History

The journal *Rendiconti dell'Istituto di Matematica dell'Università di Trieste* was founded in 1969 by Arno Predonzan, with the aim of publishing original research articles in all fields of mathematics.

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 published articles are available on-line.

In 2008 the *Dipartimento di Matematica e Informatica*, the owner of the journal, decided to renew it. The name of the journal however remained unchanged, but the subtitle *An International Journal of Mathematics* was added. The journal can be obtained by subscription, or by reciprocity with other similar journals. Currently more than 100 exchange agreements with mathematics departments and institutes around the world have been entered in.

The articles published by *Rendiconti dell'Istituto di Matematica dell'Università di Trieste* are reviewed/indexed by MathSciNet, Zentralblatt Math, Scopus, OpenStarTs.

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

Rendiconti dell'Istituto di Matematica dell'Università di Trieste
An International Journal of Mathematics

Volume 50 (2018)
Dipartimento
di Matematica
e Geoscienze

ISSN 0049-4704 (Print)
ISSN 2464-8728 (Online)

EUT EDIZIONI UNIVERSITÀ DI TRIESTE

ISSN 0049-4704 (Print)
ISSN 2464-8728 (Online)

EUT Edizioni Università di Trieste
via E. Weiss, 21 – 34128 Trieste
http://eut.units.it
eut@units.it
https://www.facebook.com/EUTEdizioniUniversitaTrieste

Rend. Istit. Mat. Univ. Trieste
Volume 50 (2018), 1–36
DOI: 10.13137/2464-8728/20577

Theory of the (m,σ)-general functions over infinite-dimensional Banach spaces

Claudio Asci

Abstract. *In this paper, we introduce some functions, called (m,σ)-general, that generalize the (m,σ)-standard functions and are defined in the infinite-dimensional Banach space E_I of the bounded real sequences $\{x_n\}_{n\in I}$, for some subset I of $\mathbf{N}^*$. Moreover, we recall the main results about the differentiation theory over E_I, and we expose some properties of the (m,σ)-general functions. Finally, we study the linear (m,σ)-general functions, by introducing a theory that generalizes the standard theory of the $m\times m$ matrices.*

Keywords: Infinite-dimensional Banach spaces, infinite-dimensional differentiation theory, (m,σ)-general functions.
MS Classification 2010: 28A15, 46B99.

1. Introduction

In this paper, we generalize the results of the articles [3] and [4], where, for any subset I of $\mathbf{N}^*$, we define the Banach space $E_I\subset\mathbf{R}^I$ of the bounded real sequences $\{x_n\}_{n\in I}$, the σ-algebra $\mathcal{B}_I$ given by the restriction to E_I of $\mathcal{B}^{(I)}$ (defined as the product indexed by I of the same Borel σ-algebra $\mathcal{B}$ on $\mathbf{R}$), and a class of functions over an open subset of E_I, with values on E_I, called (m,σ)-standard. The properties of these functions generalize the analogous ones of the standard finite-dimensional diffeomorphisms; moreover, these functions are introduced in order to provide a change of variables' formula for the integration of the measurable real functions on $\left(\mathbf{R}^I,\mathcal{B}^{(I)}\right)$. For any strictly positive integer k, this integration is obtained by using an infinite-dimensional measure $\lambda_{N,a,v}^{(k,I)}$, over the measurable space $\left(\mathbf{R}^I,\mathcal{B}^{(I)}\right)$, that in the case $I=\{1,...,k\}$ coincides with the k-dimensional Lebesgue measure on $\mathbf{R}^k$.

In the mathematical literature, some articles introduced infinite-dimensional measures analogue of the Lebesgue one (see for example the paper of Léandre [8], in the context of the noncommutative geometry, that one of Tsilevich et al. [10], which studies a family of σ-finite measures on $\mathbf{R}^+$, and that one of Baker [5], which defines a measure on $\mathbf{R}^{\mathbf{N}^*}$ that is not σ-finite).

In the paper [3], we define the linear (m,σ)-standard functions. The motivation of this paper follows from the natural extension to the infinite-dimensional case of the results of the article [2], where we estimate the rate of convergence of some Markov chains in $[0,p)^k$ to a uniform random vector. In order to consider the analogue random elements in $[0,p)^{\mathbf{N}^*}$, it is necessary to overcome some difficulties: for example, the lack of a change of variables formula for the integration in the subsets of $\mathbf{R}^{\mathbf{N}^*}$. A related problem is studied in the paper of Accardi et al [1], where the authors describe the transformations of generalized measures on locally convex spaces under smooth transformations of these spaces. In the paper [4], we expose a differentiation theory for the functions over an open subset of E_I, and in particular we define the functions C^1 and the diffeomorphisms; moreover, we remove the assumption of linearity for the (m,σ)-standard functions, and we present a change of variables' formula for the integration of the measurable real functions on $\left(\mathbf{R}^I,\mathcal{B}^{(I)}\right)$; this change of variables is defined by the (m,σ)-standard diffeomorphisms, with further properties. This result agrees with the analogous finite-dimensional result.

In this paper, we introduce a class of functions, called (m,σ)-general, that generalizes the set of the (m,σ)-standard functions given in [4]. In Section 2, we recall the main results about the differentiation theory over the infinite-dimensional Banach space E_I. Moreover, we expose some properties of the (m,σ)-general functions. In Section 3, we study the linear (m,σ)-general functions and we expose a theory that generalizes the standard theory of the $m\times m$ matrices and the results about the linear (m,σ)-standard functions, given in [3]. The main result is the definition of the determinant of a linear (m,σ)-general function, as the limit of a sequence of the determinants of some standard matrices (Theorem 3.6 and Definition 3.7). Moreover, we study some properties of this determinant, and we provide an example (Example 3.19). In Section 4, we expose some ideas for further study in the probability theory.

2. Theory of the (m,σ)-general functions

Let $I\neq\emptyset$ be a set and let $k\in\mathbf{N}^*$; indicate by τ, by $\tau^{(k)}$, by $\tau^{(I)}$, by $\mathcal{B}$, by $\mathcal{B}^{(k)}$, by $\mathcal{B}^{(I)}$, and by *Leb*, respectively, the euclidean topology on $\mathbf{R}$, the euclidean topology on $\mathbf{R}^k$, the topology $\bigotimes_{i\in I}\tau$, the Borel σ-algebra on $\mathbf{R}$, the Borel σ-algebra on $\mathbf{R}^k$, the σ-algebra $\bigotimes_{i\in I}\mathcal{B}$, and the Lebesgue measure on $\mathbf{R}$. Moreover, for any set $A\subset\mathbf{R}$, indicate by $\mathcal{B}(A)$ the σ-algebra induced by $\mathcal{B}$ on A, and by $\tau(A)$ the topology induced by τ on A; analogously, for any set $A\subset\mathbf{R}^I$, define the σ-algebra $\mathcal{B}^{(I)}(A)$ and the topology $\tau^{(I)}(A)$. Finally, if $S=\prod_{i\in I}S_i$ is a Cartesian product, for any $(x_i:i\in I)\in S$ and for any $\emptyset\neq H\subset I$, define

$x_H = (x_i : i \in H) \in \prod_{i\in H} S_i$, and define the projection $\pi_{I,H}$ on $\prod_{i\in H} S_i$ as the function $\pi_{I,H} : S \longrightarrow \prod_{i\in H} S_i$ given by $\pi_{I,H}(x_I) = x_H$.

Henceforth, we will suppose that I, J are sets such that $\emptyset \neq I, J \subset \mathbf{N}^*$; moreover, for any $k \in \mathbf{N}^*$, we will indicate by I_k the set of the first k elements of I (with the natural order and with the convention $I_k = I$ if $|I| < k$); furthermore, for any $i \in I$, set $|i| = |I \cap (0, i]|$. Analogously, define J_k and $|j|$, for any $k \in \mathbf{N}^*$ and for any $j \in J$.

Definition 2.1. *For any set $I \neq \emptyset$, define the function $\|\cdot\|_I : \mathbf{R}^I \longrightarrow [0, +\infty]$ by*

$$\|x\|_I = \sup_{i\in I}|x_i|,\ \forall\, x = (x_i : i \in I) \in \mathbf{R}^I,$$

and define the vector space

$$E_I = \{x \in \mathbf{R}^I : \|x\|_I < +\infty\}.$$

Moreover, indicate by $\mathcal{B}_I$ the σ-algebra $\mathcal{B}^{(I)}(E_I)$, by τ_I the topology $\tau^{(I)}(E_I)$, and by $\tau_{\|\cdot\|_I}$ the topology induced on E_I by the the distance $d : E_I \times E_I \longrightarrow [0, +\infty)$ defined by $d(x, y) = \|x - y\|_I$, $\forall\, x, y \in E_I$; furthermore, for any set $A \subset E_I$, indicate by $\tau_{\|\cdot\|_I}(A)$ the topology induced by $\tau_{\|\cdot\|_I}$ on A. Finally, for any $x_0 \in E_I$ and for any $\delta > 0$, indicate by $B(x_0, \delta)$ the set $\{x \in E_I : \|x - x_0\|_I < \delta\}$.

Remark 2.2: For any $A \subset E_I$, one has $\tau^{(I)}(A) \subset \tau_{\|\cdot\|_I}(A)$; moreover, E_I is a Banach space, with the norm $\|\cdot\|_I$.

Proof. The proof that $\tau^{(I)}(A) \subset \tau_{\|\cdot\|_I}(A)$, $\forall\, A \subset E_I$, follows from the definitions of $\tau^{(I)}$ and $\tau_{\|\cdot\|_I}$; moreover, the proof that E_I is a Banach space can be found, for example, in [3] (Remark 2). □

The following concept generalizes the definition 6 in [3] (see also the theory in the Lang's book [7] and that in the Weidmann's book [11]).

Definition 2.3. *Let $A = (a_{ij})_{i\in I, j\in J}$ be a real matrix $I \times J$ (eventually infinite); then, define the linear function $A = (a_{ij})_{i\in I, j\in J} : E_J \longrightarrow \mathbf{R}^I$, and write $x \longrightarrow Ax$, in the following manner:*

$$(Ax)_i = \sum_{j\in J} a_{ij} x_j,\ \forall\, x \in E_J,\ \forall\, i \in I, \tag{1}$$

on condition that, for any $i \in I$, the sum in (1) converges to a real number. In *particular, if* $|I| = |J|$, *indicate by* $\mathbf{I}_{I,J} = \left(\overline{\delta}_{ij}\right)_{i\in I, j\in J}$ *the real matrix defined by*

$$\overline{\delta}_{ij} = \begin{cases} 1 & \text{if } |i| = |j| \\ 0 & \text{otherwise} \end{cases},$$

and call $\overline{\delta}_{ij}$ *generalized Kronecker symbol. Moreover, indicate by* $A^{(L,N)}$ *the real matrix* $(a_{ij})_{i\in L, j\in N}$, *for any* $L \subset I$, *for any* $N \subset J$, *and indicate by* ${}^{t}A = (b_{ji})_{j\in J, i\in I} : E_I \longrightarrow \mathbf{R}^J$ *the linear function defined by* $b_{ji} = a_{ij}$, *for any* $j \in J$ *and for any* $i \in I$. *Furthermore, if* $I = J$ *and* $A = {}^{t}A$, *we say that* A *is a symmetric function. Finally, if* $B = (b_{jk})_{j\in J, k\in K}$ *is a real matrix* $J \times K$, *define the* $I \times K$ *real matrix* $AB = ((AB)_{ik})_{i\in I, k\in K}$ *by*

$$(AB)_{ik} = \sum_{j\in J} a_{ij} b_{jk}, \tag{2}$$

on condition that, for any $i \in I$ *and for any* $k \in K$, *the sum in (2) converges to a real number.*

Proposition 2.4. *Let* $A = (a_{ij})_{i\in I, j\in J}$ *be a real matrix* $I \times J$; *then:*

1. *The linear function* $A = (a_{ij})_{i\in I, j\in J} : E_J \longrightarrow \mathbf{R}^I$ *given by (1) is defined if and only if, for any* $i \in I$, $\sum_{j\in J} |a_{ij}| < +\infty$.
2. *One has* $A(E_J) \subset E_I$ *if and only if* A *is continuous and if and only if* $\sup_{i\in I} \sum_{j\in J} |a_{ij}| < +\infty$; *moreover,* $\|A\| = \sup_{i\in I} \sum_{j\in J} |a_{ij}|$.
3. *If* $B = (b_{jk})_{j\in J, k\in K} : E_K \longrightarrow E_J$ *is a linear function, then the linear function* $A \circ B : E_K \longrightarrow \mathbf{R}^I$ *is defined by the real matrix* AB.

Proof. The proofs of points 1 and 2 are analogous to the proof of Proposition 7 in [3]. Moreover, the proof of point 3 is analogous to that one true in the particular case $|I|, |J|, |K| < +\infty$ (see, e.g., the Lang's book [7]). $\square$

The following definitions and results (from Definition 2.5 to Proposition 2.19) can be found in [4] and generalize the differentiation theory in the finite case (see, e.g., the Lang's book [6]).

Definition 2.5. *Let* $U \in \tau_{\|\cdot\|_J}$; *a function* $\varphi : U \subset E_J \longrightarrow E_I$ *is called differentiable in* $x_0 \in U$ *if there exists a linear and continuous function* $A : E_J \longrightarrow E_I$ *defined by a real matrix* $A = (a_{ij})_{i\in I, j\in J}$, *and one has*

$$\lim_{h\to 0} \frac{\|\varphi(x_0 + h) - \varphi(x_0) - Ah\|_I}{\|h\|_J} = 0. \tag{3}$$

If φ is differentiable in x_0 for any $x_0 \in U$, φ is called differentiable in U. The function A is called differential of the function φ in x_0, and it is indicated by the symbol $d\varphi(x_0)$.

Remark 2.6: Let $U \in \tau_{\|\cdot\|_J}$ and let $\varphi, \psi : U \subset E_J \longrightarrow E_I$ be differentiable functions in $x_0 \in U$; then, for any $\alpha, \beta \in \mathbf{R}$, the function $\alpha\varphi + \beta\psi$ is differentiable in x_0, and $d(\alpha\varphi + \beta\psi)(x_0) = \alpha d\varphi(x_0) + \beta d\psi(x_0)$.

Remark 2.7: A linear and continuous function $A = (a_{ij})_{i\in I, j\in J} : E_J \longrightarrow E_I$, defined by

$$(Ax)_i = \sum_{j\in J} a_{ij}x_j, \; \forall\, x \in E_J, \; \forall\, i \in I,$$

is differentiable and $d\varphi(x_0) = A$, for any $x_0 \in E_J$.

