 4.10. *If, in the definition of $F_{ij}, i > k$, $\mathbb{C}_i$ is replaced by $\tilde{C}_i = H_1 \dots \hat{H}_k \dots H_{i-1} U^{\nu_i+1}$ (that is, if H_k is replaced with U), then $\tilde{\mathcal{B}}$, obtained from $\mathcal{B}$ by replacing F_{ij} with $\tilde{F}_{ij} = A_{ij}\tilde{C}_i$, is still a standard basis, whose Hilbert matrix $\tilde{M}(\mathfrak{I})$ differs from the $M(\mathfrak{I})$ described in Proposition 4.6 just in the column corresponding to $\tilde{F}_{k\mu_k}$, which becomes $\tilde{C}_k =^t (0, ..., -H_k, U, 0, ...0)$. The consequence is that the generator $\tilde{F}_{k\mu_k} = F_{k\mu_k}$ now is inessential, while the other generators are changed but remain with unchanged nature. $\tilde{\mathcal{B}}$ is not an e-maximal basis, but it will erase in a splitting (see Remark 4.18).*

REMARK 4.11. *Let us observe that the F_{ij}'s have U^{t+1} as a common factor. If we replace U^{t+1} by any form η, of degree $t+1$, such that $G.C.D.(\eta, \Phi) = 1$, the matrix $M^*(\mathfrak{I})$ corresponding to the new basis $\mathcal{B}^*$ differs from $M(\mathfrak{I})$ only in the first column. In particular, $\mathcal{B}^*$ is still an e-maximal basis.*

REMARK 4.12. *Let us produce other H.B. canonical matrices of $\mathfrak{I}$, relative to standard bases different from the one described in Proposition 4.6 . They are defined as follows:*

$$M'(\mathfrak{I}) = \begin{pmatrix} U^{t+1} & \\ & \mathcal{A}' \\ \mathcal{O} & \end{pmatrix},$$

where $\mathcal{A}' = (a'_{ij})$ *is a square* $\delta \times \delta$ *matrix, whose elements different from* $a'_{ii}, a'_{(i+1)i}$ *are zero, and*

- $(a'_{11}, \dots, a'_{\delta\delta}) = (-H_{\sigma(1)}, \dots, -H_{\sigma(\delta)})$, *with* σ *any permutation of the sequence* $([1]^{\mu_1}, ([2]^{\mu_2}, \dots, ([v]^{\mu_v})$,

- $a'_{(i+1),i} = -a'_{ii}$ *if* $a'_{ii} \neq a'_{jj}, j > i$ *and* $a'_{(i+1),i} = U$ *otherwise.*

In fact, Lemma 4.8 guaranties that all the inessential columns of $M'(\mathfrak{I})$ *are s.i. and it is a matter of computation to check that the maximal minors of the new matrix are still the basis of a subspace* T *such that* $\Phi S_t \bigoplus T = S_b$. *The maximal minors of* $M'(\mathfrak{I})$, *different from* Φ, *apart from a sign are:* $(\mathcal{B}'_i = U^{t+1} G_1 \dots G_i \hat{H}_{\sigma(i)} H_{\sigma(i+1)} \dots H_{\sigma(\delta)})$, $i = 1, \dots, \delta$. *A reasoning analogous to the one in the proof of Lemma 4.3 shows that they are linearly independent. In fact the relation* $\lambda_1 H_{\sigma(1)} \dots H_{\sigma(\delta)} + \sum_{i=2}^{\delta-1} = 0$, $(\lambda_1, \dots, \lambda_\delta) \neq (0, \dots, 0)$ *implies that* G_1 *divides* $H_{\sigma(1)} \dots H_{\sigma(\delta)}$, *against the hypothesis.*

Moreover, let us denote T' *the* K*-space generated by* $(\mathcal{B}'_1, \dots, \mathcal{B}'_\delta)$. *Then* $\Phi S_t \bigcap T' = (0)$, *because* $\Lambda\Phi = \sum a_i U^{t+1} G_1 \dots G_i \hat{H}_{\sigma(i)} H_{\sigma(i+1)} \dots H_{\sigma(\delta)})$, $\Lambda \neq 0$, *implies that* U *must divide* Φ, *for degree reason, against the hypothesis.*

EXAMPLE 4.13. *Let us consider the ideal*

$$\mathfrak{I} = (H_1^3 H_2^2 H_3) S + S_8 S,$$

where H_1, H_2, H_3 *are linearly independent linear forms. The basis considered in Proposition 4.6 is* $\mathcal{B}(\mathfrak{I}) = (\Phi, F_{11}, F_{12}, F_{13}, F_{21}, F_{22}, F_{31})$, *where* $\Phi = H_1^3 H_2^2 H_3$, $F_{11} = H_1^2 H_2^2 H_3 U^3$, $F_{12} = H_1 H_2^2 H_3 U^4$, $F_{13} = H_2^2 H_3 U^5$, $F_{21} = H_1 H_2 H_3 U^5$, $F_{22} = H_1 H_3 U^6$, $F_{31} = H_1 H_2 U^6$.

The corresponding H.B. matrix is

$$M(\mathfrak{I}) = \begin{pmatrix} U^3 & -H_1 & & & & & \\ & U & -H_1 & & & & \\ & & U & -H_1 & & & \\ & & & H_1 & -H_2 & & \\ & & & & U & -H_2 & \\ & & & & & H_2 & -H_3 \end{pmatrix},$$

where the unwritten entries are zero forms. The essential elements are: Φ, F_{13}, F_{22}, F_{31}. *All the other elements are s.i..*

- *If in each generator of degree* 8 *we replace* U^3 *by any degree* 3 *form* η, *with* $G.C.D.(\eta, \Phi) = 1$, *we obtain a new e-maximal basis.*

- If we replace F_{21}, F_{22}, F_{31} respectively by $\tilde{F}_{21} = H_2H_3U^6$, $\tilde{F}_{22} = H_3U^7$, $\tilde{F}_{31} = H_2U^7$, then in the new matrix $\tilde{M}(\mathfrak{I})$ the H_1 in $(4,4)$ position is replaced by U. As a consequence, $\tilde{F}_{11} = F_{11}$ and $\tilde{F}_{12} = F_{12}$ are s.i., while $\tilde{F}_{13} = F_{13}$ is inessential, but not strongly and $\tilde{F}_{21} \neq F_{21}$ is s.i..

- If we replace F_{31} by $\tilde{F}_{31} = H_1U^7$ (or, equivalently, in $M(\mathfrak{I})$ the form H_2 in position $(6,6)$ is replaced by U), then F_{22} becomes inessential (but not strongly), while the nature of the other generators does not change.

The two special cases just considered suggest us to afford the general case pointing our attention on the H.B. matrix, more than on the standard basis. We need a decomposition of the Φ's appearing in (3) into pairwise independent linear forms, as follows

$$\Phi_k = H_{k1}^{\mu_{k1}} \dots H_{k2}^{\mu_{k2}} \dots H_{kv_k}^{\mu_{kv_k}},\ k = 1, \dots, r. \tag{14}$$

Moreover, it is convenient to choose a set of linear forms $\{U, L_0, \dots, L_{\beta_0}\}$ such that the elements of the set $\{U, L_i, H_{kj}\},\ i = 0, \dots, \beta_0,\ k = 1, \dots, r,\ j = 1, \dots, v_k$ are pairwise linearly independent and define

$$\Phi_0 = L_0 \dots L_{\beta_0}. \tag{15}$$

With this notation we can state the following proposition.

PROPOSITION 4.14. *A canonical matrix of the ideal $\mathfrak{I}$ of (3) is the following one*

$$M(\mathfrak{I}) = \begin{pmatrix} \mathcal{B} & \mathcal{O} \\ \mathcal{C} & \mathcal{A} \end{pmatrix},$$

where:

i) $\mathcal{B} \in S^{\beta_0 \times (\beta_0+1)}$, $\mathcal{A} \in S^{\delta \times \delta}$, $\mathcal{C} \in S^{\delta \times (\beta_0+1)}$, $\mathcal{O}$ is a zero matrix, whose elements are of degree ≤ 0.

ii) $\mathcal{B} = (b_{ij})$, where: $b_{ii} = L_{i-1}$; $b_{i(i+1)} = -L_i$; $b_{ij} = 0$ if $j \neq i, i+1$.

iii) $\mathcal{C} = (c_{ij})$, where: $c_{1(\beta_0+1)} = U^{t_1+1}$; $c_{ij} = 0$ if $(ij) \neq (1(\beta_0+1))$

iv) $\mathcal{A} = (a_{ij})$, where:

- $a_{ij} = 0$ *if* $j \neq i, j \neq i-1$,
- $(a_{11}, \dots, a_{\delta,\delta}) = ((-H_{11})^{[\mu_{11}]}, (-H_{12})^{[\mu_{12}]}, \dots, (-H_{1v_1})^{[\mu_{1v_1}]}, \dots$
  $\dots, (-H_{k1})^{[\mu_{k1}]}, (-H_{k2})^{[\mu_{k2}]}, \dots, ((-H_{kv_k})^{[\mu_{kv_k}]}, \dots$
  $\dots, (-H_{r1})^{[\mu_{r1}]}, (-H_{r2})^{[\mu_{r2}]}, \dots, (-H_{rv_r})^{[\mu_{rv_r}]}),$

- $a_{(i+1)i} = -a_{ii}$ *if* $a_{ii} \neq a_{jj}, j > i,\ i \neq \Delta_k,$
  $a_{(i+1)i} = -a_{ii}U^{t_k}$ *if* $a_{ii} \neq a_{jj}, j > i,\ i = \Delta_k,$
  $a_{(i+1)i} = -U$ *if* $(\exists j > i)\ a_{ii} = a_{jj},\ j \neq \Delta_k,\ k < r,$
  $a_{(i+1)i} = U^{t_k+1}$ *if* $(\exists j > i)\ a_{ii} = a_{jj},\ i = \Delta_k,\ k < r.$

Moreover, the inessential columns of $M(\mathfrak{I})$ *are s.i..*

Proof. We first observe that the degree matrix of $M(\mathfrak{I})$ is the expected one. Let us denote $\mathfrak{I}$ the ideal generated by the maximal minors of $M(\mathfrak{I})$ and prove that $\mathfrak{I}$ is the one described in (3). As $\mathcal{B}$ is the matrix considered in Proposition 4.2, it is immediate to see that $\mathfrak{I}_{\alpha_0} = \Phi S_{\beta_0}$ and that the minors of $\mathcal{B}$ are linearly independent.

The minors in degree α_1 have as a common factor Φ/Φ_1. So, it is enough to prove that, divided by their common factor, they are a basis of a subspace T_1 of S_{β_1} such that $S_{\beta_1} = \Phi_1 S_{\beta_1-\delta_1} \bigoplus T_1$. But we are in the situation described in Lemma 4.3, where:

$$t = t_1,\ b = \beta_1,\ \Phi = \Phi_1,\ H_i = H_{1i}\,,$$
$$C_i = \Phi_0 L_{\beta_0}^{-1} U^{t_1+1} a_{21} \dots a_{(j+1)j},\ j = \mu_{11} + \mu_{12} + \dots + \mu_{1(i-1)},\ i = 1, \dots, v_1\,.$$

So, let us suppose the statement true until the degree α_{k-1} and prove it for α_k. Just as in the case $k = 1$, we see that all the minors have as a common factor $\Phi_{k+1} \dots \Phi_r = \Phi/\Phi_1 \dots \Phi_k$. So, it is enough to show that, divided by this factor, they are a basis of a subspace T_k of S_{β_k} such that $S_{\beta_k} = \Phi_k S_{\beta_k-\delta_k} \bigoplus T_k$. We are again in the situation of Lemma 4.3, with:

$$t = t_k,\ b = \beta_k,\ \Phi = \Phi_k,\ H_i = H_{ki}$$
$$C_i = \Phi_0 L_{\beta_0}^{-1} U^{t_1+1} a_{21} \dots a_{(\Delta_{k-1}+j+1)(\Delta_{k-1}+j)},$$
$$j = \mu_{k1} + \mu_{k2} + \dots + \mu_{k(i-1)},\ i = 1, \dots, v_k.$$

Thanks to Proposition 3.2, we immediately see that the inessential columns are exactly the ones in which $a_{(i+1)i}$ is not a multiple of a_{ii} or, equivalently, the ones whose element a_{ii} is equal to some a_{jj}, with $j > i$. The proof that they are s.i. is a consequence of Lemma 4.8. □

Extending the notation used in Proposition 4.6, we denote the basis linked to the canonical matrix of Proposition 4.14 as follows:

$$\mathcal{B}(\mathfrak{I}) = (\mathcal{B}^0, \mathcal{B}^1, \dots, \mathcal{B}^k \dots, \mathcal{B}^r), \quad \text{where}$$
$$\mathcal{B}^0 = (F_j^0),\ j = 0, \dots, \beta_0,\ \mathcal{B}^k = (F_{ij}^k),$$
$$k = 1, \dots, r,\ i = 1, \dots, v_k,\ j = 1, \dots, \mu_{ki}.$$

With this notation we can state the following corollary.

COROLLARY 4.15. *i) All the elements of $\mathcal{B}^0$ are essential. The generator $F^k_{ij} \in \mathcal{B}^k$ is essential iff $j = \mu_{ki}$ and H_{ki} is not a factor of it.*

ii) $\mathcal{B}(\mathfrak{I})$ is an e-maximal basis and the number of its essential elements in degree bigger then $\alpha(\mathfrak{I})$ is equal to the number v of the distinct linear factors appearing in a factorization of Φ.

iii) In any e-maximal basis the essential generators appearing in degree α_k are as many as the linear factors of Φ_k that do not divide $\Phi_{k+1} \dots \Phi_r$.

iv) $\mathfrak{I}$ admits a basis of essential elements iff Φ is a product of distinct linear factors.

Proof. i) From Proposition 4.14 iv) we easily see that the essential columns of $\mathcal{A}$ are the ones whose entry a_{hh} is different from every $a_{jj}, j > h$. This happens iff $a_{hh} = -H_{ki}$, where H_{ki} does not appear any more in the diagonal of $\mathcal{A}$, in position $(j, j), j > h$. A necessary condition for such a situation is that the generator corresponding to that column is of the kind $F^k_{i\mu_{ki}}$. In this case we have: $\prod_{j>h} a_{jj} = R\Phi_{k+1} \dots \Phi_r$, where H_{ki} is not a factor of R. So, the condition characterizing the essential $F^k_{i\mu_{ki}}$'s is that $\Phi_{k+1} \dots \Phi_r$ is not a multiple of H_{ki}. From the equality $F^k_{i\mu_{ki}} = \prod_{j>h} a_{jj} \prod_{j<h} a_{(j+1)j}$ we see that the previous condition is equivalent to say that H_{ki} does not divide $F^k_{i\mu_{ki}}$.

ii) $\mathcal{B}(\mathfrak{I})$ is an e-maximal basis, because its inessential elements are s.i. (Theorem 3.5). Moreover, the H_{ki} appearing in an essential column corresponding to $F^k_{i\mu_{ki}}$ is a linear factor of Φ, making there its last appearing as an element of the diagonal of $\mathcal{A}$. So, the essential columns of $\mathcal{A}$ are as many as the distinct linear factors of Φ.

iii) As the number of essential elements in an e-maximal basis does not depend on the e-maximal basis chosen, it is enough to verify the statement on the basis $\mathcal{B}(\mathfrak{I})$ of Proposition 4.14. In the proof of i) we observed that the essential elements of $\mathcal{B}^k$ are as many as the linear factors H_{ki} of Φ_k that are not divisors of $\Phi_{k+1} \dots \Phi_r$.

iv) This is an obvious consequence of ii).

□

Corollary 4.15 completes the proof of Theorem 4.1.

COROLLARY 4.16. *Let $\mathfrak{I}$ be represented as in (3), with $\Phi_k = H_{k1}^{\mu_{k1}} \dots H_{kv_k}^{\mu_{kv_k}}$. If τ_k is the number of distinct linear factors that Φ_k has in common with $\Phi_{k+1} \dots \Phi_r$, then any e-maximal basis of $\mathfrak{I}$ has exactly $\sum_{j=1}^{v_k}(\mu_{kj} - 1) + \tau_k$ strongly inessential generators in degree α_k.*

Corollary 4.16 implies that it is possible to find $\mathfrak{I} \in F[2]$ with a prescribed number of strongly inessential elements in a prescribed number of sufficiently high degree, as we see in the following proposition.

PROPOSITION 4.17. *Let $(d_1 < d_2 < \ldots < d_s)$ and $(r_1, r_2, \ldots, r_s)$ be sequences of natural numbers. There exist ideals $\mathfrak{I} \in F[2]$ with exactly r_i s.i. elements in degree d_i, $i = 1, \ldots, s$, iff*

$$d_1 > \sum_{i=1}^{s} r_i + 1. \tag{16}$$

Proof. Let us observe that the minimal degree δ of a form Φ satisfying the condition $\delta - v = \sum_{i=1}^{s} r_i$ is obtained with $v = 1$, so that Φ looks as $\Phi = H^{m+1}$, where $m = \sum_{i=1}^{s} r_i$ and H is any linear form. So, condition (16) is necessary. It is also sufficient, because the ideal

$$\begin{aligned} \mathfrak{I} \;=\; & H^{m+1}S + H^{m+1-r_1}S_{d_1-(m+1-r_1)} + \cdots \\ & \ldots + H^{m+1-\sum_{i=1}^{j} r_i}S_{d_j-(m+1-\sum_{i=1}^{j} r_i)} + \ldots + S_{d_s} \end{aligned} \tag{17}$$

obtained with the choice $\Phi_j = H^{r_j}, j = 1, \ldots, (s-1)$, $\Phi_s = H^{r_s+1}$, satisfies the required condition. If $d_1 = \sum_{i=1}^{s} r_i + 2$, then (17) is the unique ideal satisfying the condition. If $d_1 > \sum_{i=1}^{s} r_i + 2$, there are many other possibilities. In fact, the set of the ideals satisfying the required condition increases with the degree $\delta = v + m$, or, equivalently, with the number v of different linear factors of Φ. Let us observe that the degree vector of the ideal $\mathfrak{I}$ considered in (17) is the least compatible with the required condition. □

REMARK 4.18. *Every $\mathfrak{I} \in F[2]$ satisfies condition (2) (maximality with respect to Dubreil-Campanella inequality) in each degree α_i. So, for every j, $\mathfrak{I}$ splits into two ideals, $\mathfrak{I}' = (\mathfrak{I} : (\Phi_{j+1} \ldots \Phi_r))$ and $\mathfrak{I}'' = (\mathfrak{I}, \Phi_{j+1} \ldots \Phi_r)$, both elements of $F[2]$. The first $\beta_0 + 1 + \Delta_j$ rows and $\beta_0 + \Delta_j$ columns of the matrix $M(\mathfrak{I})$ produced in Proposition 4.14 form a H.B. matrix of $\mathfrak{I}'$, whose inessential columns are not necessarily s.i.. In fact, it may happen that a linear factor of $\Phi_i, i \leq j$ does not divide $\Phi_{i+1} \ldots \Phi_j$ but divides $\Phi_{j+1} \ldots \Phi_r$; so the assertion of Remark 3.11 is justified.*

EXAMPLES 4.19. *In the following examples U, H, K, L_0, L_1, L_2 are linear forms, pairwise linearly independent.*

1- $\mathfrak{I} = H^3K^2S_2S + K^2S_6S + S_{10}S.$

In this case we have: $\Phi = H^3K^2, \Phi_1 = H^3, \Phi_2 = K^2$, $GCD(\Phi_1, \Phi_2) = 1$. According to Proposition 4.14, we get

$$M(\mathfrak{I}) = \begin{pmatrix} L_0 & -L_1 & & & & & & \\ & L_1 & -L_2 & & & & & \\ & & L_2U & -H & & & & \\ & & & U & -H & & & \\ & & & & U & -H & & \\ & & & & & HU^2 & -K & \\ & & & & & & U & -K \end{pmatrix}.$$

The corresponding canonical basis is

$$\mathcal{B}(\mathfrak{I}) = (H^3K^2(L_1L_2, L_0L_2, L_0L_1); K^2L_0L_1L_2U(H^2, HU, U^2); \\ L_0L_1L_2HU^5(K, U)).$$

There are 3 *s.i. generators, according to the fact that* $\delta = 5, v = 2$. *Let us observe that in this example a s.i. generator gives rise to a s.i. generator in any splitting.*

2- $\mathfrak{I} = H^3K^2S_2S + HKS_6S + S_{10}S$.