Remark 2.8: Let $U \in \tau_{\|\cdot\|_J}$ and let $\varphi : U \subset E_J \longrightarrow E_I$ be a function differentiable in $x_0 \in U$; then, for any $i \in I$, the component $\varphi_i : U \longrightarrow \mathbf{R}$ is differentiable in x_0, and $d\varphi_i(x_0)$ is the matrix A_i given by the i-th row of $A = d\varphi(x_0)$. Moreover, if $|I| < +\infty$ and $\varphi_i : U \subset E_J \longrightarrow \mathbf{R}$ is differentiable in x_0, for any $i \in I$, then $\varphi : U \subset E_J \longrightarrow E_I$ is differentiable in x_0.

Remark 2.9: Let $U \in \tau_{\|\cdot\|_J}$ and let $\varphi : U \subset E_J \longrightarrow E_I$ be a function differentiable in $x_0 \in U$; then, φ is continuous in x_0.

Definition 2.10. *Let $U \in \tau_{\|\cdot\|_J}$, let $v \in E_J$ such that $\|v\|_J = 1$ and let a function $\varphi : U \subset E_J \longrightarrow \mathbf{R}^I$; for any $i \in I$, the function φ_i is called differentiable in $x_0 \in U$ in the direction v if there exists the limit*

$$\lim_{t\to 0} \frac{\varphi_i(x_0 + tv) - \varphi_i(x_0)}{t}.$$

This limit is indicated by $\frac{\partial \varphi_i}{\partial v}(x_0)$, and it is called derivative of φ_i in x_0 in the direction v. If, for some $j \in J$, one has $v = e_j$, where $(e_j)_k = \delta_{jk}$, for any $k \in J$, indicate $\frac{\partial \varphi_i}{\partial v}(x_0)$ by $\frac{\partial \varphi_i}{\partial x_j}(x_0)$, and call it partial derivative of φ_i in x_0, with respect to x_j. Moreover, if there exists the linear function defined by the matrix $J_\varphi(x_0) = \Big((J_\varphi(x_0))_{ij} \Big)_{i\in I, j\in J} : E_J \longrightarrow \mathbf{R}^I$, where $(J_\varphi(x_0))_{ij} = \frac{\partial \varphi_i}{\partial x_j}(x_0)$, for any $i \in I$, $j \in J$, then $J_\varphi(x_0)$ is called Jacobian matrix of the function φ in x_0.

Remark 2.11: Let $U \in \tau_{\|\cdot\|_J}$ and suppose that a function $\varphi : U \subset E_J \longrightarrow E_I$ is differentiable in $x_0 \in U$; then, for any $v \in E_J$ such that $\|v\|_J = 1$ and for any $i \in I$, the function $\varphi_i : U \subset E_J \longrightarrow \mathbf{R}$ is differentiable in x_0 in the direction v, and one has

$$\frac{\partial \varphi_i}{\partial v}(x_0) = d\varphi_i(x_0)v.$$

Corollary 2.12. *Let* $U \in \tau_{\|\cdot\|_J}$ *and let* $\varphi : U \subset E_J \longrightarrow E_I$ *be a function differentiable in* $x_0 \in U$*; then, there exists the function* $J_\varphi(x_0) : E_J \longrightarrow \mathbf{R}^I$*, and it is continuous; moreover, for any* $h \in E_J$*, one has* $d\varphi(x_0)(h) = J_\varphi(x_0)h$*.*

Theorem 2.13. *Let* $U \in \tau_{\|\cdot\|_J}$*, let* $\varphi : U \subset E_J \longrightarrow E_I$ *be a function differentiable in* $x_0 \in U$*, let* $V \in \tau_{\|\cdot\|_I}$ *such that* $V \supset \varphi(U)$*, and let* $\psi : V \subset E_I \longrightarrow E_H$ *a function differentiable in* $y_0 = \varphi(x_0)$*. Then, the function* $\psi \circ \varphi$ *is differentiable in* x_0*, and one has* $d(\psi \circ \varphi)(x_0) = d\psi(y_0) \circ d\varphi(x_0)$*.*

Definition 2.14. *Let* $U \in \tau_{\|\cdot\|_J}$*, let* $i, j \in J$ *and let* $\varphi : U \subset E_J \longrightarrow \mathbf{R}$ *be a function differentiable in* $x_0 \in U$ *with respect to* x_i*, such that the function* $\frac{\partial \varphi}{\partial x_i}$ *is differentiable in* x_0 *with respect to* x_j*. Indicate* $\frac{\partial}{\partial x_j}\left(\frac{\partial \varphi}{\partial x_i}\right)(x_0)$ *by* $\frac{\partial^2 \varphi}{\partial x_j \partial x_i}(x_0)$ *and call it second partial derivative of* φ *in* x_0 *with respect to* x_i *and* x_j*. If* $i = j$*, it is indicated by* $\frac{\partial^2 \varphi}{\partial x_i^2}(x_0)$*. Analogously, for any* $k \in \mathbf{N}^*$ *and for any* $j_1, ..., j_k \in J$*, define* $\frac{\partial^k \varphi}{\partial x_{j_k} ... \partial x_{j_1}}(x_0)$ *and call it* k*-th partial derivative of* φ *in* x_0 *with respect to* $x_{j_1}, ... x_{j_k}$*.*

Definition 2.15. *Let* $U \in \tau_{\|\cdot\|_J}$ *and let* $k \in \mathbf{N}^*$*; a function* $\varphi : U \subset E_J \longrightarrow E_I$ *is called* C^k *in* $x_0 \in U$ *if, in a neighbourhood* $V \in \tau_{\|\cdot\|_J}(U)$ *of* x_0*, for any* $i \in I$ *and for any* $j_1, ..., j_k \in J$*, there exists the function defined by* $x \longrightarrow \frac{\partial^k \varphi_i}{\partial x_{j_k} ... \partial x_{j_1}}(x)$*, and this function is continuous in* x_0*;* φ *is called* C^k *in* U *if, for any* $x_0 \in U$*,* φ *is* C^k *in* x_0*. Moreover,* φ *is called strongly* C^1 *in* $x_0 \in U$ *if, in a neighbourhood* $V \in \tau_{\|\cdot\|_J}(U)$ *of* x_0*, there exists the function defined by* $x \longrightarrow J_\varphi(x)$*, this function is continuous in* x_0*, and one has* $\|J_\varphi(x_0)\| < +\infty$*. Finally,* φ *is called strongly* C^1 *in* U *if, for any* $x_0 \in U$*,* φ *is strongly* C^1 *in* x_0*.*

Definition 2.16. *Let* $U \in \tau_{\|\cdot\|_J}$ *and let* $V \in \tau_{\|\cdot\|_I}$*; a function* $\varphi : U \subset E_J \longrightarrow V \subset E_I$ *is called diffeomorphism if* φ *is bijective and* C^1 *in* U*, and the function* $\varphi^{-1} : V \subset E_I \longrightarrow U \subset E_J$ *is* C^1 *in* V*.*

Remark 2.17: Let $U \in \tau_{\|\cdot\|_J}$ and let $\varphi : U \subset E_J \longrightarrow E_I$ be a function C^1 in $x_0 \in U$, where $|I| < +\infty$, $|J| < +\infty$, then φ is strongly C^1 in x_0.

Theorem 2.18. *Let* $U \in \tau_{\|\cdot\|_J}$*, let* $\varphi : U \subset E_J \longrightarrow \mathbf{R}$ *be a function* C^k *in* $x_0 \in U$*, let* $i_1, ..., i_k \in J$*, and let* $j_1, ..., j_k \in J$ *be a permutation of* $i_1, ..., i_k$*. Then, one has*

$$\frac{\partial^k \varphi}{\partial x_{i_1} ... \partial x_{i_k}}(x_0) = \frac{\partial^k \varphi}{\partial x_{j_1} ... \partial x_{j_k}}(x_0).$$

Proposition 2.19. *Let* $U = \left(\prod_{j\in J} A_j\right) \cap E_J \in \tau_{\|\cdot\|_J}$, *where* $A_j \in \tau$, *for any* $j \in J$, *and let* $\varphi : U \subset E_J \longrightarrow E_I$ *be a function* C^1 *in* $x_0 \in U$, *such that*

$$\varphi_i(x) = \sum_{j\in J} \varphi_{ij}(x_j),\ \forall\, x = (x_j : j \in J) \in U,\ \forall\, i \in I, \tag{4}$$

where $\varphi_{ij} : A_j \longrightarrow \mathbf{R}$, *for any* $i \in I$ *and for any* $j \in J$; *moreover, suppose that, in a neighbourhood* $V \in \tau_{\|\cdot\|_J}(U)$ *of* x_0, *there exists the function defined by* $x \longrightarrow J_\varphi(x)$ *and one has* $\sup_{x\in V} \|J_\varphi(x)\| < +\infty$. *Then,* φ *is continuous in* x_0; *in particular, if* φ *is strongly* C^1 *in* x_0 *and* $|I| < +\infty$, φ *is differentiable in* x_0.

Definition 2.20. *Let* $m \in \mathbf{N}^*$ *and let* $U = \left(U^{(m)} \times \prod_{j\in J\setminus J_m} A_j\right) \cap E_J \in \tau_{\|\cdot\|_J}$, *where* $U^{(m)} \in \tau^{(m)}$, $A_j \in \tau$, *for any* $j \in J\setminus J_m$. *A function* $\varphi : U \subset E_J \longrightarrow E_I$ *is called m-general if, for any* $i \in I$ *and for any* $j \in J\setminus J_m$, *there exist some functions* $\varphi_i^{(I,m)} : U^{(m)} \longrightarrow \mathbf{R}$ *and* $\varphi_{ij} : A_j \longrightarrow \mathbf{R}$ *such that*

$$\varphi_i(x) = \varphi_i^{(I,m)}(x_{J_m}) + \sum_{j\in J\setminus J_m} \varphi_{ij}(x_j),\ \forall\, x \in U.$$

Moreover, for any $\emptyset \neq L \subset I$ *and for any* $J_m \subset N \subset J$, *indicate by* $\varphi^{(L,N)}$ *the function* $\varphi^{(L,N)} : \pi_{J,N}(U) \longrightarrow \mathbf{R}^L$ *defined by*

$$\varphi_i^{(L,N)}(x_N) = \varphi_i^{(I,m)}(x_{J_m}) + \sum_{j\in N\setminus J_m} \varphi_{ij}(x_j),\ \forall\, x_N \in \pi_{J,N}(U),\ \forall\, i \in L. \tag{5}$$

Furthermore, for any $\emptyset \neq L \subset I$ *and for any* $\emptyset \neq N \subset J\setminus J_m$, *indicate by* $\varphi^{(L,N)}$ *the function* $\varphi^{(L,N)} : \pi_{J,N}(U) \longrightarrow \mathbf{R}^L$ *given by*

$$\varphi_i^{(L,N)}(x_N) = \sum_{j\in N} \varphi_{ij}(x_j),\ \forall\, x_N \in \pi_{J,N}(U),\ \forall\, i \in L. \tag{6}$$

In particular, suppose that $m = 1$; *then, let* $j \in J$ *such that* $\{j\} = J_1$ *and indicate* $U^{(1)}$ *by* A_j *and* $\varphi_i^{(I,1)}$ *by* φ_{ij}, *for any* $i \in I$; *moreover, for any* $\emptyset \neq L \subset I$ *and for any* $\emptyset \neq N \subset J$, *indicate by* $\varphi^{(L,N)}$ *the function* $\varphi^{(L,N)} : \pi_{J,N}(U) \longrightarrow \mathbf{R}^L$ *defined by formula (6).*

Furthermore, for any $l, n \in \mathbf{N}^*$, *indicate* $\varphi^{(I_l,N)}$ *by* $\varphi^{(l,N)}$, $\varphi^{(L,J_n)}$ *by* $\varphi^{(L,n)}$, *and* $\varphi^{(I_l,J_n)}$ *by* $\varphi^{(l,n)}$.

Definition 2.21. *Let* $m \in \mathbf{N}^*$, *let* $U = \left(U^{(m)} \times \prod_{j \in J \backslash J_m} A_j\right) \cap E_J \in \tau_{\|\cdot\|_J}$, *where* $U^{(m)} \in \tau^{(m)}$, $A_j \in \tau$, *for any* $j \in J \backslash J_m$, *and let* $\sigma : I \backslash I_m \longrightarrow J \backslash J_m$ *be an increasing function; a function* $\varphi : U \subset E_J \longrightarrow E_I$ m*-general and such that* $|J| = |I|$ *is called* (m, σ)*-general if:*

1. $\forall\, i \in I \backslash I_m$, $\forall\, j \in J \backslash (J_m \cup \{\sigma(i)\})$, $\forall\, t \in A_j$, *one has* $\varphi_{ij}(t) = 0$*; moreover*
$$\varphi^{(I \backslash I_m, J \backslash J_m)}\left(\pi_{J, J \backslash J_m}(U)\right) \subset E_{I \backslash I_m}.$$
2. $\forall\, i \in I \backslash I_m$, $\forall\, x \in U$, *there exists* $J_{\varphi_i}(x) : E_J \longrightarrow \mathbf{R}$*; moreover,* $\forall\, x_{J_m} \in U^{(m)}$, *one has* $\sum_{i \in I \backslash I_m} \left\| J_{\varphi_i^{(I,m)}}(x_{J_m}) \right\| < +\infty$.
3. $\forall\, i \in I \backslash I_m$, *the function* $\varphi_{i, \sigma(i)} : A_{\sigma(i)} \longrightarrow \mathbf{R}$ *is constant or injective; moreover,* $\forall\, x_{\sigma(I \backslash I_m)} \in \prod_{j \in \sigma(I \backslash I_m)} A_j$, *one has* $\sup_{i \in I \backslash I_m} \left| \varphi'_{i, \sigma(i)}\left(x_{\sigma(i)}\right) \right| < +\infty$ *and* $\inf_{i \in \mathcal{I}_\varphi} \left| \varphi'_{i, \sigma(i)}\left(x_{\sigma(i)}\right) \right| > 0$, *where* $\mathcal{I}_\varphi = \{i \in I \backslash I_m : \varphi_{i, \sigma(i)}$ *is injective*$\}$.
4. *If, for some* $h \in \mathbf{N}$, $h \geq m$, *one has* $|\sigma(i)| = |i|$, $\forall\, i \in I \backslash I_h$, *then,* $\forall\, x_{\sigma(I \backslash I_m)} \in \prod_{j \in \sigma(I \backslash I_m)} A_j$, *there exists* $\prod_{i \in \mathcal{I}_\varphi} \varphi'_{i, \sigma(i)}\left(x_{\sigma(i)}\right) \in \mathbf{R}^*$.