In this case we have: $\Phi = H^3K^2, \Phi_1 = H^2K, \Phi_2 = HK$, *so that all the linear factors of* Φ_1 *are also divisors of* Φ_2. *According to Proposition 4.14, we get*

$$M(\mathfrak{I}) = \begin{pmatrix} L_0 & -L_1 & & & & & & \\ & L_1 & -L_2 & & & & & \\ & & L_2U & -H & & & & \\ & & & U & -H & & & \\ & & & & U & -K & & \\ & & & & & U^3 & -H & \\ & & & & & & H & -K \end{pmatrix}.$$

The corresponding canonical basis is

$$\mathcal{B}(\mathfrak{I}) = (H^3K^2(L_1L_2, L_0L_2, L_0L_1); HKL_0L_1L_2U(HK, UK, U^2); \\ L_0L_1L_2U^6(K, H)).$$

There are 3 *s.i. generators, according to the fact that* $\delta = 5, v = 2$.

Let us observe that in this case the splitting in degree $p = 8$ *gives rise to the matrices*

$$M(\mathfrak{I}') = \begin{pmatrix} L_0 & -L_1 & & & & \\ & L_1 & -L_2 & & & \\ & & L_2U & -H & & \\ & & & U & -H & \\ & & & & U & -K \end{pmatrix},$$

and, respectively,

$$M(\mathfrak{I}'') = \begin{pmatrix} U^9 & -H & \\ & H & -K \end{pmatrix}.$$

The 6-th column of $M(\mathfrak{I}')$ is inessential, but not s.i.. A H.B. matrix of $\mathfrak{I}'$, whose corresponding inessential generators are s.i., can be obtained from $M(\mathfrak{I}')$ just by replacing U with H in its 6-th column.

3- $\mathfrak{I} = H^3S_2 + H^2S_4 + HS_8 + S_{10}$.

In this case we have: $\Phi = H^3, \Phi_1 = \Phi_2 = \Phi_3 = H$. According to Proposition 4.14, we get:

$$M(\mathfrak{I}) = \begin{pmatrix} L_0 & -L_1 & & & \\ & L_1 & -L_2 & & \\ & & L_2U & -H & \\ & & & U^4 & -H \\ & & & & U^2 & -H \end{pmatrix}.$$

There are 2 s.i. generators and the corresponding canonical basis is:

$\mathcal{B}(\mathfrak{I}) = (H^3(L_1L_2, L_0L_2, L_0L_1); H^2L_0L_1L_2U; HL_0L_1L_2U^5; L_0L_1L_2U^7)$

The splittings in degrees 8 and 9 produce a new inessential, but not s.i., element in $\mathfrak{I}'$.

5. Behaviour of $\mathfrak{I} \in F[3]$ with respect to essentiality: special cases and examples

According to Theorem 1.5 of [3], every element $\mathfrak{I} \in F[3]$ has a shape very similar to the one described in (3) for the elements of $F[2]$. The difference is that S_{β_i} is replaced by a linear subspace $T_{\beta_i} \subseteq S_{\beta_i}$ of $S = K[X, Y, Z]$, where $\dim T_{\beta_i} = \beta_i + 1$. The subspaces T_{β_i} are characterized by Theorem 3.4 of [3]. That theorem says that, up to a change of cohordinates, every element $\mathfrak{I} \in F[3]$ is generated by the maximal minors of an $\alpha \times (\alpha+1)$ matrix, obtained by lifting to $K[X, Y, Z]$ a H.B. matrix of its image $\bar{\mathfrak{I}} \subset K[X, Y] = \bar{S}$, modulo a regular linear form Z. So, $\mathfrak{I}$ looks like

$$\begin{aligned} \mathfrak{I} \;=\; & \Phi_1 \dots \Phi_r T_{\beta_0} S + \dots + \Phi_i \dots \Phi_r T_{\beta_{i-1}} S + \Phi_{i+1} \dots \Phi_r T_{\beta_i} S + \dots \\ & \dots + \Phi_r T_{\beta_{r-1}} S + T_{\beta_r} S = \sum_{i=0}^{r} \Phi_{i+1} \dots \Phi_{r+1} T_{\beta_i} S, , \quad (18) \end{aligned}$$

where Φ_i is a form in S and $\Phi_{r+1} = 1$, and its image modulo Z becomes

$$\bar{\mathfrak{I}} = \bar{\Phi}_1 \dots \bar{\Phi}_r \bar{S}_{\beta_0} \bar{S} + \dots + \bar{\Phi}_i \dots \bar{\Phi}_r \bar{S}_{\beta_{i-1}} \bar{S} + \bar{\Phi}_{i+1} \dots \bar{\Phi}_r \bar{S}_{\beta_i} \bar{S} + \dots$$
$$\dots + \bar{\Phi}_r \bar{S}_{\beta_{r-1}} \bar{S} + \bar{S}_{\beta_r} \bar{S} = \sum_{i=0}^{r} \bar{\Phi}_{i+1} \dots \bar{\Phi}_{r+1} \bar{S}_{\beta_i} \bar{S},\ \bar{\Phi}_{r+1} = 1. \quad (19)$$

The problem of stating if a lifting of $\bar{\mathfrak{I}} \in F[2]$ to $\mathfrak{I} \in F[3]$ preserves the strong inessentiality of the entries of $\mathcal{B}(\bar{\mathfrak{I}})$ becomes a lifting problem of H.B. matrices, which seems not easy to be solved. So, we start to consider a very special case. More precisely, we focus our attention on the ideals $\mathfrak{I} \in F[3]$ with the largest number of s.i. generators in any e-maximal basis. If $\alpha = \alpha(\mathfrak{I}) = \alpha(\bar{\mathfrak{I}})$ is the minimal degree of the generators of $\mathfrak{I}$, we will see that the maximal expected number is $\alpha - 2$; we'll prove that such a number is reached. Let us first state a property for every homogeneous saturated ideal $\mathfrak{I}$ of $S = K[X_1, \dots, X_n]$.

PROPOSITION 5.1. *Let $\mathcal{B} = (b_1, \dots, b_h, c_1, \dots, c_k)$, $k \geq 1$, be an e-maximal basis of the saturated homogeneous ideal $\mathfrak{I} \subset S = K[X_1, \dots, X_n]$, where $b_1, \dots, b_h$ are essential and $c_1, \dots, c_k$ are s.i. elements. The condition depth $\mathfrak{I} = r$ implies $h > r$.*

Proof. Thanks to Corollary 5.3 of [4], $(c_1, \dots, c_k)$ is an inessential set ([4], Definition 5.2), so that $\mathfrak{I} = (b_1, \dots, b_h)^{sat}$. As depth $\mathfrak{I}$ = depth $(b_1, \dots, b_h)$, the hypothesis implies $h \geq r$; however, the equality holds iff $(b_1, \dots, b_h)$ is a c.i. and, as a consequence, a saturated ideal, against the hypothesis $k \geq 1$. □

Choosing $h = 2$, we get immediately the following statement.

COROLLARY 5.2. *The largest possible number of s.i. generators in an e-maximal basis of an ideal $\mathfrak{I} \in F[3]$ is $\alpha(\mathfrak{I}) - 2$.*

If $\alpha(\mathfrak{I}) = 2$, then $\mathfrak{I} \in F[3]$ has 3 generators and Corollary 5.2 says that in every e-maximal basis they are essential. Let us point our attention on the case $\alpha(\mathfrak{I}) > 2$.

We will use the following Notation

$$\mathcal{S} = \{\mathfrak{I} \in F[3] : \nu_e(\mathfrak{I}) = 3,\ \alpha(\mathfrak{I}) > 2\},$$

where $\nu_e(\mathfrak{I})$ denotes the number of essential elements of any e-maximal basis of $\mathfrak{I}$ (see [4], Definition 5.1).

Let us observe that any dehomogenization $\mathfrak{I}_*$ with respect to a regular linear form of an ideal $\mathfrak{I} \in \mathcal{S}$ has just 3 generators (that is the least number for a non complete intersection), while the number of generators of $\mathfrak{I}$ is the maximum allowed by Dubreil's inequality.

PROPOSITION 5.3. *For every $\mathfrak{I} \in \mathcal{S}$ the form Φ appearing in (18) is of degree $\delta = \alpha(\mathfrak{I})$ and Φ has necessarily one of the following shapes:*

i) $\Phi = H^\delta$,

ii) $\Phi = H^r K^s,\ r + s = \delta$,

iii) $\Phi = C^\gamma, 2\gamma = \delta$,

where H and K are independent linear forms and C is a quadratic irreducible form in $K[X, Y, Z]$.

Proof. An immediate consequence of Proposition 3.8 is that if $\mathfrak{I} \subset K[X, Y, Z]$ has $\alpha - 2$ s.i. generators, then the number of s.i. generators of its quotient $\bar{\mathfrak{I}}$ modulo a regular linear form is either $\alpha - 2$ or $\alpha - 1$. Applying Theorem 4.1 to $\bar{\mathfrak{I}}$, we immediately get

$$\delta - v = \alpha - 1,\ \delta \leq \alpha, \tag{20}$$

or

$$\delta - v = \alpha - 2,\ \delta \leq \alpha. \tag{21}$$

Relation (20) is equivalent to $\delta = \alpha,\ v = 1$, while relation (21) gives two possible situations:

$$\delta = \alpha,\ v = 2 \tag{22}$$

and

$$\delta = \alpha - 1,\ v = 1. \tag{23}$$

Let us verify that (23) is not realized. In fact in this case we have $(\bar{\mathfrak{I}})_\alpha = \bar{H}^{\alpha-1}\bar{S}_1$ and the s.i. generators lie all in degree bigger than α; so, a splitting in degree α (see Theorem 3.9) gives rise to an ideal $\mathfrak{I}''$, with $\alpha(\mathfrak{I}'') = \alpha - 1$ and $\alpha - 2$ s.i. generators, against Corollary 5.2.