Moreover, set
$$\mathcal{A} = \mathcal{A}(\varphi) = \{h \in \mathbf{N},\, h \geq m : |\sigma(i)| = |i|\,,\ \forall\, i \in I \backslash I_h\}\,.$$

If the sequence $\left\{ J_{\varphi_i^{(I,m)}}(x_{J_m}) \right\}_{i \in I \backslash I_m}$ *converges uniformly on* $U^{(m)}$ *to the matrix* $(0 \ldots 0)$ *and there exists* $a \in \mathbf{R}$ *such that, for any* $\varepsilon > 0$, *there exists* $i_0 \in \mathbf{N}$, $i_0 \geq m$, *such that, for any* $i \in \mathcal{I}_\varphi \cap (I \backslash I_{i_0})$ *and for any* $t \in A_{\sigma(i)}$, *one has* $\left| \varphi'_{i, \sigma(i)}(t) - a \right| < \varepsilon$, *then* φ *is called strongly* (m, σ)*-general.*

Furthermore, for any $I_m \subset L \subset I$ *and for any* $J_m \subset N \subset J$, *define the function* $\overline{\varphi}^{(L,N)} : U \subset E_J \longrightarrow \mathbf{R}^I$ *in the following manner:*
$$\overline{\varphi}_i^{(L,N)}(x) = \begin{cases} \varphi_i^{(L,N)}(x_N) & \forall\, i \in I_m,\, \forall\, x \in U \\ \varphi_i(x) & \forall\, i \in L \backslash I_m,\, \forall\, x \in U \\ \varphi_{i, \sigma(i)}(x_{\sigma(i)}) & \forall\, i \in I \backslash L,\, \forall\, x \in U \end{cases}.$$

Finally, for any $l, n \in \mathbf{N}$, $l, n \geq m$, *indicate* $\overline{\varphi}^{(I_l, N)}$ *by* $\overline{\varphi}^{(l,N)}$, $\overline{\varphi}^{(L, J_n)}$ *by* $\overline{\varphi}^{(L,n)}$, $\overline{\varphi}^{(I_l, J_n)}$ *by* $\overline{\varphi}^{(l,n)}$, *and* $\overline{\varphi}^{(m,m)}$ *by* $\overline{\varphi}$.

Definition 2.22. *A function $\varphi : U \subset E_J \longrightarrow E_I$ (m,σ)-general is called (m,σ)-standard (or (m,σ) of the first type) if, for any $i \in I\backslash I_m$ and for any $x_{J_m} \in U^{(m)}$, one has $\varphi_i^{(I,m)}(x_{J_m}) = 0$. Moreover, a function $\varphi : U \subset E_J \longrightarrow E_I$ (m,σ)-standard and strongly (m,σ)-general is called strongly (m,σ)-standard (see also Definition 28 in [4]).*

Remark 2.23: Let $\varphi : U \subset E_J \longrightarrow E_I$ be a m-general function; then:

1. Let $\emptyset \neq L \subset I$ and let $J_m \subset N \subset J$ such that $\varphi^{(L,N)}(\pi_{J,N}(U)) \subset E_L$; then, for any $n \in \mathbf{N}$, $n \geq m$, the function $\varphi^{(L,N)} : \pi_{J,N}(U) \longrightarrow E_L$ is n-general.

2. Let $\emptyset \neq L \subset I$ and let $\emptyset \neq N \subset J\backslash J_m$ such that $\varphi^{(L,N)}(\pi_{J,N}(U)) \subset E_L$; then, for any $n \in \mathbf{N}^*$, the function $\varphi^{(L,N)}(\pi_{J,N}(U)) \longrightarrow E_L$ is n-general.

3. If $m = 1$, let $\emptyset \neq L \subset I$ and let $\emptyset \neq N \subset J$ such that $\varphi^{(L,N)}(\pi_{J,N}(U)) \subset E_L$; then, for any $n \in \mathbf{N}^*$, the function $\varphi^{(L,N)} : \pi_{J,N}(U) \longrightarrow E_L$ is n-general.

Proof. The proof follows from the definition of $\varphi^{(L,N)}$. □

Proposition 2.24. *Let $\varphi : U \subset E_J \longrightarrow E_I$ be a (m,σ)-general function; then:*

1. *σ is bijective if and only if $|\sigma(i)| = |i|$, $\forall\, i \in I\backslash I_m$.*

2. *$\prod_{j\in J\backslash J_m} A_j \subset E_{J\backslash J_m}$ if and only if there exist $a \in \mathbf{R}^+$ and $m_0 \in \mathbf{N}$, $m_0 \geq m$, such that, for any $j \in J\backslash J_{m_0}$, one has $A_j \subset (-a,a)$.*

3. *Let $I_m \subset L \subset I$ and let $J_m \subset N \subset J$; then, one has $\varphi^{(L,N)}(\pi_{J,N}(U)) \subset E_L$ and $\overline{\varphi}^{(L,N)}(U) \subset E_I$; moreover, the function $\overline{\varphi}^{(L,N)} : U \subset E_J \longrightarrow E_I$ is (m,σ)-general.*

4. *For any $x \in U$, there exists the function $J_{\varphi^{(I\backslash I_m,J)}}(x) : E_J \longrightarrow E_{I\backslash I_m}$, and it is continuous.*

5. *If, for any $j \in J\backslash J_m$ and for any $t \in A_j$, one has $\sum_{i\in I\backslash I_m} \left|\varphi'_{i,j}(t)\right| < +\infty$, then, for any $n \in \mathbf{N}$, $n \geq m$, φ is (n,ξ)-general, where the increasing function $\xi : I\backslash I_n \longrightarrow J\backslash J_n$ is defined by:*

$$\xi(i) = \begin{cases} \sigma(i) & \text{if } \sigma(i) \in J\backslash J_n \\ \min(J\backslash J_n) & \text{if } \sigma(i) \notin J\backslash J_n \end{cases}, \forall\, i \in I\backslash I_n. \tag{7}$$

6. *Suppose that σ is injective; moreover, for any $I_m \subset L \subset I$ such that $|L| < +\infty$ and for any $J_m \subset N \subset J$, let $\widehat{m} = |\max L| \in \mathbf{N}\backslash\{0, ...m-1\}$; then, for any $n \in \mathbf{N}$, $n \geq \widehat{m}$, the function $\overline{\varphi}^{(L,N)}$ is $\left(n, \sigma|_{I\backslash I_n}\right)$-standard.*

Proof.

1. The proof follows from the fact that σ is increasing.
2. The proof follows from the definition of $E_{J\backslash J_m}$.
3. $\forall\, x \in \pi_{J,N}(U)$, let $y \in U$ such that $y_N = x$; then, $\forall\, i \in L\backslash I_m$, we have $\varphi_i^{(L,m)}(x_{J_m}) = \varphi_i(y) - \varphi_{i,\sigma(i)}\left(y_{\sigma(i)}\right)$, and so
$$\sup_{i\in L\backslash I_m}\left|\varphi_i^{(L,m)}(x_{J_m})\right| \leq \sup_{i\in L\backslash I_m}|\varphi_i(y)| + \sup_{i\in L\backslash I_m}\left|\varphi_{i,\sigma(i)}\left(y_{\sigma(i)}\right)\right| < +\infty;$$
then, we obtain
$$\sup_{i\in L\backslash I_m}\left|\varphi_i^{(L,N)}(x)\right| \leq \sup_{i\in L\backslash I_m}\left|\varphi_i^{(L,m)}(x_{J_m})\right| + \sup_{i\in L\backslash I_m}\left|\varphi_{i,\sigma(i)}\left(y_{\sigma(i)}\right)\right| < +\infty,$$
from which $\varphi^{(L,N)}\left(\pi_{J,N}(U)\right) \subset E_L$. Moreover, $\forall\, z \in U$, $\forall\, i \in I\backslash I_m$, we have
$$\left|\overline{\varphi}_i^{(L,N)}(z)\right| \leq \left|\varphi_i^{(I,m)}(z_{J_m})\right| + \left|\varphi_{i,\sigma(i)}\left(z_{\sigma(i)}\right)\right|,$$
and so $\sup\limits_{i\in I\backslash I_m}\left|\overline{\varphi}_i^{(L,N)}(z)\right| < +\infty$; then, $\overline{\varphi}^{(L,N)}(U) \subset E_I$. Finally, from the definition of $\overline{\varphi}^{(L,N)}$, the function $\overline{\varphi}^{(L,N)} : U \subset E_J \longrightarrow E_I$ is (m,σ)-general.
4. $\forall\, x \in U$, $\forall\, i \in I\backslash I_m$, we have
$$\|J_{\varphi_i}(x)\| = \left\|J_{\varphi_i^{(I,m)}}(x_{J_m})\right\| + \left|\varphi'_{i,\sigma(i)}\left(x_{\sigma(i)}\right)\right|;$$
furthermore, since $\sum\limits_{i\in I\backslash I_m}\left\|J_{\varphi_i^{(I,m)}}(x_{J_m})\right\| < +\infty$, we have
$$\sup_{i\in I\backslash I_m}\left\|J_{\varphi_i^{(I,m)}}(x_{J_m})\right\| < +\infty,$$
and so
$$\sup_{i\in I\backslash I_m}\|J_{\varphi_i}(x)\| \leq \sup_{i\in I\backslash I_m}\left\|J_{\varphi_i^{(I,m)}}(x_{J_m})\right\| + \sup_{i\in I\backslash I_m}\left|\varphi'_{i,\sigma(i)}\left(x_{\sigma(i)}\right)\right| < +\infty;$$
then, from Proposition 2.4, there exists the function $J_{\varphi^{(I\backslash I_m,J)}}(x) : E_J \longrightarrow E_{I\backslash I_m}$, and it is continuous.

5. $\forall n \in \mathbf{N}$, $n \geq m$, and $\forall x_{J_n} \in \pi_{J,J_n}(U)$, we have

$$\sum_{i\in I\setminus I_n} \left\| J_{\varphi_i^{(I,n)}}(x_{J_n}) \right\|$$
$$= \sum_{i\in I\setminus I_n} \left\| J_{\varphi_i^{(I,m)}}(x_{J_m}) \right\| + \sum_{j\in J_n\setminus J_m} \left(\sum_{i\in I\setminus I_n} \left|\varphi'_{i,j}\left(x_j\right)\right| \right) < +\infty;$$

then, by Definition 2.21 and by definition of ξ, φ is (n,ξ)-general.

6. From points 3 and 5 and since σ is injective, $\forall n \in \mathbf{N}$, $n \geq \widehat{m}$, $\overline{\varphi}^{(L,N)}$ is $\left(n, \sigma|_{I\setminus I_n}\right)$-general; moreover, since σ is increasing, $\forall i \in I\setminus I_n$ and $\forall x_{J_n} \in \pi_{J,J_n}(U)$, we have $\varphi_i^{(I,n)}(x_{J_n}) = 0$; then, we have the statement.

□

Remark 2.25: Let $\varphi : U \subset E_J \longrightarrow E_I$ be a (m,σ)-general function such that $U^{(m)} = \prod_{j\in J_m} A_j$, where $A_j \in \tau$, for any $j \in J_m$, and

$$\varphi_i^{(I,m)}(x_{J_m}) = \sum_{j\in J_m} \varphi_{ij}(x_j),\ \forall\, x_{J_m} \in U^{(m)},\ \forall\, i \in I,$$

where $\varphi_{ij} : A_j \longrightarrow \mathbf{R}$, for any $i \in I$ and for any $j \in J_m$; moreover, suppose that, for any $j \in J_m$, for any $t \in A_j$, one has $\sup_{i\in I\setminus I_m} \left|\varphi_{i,j}\left(t\right)\right| < +\infty$, and, for any $j \in J\setminus J_m$, for any $t \in A_j$, one has $\sum_{i\in I\setminus I_m} \left|\varphi'_{i,j}\left(t\right)\right| < +\infty$; furthermore, let $\emptyset \neq L \subset I$ and let $\emptyset \neq N \subset J$ such that $|I\setminus L| = |J\setminus N| < +\infty$. Then, for any $n \in \mathbf{N}$, $n \geq m$, the function $\varphi^{(L,N)} : \pi_{J,N}(U) \longrightarrow \mathbf{R}^L$ is (n,ρ)-general, where the function $\rho : L\setminus L_n \longrightarrow N\setminus N_n$ is defined by

$$\rho(i) = \begin{cases} \sigma(i) & \text{if } \sigma(i) \in N\setminus N_n \\ \min\left\{j > \sigma(i) : j \in N\setminus N_n\right\} & \text{if } \sigma(i) \notin N\setminus N_n \end{cases},\ \forall\, i \in L\setminus L_n.$$

Proof. We have $|L| = |N|$; moreover, $\forall n \in \mathbf{N}$, $n \geq m$, $\forall i \in L\setminus L_n$ and $\forall x \in \pi_{J,N}(U)$, let $y \in U$ such that $y_N = x$; we have

$$\left|\varphi_i\left(x\right)\right| \leq \sum_{j\in N\cap J_m} \left|\varphi_{i,j}\left(x_j\right)\right| + \left|\varphi_{i,\sigma(i)}\left(y_{\sigma(i)}\right)\right|$$
$$\Rightarrow \left\|\varphi\left(x\right)\right\|_{L\setminus L_n} \leq \sum_{j\in N\cap J_m} \sup_{i\in L\setminus I_n} \left|\varphi_{i,j}\left(x_j\right)\right| + \sup_{i\in L\setminus I_n} \left|\varphi_{i,\sigma(i)}\left(y_{\sigma(i)}\right)\right| < +\infty,$$

from which $\varphi(\pi_{J,N}(U)) \subset E_L$. Analogously, $\forall\, n \in \mathbf{N}$, $n \geq m$, and $\forall\, x_{N_n} \in \pi_{J,N_n}(U)$, we have

$$\sum_{i \in L \backslash L_n} \left\| J_{\varphi_i^{(L,N_n)}}(x_{N_n}) \right\|$$
$$= \sum_{i \in L \backslash L_n} \left\| J_{\varphi_i^{(L,N_n \cap J_m)}}(x_{N_m \cap J_m}) \right\| + \sum_{j \in N_n \backslash J_m} \left(\sum_{i \in L \backslash L_n} \left| \varphi'_{i,j}(x_j) \right| \right) < +\infty;$$

then, by definition of ρ, $\varphi^{(L,N)}$ is (n,ρ)-general. □

Proposition 2.26. *Let $\varphi : U \subset E_J \longrightarrow E_I$ be a (m,σ)-general function such that there exists $m_0 \in \mathbf{N}$, $m_0 \geq m$, such that, for any $j \in J \backslash J_{m_0}$, A_j is bounded; moreover, suppose that $\sigma(I \backslash I_m) \cap (J \backslash J_{m_0}) \neq \emptyset$ and, for any $i \in I \backslash I_m$, $\varphi_i^{(I,m)}$ is bounded; then, there exists $m_1 \in \mathbf{N}$, $m_1 \geq m$, such that, for any $i \in I \backslash I_{m_1}$, φ_i is bounded. In particular, if $|I| = +\infty$, φ is not surjective.*