So, Φ must be a form, of degree α, whose quotient modulo any regular linear form splits into a product of powers of at most two different linear factors. This means that the curve $\Phi = 0$ meets a generic line in at most two different points, so that Φ is necessarily as described in i), ii), iii). □

REMARK 5.4. 1- *We do not have examples in which the situation iii) appears. Let us observe that it requires every $\nu(\mathfrak{I}, j),\ j = 1, \ldots, r$, to be a power of 2.*

2- *Proposition 5.3 says that the schemes corresponding to ideals with $\alpha - 2$ s.i. generators lie necessarily either on a multiple line or on two multiple lines or (may be) on a multiple irreducible conic. However this condition is not sufficient. For instance, by lifting the canonical matrix M of an ideal $\mathfrak{J}$ of $K[X, Y]$ with $\alpha - 1$ s.i. generators with M itself, we obtain a basis for an ideal $\mathfrak{I} \subset K[X, Y, Z]$ without inessential elements.*

3- *In case i), $\bar{\Phi}$ has the same structure of Φ, for every regular form L, while in case ii) and iii) we have generically $\bar{\Phi} = \bar{H}^r \bar{K}^s$, but we get $\bar{\Phi} = \bar{M}^{r+s}$, M a linear form in $K[X,Y]$, iff the line $L = 0$ either is tangent to $\Phi = 0$ (case iii)) or meets it in its singular point (case ii)). So, it is possible to represent an element $\mathfrak{I} \in \mathcal{S}$ as a lifting of ideals $\bar{\mathfrak{I}} \subset k[X,Y]$ such that $\bar{\Phi} = \bar{M}^\delta$, except for the case in which $\Phi = H^r K^s$ and the intersection between the two lines $H = 0$ and $K = 0$ is in the support of the corresponding scheme.*

PROPOSITION 5.5. *If $\mathfrak{I}$ is an element of $\mathcal{S}$, let us consider its splitting into $\mathfrak{I}'$ and $\mathfrak{I}''$ (see Theorem 2.3), in degree α_k, with $k < r$ if $\deg \Phi_r \geq 2$ and $k < r-1$ if $\deg\ \Phi_r = 1$. Then $\mathfrak{I}''$ is still an element of $\mathcal{S}$.*

Proof. Theorem 3.9 says that if $\mathcal{B}(\mathfrak{I}) = (\Phi, \mathcal{B}^1, \mathcal{B}^2, \ldots, \mathcal{B}^r)$ is an e-maximal basis of $\mathfrak{I}$, then $\mathcal{B}(\mathfrak{I}'') = (\Phi'', \mathcal{B}^{k+1}, \ldots, \mathcal{B}^r)$, where $\Phi'' = \Phi_{k+1} \ldots \Phi_r$, is an e-maximal basis of $\mathfrak{I}''$ and the forms of $\mathcal{B}(\mathfrak{I}'')$, different from Φ'', maintain the same nature they had in $\mathcal{B}(\mathfrak{I})$. So, as $\mathcal{B}(\mathfrak{I})$ has two essential elements different from Φ, $\mathcal{B}(\mathfrak{I}'')$ cannot contain more then two essential elements different from Φ''; the hypothesis on the choice of k guaranties that it has at least 3 elements; then, Corollary 5.2 says that $\mathcal{B}(\mathfrak{I}'')$ must have exactly 3 essential elements, so that $\mathfrak{I}'' \in \mathcal{S}$. □

COROLLARY 5.6. *In any e-maximal basis of an ideal $\mathfrak{I} \in \mathcal{S}$ the degree of the three essential generators are: $\alpha = \alpha(\mathfrak{I})$, $\alpha_r(\mathfrak{I})$ and either $\alpha_r(\mathfrak{I})$ or $\alpha_{r-1}(\mathfrak{I})$. The latter possibility takes place iff in degree $\alpha_r(\mathfrak{I})$ there is just one generator.*

Proof. It is enough to apply Proposition 5.5, with $k = r-1$ if in degree α_r the basis $\mathcal{B}(\mathfrak{I})$ contains at least two forms and with $k = r-2$ otherwise. In fact the two essential elements of $\mathcal{B}(\mathfrak{I}'')$, not in minimal degree, must be essential also in $\mathcal{B}(\mathfrak{I})$. □

Now, let us produce examples of ideals of $\mathcal{S}$.

PROPOSITION 5.7. *Every $\bar{\mathfrak{I}} \subset K[X,Y]$ with $\alpha - 1$ s.i. generators in any e-maximal basis has at least a lifting in $\mathcal{S}$.*

Proof. After a possible change of coordinates, a H.B. matrix of $\bar{\mathfrak{I}}$ is

$$M(\bar{\mathfrak{I}}) = \begin{pmatrix} Y^{t_0} & -X & & & & & \\ & Y^{t_1} & -X & & & & \\ & & Y^{t_2} & -X & & & \\ \cdots & \cdots & \cdots & \cdots & \cdots & \cdots & \\ & & & & Y^{t_{\delta-2}} & -X & \\ & & & & & Y^{t_{\delta-1}} & -X \end{pmatrix} \tag{24}$$

A lifting $M(\mathfrak{I})$ of $M(\bar{\mathfrak{I}})$, with $\alpha-2$ s.i. columns, is

$$\begin{pmatrix} Y^{t_0} & -X & & & & & \\ ZP_1 & Y^{t_1} & -X & & & & \\ ZP_2 & Z^{t_1+t_2-1} & Y^{t_2} & -X & & & \\ ZP_3 & & Z^{t_2+t_3-1} & Y^{t_3} & & & \\ & \cdots & \cdots & \cdots\ \cdots & \cdots & \cdots & \\ ZP_{\delta-2} & & & & Y^{t_{\delta-2}} & -X & \\ ZP_{\delta-1} & & & & Z^{t_{\delta-2}+t_{\delta-1}-1} & Y^{t_{\delta-1}} & -X \end{pmatrix}, \tag{25}$$

where $P_i \in K[X,Y,Z]$ is a form of degree $\sum_{j=0}^{i} t_j - i - 1$ and, as usual, the unwritten entries are zero. $M(\mathfrak{I})$ is obtained from $M(\bar{\mathfrak{I}})$ by leaving unchanged the last two columns and replacing the zero entries in position $(i+1,i), i = 2,\ldots,\delta-1$, with $Z^{t_{i-1}+t_i-1}$ and the ones in position $(i,1)$ with ZP_{i-1}. It is immediate to verify that the second column is s.i. and Lemma 5.3 of [4] says that the same reasoning can be repeated for the following ones with three non zero entries. □

Finally, let us point our attention on the ideals $\mathfrak{I} \in \mathcal{S}$ with the smallest $\alpha(\mathfrak{I})$ allowing the presence of s.i. generators. Corollary 5.2 implies that if $\mathfrak{I}$ has a s.i. generator then $\alpha(\mathfrak{I}) \geq 3$. So, let us look for the ideals with $\alpha(\mathfrak{I}) = 3$ and just one s.i. generator in every e-maximal basis; they are the elements of $\mathcal{S}$ with the smallest number of generators. Let us consider the special case of generators in two different degrees. Proposition 5.3 says that, apart from a coordinates change, they can be obtained by lifting an ideal of one of the following types

$$\bar{\mathfrak{I}}_1 = X^3 S + S_\beta S,\ \beta > 3, \tag{26}$$

$$\bar{\mathfrak{I}}_2 = X^2 Y S + S_\beta S,\ \beta > 3. \tag{27}$$

Let us first consider all the required liftings of $\bar{\mathfrak{I}}_1$ or, equivalently, all the liftings $M(\mathfrak{I}_1)$ of the matrix

$$M(\bar{\mathfrak{I}}_1) = \begin{pmatrix} Y^t & -X & 0 & 0 \\ 0 & Y & -X & 0 \\ 0 & 0 & Y & -X \end{pmatrix}, \quad t = \beta - 2, \tag{28}$$

having a s.i. column. $M(\mathfrak{I})$ has the following shape (see [3])

$$M(\mathfrak{I}_1) = \begin{pmatrix} Y^t + ZP_1(X,Y,Z) & -X + a_{11}Z & a_{12}Z & a_{13}Z \\ ZP_2(X,Y,Z) & Y + a_{21}Z & -X + a_{22}Z & a_{23}Z \\ ZP_3(X,Y,Z) & a_{31}Z & Y + a_{32}Z & -X + a_{33}Z \end{pmatrix}, \tag{29}$$

where $a_{ij} \in K$, $\deg P_i = t-1$. The forms P_1, P_2, P_3 can be chosen arbitrarily among the ones of degree $t-1$, so that we just have to characterize the matrix $\mathcal{A} = (a_{ij})$, $i,j = 1,2,3$. As the first column of $M(\mathfrak{I}_1)$ is essential for every choice of the a_{ij}'s, let us consider the second and third columns. The second column is s.i. iff the forms $L_1 = -X + a_{11}Z + \lambda_2 a_{12}Z + \lambda_3 a_{13}Z$, $L_2 = Y + a_{21}Z + \lambda_2(-X + a_{22}Z) + \lambda_3 a_{23}Z$, $L_3 = a_{31}Z + \lambda_2(Y + a_{23}Z) + \lambda_3(-X + a_{33}Z)$ are linearly independent, for every choice of λ_2, λ_3 or, equivalently, iff the matrix

$$\mathcal{B} = \begin{pmatrix} -1 & 0 & a_{11} + \lambda_2 a_{12} + \lambda_3 a_{13} \\ -\lambda_2 & 1 & a_{21} + \lambda_2 a_{22} + \lambda_3 a_{23} \\ -\lambda_3 & \lambda_2 & a_{31} + \lambda_2 a_{32} + \lambda_3 a_{33} \end{pmatrix}$$

has determinant different from zero. Such a condition gives the relation

$$-a_{12}\lambda_2^3 - a_{13}\lambda_2^2\lambda_3 + (a_{22} - a_{11})\lambda_2^2 + a_{13}\lambda_3^2 + (a_{12} + a_{23})\lambda_2\lambda_3 + (-a_{32} + a_{21})\lambda_2 + (-a_{33} + a_{11})\lambda_3 - a_{31} \neq 0. \quad (30)$$