Proof. Let $j_0 = \min(\sigma(I \backslash I_m) \cap (J \backslash J_{m_0}))$, let $i_0 = \min(\sigma^{-1}(j_0)) \in I$, let $\widehat{m} = |i_0| - 1$ and let $\mathcal{H} = \left\{ i \in I \backslash I_{\widehat{m}} : \varphi_{i,\sigma(i)} \text{ is not bounded} \right\}$; we have $|\mathcal{H}| < +\infty$; indeed, $\forall\, i \in \mathcal{H}$, the set $A_{\sigma(i)}$ is bounded, and so there exists $t_i \in A_{\sigma(i)}$ such that $\left| \varphi'_{i,\sigma(i)}(t_i) \right| > |i|$; then, $\forall\, x_{\sigma(I \backslash I_m)} \in \prod_{j \in \sigma(I \backslash I_m)} A_j$ such that $\left(x_{\sigma(i)} : i \in \mathcal{H} \right) = (t_i : i \in \mathcal{H})$, by supposing by contradiction $|\mathcal{H}| = +\infty$, we would obtain

$$\sup_{i \in I \backslash I_m} \left| \varphi'_{i,\sigma(i)}\left(x_{\sigma(i)}\right) \right| \geq \sup_{i \in \mathcal{H}} \left| \varphi'_{i,\sigma(i)}\left(x_{\sigma(i)}\right) \right| = \sup_{i \in \mathcal{H}} \left| \varphi'_{i,\sigma(i)}(t_i) \right| = +\infty$$

(a contradiction). Then, there exists $m_1 \in \mathbf{N}$, $m_1 \geq m$, such that, $\forall\, i \in I \backslash I_{m_1}$, $\varphi_{i,\sigma(i)}$ is bounded, and so φ_i is bounded. In particular, $\forall\, i \in I \backslash I_{m_1}$, φ_i is not surjective; then, if $|I| = +\infty$, φ is not surjective. □

Proposition 2.27. *Let $\varphi : U \subset E_J \longrightarrow E_I$ be a (m,σ)-general function such that $\varphi_{ij}(x_j) = 0$, for any $i \in I_m$, for any $j \in J \backslash J_m$ and for any $x_j \in A_j$; then:*

1. *If the functions $\varphi_{i,\sigma(i)}$, for any $i \in I \backslash I_m$, and $\varphi^{(m,m)}$ are injective, and σ is surjective, then φ is injective.*
2. *If the functions $\varphi_{i,\sigma(i)}$, for any $i \in I \backslash I_m$, and $\varphi^{(m,m)}$ are surjective, and σ is injective, then φ is surjective.*

Proof.

1. Let $x,\ y\in U$ be such that $\varphi(x)=\varphi(y)$; we have $\varphi^{(m,m)}(x_{J_m})=(\varphi(x))_{I_m}=(\varphi(y))_{I_m}=\varphi^{(m,m)}(y_{J_m})$; then, if $\varphi^{(m,m)}$ is injective, we have $x_{J_m}=y_{J_m}$; moreover, $\forall\, i\in I\backslash I_m$:
$$\begin{aligned}\varphi^{(\{i\},m)}(x_{J_m})&+\varphi_{i,\sigma(i)}(x_{\sigma(i)})\\ &=\varphi_i(x)=\varphi_i(y)=\varphi^{(\{i\},m)}(y_{J_m})+\varphi_{i,\sigma(i)}(y_{\sigma(i)}),\end{aligned}$$
from which $\varphi_{i,\sigma(i)}(x_{\sigma(i)})=\varphi_{i,\sigma(i)}(y_{\sigma(i)})$; then, if $\varphi_{i,\sigma(i)}$ is injective, we have $x_{\sigma(i)}=y_{\sigma(i)}$; finally, if σ is surjective, we obtain $x_{J\backslash J_m}=y_{J\backslash J_m}$, and so $x=y$; then, φ is injective.

2. Let $y\in E_I$; moreover, if the functions $\varphi_{i,\sigma(i)}$, for any $i\in I\backslash I_m$, and $\varphi^{(m,m)}$ are surjective, and σ is injective, define $x\in U^{(m)}\times\prod_{j\in J\backslash J_m}A_j$ in the following manner:
$$\begin{aligned}x_{J_m}&=\left(\varphi^{(m,m)}\right)^{-1}(y_{I_m})\in U^{(m)},\\ x_j&=\varphi^{-1}_{\sigma^{-1}(j),j}(z_i)\in A_j,\ \forall\, j\in\sigma\left(I\backslash I_m\right),\\ x_j&=0,\ \forall\, j\in J\backslash\sigma\left(I\backslash I_m\right),\end{aligned}$$
where
$$z_i=y_i-\varphi_i^{(I,m)}(x_{J_m}),\forall\, i\in I\backslash I_m. \tag{8}$$
Let $x_0=(x_{0,j}:j\in J)\in U$; $\forall\, i\in I\backslash I_m$, we have
$$\begin{aligned}\left|x_{\sigma(i)}\right|&=\left|\varphi^{-1}_{i,\sigma(i)}(z_i)-x_{0,\sigma(i)}+x_{0,\sigma(i)}\right|\\ &\le\left|\varphi^{-1}_{i,\sigma(i)}(z_i)-\varphi^{-1}_{i,\sigma(i)}(\varphi_{i,\sigma(i)}(x_{0,\sigma(i)}))\right|+\left|x_{0,\sigma(i)}\right|;\end{aligned}\tag{9}$$
moreover, the function $\varphi^{-1}_{i,\sigma(i)}:\mathbf{R}\longrightarrow A_{\sigma(i)}$ is derivable, and
$$\left(\varphi^{-1}_{i,\sigma(i)}\right)'(t)=\frac{1}{\varphi'_{i,\sigma(i)}(\varphi^{-1}_{i,\sigma(i)}(t))}\in\mathbf{R}^*,\ \forall\, i\in I\backslash I_m,\ \forall\, t\in\mathbf{R}; \tag{10}$$
then, the Lagrange theorem implies that, for some
$$\xi_i\in\left(\min\{z_i,\varphi_{i,\sigma(i)}(x_{0,\sigma(i)})\},\max\{z_i,\varphi_{i,\sigma(i)}(x_{0,\sigma(i)})\}\right),$$
we have
$$\begin{aligned}&\left|\varphi^{-1}_{i,\sigma(i)}(z_i)-\varphi^{-1}_{i,\sigma(i)}(\varphi_{i,\sigma(i)}(x_{0,\sigma(i)}))\right|\\ &\qquad=\left|\left(\varphi^{-1}_{i,\sigma(i)}\right)'(\xi_i)\right|\left|z_i-\varphi_{i,\sigma(i)}(x_{0,\sigma(i)})\right|;\end{aligned}$$

thus, from (9) and (10), we obtain

$$\left|x_{\sigma(i)}\right| \leq \frac{\left|z_i - \varphi_{i,\sigma(i)}(x_{0,\sigma(i)})\right|}{\left|\varphi'_{i,\sigma(i)}(\varphi^{-1}_{i,\sigma(i)}(\xi_i))\right|} + \left|x_{0,\sigma(i)}\right|. \tag{11}$$

Furthermore, from point 3 of Proposition 2.24, we have $\varphi^{(I,m)}(U^{(m)}) \subset E_I$, and so, from (8), we have

$$\|z\|_{I\backslash I_m} \leq \|y\|_{I\backslash I_m} + \sup_{i\in I\backslash I_m} \left|\varphi_i^{(I,m)}\left(x_{J_m}\right)\right| < +\infty, \tag{12}$$

and analogously

$$\begin{aligned}
&\varphi_{i,\sigma(i)}(x_{0,\sigma(i)}) = \varphi_i\left(x_0\right) - \varphi_i^{(I,m)}\left(\left(x_0\right)_{J_m}\right), \ \forall\, i \in I\backslash I_m \\
&\quad\Longrightarrow \sup_{i\in I\backslash I_m} \left|\varphi_{i,\sigma(i)}(x_{0,\sigma(i)})\right| \\
&\qquad\qquad \leq \|\varphi(x_0)\|_{I\backslash I_m} + \sup_{i\in I\backslash I_m} \left|\varphi_i^{(I,m)}\left(\left(x_0\right)_{J_m}\right)\right| < +\infty. \qquad (13)
\end{aligned}$$

Moreover, we have $\inf_{i\in\mathcal{I}_\varphi} \left|\varphi'_{i,\sigma(i)}(\varphi^{-1}_{i,\sigma(i)}(\xi_i)\right| > 0$; furthermore, since, $\forall\, i \in I\backslash I_m$, $\varphi_{i,\sigma(i)}$ is surjective, then $\varphi_{i,\sigma(i)}$ is injective too, and so $\mathcal{I}_\varphi = I\backslash I_m$; then, there exists $c \in \mathbf{R}^+$ such that $\sup_{i\in I\backslash I_m} \left|\varphi'_{i,\sigma(i)}(\varphi^{-1}_{i,\sigma(i)}(\xi_i))\right|^{-1} \leq c$, and so formulas (11), (12) and (13) imply

$$\sup_{i\in I\backslash I_m} \left|x_{\sigma(i)}\right| \leq c\left(\|z\|_{I\backslash I_m} + \sup_{i\in I\backslash I_m} \left|\varphi_{i,\sigma(i)}(x_{0,\sigma(i)})\right|\right) + \|x_0\|_J < +\infty;$$

then, we have $x \in E_J$, from which $x \in U$. Finally, it is easy to prove that $\varphi(x) = y$, and so φ is surjective.

□

Proposition 2.28. *Let $m \in \mathbf{N}^*$, let $\emptyset \neq L \subset I$, let $J_m \subset N \subset J$ and let $\varphi : U \subset E_J \longrightarrow E_I$ be a function m-general and C^1 in $x_0 = (x_{0,j} : j \in J) \in U$; then:*

1. *If $\varphi^{(L,N)}\left(\pi_{J,N}(U)\right) \subset E_L$, then the function $\varphi^{(L,N)} : \pi_{J,N}(U) \longrightarrow E_L$ is C^1 in $(x_{0,j} : j \in N)$.*
2. *If φ is (m,σ)-general and $I_m \subset L$, then the function $\overline{\varphi}^{(L,N)} : U \subset E_J \longrightarrow E_I$ is C^1 in x_0.*

3. *If φ is (m,σ)-general, $I_m \subset L$ and $|N| < +\infty$, then there exists the function $J_{\overline{\varphi}^{(L,N)}}(x_0) : E_J \longrightarrow E_I$, and it is continuous.*

4. *If φ is strongly (m,σ)-general, $I_m \subset L$ and $|N| < +\infty$, then $\overline{\varphi}^{(L,N)}$ is differentiable in x_0.*

5. *If φ is strongly C^1 in x_0 and strongly (m,σ)-general, then φ is differentiable in x_0.*

Proof.

1. By assumption, there exists a neighbourhood $V = \prod_{j\in J} V_j \in \tau_{\|\cdot\|_J}(U)$ of x_0 such that, $\forall\, i \in I$, $\forall\, j \in J$, there exists the function $x \longrightarrow \frac{\partial\varphi_i(x)}{\partial x_j}$ on V, and this function is continuous in x_0; then, $\forall\, x \in \prod_{j\in N} V_j$, let $\overline{x} = (\overline{x}_j : j \in J) \in V$ such that $(\overline{x}_j : j \in N) = x$; since φ is a m-general function, $\forall\, i \in L$, $\forall\, j \in N$, we have
$$\frac{\partial\varphi_i^{(L,N)}(x)}{\partial x_j} = \frac{\partial\varphi_i(\overline{x})}{\partial x_j},$$
from which $\varphi^{(L,N)}$ is C^1 in $(x_{0,j} : j \in N)$.

2. Let $V \in \tau_{\|\cdot\|_J}(U)$ be the neighbourhood of x_0 defined in the proof of point 1; if φ is (m,σ)-general and $I_m \subset L$, $\forall\, x \in V$, we have
$$\frac{\partial\overline{\varphi}_i^{(L,N)}(x)}{\partial x_j} = \begin{cases} \frac{\partial\varphi_i(x)}{\partial x_j} & \text{if } (i,j) \notin (I_m \times (J\backslash N)) \cup ((I\backslash L) \times J_m) \\ 0 & \text{if } (i,j) \in (I_m \times (J\backslash N)) \cup ((I\backslash L) \times J_m) \end{cases},$$
and so $\overline{\varphi}^{(L,N)}$ is C^1 in x_0.

3. If φ is C^1 in x_0 and (m,σ)-general, $I_m \subset L$ and $|N| < +\infty$, then, from point 2, $\forall\, i \in I_m$, the function $\overline{\varphi}_i^{(L,N)} : U \subset E_J \longrightarrow \mathbf{R}$ is C^1 in x_0 and depends only on a finite number of variables; then, we have $\left\| J_{\overline{\varphi}_i^{(L,N)}}(x_0) \right\| < +\infty$; moreover, $\forall\, i \in I\backslash I_m$, we have
$$\left\| J_{\overline{\varphi}_i^{(L,N)}}(x_0) \right\| \leq \|J_{\varphi_i}(x_0)\|;$$
then, from point 4 of Proposition 2.24:
$$\sup_{i\in I\backslash I_m} \left\| J_{\overline{\varphi}_i^{(L,N)}}(x_0) \right\| \leq \sup_{i\in I\backslash I_m} \|J_{\varphi_i}(x_0)\| < +\infty;$$
then, from Proposition 2.4, there exists the function $J_{\overline{\varphi}^{(L,N)}}(x_0) : E_J \longrightarrow E_I$, and it is continuous.

4. If φ is strongly (m,σ)-general, there exists $a \in \mathbf{R}$ such that, $\forall \varepsilon > 0$, there exists $\widehat{i} \in \mathbf{N}$, $\widehat{i} \geq m$, such that

$$\left\| J_{\varphi_i^{(I,m)}}\left(x_{J_m}\right) \right\| < \frac{\varepsilon}{4}, \forall i \in I \backslash I_{\widehat{i}}, \forall x_{J_m} \in U^{(m)};$$
$$\left| \varphi'_{i,\sigma(i)}\left(t\right) - a \right| < \frac{\varepsilon}{4}, \forall i \in \mathcal{I}_\varphi \cap I \backslash I_{\widehat{i}}, \forall t \in A_{\sigma(i)}. \tag{14}$$

Moreover, if $I_m \subset L$ and $|N| < +\infty$, $\forall i \in I$, the function $\overline{\varphi}_i^{(L,N)} : U \subset E_J \longrightarrow \mathbf{R}$ is C^1 in x_0 and depends only on a finite number of variables; then, $\overline{\varphi}_i^{(L,N)}$ is differentiable in x_0, and so there exists a neighbourhood $D = \prod_{j \in J} D_j \in \tau_{\|\cdot\|_J}(U)$ of x_0, where D_j is an open interval, $\forall j \in J$, such that, $\forall x = (x_j : j \in J) \in D \backslash \{x_0\}$, we have

$$\sup_{i \in I_{\widehat{i}}} \frac{\left| \overline{\varphi}_i^{(L,N)}(x) - \overline{\varphi}_i^{(L,N)}(x_0) - J_{\overline{\varphi}_i^{(L,N)}}(x_0)(x - x_0) \right|}{\|x - x_0\|_J} < \varepsilon. \tag{15}$$

Observe that, $\forall i \in \left(I \backslash I_{\widehat{i}}\right) \backslash L$, $\forall y = (y_j : j \in J) \in U$, we have $\overline{\varphi}_i^{(L,N)}(y) = \varphi_{i,\sigma(i)}(y_{\sigma(i)})$; moreover, $\varphi_{i,\sigma(i)}$ is derivable in $A_{\sigma(i)}$ and so, from the Lagrange theorem, $\forall x \in D \backslash \{x_0\}$, there exists $\theta_i \in (\min\{x_{0,\sigma(i)}, x_{\sigma(i)}\}, \max\{x_{0,\sigma(i)}, x_{\sigma(i)}\})$ such that

$$\varphi_{i,\sigma(i)}\left(x_{\sigma(i)}\right) - \varphi_{i,\sigma(i)}\left(x_{0,\sigma(i)}\right) = \varphi'_{i,\sigma(i)}\left(\theta_i\right)\left(x_{\sigma(i)} - x_{0,\sigma(i)}\right),$$

from which

$$\begin{aligned}
&\frac{\left| \overline{\varphi}_i^{(L,N)}(x) - \overline{\varphi}_i^{(L,N)}(x_0) - J_{\overline{\varphi}_i^{(L,N)}}(x_0)(x - x_0) \right|}{\|x - x_0\|_J} \\
&= \frac{\left| \varphi_{i,\sigma(i)}\left(x_{\sigma(i)}\right) - \varphi_{i,\sigma(i)}\left(x_{0,\sigma(i)}\right) - \varphi'_{i,\sigma(i)}\left(x_{0,\sigma(i)}\right)\left(x_{\sigma(i)} - x_{0,\sigma(i)}\right) \right|}{\|x - x_0\|_J} \\
&= \frac{\left| \varphi'_{i,\sigma(i)}\left(\theta_i\right) - \varphi'_{i,\sigma(i)}\left(x_{0,\sigma(i)}\right) \right| \left| x_{\sigma(i)} - x_{0,\sigma(i)} \right|}{\|x - x_0\|_J} \\
&\leq \left(\left| \varphi'_{i,\sigma(i)}\left(\theta_i\right) - a \right| + \left| \varphi'_{i,\sigma(i)}\left(x_{0,\sigma(i)}\right) - a \right| \right) 1_{\mathcal{I}_\varphi}(i) < \frac{\varepsilon}{2}.
\end{aligned} \tag{16}$$

Conversely, $\forall i \in \left(I \backslash I_{\widehat{i}}\right) \cap L$, $\forall y = (y_j : j \in J) \in U$, we have $\overline{\varphi}_i^{(L,N)}(y) = \varphi_i(y)$; moreover, from point 3 of Proposition 2.24 and from point 1, $\varphi_i^{(I,m)}$ is C^1 in $(x_0)_{J_m}$ and so $\varphi_i^{(I,m)}$ is C^1 in a neighbourhood $M = \prod_{j \in J_m} M_j \in \tau_{\|\cdot\|_{J_m}}(U^{(m)})$ of $(x_0)_{J_m}$ such that M_j is an open interval,

$\forall\, j \in J_m$, and $M \subset \prod_{j\in J_m} D_j$; then, from the Taylor theorem, $\forall\, x \in \left(M \times \prod_{j\in J\setminus J_m} D_j\right)\setminus\{x_0\}$, there exists $\xi_{J_m} \in (M\setminus\{(x_0)_{J_m}\})$ such that

$$\varphi_i^{(I,m)}\left(x_{J_m}\right) - \varphi_i^{(I,m)}\left((x_0)_{J_m}\right) = J_{\varphi_i^{(I,m)}}\left(\xi_{J_m}\right)\left(x_{J_m} - (x_0)_{J_m}\right),$$

and so

$$\begin{aligned}
&\frac{\left|\overline{\varphi}_i^{(L,N)}(x) - \overline{\varphi}_i^{(L,N)}(x_0) - J_{\overline{\varphi}_i^{(L,N)}}(x_0)(x-x_0)\right|}{\|x-x_0\|_J}\\
&= \frac{|\varphi_i(x) - \varphi_i(x_0) - J_{\varphi_i}(x_0)(x-x_0)|}{\|x-x_0\|_J}\\
&\le \frac{\left|\varphi_i^{(I,m)}(x_{J_m}) - \varphi_i^{(I,m)}((x_0)_{J_m}) - J_{\varphi_i^{(I,m)}}((x_0)_{J_m})(x_{J_m} - (x_0)_{J_m}))\right|}{\|x-x_0\|_J}\\
&\quad + \frac{\left|\varphi_{i,\sigma(i)}\left(x_{\sigma(i)}\right) - \varphi_{i,\sigma(i)}\left(x_{0,\sigma(i)}\right) - \varphi'_{i,\sigma(i)}\left(x_{0,\sigma(i)}\right)\left(x_{\sigma(i)} - x_{0,\sigma(i)}\right)\right|}{\|x-x_0\|_J}\\
&\le \frac{\left\|J_{\varphi_i^{(I,m)}}\left(\xi_{J_m}\right) - J_{\varphi_i^{(I,m)}}\left((x_0)_{J_m}\right)\right\| \left\|(x_{J_m} - (x_0)_{J_m})\right\|_{J_m}}{\|x-x_0\|_J}\\
&\quad + \frac{\left|\varphi'_{i,\sigma(i)}\left(\theta_i\right) - \varphi'_{i,\sigma(i)}\left(x_{0,\sigma(i)}\right)\right| \left|x_{\sigma(i)} - x_{0,\sigma(i)}\right|}{\|x-x_0\|_J}\\
&\le \left\|J_{\varphi_i^{(I,m)}}\left(\xi_{J_m}\right) - J_{\varphi_i^{(I,m)}}\left((x_0)_{J_m}\right)\right\| + \left|\varphi'_{i,\sigma(i)}\left(\theta_i\right) - \varphi'_{i,\sigma(i)}\left(x_{0,\sigma(i)}\right)\right|\\
&\le \left\|J_{\varphi_i^{(I,m)}}\left(\xi_{J_m}\right)\right\| + \left\|J_{\varphi_i^{(I,m)}}\left((x_0)_{J_m}\right)\right\|\\
&\quad + \left(\left|\varphi'_{i,\sigma(i)}\left(\theta_i\right) - a\right| + \left|\varphi'_{i,\sigma(i)}\left(x_{0,\sigma(i)}\right) - a\right|\right) 1_{\mathcal{I}_\varphi}(i) < \varepsilon.
\end{aligned} \tag{17}$$

Then, from (15), (16) and (17), $\forall\, x \in \left(M \times \prod_{j\in J\setminus J_m} D_j\right)\setminus\{x_0\}$, we have

$$\frac{\left\|\overline{\varphi}^{(L,N)}(x) - \overline{\varphi}^{(L,N)}(x_0) - J_{\overline{\varphi}^{(L,N)}}(x_0)(x-x_0)\right\|_I}{\|x-x_0\|_J} < \varepsilon; \tag{18}$$

thus, $\overline{\varphi}^{(L,N)}$ is differentiable in x_0.

5. If φ is strongly C^1 in x_0 and (m,σ)-general, the function $\psi = \varphi - \overline{\varphi}^{(I,m)}$:

$U \subset E_J \longrightarrow E_I$ given by

$$\psi_i(x) = \begin{cases} \sum\limits_{j \in J \setminus J_m} \varphi_{ij}(x_j) & \forall\, i \in I_m,\, \forall\, x \in U \\ 0 & \forall\, i \in I \setminus I_m,\, \forall\, x \in U \end{cases} \tag{19}$$

is strongly C^1 in x_0, and so it is differentiable in x_0 from Proposition 2.19, since $|I_m| < +\infty$; then, if φ is strongly (m, σ)-general, from point 4 $\overline{\varphi}^{(I,m)}$ is differentiable in x_0, and so this is true for $\varphi = \psi + \overline{\varphi}^{(I,m)}$ too, from Remark 2.6.

□

PROPOSITION 2.29. *Let $\varphi : U \subset E_J \longrightarrow E_I$ be a function C^1 and m-general; then, $\varphi : (U, \mathcal{B}^{(J)}(U)) \longrightarrow \left(\mathbf{R}^I, \mathcal{B}^{(I)}\right)$ is measurable.*

Proof. From point 1 of Proposition 2.28, $\forall\, i \in I$ and $\forall\, n \in \mathbf{N}$, $n \geq m$, the function $\varphi^{(\{i\},n)} : \pi_{J,J_n}(U) \longrightarrow \mathbf{R}$ is C^1; thus, $\forall\, C \in \tau$, we have $\left(\varphi^{(\{i\},n)}\right)^{-1}(C) \in \tau^{(n)}(\pi_{J,J_n}(U)) \subset \mathcal{B}^{(n)}(\pi_{J,J_n}(U))$; then, since $\sigma(\tau) = \mathcal{B}$, $\forall\, C \in \mathcal{B}$, we obtain $\left(\varphi^{(\{i\},n)}\right)^{-1}(C) \in \mathcal{B}^{(n)}(\pi_{J,J_n}(U))$. Moreover, $\forall\, i \in I$, consider the function $\widehat{\varphi}^{(\{i\},n)} : U \longrightarrow \mathbf{R}$ defined by

$$\widehat{\varphi}^{(\{i\},n)}(x) = \varphi^{(\{i\},n)}(x_{J_n}),\ \forall\, x \in U;$$

$\forall\, C \in \mathcal{B}$, we have

$$\left(\widehat{\varphi}^{(\{i\},n)}\right)^{-1}(C) = \left(\varphi^{(\{i\},n)}\right)^{-1}(C) \times \pi_{J,J\setminus J_n}(U) \in \mathcal{B}^{(J)}(U),$$

and so $\widehat{\varphi}^{(\{i\},n)}$ is $\left(\mathcal{B}^{(J)}(U), \mathcal{B}\right)$-measurable; then, since $\lim\limits_{n \longrightarrow +\infty} \widehat{\varphi}^{(\{i\},n)} = \varphi_i$, the function φ_i is $\left(\mathcal{B}^{(J)}(U), \mathcal{B}\right)$-measurable too. Furthermore, let

$$\Sigma(I) = \left\{ B = \prod_{i \in I} B_i : B_i \in \mathcal{B},\ \forall\, i \in I \right\};$$

$\forall\, B = \prod\limits_{i \in I} B_i \in \Sigma(I)$, we have

$$\varphi^{-1}(B) = \bigcap_{i \in I} (\varphi_i)^{-1}(B_i) \in \mathcal{B}^{(J)}(U).$$

Finally, since $\sigma(\Sigma(I)) = \mathcal{B}^{(I)}$, $\forall\, B \in \mathcal{B}^{(I)}$, we obtain $\varphi^{-1}(B) \in \mathcal{B}^{(J)}(U)$. □

3. Linear (m,σ)-general functions

DEFINITION 3.1. *Let* $A = (a_{ij})_{i\in I, j\in J} : E_J \longrightarrow E_I$ *be a linear* (m,σ)*-general function;* $\forall\, i \in I\backslash I_m$*, set* $\lambda_i = \lambda_i(A) = a_{i,\sigma(i)}$.

REMARK 3.2: For any $m \in \mathbf{N}^*$, a linear function $A = (a_{ij})_{i\in I, j\in J} : E_J \longrightarrow E_I$ is m-general; moreover, if $|J| = |I|$ and $\sigma : I\backslash I_m \longrightarrow J\backslash J_m$ is an increasing function, A is (m,σ)-general if and only if:

1. $\forall\, i \in I\backslash I_m$, $\forall\, j \in J\backslash \left(J_m \cup \{\sigma(i)\}\right)$, one has $a_{ij} = 0$.
2. $\forall\, j \in J_m$, $\displaystyle\sum_{i\in I\backslash I_m} |a_{ij}| < +\infty$; moreover, one has $\displaystyle\sup_{i\in I\backslash I_m} |\lambda_i| < +\infty$ and $\displaystyle\inf_{i\in I\backslash I_m : \lambda_i \neq 0} |\lambda_i| > 0$.
3. If $\mathcal{A} \neq \emptyset$, there exists $\displaystyle\prod_{i\in I\backslash I_m : \lambda_i \neq 0} \lambda_i \in \mathbf{R}^*$.

Furthermore, A is strongly (m,σ)-general if and only if A is (m,σ)-general and there exists $a \in \mathbf{R}$ such that the sequence $\{\lambda_i\}_{i\in I\backslash I_m : \lambda_i \neq 0}$ converges to a.

Finally, A is (m,σ)-standard if and only if A is (m,σ)-general and $a_{ij} = 0$, for any $i \in I\backslash I_m$, for any $j \in J_m$.

COROLLARY 3.3. *Let* $A = (a_{ij})_{i\in I, j\in J} : E_J \longrightarrow E_I$ *be a linear function; then,* $A : (E_J, \mathcal{B}_J) \longrightarrow \left(\mathbf{R}^I, \mathcal{B}^{(I)}\right)$ *is measurable.*

Proof. The statement follows from Remark 3.2 and Proposition 2.29. □

PROPOSITION 3.4. *Let* $A = (a_{ij})_{i\in I, j\in J} : E_J \longrightarrow E_I$ *be a linear* (m,σ)*-general function. Then:*

1. *A is continuous.*
2. *Let* $\mathcal{C} = \left\{h \in \mathbf{N},\, h \geq m : \sigma|_{I\backslash I_h} \text{ is injective}\right\}$*; if* $\mathcal{C} \neq \emptyset$*, by setting* $\widetilde{m} = \min \mathcal{C}$*, let* $i_{\widetilde{m}} \in I$ *such that* $|i_{\widetilde{m}}| = \widetilde{m}$ *and let*

$$\widetilde{\widetilde{m}} = \begin{cases} \min\{\widetilde{m}, |\sigma(i_{\widetilde{m}})|\} & \text{if } \widetilde{m} > m \\ m & \text{if } \widetilde{m} = m \end{cases} ; \tag{20}$$

then, for any $n \in \mathbf{N}$, $n \geq \widetilde{\widetilde{m}}$*, the linear function* ${}^tA : E_I \longrightarrow \mathbf{R}^J$ *is* (n,τ)*-general, where* $\tau : J\backslash J_n \longrightarrow I\backslash I_n$ *is the increasing function defined by*

$$\tau(j) = \min\left\{\ \sigma^{-1}(k) : k \geq j,\ k \in \sigma\left(I\backslash I_n\right)\ \right\},\ \forall\, j \in J\backslash J_n. \tag{21}$$

Proof.