An easy computation shows that the matrices $\mathcal{A}$ for which this condition is satisfied are

$$\mathcal{A} = \begin{pmatrix} a_{11} & 0 & 0 \\ a_{21} & a_{11} & 0 \\ a_{31} & a_{21} & a_{11} \end{pmatrix}, \qquad a_{31} \neq 0. \quad (31)$$

Considerations very similar to the previous ones lead to the conclusion that the second column is s.i. iff the matrix $\mathcal{A}$ has the following shape

$$\mathcal{A} = \begin{pmatrix} a_{11} & a_{12} & 0 \\ a_{21} & a_{11} & -a_{12} \\ 0 & a_{21} & a_{11} \end{pmatrix}, \qquad a_{12} \neq 0. \quad (32)$$

In case (31), after the coordinate change $-a_{11}Z + X = X'$, $a_{21}Z + Y = Y'$, $a_{31}Z = Z'$ and dropping the apostrophes, the required matrix can be written

$$M(\mathfrak{I}_{11}) = \begin{pmatrix} Y^t + ZQ_1 & -X & 0 & 0 \\ ZQ_2 & Y & -X & 0 \\ ZQ_3 & Z & Y & -X \end{pmatrix} \quad (33)$$

In case (32), after the coordinate change $-a_{11}Z + X = X'$, $-a_{21}Z + Y = Y'$, $a_{12}Z = Z'$ and dropping the apostrophes, the required matrix can be written

$$M(\mathfrak{I}_{12}) = \begin{pmatrix} Y^t + ZQ_1 & -X & Z & 0 \\ ZQ_2 & Y & -X & -Z \\ ZQ_3 & 0 & Y & -X \end{pmatrix}, \quad t = \beta - 2. \quad (34)$$

Let us observe that both schemes relative to $\mathfrak{I}_{11}$ and $\mathfrak{I}_{12}$ are supported at at most $t+1$ points lying on a triple line ($X = 0$ in our basis) and that their multiplicity is $e(\mathfrak{I}) = 3 + 3t$ ([6]).

With a very similar computation it is possible to see that, apart from a coordinate change, a lifting $\mathfrak{I}_2$ of $\bar{\mathfrak{I}}_2$ belongs to $\mathcal{S}$ iff its H.B. matrix has the following shape

$$M(\mathfrak{I}_2) = \begin{pmatrix} (X+Y)^t + ZQ_1 & -X & 0 & 0 \\ ZQ_2 & X+Y & -X & 0 \\ ZQ_3 & Z & X & -Y \end{pmatrix}, \tag{35}$$

where P_1, P_2, P_3 are forms of degree $t-1$ in $K[X,Y,Z]$. The corresponding schemes, still of multiplicity $e(\mathfrak{I}) = 3t+3$, are all supported at two different lines ($X=0$ and $Y=0$). The intersection of the two lines is one of the points in the support of the scheme; as a consequence, the ideals cannot be obtained by lifting an ideal of type (28).

The characterization of the elements of $\mathcal{S}$ with $\alpha(\mathfrak{I}) > 3$ is more difficult to be faced, also for ideals generated in two degrees. In fact, the request of (28) (and the analogous for the lifting of $\bar{\mathfrak{I}}_2$) are replaced by the requirement that a system of non linear equations $\{E_u(a_{ij}, \lambda_v) = 0\}$, in a set $\{\lambda_v, v = 1, \ldots, \alpha-1\}$ of variables, admits no solutions. Such a condition defines the entries a_{ij}'s of the matrix $\mathcal{A}$ as the elements for which the ideal generated by the E_u's is the whole ring $K[\lambda_1, \ldots, \lambda_{\alpha-1}]$.

References

[1] G. Beccari, E. Davis, and C. Massaza, *Extremality with respect to the estimates of Dubreil-Campanella: the Hilbert matrix*, Rend. Sem. Mat. Univ. Politec. Torino **48** (1990), 457–48.

[2] G. Beccari, E. Davis, and C. Massaza, *Extremality with respect to the estimates of Dubreil-Campanella: splitting theorems*, J. Pure Appl. Algebra **70** (1991), 211–225.

[3] G. Beccari and C. Massaza, *Perfect homogeneous ideals of height two, with standard basis of maximal cardinality*, Boll. Un. Mat. Ital. D (6) **5** (1986), 201–223.

[4] G. Beccari and C. Massaza, *Essential and inessential elements of a standard basis*, J. Pure Appl. Algebra **215** (2011), 1726–1736.

[5] G. Campanella, *Standard bases of perfect homogeneous polynomial ideals of height 2*, J. Algebra **101** (1986), 47–60.

[6] C. Ciliberto, A.V. Geramita, and F. Orecchia, *Remarks on a theorem of Hilbert-Burch*, Boll. Un. Mat. Ital. B (7) **2** (1988), 463–488.

[7] E. Davis, *Complements to a paper of P.Dubreil*, Ricerche Mat. **37** (1988), 347–357.

[8] E. Davis, A. Geramita, and F. Maroscia, *Perfect homogeneous ideals: Dubreil's theorem revisited*, Bull. Sci. Math. **108** (1984), 143–185.

[9] P. Dubreil, *Sur quelques propriétés des systèmes de points dans le plan des courbes gauches algébriques*, Bull. Soc. Math. France **61** (1933), 258–283.

[10] W. Fulton, *Algebraic curves. An introduction to Algebraic Geometry*, W.A., 1974.
[11] R Hartshorne, *Algebraic Geometry*, Graduate Texts in Mathematics, no. 52, Springer-Verlag, New York, 1977.
[12] F.S. Macaulay, *Some properties of enumeration in the theory of modular systems*, Proc. London Math. Soc. **26** (1927), 531–555.
[13] D.G. Northcott, *Finite free resolutions*, Cambridge University Press, 1976.
[14] R.P. Stanley, *Hilbert functions of graded algebras*, Advances in Math. **28** (1978), 57–83.
[15] O. Zariski and P. Samuel, *Commutative Algebra, vol. 2*, Van Nostrand Company, 1963.

Authors' addresses:

Giannina Beccari
Dipartimento di Scienze Matematiche
Politecnico di Torino
Corso Duca degli Abruzzi 24, 10129 Torino, Italy
E-mail: giannina.beccari@polito.it

Carla Massaza
Dipartimento di Scienze Matematiche
Politecnico di Torino
Corso Duca degli Abruzzi 24, 10129 Torino, Italy
E-mail: carla.massaza@polito.it

Received July 20, 2012
Revised November 13, 2012

Rend. Istit. Mat. Univ. Trieste
Volume 44 (2012), 503–511

Katětov order, Fubini property and Hausdorff ultrafilters[1]

Michael Hrušák and David Meza-Alcántara

Abstract. *We study the Fubini property of ideals on ω and prove that the Solecki's ideal $\mathcal{S}$ is critical for this property in the Katětov order. We show that a well-known F_σ-ideal is critical for Hausdorff ultrafilters in the Katětov order and, by investigating the position of this ideal in the Katětov order, we show some of the known properties of this class of ultrafilters, including the Fubini property.*

Keywords: Hausdorff ultrafilter, Katětov order, Fubini property
MS Classification 2010: 03E15, 03C20, 03H15

1. Introduction

An ultrafilter $\mathcal{U}$ on an infinite set is *Hausdorff* if the ultrapower of $\mathbb{N}$ modulo $\mathcal{U}$, equipped with the S-topology, is Hausdorff. The S-topology is defined for non-standard models ${}^*X$ of a topological space X, as the generated by the ${}^*A$ sets, for open sets $A \subseteq X$. In the particular case of the ultrapower $\mathbb{N}^{\mathbb{N}}/\mathcal{U}$ as a non-standard model for the first-order arithmetic, we consider $\mathbb{N}$ equipped with the discrete topology, and then, the S-topology on $\mathbb{N}^{\mathbb{N}}/\mathcal{U}$ is Hausdorff if and only if, for every $f, g \in \mathbb{N}^{\mathbb{N}}$ there exists $U \in \mathcal{U}$ such that either $f \restriction U = g \restriction U$ or $f''U \cap g''U = \emptyset$ (see Proposition 2.1).

Hausdorff ultrafilters have been studied recently by several authors, see e.g. by M. di Nasso and M. Forti [6]. The main question about them is their existence, that is, does ZFC prove the existence of a Hausdorff ultrafilter? In this note we characterize this class of ultrafilters by using the Katětov order and an F_σ-ideal on the integers that we call $\mathcal{G}_{fc}$.

The *Katětov order* is defined as follows: for any two ideals I, J on countable sets X and Y respectively, $\mathsf{I} \leq_K \mathsf{J}$ if there is a function f from Y to X so that $f^{-1}[I] \in \mathsf{J}$ for all $I \in \mathsf{I}$. We write $\mathsf{I} \leq_{KB} \mathsf{J}$ (the *Katětov-Blass order*) when f is a finite-to-one function. An introduction to the Katětov order can be found in [8].

[1]The research of first and second authors was partially supported by PAPIIT grant IN101608 and CONACYT grant 80355. Second author was supported by grants PROMEP-UMSNH-NPTC-284 and UMSNH-CIC-9.30.

Katětov order is closely connected to Baumgartner's notion of I-ultrafilter. An ultrafilter $\mathcal{U}$ is an I-*ultrafilter* if and only if $\mathsf{I} \not\leq_K \mathcal{U}^*$. Several classes of ultrafilters are characterized as I-ultrafilters, for example, selective ideals are exactly the $\mathcal{ED}$-ultrafilters (see [7, 10, 14]).

Information about the position of ideals in the Katětov order provides information about belonging to classical families of ultrafilters, like P-points, Q-points and selective ultrafilters, since the I-ultrafilters (in the sense of Baumgartner [1]) are exactly the ultrafilters $\mathcal{U}$ such that $I \not\leq_K \mathcal{U}^*$.