1. Since $A(E_J) \subset E_I$, the statement follows from Proposition 2.4.
2. We have

$$\begin{aligned}\sup_{j\in J}\sum_{i\in I}\left|\left({}^tA\right)_{ji}\right| &= \sup_{j\in J}\sum_{i\in I}|a_{ij}| \\ &= \sup\left\{\sup_{j\in J_m}\sum_{i\in I}|a_{ij}|, \sup_{j\in J_{\widetilde{\widetilde{m}}}\backslash J_m}\sum_{i\in I}|a_{ij}|, \sup_{j\in J\backslash J_{\widetilde{\widetilde{m}}}}\sum_{i\in I}|a_{ij}|\right\}. \quad (22)\end{aligned}$$

Moreover, from point 2 of Remark 3.2, we have $\sup_{j\in J_m}\sum_{i\in I}|a_{ij}| < +\infty$; furthermore, by definition of $\widetilde{m}$ and $\widetilde{\widetilde{m}}$, $\forall\, j \in J_{\widetilde{\widetilde{m}}}\backslash J_m$, we have $\sum_{i\in I}|a_{ij}| = \sum_{i\in I_{\widetilde{m}+1}}|a_{ij}| < +\infty$; finally, observe that

$$\begin{aligned}\sup_{j\in J\backslash J_{\widetilde{\widetilde{m}}}}\sum_{i\in I}|a_{ij}| &\leq \sum_{i\in I}\left(\sup_{j\in J\backslash J_{\widetilde{\widetilde{m}}}}|a_{ij}|\right) \\ &= \sum_{i\in I_{\widetilde{\widetilde{m}}}}\left(\sup_{j\in J\backslash J_{\widetilde{\widetilde{m}}}}|a_{ij}|\right) + \sum_{i\in I\backslash I_{\widetilde{\widetilde{m}}}}\left(\sup_{j\in J\backslash J_{\widetilde{\widetilde{m}}}}|a_{ij}|\right) \\ &\leq \sum_{i\in I_{\widetilde{\widetilde{m}}}}\left(\sup_{j\in J\backslash J_{\widetilde{\widetilde{m}}}}|a_{ij}|\right) + \sup_{i\in I\backslash I_m}|\lambda_i|. \quad (23)\end{aligned}$$

From Proposition 2.4, $\forall\, i \in I_{\widetilde{\widetilde{m}}}$, we have $\sup_{j\in J\backslash J_{\widetilde{\widetilde{m}}}}|a_{ij}| \leq \sum_{j\in J\backslash J_{\widetilde{\widetilde{m}}}}|a_{ij}| < +\infty$; moreover, we have $\sup_{i\in I\backslash I_m}|\lambda_i| < +\infty$; then, from (23), we obtain $\sup_{j\in J\backslash J_{\widetilde{\widetilde{m}}}}\sum_{i\in I}|a_{ij}| < +\infty$, from which $\sup_{j\in J}\sum_{i\in I}\left|\left({}^tA\right)_{ji}\right| < +\infty$, from formula (22), and so ${}^tA(E_I) \subset E_J$ from Proposition 2.4. Finally, from Remark 3.2, $\forall\, n \in \mathbf{N}$, $n \geq \widetilde{\widetilde{m}}$, the function ${}^tA : E_I \longrightarrow E_J$ is (n,τ)-general, where $\tau : J\backslash J_n \longrightarrow I\backslash I_n$ is the increasing function defined by

$$\tau(j) = \min\left\{\ \sigma^{-1}(k) : k \geq j,\ k \in \sigma\left(I\backslash I_n\right)\ \right\},\ \forall\, j \in J\backslash J_n.$$

□

Henceforth, we will suppose that $|I| = +\infty$.

Definition 3.5. *Let $A=(a_{ij})_{i\in I,j\in J}:E_J\longrightarrow E_I$ be a linear (m,σ)-general function; indicate by $N(A)\in\{0,1,...,m\}$ the number of zero columns of the matrix $A^{(I\backslash I_m,J_m)}$.*

Theorem 3.6. *Let $A=(a_{ij})_{i\in I,j\in J}:E_J\longrightarrow E_I$ be a linear (m,σ)-general function; then, the sequence $\left\{\det A^{(n,n)}\right\}_{n\geq m}$ converges to a real number. Moreover, if $\mathcal{A}\neq\emptyset$, by setting $\overline{m}=\min\mathcal{A}$, we have*

$$\lim_{n\longrightarrow+\infty}\det A^{(n,n)}=\sum_{p\in I\backslash I_{\overline{m}}}\left(\prod_{q\in I\backslash I_{|p|}}\lambda_q\right)\sum_{j\in J_m}a_{p,j}\left(cof\,A^{(|p|,|p|)}\right)_{p,j}+\det A^{(\overline{m},\overline{m})}\left(\prod_{q\in I\backslash I_{\overline{m}}}\lambda_q\right).\quad(24)$$

Conversely, if $\mathcal{A}=\emptyset$, we have $\lim\limits_{n\longrightarrow+\infty}\det A^{(n,n)}=0$.

Proof. $\forall\, l\in\mathbf{Z}$, set $\mathcal{D}_l=\mathcal{D}_l(A)=\{h\in\mathbf{N},\,h\geq m:|\sigma(i)|=|i|+l,\,\forall\, i\in I\backslash I_h\}$; moreover, if $\mathcal{D}_l\neq\emptyset$, set $\overline{m}_l=\min\mathcal{D}_l$; furthermore, set $\mathcal{D}=\mathcal{D}(A)=\bigcup\limits_{l\in\mathbf{Z}}\mathcal{D}_l$. If there exists $l\in\mathbf{N}$ such that $\mathcal{D}_l\neq\emptyset$, we will prove the statement by recursion on $N(A)=k\in\{0,1,...,m\}$. Suppose that $N(A)=0$ and observe that, if $\mathcal{A}\neq\emptyset$, we have $\overline{m}_0=\overline{m}$, since $\mathcal{D}_0=\mathcal{A}$; then, $\forall\, n\in\mathbf{N}$, $n>\overline{m}_l$, we have

$$\det A^{(n,n)}=\begin{cases}\det A^{(\overline{m},\overline{m})}\left(\prod\limits_{q\in I_n\backslash I_{\overline{m}}}\lambda_q\right) & \text{if } l=0\\[2ex] 0 & \text{if } l\in\mathbf{N}^*\end{cases},$$

from which

$$\lim_{n\longrightarrow+\infty}\det A^{(n,n)}=\begin{cases}\det A^{(\overline{m},\overline{m})}\left(\prod\limits_{q\in I\backslash I_{\overline{m}}}\lambda_q\right)\in\mathbf{R} & \text{if } l=0\\[2ex] 0 & \text{if } l\in\mathbf{N}^*\end{cases};$$

then, since we have $a_{p,j}=0$, $\forall\, p\in I\backslash I_{\overline{m}}$, $\forall\, j\in J_m$, the statement is true. Suppose that the statement is true for $N(A)=k$, where $0\leq k\leq m-1$, and suppose that $N(A)=k+1$; $\forall\, n\in\mathbf{N}$, $n>\overline{m}_l$, let $i_n\in I$ such that $|i_n|=n$; we have

$$\det A^{(n,n)}=\sum_{j\in J_n}a_{i_n,j}\left(cof\,A^{(n,n)}\right)_{i_n,j};\quad(25)$$

moreover, let $\{j_1,...,j_{k+1}\}\subset J_m$ such that $a_{i_n,j}=0$, $\forall\, j\in J_m\backslash\{j_1,...,j_{k+1}\}$.

If $l=0$, from (25), we have

$$\det A^{(n,n)}=\sum_{h=1}^{k+1}a_{i_n,j_h}\left(cofA^{(n,n)}\right)_{i_n,j_h}+\lambda_{i_n}\det A^{(n-1,n-1)};$$

then, by induction on n, we obtain

$$\det A^{(n,n)}=a_n+\det A^{(\overline{m},\overline{m})}\left(\prod_{q\in I_n\backslash I_{\overline{m}}}\lambda_q\right),\ \forall\, n>\overline{m}, \tag{26}$$

where

$$a_n=\sum_{p\in I_n\backslash I_{\overline{m}}}\left(\prod_{q\in I_n\backslash I_{|p|}}\lambda_q\right)\sum_{h=1}^{k+1}a_{p,j_h}\left(cofA^{(|p|,|p|)}\right)_{p,j_h}. \tag{27}$$

Moreover, $\forall\, h=1,...,k+1$, $\forall\, p\in I\backslash I_{\overline{m}}$, we have

$$\left|\left(cofA^{(|p|,|p|)}\right)_{p,j_h}\right|=\left|\det A^{(I_{|p|-1},I_{|p|}\backslash\{j_h\})}\right|=\left|\det B_{j_h,p}^{(|p|-1,|p|-1)}\right|, \tag{28}$$

where $B_{j_h,p}: E_J\longrightarrow E_I$ is the linear function obtained by exchanging the $|j_h|$-th column of A for the $|p|$-th column of A; furthermore

$$\begin{aligned}\left|\det B_{j_h,p}^{(|p|-1,|p|-1)}\right|&=\left|\sum_{i\in I_m}a_{i,p}\left(cofB_{j_h,p}^{(|p|-1,|p|-1)}\right)_{i,j_h}\right|\\&\leq\sum_{i\in I_m}|a_{i,p}|\left|\det\left(A^{(I\backslash\{i\},J\backslash\{j_h\})}\right)^{(|p|-2,|p|-2)}\right|.\end{aligned} \tag{29}$$

Observe that, $\forall\, i\in I_m$, $A^{(I\backslash\{i\},J\backslash\{j_h\})}: E_{J\backslash\{j_h\}}\longrightarrow E_{I\backslash\{i\}}$ is a linear $(m-1,\sigma)$-general function such that $\mathcal{D}_0\left(A^{(I\backslash\{i\},J\backslash\{j_h\})}\right)\neq\emptyset$, $N\left(A^{(I\backslash\{i\},J\backslash\{j_h\})}\right)=k$; then, from the recursive assumption, there exists

$$\lim_{|p|\longrightarrow+\infty}\det\left(A^{(I\backslash\{i\},J\backslash\{j_h\})}\right)^{(|p|-2,|p|-2)}\in\mathbf{R},$$

and so

$$\lim_{|p|\longrightarrow+\infty}\sum_{i\in I_m}|a_{i,p}|\left|\det\left(A^{(I\backslash\{i\},J\backslash\{j_h\})}\right)^{(|p|-2,|p|-2)}\right|=0,\ \forall\, h=1,...,k+1;$$

consequently, from (28) and (29), there exists $b\in\mathbf{R}^+$ such that

$$\sup\left\{\left|\left(cofA^{(|p|,|p|)}\right)_{p,j_h}\right|: h\in\{1,...,k+1\},p\in I\backslash I_{\overline{m}}\right\}\leq b. \tag{30}$$

Moreover, since $\prod_{q\in I\setminus I_m:\lambda_q\neq 0}\lambda_q\in\mathbf{R}^*$, we have $\prod_{q\in I\setminus I_m}\overline{\lambda}_q\equiv c\in\mathbf{R}^+$, where

$$\overline{\lambda}_q=\begin{cases}1 & \text{if } \lambda_q=0\\ \frac{1}{|\lambda_q|} & \text{if } 0<|\lambda_q|<1\\ |\lambda_q| & \text{if } |\lambda_q|\geq 1\end{cases},$$

and so

$$\left|\prod_{q\in H}\lambda_q\right|\leq c,\ \forall\, H\subset I\setminus I_{\overline{m}}. \tag{31}$$

Observe that

$$\lim_{n\longrightarrow+\infty}\det A^{(\overline{m},\overline{m})}\left(\prod_{q\in I_n\setminus I_{\overline{m}}}\lambda_q\right)=\det A^{(\overline{m},\overline{m})}\left(\prod_{q\in I\setminus I_{\overline{m}}}\lambda_q\right)\in\mathbf{R}; \tag{32}$$

moreover, set

$$a=\sum_{p\in I\setminus I_{\overline{m}}}\left(\prod_{q\in I\setminus I_{|p|}}\lambda_q\right)\sum_{h=1}^{k+1}a_{p,j_h}\left(cof A^{(|p|,|p|)}\right)_{p,j_h}; \tag{33}$$

then, $\forall\, n>\overline{m}$, we have

$$\begin{aligned}a-a_n=&\sum_{p\in I\setminus I_n}\left(\prod_{q\in I\setminus I_{|p|}}\lambda_q\right)\sum_{h=1}^{k+1}a_{p,j_h}\left(cof A^{(|p|,|p|)}\right)_{p,j_h}\\&+\sum_{p\in I_n\setminus I_{\overline{m}}}\left(\prod_{q\in I_n\setminus I_{|p|}}\lambda_q\right)\left(\left(\prod_{r\in I\setminus I_n}\lambda_r\right)-1\right)\sum_{h=1}^{k+1}a_{p,j_h}\left(cof A^{(|p|,|p|)}\right)_{p,j_h}.\end{aligned} \tag{34}$$

If there exists $n_0\in\mathbf{N}$, $n_0\geq\overline{m}$, such that $\lambda_q\neq 0\ \forall\, q\in I\setminus I_{n_0}$, we have $\prod_{q\in I\setminus I_{n_0}}\lambda_q\in\mathbf{R}^*$; then $\forall\,\varepsilon\in\mathbf{R}^+$, there exists $n_1\in\mathbf{N}$, $n_1\geq n_0$, such that, $\forall\, n\in\mathbf{N}$, $n>n_1$, we have $\left|\left(\prod_{r\in I\setminus I_n}\lambda_r\right)-1\right|<\varepsilon$; thus, from formulas (34), (30) and (31), we obtain

$$|a-a_n|\leq bc\sum_{p\in I\setminus I_n}\sum_{h=1}^{k+1}|a_{p,j_h}|+bc\varepsilon\sum_{p\in I_n\setminus I_{\overline{m}}}\sum_{h=1}^{k+1}|a_{p,j_h}|,\ \forall\, n>n_1. \tag{35}$$