We also study a property that Kanovei and Reeken [12] call the Fubini property. It concerns ideals (and filters) in general. For simplicity, we use a common notation: for any $A \subseteq \omega \times 2^\omega$, $n \in \omega$ and $x \in 2^\omega$ we denote $(A)_n = \{y \in 2^\omega : (n, y) \in A\}$ and $(A)^x = \{k \in \omega : (k, x) \in A\}$.

Definition 1.1. I *satisfies the* Fubini property *if for any Borel subset A of $\omega \times 2^\omega$ and any $\varepsilon > 0$, $\{n < \omega : \lambda((A)_n) > \varepsilon\} \in \mathsf{I}^+$ implies $\lambda^*(\{x \in 2^\omega : (A)^x \in \mathsf{I}^+\}) \geq \varepsilon$ (here λ^* means the Lebesgue outer measure on 2^ω).*

Particularly relevant for this work are the following ideals:

1. $\mathcal{ED} = \{A \subseteq \mathbb{N}^2 : \exists n \forall m > n |A \cap (\{m\} \times \mathbb{N})| \leq n\}$ is critical for selective ultrafilters in the Katětov order.

2. Let us denote by Δ the set $\{(n, m) : m \leq n\}$. Then, the ideal $\mathcal{ED}_{fin} = \{I \cap \Delta : I \in \mathcal{ED}\}$ on Δ is critical for Q-point ultrafilters in the Katětov-Blass order.

3. The *Solecki's ideal* $\mathcal{S}$ on the countable set Ω of all the clopen subsets of $2^\mathbb{N}$ with Lebesgue-measure equal to $\frac{1}{2}$, is generated by the family $\{A \subseteq \Omega : \bigcap A \neq \emptyset\}$. It was defined in [16], where the author proved that $\mathcal{S}$ is critical for the Fatou's property.

4. $\mathcal{G}_{fc} = \{A \subseteq [\mathbb{N}]^2 : ch(A) < \infty\}$, the ideals of graphs with finite chromatic number,[1] was used by Solecki in [16], where he asked if this ideal is critical for the Fatou property. This question was answered in the negative in [11].

5. $\mathcal{G}_c = \{A \subseteq [\mathbb{N}]^2 : \forall B \in [\mathbb{N}]^{\aleph_0}([B]^2 \setminus A \neq \emptyset)\}$, the ideal of graphs without infinite complete subgraphs.

The first four ideals are F_σ while the last is co-analytic.

---

[1]The *chromatic number* $ch(A)$ of a graph A on ω is defined as the minimal cardinal number κ for which there is a *coloring* $c : \omega \to \kappa$ so that $c(a) \neq c(b)$ for all $\{a, b\} \in A$.

2. Hausdorff ultrafilters and $\mathcal{G}_{fc}$

We now prove that $\mathcal{G}_{fc}$ is critical for Hausdorff ultrafilters in the Katětov order, i.e. $\mathcal{U}$ is Hausdorff if and only if $\mathcal{G}_{fc} \not\leq_K \mathcal{U}$. First we prove the following easy characterizations of Hausdorff ultrafilters. Note that f and g are $\mathcal{U}$-equivalent if and only if there is $U \in \mathcal{U}$ such that $f \restriction U = g \restriction U$.

PROPOSITION 2.1 ([6]). *The following conditions are equivalent, for any ultrafilter $\mathcal{U}$ on $\mathbb{N}$.*

1. *$\mathcal{U}$ is Hausdorff,*
2. *for every $f, g \in \mathbb{N}^{\mathbb{N}}$, f and g are $\mathcal{U}$-equivalent or $f''U \cap g''U = \emptyset$ for some $U \in \mathcal{U}$, and*
3. *for every $f, g \in \mathbb{N}^{\mathbb{N}}$, if $f(\mathcal{U}) = g(\mathcal{U})$ then there is $U \in \mathcal{U}$ such that $f \restriction U = g \restriction U$.*

Proof. We denote by $[h]$ the equivalence class of $h \in \mathbb{N}^{\mathbb{N}}$ modulo $\mathcal{U}$. $(1 \Rightarrow 2)$ If f and g are not $\mathcal{U}$-equivalent then there is $A \subseteq \mathbb{N}$ such that $[f] \in {}^*A$ and $[g] \in {}^*(N \setminus A)$, which means that there are V and W in $\mathcal{U}$ so that $f''V \subseteq A$ and $g''W \subseteq \mathbb{N} \setminus A$. Let $U = V \cap W$.
$(2 \Rightarrow 3)$ Assume $f \restriction X \neq g \restriction X$ for all $X \in \mathcal{U}$, and take U as in (2). From $f''(U) \in f(\mathcal{U})$ and $g''(U) \in g(\mathcal{U})$ follows $f(\mathcal{U}) \neq g(\mathcal{U})$.
$(3 \Rightarrow 1)$ If f and g are non-$\mathcal{U}$-equivalent then by (3) there is $A \in f(\mathcal{U}) \setminus g(\mathcal{U})$, and then $[f] \in {}^*A$ and $[g] \in {}^*(\mathbb{N} \setminus A)$. □

Now we describe a useful characterization of the ideal $\mathcal{G}_{fc}$. For each ordered pair $\langle A, B \rangle$ of nonempty disjoint subsets of $\mathbb{N}$, we define de set

$$I_{\langle A,B \rangle} = \{\{n, m\} : n \in A,\ m \in B,\ n < m\}$$

PROPOSITION 2.2. *$\mathcal{G}_{fc}$ is generated by the sets $I_{\langle A,B \rangle}$.*

Proof. On the one hand, it is clear that $ch(I_{\langle A,B \rangle}) \leq 2$. On the other hand, note that bipartite graphs are a base for $\mathcal{G}_{fc}$, since if $ch(G) = n$ then pick a coloring $c : \omega \to n$ so that $\{a, b\} \in G$ implies $c(a) \neq c(b)$, and for each pair $0 \leq i < j < n$ define $G_{i,j} = \{\{a, b\} : c(a) = i, c(b) = j\}$. Then, $G \subseteq \bigcup_{0 \leq i < j < n} G_{i,j}$. Finally, note that $I_{\langle A,B \rangle} \cup I_{\langle B,A \rangle}$ is the bipartite graph defined by A and B. □

We now prove the characterization of Hausdorff ultrafilters in the Katětov order, and additionally two graph-theoretic characterizations.

THEOREM 2.3. *The following conditions are equivalent for any ultrafilter $\mathcal{U}$ on $\mathbb{N}$*

1. *$\mathcal{U}$ is Hausdorff,*

2. *for every graph (G, E) and every $\varphi : \mathbb{N} \to E$, there exists $U \in \mathcal{U}$ such that $\varphi''U$ is contained in a bipartite graph.*

3. *for every graph (G, E) on $\mathbb{N}$ and every $\varphi : \mathbb{N} \to E$, there exists $U \in \mathcal{U}$ such that $ch(\varphi''U) < \infty$, and*

4. $\mathcal{G}_{fc} \not\leq_K \mathcal{U}^*$,

Proof. $(1 \to 2)$ Let us assume $\mathcal{U}$ is Hausdorff, and let φ be a function from $\mathbb{N}$ to $[\mathbb{N}]^2$. Define $f(n) = \min(\varphi(n))$ and $g(n) = \max(\varphi(n))$. It is clear that $f \neq g \mod \mathcal{U}$. By 2.1 there is $U \in \mathcal{U}$ such that $f''U \cap g''U = \emptyset$. Clearly, $I_{\langle f''U, g''U\rangle}$ is contained in a bipartite graph, and $\varphi''U \subseteq I_{\langle f''U, g''U\rangle})$.

$(2 \to 3)$ and $(3 \to 4)$ are immediate.

$(4 \to 1)$ Let us assume $\mathcal{G}_{fc} \not\leq \mathcal{U}^*$, and let f and g two non $\mathcal{U}$-equivalent functions. Since $\{n : f(n) = g(n)\} \notin \mathcal{U}$, either $\{n : f(n) > g(n)\} \in \mathcal{U}$ or $\{n : f(n) < g(n)\} \in \mathcal{U}$. Let us assume the first case (the other one is analogous), and define $\varphi(n) = \{g(n), f(n)\}$ if $g(n) < f(n)$, and $\varphi(n) = \{0, 1\}$ if not. Since there is $V \in \mathcal{U}$ such that $\varphi''V \in \mathcal{G}_{fc}$, and each element in $\mathcal{G}_{fc}$ is covered by a finite family of basic sets, there exist disjoint sets A and B so that for some $W \subseteq \{n \in V : g(n) < f(n)\}$ in $\mathcal{U}$, $\varphi''W \subseteq I_{\langle A,B\rangle}$, but this implies $f''W \subseteq A$ and $g''W \subseteq B$. □

About the position of $\mathcal{G}_{fc}$ some results are known: The identity function in $[\mathbb{N}]^2$ witnesses $\mathcal{G}_{fc} \leq_K \mathcal{G}_c$. Solecki proved in [16] that $\mathcal{S} \leq_K \mathcal{G}_{fc}$.

LEMMA 2.4. *[14]* $\mathcal{G}_{fc} \geq_{KB} \mathcal{ED}_{fin}$

Proof. Define f from $[\mathbb{N}]^2$ to $\mathbb{N} \times \mathbb{N}$ by

$$f(\{n, m\}) = (\max\{m, n\}, \min\{m, n\}).$$

This f witnesses the Katětov relation since the chromatic numbers of the f-preimages of sets $\{k\} \times \mathbb{N}$ are equal to 2, and the chromatic numbers the f-preimages of sets $H = \{(n, h(n)) : n \in \omega\}$ $(h \in \mathbb{N}^{\mathbb{N}})$ are also equal to 2, since we can construct recursively a coloring c by letting $c(0) = 0$, $c(1) = 1$ and for $n \geq 2$, $c(n) = 1 - c(h(n))$ if $h(n) < n$. Hence, if $\{n < m\} \in f^{-1}[H]$ then $n = h(m)$ and then $c(n) \neq c(m)$. □

Since $\mathcal{ED} \leq_{KB} \mathcal{ED}_{fin}$ (inclusion of Δ into $\omega \times \omega$ witnesses the Katětov-Blass relation), we get immediately the following corollary.

COROLLARY 2.5 (Daguenet-Teissier [5]). *Every selective ultrafilter is Hausdorff.*

3. Fubini property

In [12, Proposition 24], Kanovei and Reeken claimed without a proof that Fubini property is equivalent to the validity of Fatou's lemma. We will prove this as a corollary of the following Theorem, which is obtained by mimicking Solecki's proof of [16, Theorem 2.1].

THEOREM 3.1. *Let* I *be an ideal on* ω*. Then, there exists an* I*-positive set* X *such that* $\mathsf{I} \restriction X \geq_K \mathcal{S}$ *if and only if* I *does not satisfy the Fubini property.*

Proof. Let $f : X \to \Omega$ be a witness of $\mathsf{I} \restriction X \geq_K \mathcal{S}$, and define $A = \{(n, x) : x \in f(n)\}$. Note that $(A)_n = f(n)$ and then $\lambda((A)_n) = \frac{1}{2}$ for all $n \in X$. For any $x \in 2^\omega$, $\{S \in \Omega : x \in S\} \in \mathcal{S}$ and then $\{n < \omega : x \in (A)_n\} \in \mathsf{I}$ for all $x \in 2^\omega$.

On the other hand, assume that I does not satisfy the Fubini property, and take a Borel set $A \subseteq \omega \times 2^\omega$ such that for some $\varepsilon > 0$, the set $X := \{n < \omega : \lambda((A)_n) > \varepsilon\}$ is I-positive, and if $R := \{x \in 2^\omega : (A)^x \in \mathsf{I}^+\}$ then $\lambda^*(R) < \varepsilon$.

First, we can assume that (1) $R = \emptyset$, (2) for any $n \in X$, A_n is closed and (3) for any $n \in X$, $\lambda(A_n) = \varepsilon$. If it is not the case, we could replace (a) ε with $\varepsilon' = \varepsilon - \lambda^*(R)$ and (b) for each n, A_n with a closed subset A'_n of $A_n \setminus R'$, so that $\lambda(A'_n) = \varepsilon'$, where R' is a G_δ-set so that $R' \supseteq R$ and $\lambda(R') = \lambda^*(R)$.

Let $k < \omega$ be so that $(1 - \varepsilon)^k < \frac{1}{3}$. Recall that the power of Cantor space $(2^\omega)^k$ endowed with the product measure λ^k is isomorphic to the Cantor space 2^ω endowed with the Lebesgue measure λ, via a homeomorphism between those spaces. For any $n < \omega$, we define a subset B_n of $(2^\omega)^k$ by $B_n = \bigcup_{i=1}^k proj_i^{-1}[A_n]$. Then $(2^\omega)^k \setminus B_n = \prod_{i=1}^k (2^\omega \setminus A_n)$ and then $\lambda^k(B_n) > \frac{2}{3}$. Note that the family $\{B_n : n \in X\}$ fulfils that $R'' := \{x \in (2^\omega)^k : \{n < \omega : x \in B_n\} \in \mathsf{I}^+\} = \emptyset$, since if $x = \langle x_i : 1 \leq i \leq k\rangle$ then $\{n < \omega : x \in B_n\} = \bigcup_{i=1}^k \{n < \omega : x_i \in A_n\} \in \mathsf{I}$.

Now, for $n \in X$ choose a clopen subset U_n of $(2^\omega)^k$ such that $\lambda^k(U_n) \geq \frac{7}{12}$ and $\lambda^k(U_n \setminus B_n) < \frac{1}{3 \cdot 2^{n+2}}$. If $S := \{x \in (2^\omega)^k : \{n \in \omega : x \in U_n\} \in \mathsf{I}^+\}$ then $S \subseteq \bigcap_{m<\omega} \bigcup_{n \geq m} (U_n \setminus B_n)$, proving that $\lambda^k(S) = 0$. Let $\{C_n : n < \omega\}$ be an increasing family of clopen sets such that $S \subseteq \bigcup_{n<\omega} C_n$ and $\lambda^k(\bigcup_{n<\omega} C_n) \leq \frac{1}{12}$. Finally, by taking for any $n \in X$ a clopen subset $f(n)$ of $U_n \setminus C_n$ with $\lambda^k(f(n)) = \frac{1}{2}$ we get the Katětov function f wanted, since for any $x \in 2^\omega = (2^\omega)^k$, if $\{n \in X : x \in f(n)\}$ is infinite then $x \notin \bigcup C_n$ and then $x \notin S$. Hence $\{n \in X : x \in f(n)\} \in \mathsf{I}$ for all $x \in 2^\omega$. □

From Solecki's [16, Theorem 2.1] and the previous theorem we get:

COROLLARY 3.2. *If* I *is a universally measurable ideal on* ω *then* I *has the Fubini property if and only if* I *fulfils Fatou's lemma.* □

EXAMPLE 3.3. **Fin** *and* $\mathcal{Z}$ *have the Fubini property.*

Proof. (**Fin**) Since $\mathcal{S}$ is a tall ideal and **Fin** is K-uniform we have that $\mathcal{S} \not\leq_K \mathbf{Fin} \restriction X$, for all infinite subset X of ω.

($\mathcal{Z}$) Let $f : \omega \to \Omega$ be a function. By the classical Fubini's Theorem, for every $n < \omega$, there is $A_n \in \Omega$ such that for all $x \in A_n$,

$$|\{m \in [2^n, 2^{n+1}) : x \in f(m)\}| \geq 2^{n-1}.$$

Since **Fin** has the Fubini property, there is $x \in 2^\omega$ and there is an increasing sequence $\langle n_k : k \in \omega \rangle$ such that $x \in A_{n_k}$. Then, for any $k < \omega$,

$$\limsup_{n\to\infty} \frac{|f^{-1}[I_x] \cap [2^n, 2^{n+1})|}{2^n} \geq \lim_{k\to\infty} \frac{|f^{-1}[I_x] \cap [2^{n_k}, 2^{n_k+1})|}{2^{n_k}} \geq \frac{1}{2}$$

proving that f is not a witness for $\mathcal{S} \leq_K \mathcal{Z}$. □

4. Fubini and Hausdorff ultrafilters

Let $\mathcal{U}$ be an ultrafilter on ω, and A_n a Borel subset of Cantor space 2^ω, for all $n < \omega$. The $\mathcal{U}$-limit of the sequence of sets is the set defined by

$$\mathcal{U}\text{-}\lim A_n = \{x \in 2^\omega : \{n \in \omega : x \in A_n\} \in \mathcal{U}\}.$$

If $\langle x_n : n < \omega \rangle$ is a sequence of real numbers then $l \in \mathbb{R}$ is the $\mathcal{U}$-limit of the sequence provided that $\{n < \omega : |x_n - l| < \varepsilon\} \in \mathcal{U}$ for all $\varepsilon > 0$.

As usual, an $\mathcal{S}$-ultrafilter is a free ultrafilter $\mathcal{U}$ whose dual ideal is not Katětov above the Solecki's ideal $\mathcal{S}$.

THEOREM 4.1. *Let $\mathcal{U}$ be a free ultrafilter. Then the following are equivalent:*

1. *$\mathcal{U}$ is an $\mathcal{S}$-ultrafilter,*
2. *$\mathcal{U}^*$ satisfies the Fubini property and*
3. *for any sequence $\langle A_n : n < \omega \rangle$ of Borel subsets of 2^ω, if $\mathcal{U}\text{-}\lim \lambda(A_n) > 0$ then $\mathcal{U}\text{-}\lim A_n \neq \emptyset$.*

Proof. Theorem 3.1 claims that the ideals I which do not have I-positive sets X such that $\mathsf{I} \restriction X \geq_K \mathcal{S}$, are exactly those ideals satisfying the Fubini property, and since every maximal ideal is Katětov equivalent to all its restrictions to positive sets, we have that dual ideals of $\mathcal{S}$-ultrafilters are exactly the maximal ideals with the Fubini property. Now, Fubini property among maximal ideals (or ultrafilters) means: for any sequence $\langle A_n : n < \omega \rangle$ of Borel subsets of 2^ω and any $\varepsilon > 0$, if $\{n < \omega : \lambda(A_n) > \varepsilon\} \in \mathcal{U}$ then $\lambda^*(\{x \in 2^\omega : \{n < \omega : x \in A_n\} \in \mathcal{U}\}) \geq \varepsilon$. Hence, if $\mathcal{S} \not\leq_K \mathcal{U}^*$ and $\mathcal{U}\text{-}\lim \lambda(A_n) > 0$ then $\lambda^*(\mathcal{U}\text{-}\lim A_n) > 0$ and then $\mathcal{U}\text{-}\lim A_n \neq \emptyset$. On the other hand, let suppose that

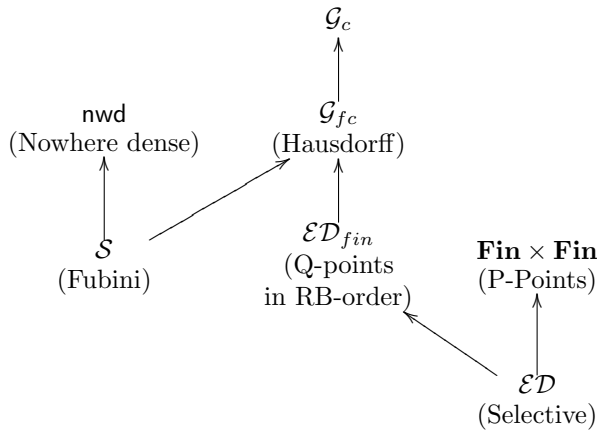
$\mathcal{U}\text{-}\lim \lambda(A_n) > \varepsilon$ and $\lambda^*(\mathcal{U}\lim A_n) = \delta < \varepsilon$, for some sequence $\langle A_n : n < \omega\rangle$ and some $\varepsilon > 0$. For any $k < \omega$ let us choose a Borel set $A'_k \subseteq A_k \setminus \mathcal{U}\lim A_n$, with $\lambda(A'_k) = \varepsilon - \delta$. Then, $\mathcal{U}\text{-}\lim \lambda(A'_n) \geq \varepsilon - \delta$ but $\mathcal{U}\text{-}\lim A'_n = 0$, since for any $x \in 2^\omega$, $\{n : x \in A_n\} \in \mathcal{U}^*$. □

COROLLARY 4.2 (Benedikt). *Every Fubini ulfrafilter is a Hausdorff ultrafilter.*

Proof. Solecki proved in [16] that $\mathcal{G}_{fc} \geq_K \mathcal{S}$ and if $\mathcal{U}$ is Fubini then by 4.1 $\mathcal{U}^* \not\geq \mathcal{S}$. Hence, $\mathcal{U}^* \not\geq_K \mathcal{G}_{fc}$. □

5. Final remarks and questions

The known Katětov relations are displayed in the following diagram:[2]



Of course, the main question about Hausdorff ultrafilters is if ZFC implies their existence. As a consequence of the fact that $\mathcal{S} \leq_K \mathsf{nwd}$ ([11, Theorem 5.10]), every Fubini ultrafilter is a nowhere dense ultrafilter. This fact was proved by Shelah ([15, Proposition 26]). The same does not hold for nowhere dense and Hausdorff ultrafilters since in [11] it was proved that $\mathcal{G}_{fc} \not\leq_K \mathsf{nwd}$, which is a consequence of 2.4 and the following

PROPOSITION 5.1. $\mathcal{ED} \not\leq \mathsf{nwd}$.