Finally, there exists $d \in \mathbf{R}^{+}$ such that $\sum_{p \in I \setminus I_{\overline{m}}} \sum_{h=1}^{k+1} |a_{p,j_h}| \leq d$, and so there exists $n_2 \in \mathbf{N}$, $n_2 \geq n_1$, such that, $\forall\, n \in \mathbf{N}$, $n \geq n_2$, we have $\sum_{p \in I \setminus I_n} \sum_{h=1}^{k+1} |a_{p,j_h}| < \varepsilon$; then, from formula (35), we obtain

$$|a - a_n| \leq bc\varepsilon + bcd\varepsilon = bc\,(1+d)\,\varepsilon,\ \forall\, n \geq n_2.$$

Then, from (26) and (32), we have

$$\begin{aligned}
&\lim_{n \longrightarrow +\infty} \det A^{(n,n)} \\
&= \sum_{p \in I \setminus I_{\overline{m}}} \left(\prod_{q \in I \setminus I_{|p|}} \lambda_q \right) \sum_{h=1}^{k+1} a_{p,j_h} \left(cof\, A^{(|p|,|p|)} \right)_{p,j_h} + \det A^{(\overline{m},\overline{m})} \left(\prod_{q \in I \setminus I_{\overline{m}}} \lambda_q \right) \\
&= \sum_{p \in I \setminus I_{\overline{m}}} \left(\prod_{q \in I \setminus I_{|p|}} \lambda_q \right) \sum_{j \in J_m} a_{p,j} \left(cof\, A^{(|p|,|p|)} \right)_{p,j} + \det A^{(\overline{m},\overline{m})} \left(\prod_{q \in I \setminus I_{\overline{m}}} \lambda_q \right) \in \mathbf{R}.
\end{aligned}$$

Moreover, suppose that σ is bijective and there exists a subsequence $\{\lambda_{q_t}\}_{t \in \mathbf{N}} \subset \{\lambda_q\}_{q \in I \setminus I_{\overline{m}} : \lambda_q = 0}$; then, from formulas (27) and (33), $\forall\, t \in \mathbf{N}$, $\forall\, n \geq |q_t|$, we obtain

$$\begin{aligned}
a - a_n = &\sum_{p \in I \setminus I_n} \left(\prod_{q \in I \setminus I_{|p|}} \lambda_q \right) \sum_{h=1}^{k+1} a_{p,j_h} \left(cof\, A^{(|p|,|p|)} \right)_{p,j_h} \\
&- \sum_{p \in I_n \setminus I_{\overline{m}}} \left(\prod_{q \in I_n \setminus I_{|p|}} \lambda_q \right) \sum_{h=1}^{k+1} a_{p,j_h} \left(cof\, A^{(|p|,|p|)} \right)_{p,j_h} \\
= &- \sum_{p \in I_n \setminus I_{|q_t|-1}} \left(\prod_{q \in I_n \setminus I_{|p|}} \lambda_q \right) \sum_{h=1}^{k+1} a_{p,j_h} \left(cof\, A^{(|p|,|p|)} \right)_{p,j_h}. \qquad (36)
\end{aligned}$$

Thus, from formulas (30), (31) and (36):

$$|a - a_n| \leq bc \sum_{p \in I_n \setminus I_{|q_t|-1}} \sum_{h=1}^{k+1} |a_{p,j_h}|\,,\ \forall\, t \in \mathbf{N},\ \forall\, n \geq |q_t|\,. \qquad (37)$$

Finally, $\forall\, \varepsilon \in \mathbf{R}^{+}$, there exists $t \in \mathbf{N}$ such that $\sum_{p \in I_n \setminus I_{|q_t|-1}} \sum_{h=1}^{k+1} |a_{p,j_h}| < \varepsilon$, $\forall\, n \geq |q_t|$; then, from (37), we obtain

$$|a - a_n| \leq bc\varepsilon,\ \forall\, n \geq |q_t|\,.$$

Thus, from (26) and (32), we have formula (24).

Moreover, if $l \in \mathbf{N}^*$, from (25) we have

$$\det A^{(n,n)} = \sum_{h=1}^{k+1} a_{i_n,j_h} \left(cof A^{(n,n)}\right)_{i_n,j_h}, \forall n > \overline{m}_l; \tag{38}$$

moreover, $\forall h = 1, ..., k+1$, we have

$$\left|\left(cof A^{(n,n)}\right)_{i_n,j_h}\right| = \left|\det A^{(I_{n-1},I_n\backslash\{j_h\})}\right| = \left|\det\left(A^{(I,J\backslash\{j_h\})}\right)^{(n-1,n-1)}\right|. \tag{39}$$

Observe that $A^{(I,J\backslash\{j_h\})} : E_{J\backslash\{j_h\}} \longrightarrow E_I$ is a linear (m,τ)-general function, where $\tau : I\backslash I_m \longrightarrow J\backslash J_{m+1}$ is the function defined by $\tau(i) = \sigma(i)$, $\forall i \in I\backslash I_m$; moreover, $\mathcal{D}_{l-1}\left(A^{(I,J\backslash\{j_h\})}\right) \neq \emptyset$, $l-1 \in \mathbf{N}$, $N\left(A^{(I,J\backslash\{j_h\})}\right) = k$; then, from the recursive assumption, there exists $\lim\limits_{n\longrightarrow+\infty} \det\left(A^{(I,J\backslash\{j_h\})}\right)^{(n-1,n-1)} \in \mathbf{R}$, and so

$$\lim_{n\longrightarrow+\infty} |a_{i_n,j_h}| \left|\det\left(A^{(I,J\backslash\{j_h\})}\right)^{(n-1,n-1)}\right| = 0, \forall h = 1, ..., k+1;$$

consequently, from (38) and (39), we obtain $\lim\limits_{n\longrightarrow+\infty} \det A^{(n,n)} = 0$.

Furthermore, suppose that there exists $l \in \mathbf{Z}^-$ such that $\mathcal{D}_l \neq \emptyset$; since the function $\sigma|_{I\backslash I_{\overline{m}_l}}$ is injective, from Proposition 3.4, the linear function ${}^tA : E_I \longrightarrow E_J$ is $(\overline{m}_l, \tau)$-general, where $\tau : J\backslash J_{\overline{m}_l} \longrightarrow I\backslash I_{\overline{m}_l}$ is the increasing function defined by $\tau(j) = \sigma^{-1}(j)$, $\forall j \in J\backslash J_{\overline{m}_l}$; moreover, we have $\mathcal{D}_{-l}\left({}^tA\right) \neq \emptyset$, $-l \in \mathbf{N}^*$; then, from the previous arguments, we obtain

$$\lim_{n\longrightarrow+\infty} \det A^{(n,n)} = \lim_{n\longrightarrow+\infty} {}^tA^{(n,n)} = 0.$$

Finally, if $\mathcal{D} = \emptyset$, we have

$$|\{i \in I\backslash I_m : \sigma(i) = \sigma(h), \text{ fore some } h \in I\backslash I_m, h < i\}| = +\infty$$

or $|(J\backslash J_m)\backslash\sigma(I\backslash I_m)| = +\infty$; then, the rows or the columns of the matrix $A^{(n,n)}$ are linearly dependent, for n sufficiently large, and so we have $\det A^{(n,n)} = 0$, from which $\lim\limits_{n\longrightarrow+\infty} \det A^{(n,n)} = 0$. □

Definition 3.7. *Let $A = (a_{ij})_{i\in I, j\in J} : E_J \longrightarrow E_I$ be a linear (m,σ)-general function; define the determinant of A, and call it* $\det A$*, the real number*

$$\det A = \lim_{n\longrightarrow+\infty} \det A^{(n,n)}.$$

Corollary 3.8. *Let* $A = (a_{ij})_{i\in I, j\in J} : E_J \longrightarrow E_I$ *be a linear* (m,σ)*-general function such that* $a_{ij} = 0$, $\forall\, i \in I_m$, $\forall\, j \in J\backslash J_m$, *or* A *is* (m,σ)*-standard. Then, if* σ *is bijective, we have*

$$\det A = \det A^{(m,m)} \prod_{i\in I\backslash I_m} \lambda_i.$$

Conversely, if σ *is not bijective, we have* $\det A = 0$*. In particular, if* $A = \mathbf{I}_{I,J}$*, we have* $\det A = 1$*.*

Proof. If σ is bijective, $\forall\, i \in I\backslash I_m$, we have $|\sigma(i)| = |i|$; then, $\forall\, n \in \mathbf{N}$, $n \geq m$, we have

$$\det A^{(n,n)} = \det A^{(m,m)} \prod_{i\in I_n\backslash I_m} \lambda_i,$$

from which

$$\det A = \lim_{n\longrightarrow +\infty} \det A^{(n,n)} = \det A^{(m,m)} \prod_{i\in I\backslash I_m} \lambda_i.$$

Moreover, suppose that $\mathcal{A} \neq \emptyset$ but σ is not bijective, and set $\overline{m} = \min \mathcal{A}$; by definition of $\overline{m}$, we have $\overline{m} > m$ and the matrix $A^{(\overline{m},\overline{m})}$ is not invertible; then, $\forall\, n \in \mathbf{N}$, $n \geq \overline{m}$, we obtain

$$\det A^{(n,n)} = \det A^{(\overline{m},\overline{m})} \prod_{p\in I_n\backslash I_{\overline{m}}} \lambda_p = 0,$$

and so $\det A = \lim_{n\longrightarrow +\infty} \det A^{(n,n)} = 0$. Finally, if $\mathcal{A} = \emptyset$, from Theorem 3.6 we have $\det A = 0$ again. In particular, if $A = \mathbf{I}_{I,J}$, then A is $(1,\sigma)$-standard, where $A^{(1,1)} = (1)$, $\lambda_i = 1$, $\forall\, i \in I\backslash I_1$, and σ is bijective; then, $\det A = 1$. □

Proposition 3.9. *Let* $A = (a_{ij})_{i\in I, j\in J} : E_J \longrightarrow E_I$ *be a linear* (m,σ)*-general function such that* $a_{ij} = 0$, $\forall\, i \in I_m$, $\forall\, j \in J\backslash J_m$, *or* A *is* (m,σ)*-standard; then:*

1. *One has* $\det A \neq 0$ *if and only if* $A^{(m,m)}$ *is invertible,* $\lambda_i \neq 0$*, for any* $i \in I\backslash I_m$*, and* σ *is bijective.*
2. *If* $a_{ij} = 0$, $\forall\, i \in I_m$, $\forall\, j \in J\backslash J_m$*, and* $\det A \neq 0$*, then* A *is bijective.*
3. *If* A *is* (m,σ)*-standard, then one has* $\det A \neq 0$ *if and only if* A *is bijective.*

Proof.

1. If σ is bijective, from Corollary 3.8, we have

$$\det A = \det A^{(m,m)} \prod_{i \in I \setminus I_m} \lambda_i.$$

Moreover, if $A^{(m,m)}$ is invertible and $\lambda_i \neq 0$, $\forall\, i \in I \setminus I_m$, we have $\det A^{(m,m)} \neq 0$, $\prod_{i \in I \setminus I_m} \lambda_i = \prod_{i \in I \setminus I_m : \lambda_i \neq 0} \lambda_i \in \mathbf{R}^*$, and so $\det A \neq 0$.

Conversely, if $\det A \neq 0$, from Corollary 3.8, σ is bijective, and so

$$\det A^{(m,m)} \prod_{i \in I \setminus I_m} \lambda_i = \det A \neq 0;$$

then, $A^{(m,m)}$ is invertible and $\lambda_i \neq 0$, $\forall\, i \in I \setminus I_m$.

2. If $a_{ij} = 0$, $\forall\, i \in I_m$, $\forall\, j \in J \setminus J_m$, and $\det A \neq 0$, from point 1 and Proposition 2.27, we obtain that A is bijective.

3. The statement follows from Proposition 10 and Remark 14 in [3].

$\square$

PROPOSITION 3.10. *Let $A = (a_{ij})_{i \in I, j \in J} : E_J \longrightarrow E_I$ be a linear (m, σ)-general function such that $\left\{ h \in \mathbf{N},\, h \geq m : \sigma|_{I \setminus I_h} \text{ is injective} \right\} \neq \emptyset$; then,* $\det A = \det {}^t A$.

Proof. Since $\left\{ h \in \mathbf{N},\, h \geq m : \sigma|_{I \setminus I_h} \text{ is injective} \right\} \neq \emptyset$, from Proposition 3.4, the function ${}^t A : E_I \longrightarrow E_J$ is $\left(\tilde{\tilde{m}}, \tau \right)$-general, where $\tilde{\tilde{m}} \in \mathbf{N}^*$ is defined by formula (20), and the function $\tau : J \setminus J_{\tilde{\tilde{m}}} \longrightarrow I \setminus I_{\tilde{\tilde{m}}}$ is given by

$$\tau(j) = \min \left\{ \ \sigma^{-1}(k) : k \geq j,\, k \in \sigma \left(I \setminus I_{\tilde{\tilde{m}}} \right) \ \right\}, \forall\, j \in J \setminus J_{\tilde{\tilde{m}}}.$$

Then, we have

$$\begin{aligned} \det A &= \lim_{n \longrightarrow +\infty} \det A^{(n,n)} \\ &= \lim_{n \longrightarrow +\infty} \det {}^t \left(A^{(n,n)} \right) = \lim_{n \longrightarrow +\infty} \det \left({}^t A \right)^{(n,n)} = \det {}^t A. \end{aligned}$$

$\square$

PROPOSITION 3.11. *Let $A = (a_{ij})_{i \in I, j \in J} : E_J \longrightarrow E_I$ be a linear (m, σ)-general function such that $\sum_{i \in I \setminus I_m} |a_{i,j}| < +\infty$, for any $j \in J \setminus J_m$; moreover, let $s, t \in \mathbf{N}^*$, $s < t$, let $p = \max\{t, m\}$ and let $i_t \in I$ such that $|i_t| = t$; then:*

1. *If there exist* $u = (u_j : j \in J) \in E_J$, $v = (v_j : j \in J) \in E_J$, $c_1, c_2 \in \mathbf{R}$ *such that* $\sum_{j \in J} |u_j| < +\infty$, $\sum_{j \in J} |v_j| < +\infty$, $a_{i_t,j} = c_1 u_j + c_2 v_j$, *for any* $j \in J$, *by indicating by* $U = (u_{ij})_{i \in I, j \in J}$ *and* $V = (v_{ij})_{i \in I, j \in J}$ *the linear functions obtained by substituting the* t*-th row of* A *for* u *and* v*, respectively, then* U *and* V *are* (p, ξ)*-general, where the increasing function* $\xi : I \backslash I_p \longrightarrow J \backslash J_p$ *is defined by*

$$\xi(i) = \begin{cases} \sigma(i) & \text{if } \sigma(i) \in J \backslash J_p \\ \min(J \backslash J_p) & \text{if } \sigma(i) \notin J \backslash J_p \end{cases}, \forall\, i \in I \backslash I_p; \tag{40}$$

moreover, one has $\det A = c_1 \det U + c_2 \det V$.