---

[2] An ultrafilter $\mathcal{U}$ is:
(1) *nowhere dense* if for each function f from $\mathbb{N}$ to $\mathbb{R}$, there is $U \in \mathcal{U}$ such that $f''U$ is nowhere dense.
(2) Q-*point* if for each partition $\{A_n : n < \omega\}$ of $\mathbb{N}$ such that each A_n is finite, there is $U \in \mathcal{U}$ such that $|U \cap A_n| \leq 1$ for all n.
(3) P-*point* if for each partition $\{A_n : n < \omega\}$ of $\mathbb{N}$, there is $U \in \mathcal{U}$ such that $|U \cap A_n| < \aleph_0$ for all n.

Proof. Let f be an arbitrary function from $\mathbb{Q}$ to $\omega \times \omega$ and let $\{U_n : n < \omega\}$ be a base for the open sets of $\mathbb{Q}$. Assume that $f^{-1}(n \times \omega) \in \mathsf{nwd}$ for all $n < \omega$ (if it is not the case we finished). Choose q_0 arbitrarily and recursively, choose $q_n \in U_n$ so that $proj_1(f(q_n)) > \max\{proj_1(f(q_j)) : j < n\}$. This is possible by our assumption. Then, $\{f(q_n) : n < \omega\} \in \mathcal{ED}$ but $\{q_n : n < \omega\}$ is dense in $\mathbb{Q}$. □

Di Nasso and Forti proved that if $\mathcal{U}$ and $\mathcal{V}$ are two isomorphic ultrafilters then $\mathcal{U} \times \mathcal{V}$ is not Hausdorff. On the other hand, it is easy to prove that if $\mathcal{U}$ is nowhere dense and $\mathcal{V}$ is P-point then $\mathcal{U} \times \mathcal{V}$ is a nowhere dense ultrafilter. Since every P-point is nowhere dense, for any P-point $\mathcal{U}$ we have that $\mathcal{U} \times \mathcal{U}$ is nowhere dense but not Hausdorff. Hence, from the consistency of the existence of a P-point ultrafilter it follows that there is a nowhere dense non Hausdorff ultrafilter. Consequently, a natural question is:

PROBLEM 5.2: Are there consistently Hausdorff ultrafilters that are not nowhere dense?

It is well known that there is no P-point ultrafilter extending the filter nwd^*, however we would like to know if (consistently) there is a Hausdorff ultrafilter extending nwd^*, which is clearly a little stronger than Problem 5.2.

Di Nasso and Forti [6] asked about a set-theoretic hypothesis weaker than those providing selective ultrafilters, which implies the existence of Hausdorff ultrafilters, e.g. an equality or inequality between cardinal invariants of the continuum. We think it would be interesting to understand generic existence of Hausdorff ultrafilters[3]. For some classes of ideals this cardinal conditions are well known, for example, Canjar [3] proved that $\mathsf{cov}(\mathcal{M}) = \mathfrak{c}$ is equivalent to generic existence of selective ultrafilters, and Benedikt [2] proved that $\mathsf{cov}(\mathcal{E}) = \mathfrak{c}$ is equivalent to generic existence of Fubini ultrafilters. The natural question is

PROBLEM 5.3: Is there a suitable cardinal condition which is equivalent to generic existence of Hausdorff ultrafilters?

Finally, we want to ask about the existence of $\mathcal{G}_c$ ultrafilters.

PROBLEM 5.4: Does ZFC prove that there exists a $\mathcal{G}_c$-ultrafilter?

REFERENCES

[1] J. BAUMGARTNER, *Ultrafilters on ω*, J. Symbolic Logic **60** (1995), 624–639.
[2] M. BENEDIKT, *Hierarchies of measure-theoretic ultrafilters*, Ann. Pure Appl. Logic **97** (1999), 203–219.

[3]Let $\mathcal{C}$ be a class of ultrafilters. It is said that (under a suitable assumption) ultrafilters of the class $\mathcal{C}$ exist generically if every filter base with cardinality less than continuum can be extended to a $\mathcal{C}$ ultrafilter.

[3] M. Canjar, *On the generic existence of special ultrafilters*, Proc. Amer. Math. Soc **110** (1990), 233–241.
[4] J. P. R. Christensen, *Some results with relation to the control measure problem*, Vector Space Measures and Applications II (Richard Aron and Sen Dineen, eds.), Lecture Notes in Math., vol. 645, Springer, Berlin, 1978, pp. 27–34.
[5] M. Daguenet-Teissier, *Ultràfiltres a la façon de Ramsey*, Trans. Amer. Math. Soc. **250** (1979), 91–120.
[6] M. di Nasso and M. Forti, *Hausdorff ultrafilters*, Proc. Amer. Math. Soc. **134** (2006), 1809–1818.
[7] J. Fašková, *Description of some ultrafilters via I-ultrafilters*, RIMS Kôkyûroku **1619** (2008), 20–31.
[8] F. Hernández-Hernández and M. Hrušák, *Cardinal invariants of analytic P-ideals*, Canad. J. Math. **59** (2007), 575–595.
[9] M. Hrušák, *Katětov order on Borel ideals*, In preparation.
[10] M. Hrušák, *Combinatorics of ideals and filters on ω*, Set theory and its applications (Providence, RI), Contemp. Math., vol. 533, Amer. Math. Soc., 2011, pp. 3–13.
[11] M. Hrušák and D. Meza-Alcántara, *Pair splitting, pair reaping and cardinal invariants of F_σ-ideals*, J. Symbolic Logic (2010), 667–679.
[12] V. Kanovei and M. Reeken, *On Ulam's problem concerning the stability of approximate homomorphisms*, Proc. Steklov Inst. Math. (1987).
[13] V. Kanovei and M. Reeken, *New Radon-Nikodým ideals*, Mathematika **47** (2000), 219–227.
[14] D. Meza-Alcántara, *Ideals and filters on countable sets*, Ph.D. thesis, Universidad Nacional Autónoma de México, 2009.
[15] S. Shelah, *There may be no nowhere dense ultrafilter*, Logic Colloquium '95 (Haifa) (Berlin), Lecture Notes Logic, vol. 11, Springer, 1998, pp. 305–324.
[16] S. Solecki, *Filters and sequences*, Fund. Math. **163** (2000), 215–228.

Authors' addresses:

Michael Hrušák
Centro de Ciencias Matemáticas
UNAM, Apartado Postal 61-3
Xangari, 58089 Morelia, Michoacán, México.
E-mail: michael@matmor.unam.mx

David Meza-Alcántara
Facultad de Ciencias Físico-Matemáticas
UMSNH, Edificio ALPHA
Ciudad Universitaria, 58060 Morelia, Michoacán, México
E-mail: dmeza@fismat.umich.mx

Received May 31, 2012
Revised November 26, 2012

Contents

Editorial Policy

The journal *Rendiconti dell'Istituto di Matematica dell'Università di Trieste* publishes original articles in all areas of mathematics. Special regard is given to research papers, but attractive expository papers may also be considered for publication.

The journal usually appears in one issue per year. Additional issues may however be published. In particular, the Managing Editors may consider the publication of supplementary volumes related to some special events, like conferences, workshops, and advanced schools.

All submitted papers will be refereed. Manuscripts are accepted for review with the understanding that the work has not been published before and is not under consideration for publication elsewhere.

The journal can be obtained by exchange agreements with other similar journals.

Instructions for Authors

Authors are invited to submit their papers by e-mail directly to one of the Managing Editors in PDF format.

All the correspondence regarding the submission and the editorial process of the paper are done by e-mail.

Papers have to be written in one of the following languages: English, French, German, or Italian. Abstracts should not exceed ten printed lines, except for papers written in French, German, or Italian, for which an extended English summary is required.

After acceptance, manuscripts have to be prepared in LaTeX using the style *rendiconti.cls* which can be downloaded from the web page.

Any figure should be recorded in a single PDF, PS (PostScript), or EPS (Encapsulated PostScript) file.

INDIRIZZO DEL COMITATO DI REDAZIONE

Rendiconti dell'Istituto di Matematica dell'Università di Trieste
Dipartimento di Matematica e Geoscienze - Sezione di Matematica
Università degli Studi di Trieste - Via Valerio 12/1 - 34127 Trieste (Italy)

Tel. +39-040-5582635 — Fax +39-040-5582636 — Pagina web: http://rendiconti.dmi.units.it

Direttore Responsabile: DANIELE DEL SANTO

Periodico registrato il 25 Settembre 1968 al n. 358 del registro dei periodici del Tribunale di Trieste

Annual subscription rate: €50,00 (one issue, inclusive of any supplements).

Subscription inquiries should be sent to

EUT, Edizioni Università di Trieste, Piazzale Europa 1, 34127 Trieste, Italia.

ISSN 0049-4704