2. *If* $B = (b_{ij})_{i \in I, j \in J} : E_J \longrightarrow E_I$ *is the linear function obtained by exchanging the* s*-th row of* A *for the* t*-th row of* A*, then* B *is* (p, ξ)*-general and one has* $\det B = -\det A$.

3. *If* $C = (c_{ij})_{i \in I, j \in J} : E_J \longrightarrow E_I$ *is the linear function obtained by substituting the* t*-th row of* A *for the* s*-th row of* A*, or the* s*-th one for the* t*-th one, then* C *is* (p, ξ)*-general and one has* $\det C = 0$.

Proof.

1. Since $\sum_{i \in I \backslash I_m} |a_{i,j}| < +\infty$, $\forall\, j \in J \backslash J_m$, we have $\sum_{i \in I \backslash I_m} |u_{ij}| < +\infty$, $\sum_{i \in I \backslash I_m} |v_{ij}| < +\infty$, $\forall\, j \in J \backslash J_m$; then, from point 5 of Proposition 2.24, the functions U and V are (p, ξ)-general. Moreover, $\forall\, n \in \mathbf{N}^*$, we have $\det A^{(n,n)} = c_1 \det U^{(n,n)} + c_2 \det V^{(n,n)}$, from which

$$\det A = \lim_{n \longrightarrow +\infty} \det A^{(n,n)} = \lim_{n \longrightarrow +\infty} \left(c_1 \det U^{(n,n)} + c_2 \det V^{(n,n)} \right)$$
$$= c_1 \det U + c_2 \det V.$$

2. By proceeding as in the proof of point 1, we can prove that B is (p, ξ)-general; moreover, $\forall\, n \in \mathbf{N}$, $n \geq p$, $B^{(n,n)}$ is the matrix obtained by exchanging the s-th row of $A^{(n,n)}$ for the t-th row of $A^{(n,n)}$; then, one has $\det B^{(n,n)} = -\det A^{(n,n)}$, from which

$$\det B = \lim_{n \longrightarrow +\infty} \det B^{(n,n)} = -\lim_{n \longrightarrow +\infty} \det A^{(n,n)} = -\det A.$$

3. By proceeding as in the proof of point 1, we can prove that C is (p, ξ)-general; moreover, since the s-th row of C and the t-th row of C are equals, by exchanging these rows among themselves we obtain again the matrix C; then, from point 2, we have $\det C = -\det C$, from which $\det C = 0$.

□

PROPOSITION 3.12. *Let* $A = (a_{ij})_{i\in I, j\in J} : E_J \longrightarrow E_I$ *be a linear* (m,σ)*-general function such that* $\sum_{i\in I\setminus I_m} |a_{i,j}| < +\infty$*, for any* $j \in J\setminus J_m$*; moreover, let* $s,t \in \mathbf{N}^*$*,* $s < t$*, let* $p = \max\{t,m\}$*, let* $j_t \in J$ *such that* $|j_t| = t$*, and let the function* $\xi : I\setminus I_p \longrightarrow J\setminus J_p$ *defined by (40); then:*

1. *If there exist* $u = (u_i : i \in I) \in E_I$*,* $v = (v_i : i \in I) \in E_I$*,* $c_1, c_2 \in \mathbf{R}$ *such that* $\sum_{i\in I} |u_i| < +\infty$*,* $\sum_{i\in I} |v_i| < +\infty$*,* $a_{i,j_t} = c_1 u_i + c_2 v_i$*, for any* $i \in I$*, by indicating by* $U = (u_{ij})_{i\in I, j\in J}$ *and* $V = (v_{ij})_{i\in I, j\in J}$ *the linear functions obtained by substituting the t-th column of A for u and v, respectively, then U and V are* (p,ξ)*-general and one has* $\det A = c_1 \det U + c_2 \det V$.

2. *If* $B = (b_{ij})_{i\in I, j\in J} : E_J \longrightarrow E_I$ *is the linear function obtained by exchanging the s-th column of A for the t-th column of A, then B is* (p,ξ)*-general and one has* $\det B = -\det A$.

3. *If* $C = (c_{ij})_{i\in I, j\in J} : E_J \longrightarrow E_I$ *is the linear function obtained by substituting the t-th column of A for the s-th column of A, or the s-th one for the t-th one, then C is* (p,ξ)*-general and one has* $\det C = 0$.

Proof. The proof is analogous to that one of Proposition 3.11. □

PROPOSITION 3.13. *Let* $A = (a_{ij})_{i\in I, j\in J} : E_J \longrightarrow E_I$ *be a linear* (m,σ)*-general function such that* $\sum_{i\in I\setminus I_m} |a_{i,j}| < +\infty$*, for any* $j \in J\setminus J_m$*. If the dimension of the vector space generated by the rows or the columns of A is finite, then* $\det A = 0$.

Proof. Suppose that the dimension of the vector space generated by the rows of A is finite; then, there exist n rows $v^{(1)}$,...,$v^{(n)}$ of A, where $v^{(k)} = \left(v_j^{(k)} : j \in J\right)$, $\forall\, k \in \{1, ..., n\}$, such that, if $v = (v_j : j \in J)$ is as row of A, there exist c_1,...,$c_n \in \mathbf{R}$ such that $v = c_1 v^{(1)} + ... + c_n v^{(n)}$. From Proposition 3.11, by indicating by V_k, $\forall\, k \in \{1, ..., n\}$, the linear function obtained by substituting the row v of A for $v^{(k)}$, by recursion we have $\det A = c_1 \det V_1 + ... + c_n \det V_n$; moreover, V_k has two rows equals to $v^{(k)}$, and so $\det V_k = 0$, $\forall\, k \in \{1, ..., n\}$; then, $\det A = 0$. Analogously, if the dimension of the vector space generated by the columns of A is finite, from Proposition 3.12 we obtain $\det A = 0$. □

Remark 3.14: Let $A = (a_{ij})_{i\in I, j\in J} : E_J \longrightarrow E_I$ be a linear (m,σ)-general function such that $\sum_{i\in I\setminus I_m} |a_{i,j}| < +\infty$, for any $j \in J\setminus J_m$. Then, for any $n \in \mathbf{N}$, $n \geq m$, for any $\emptyset \neq L \subset I$ and for any $\emptyset \neq N \subset J$ such that $|I\setminus L| = |J\setminus N| < +\infty$, the linear function $A^{(L,N)} : E_N \longrightarrow E_L$ is (n,ρ)-general, where the function $\rho : L\setminus L_n \longrightarrow N\setminus N_n$ is defined by

$$\rho(i) = \begin{cases} \sigma(i) & \text{if } \sigma(i) \in N\setminus N_n \\ \min\{j > \sigma(i) : j \in N\setminus N_n\} & \text{if } \sigma(i) \notin N\setminus N_n \end{cases}, \forall\, i \in L\setminus L_n.$$

Proof. The proof follows from Remark 2.25. □

Definition 3.15. *Let $A = (a_{ij})_{i\in I, j\in J} : E_J \longrightarrow E_I$ be a linear (m,σ)-general function such that $\sum_{i\in I\setminus I_m} |a_{i,j}| < +\infty$, for any $j \in J\setminus J_m$; define the $I \times J$ matrix $cof A$ by*

$$(cof A)_{ij} = (-1)^{|i|+|j|} \det \left(A^{(I\setminus\{i\}, J\setminus\{j\})}\right), \forall\, i \in I, \forall\, j \in J.$$

Proposition 3.16. *Let $A = (a_{ij})_{i\in I, j\in J} : E_J \longrightarrow E_I$ be a linear (m,σ)-general function such that $\sum_{i\in I\setminus I_m} |a_{i,j}| < +\infty$, for any $j \in J\setminus J_m$; moreover, suppose that $a_{ij} = 0$, $\forall\, i \in I_m$, $\forall\, j \in J\setminus J_m$, or A is (m,σ)-standard; then, one has:*

$$\det A = \sum_{t\in J} a_{it} (cof A)_{it}, \forall\, i \in I; \tag{41}$$

$$\det A = \sum_{s\in I} a_{sj} (cof A)_{sj}, \forall\, j \in J. \tag{42}$$

Proof. Suppose that $\mathcal{A} \neq \emptyset$ and set $\overline{m} = \min \mathcal{A}$; $\forall\, i \in I$, $\forall\, j \in J$ and $\forall\, n \in \mathbf{N}$, $n \geq \max\{|i|, |j|, \overline{m}\}$, we have

$$\det A = \det A^{(n,n)} \prod_{p\in I\setminus I_n} \lambda_p, \tag{43}$$

from which

$$\det A = \sum_{t\in J_n} a_{it} (cof A^{(n,n)})_{it} \prod_{p\in I\setminus I_n} \lambda_p = \sum_{t\in J_n} a_{it} (cof A)_{it};$$

then

$$\det A = \lim_{n\longrightarrow+\infty} \sum_{t\in J_n} a_{it} (cof A)_{it} = \sum_{t\in J} a_{it} (cof A)_{it}.$$

Analogously, from formula (43), we have

$$\det A = \sum_{s\in I_n} a_{sj}(cofA^{(n,n)})_{sj} \prod_{p\in I\setminus I_n} \lambda_p = \sum_{s\in I_n} a_{sj}(cofA)_{sj},$$

and so

$$\det A = \sum_{s\in I} a_{sj}(cofA)_{sj}.$$

Conversely, if $\mathcal{A} = \emptyset$, $\forall\, s \in I$, $\forall\, t \in J$, we have $\mathcal{A}\left(A^{(I\setminus\{s\},J\setminus\{t\})}\right) = \emptyset$; then, from Theorem 3.6, we obtain $\det A = \det\left(A^{(I\setminus\{s\},I\setminus\{t\})}\right) = 0$, and so $(cofA)_{st} = 0$; then:

$$\det A = 0 = \sum_{t\in J} a_{it}(cofA)_{it},\ \forall\, i \in I;$$
$$\det A = 0 = \sum_{s\in I} a_{sj}(cofA)_{sj},\ \forall\, j \in J.$$

□

Corollary 3.17. *Let $A = (a_{ij})_{i\in I, j\in J} : E_J \longrightarrow E_I$ be a linear (m,σ)-general function such that $\sum_{i\in I\setminus I_m} |a_{i,j}| < +\infty$, for any $j \in J\setminus J_m$; moreover, suppose that $a_{ij} = 0$, $\forall\, i \in I_m$, $\forall\, j \in J\setminus J_m$, or A is (m,σ)-standard; then:*

1. *One has*
$$A\ {}^t\,(cofA) = (\det A)\mathbf{I}_{I,I}; \tag{44}$$
moreover, if A is bijective, the linear functions $A^{-1} : E_I \longrightarrow E_J$ and ${}^t\,(cofA) : E_I \longrightarrow E_J$ are continuous.

2. *If A is bijective, then one has $\det A \neq 0$ if and only if $cofA \neq 0$; moreover, in this case*
$$A^{-1} = \frac{1}{\det A}\ {}^t\,(cofA)\ . \tag{45}$$

3. *If A is (m,σ)-standard and bijective, then A^{-1} is $\left(m,\sigma^{-1}\right)$-standard.*

Proof.

1. From formula (41), we have

$$\sum_{t\in J} a_{it}(cofA)_{it} = \det A,\ \forall\, i \in I.$$

Moreover, we have

$$\sum_{t\in J} a_{it}(cofA)_{jt} = 0,\ \forall\, i,j \in I,\ i \neq j; \tag{46}$$

in fact, from formula (41) and Proposition 3.11, the left side of (46) is equal to $\det C$, where C is the (p,ξ)-general function obtained by substituting the j-th row of A for the i-th row of A, $p = \max\{|i|,|j|,m\}$, and the increasing function $\xi : I\backslash I_p \longrightarrow J\backslash J_p$ is defined by (40); then, from Proposition 3.11, we have $\det C = 0$. This implies that

$$\sum_{t\in J} a_{it}(cofA)_{jt} = (\det A)\delta_{ij},\ \forall\, i,j \in I,$$

where δ_{ij} is the Kronecker symbol, and so formula (44) follows, since the functions δ_{ij} and $\overline{\delta}_{ij}$ coincide on $I \times I$. Moreover, suppose that A is bijective; since A is continuous from Proposition 3.4, then the linear function $A^{-1} : E_I \longrightarrow E_J$ is continuous (see, e.g., the theory in Weidmann's book [11]); furthermore, from formula (44), we have

$$^{t}\,(cofA) = (\det A)A^{-1},$$

and so the linear function $^{t}\,(cofA) : E_I \longrightarrow E_J$ is continuous too.

2. If A is bijective, from formula (44) we have $\det A = 0$ if and only if $cofA = 0$, and so $\det A \neq 0$ if and only if $cofA \neq 0$; moreover, in this case, from formula (44) we obtain formula (45).

3. If A is (m,σ)-standard and bijective, from Proposition 3.9, we have $\det A \neq 0$, $\lambda_i \neq 0$, $\forall\, i \in I\backslash I_m$, and σ is bijective; moreover, $\forall\, y \in E_I$, we have $A\left(A^{-1}y\right) = y$, from which

$$\left(A^{-1}y\right)_i = \frac{y_i}{\lambda_i},\ \forall\, i \in I\backslash I_m; \tag{47}$$

furthermore, we have $\left\{i \in I\backslash I_m : (\lambda_i)^{-1} \neq 0\right\} = I\backslash I_m$, from which

$$\prod_{i\in I\backslash I_m:(\lambda_i)^{-1}\neq 0} (\lambda_i)^{-1} = \left(\prod_{i\in I\backslash I_m} \lambda_i\right)^{-1} = \left(\prod_{i\in I\backslash I_m:\lambda_i\neq 0} \lambda_i\right)^{-1} \in \mathbf{R}^*;$$

then, we obtain $\sup\limits_{i\in I\backslash I_m} \left|(\lambda_i)^{-1}\right| < +\infty$ and $\inf\limits_{i\in I\backslash I_m:(\lambda_i)^{-1}\neq 0} \left|(\lambda_i)^{-1}\right| > 0$. Finally, from formula (47) and since the linear function $A^{-1} : E_I \longrightarrow E_J$ is given by formula (45), then A^{-1} is $\left(m,\sigma^{-1}\right)$-standard, with $\lambda_i\left(A^{-1}\right) = (\lambda_i)^{-1}$, $\forall\, i \in I\backslash I_m$.

$\square$

PROPOSITION 3.18. *Let* $\varphi : U \subset E_J \longrightarrow E_I$ *be a* (m,σ)*-general function and let* $x_0 = (x_{0,j} : j \in J) \in U$ *such that there exists the function* $J_\varphi(x_0) : E_J \longrightarrow E_I$*; then,* $J_\varphi(x_0)$ *is a linear* (m,σ)*-general function; moreover, for any* $n \in \mathbf{N}$*,* $n \geq m$*, there exists the linear* (m,σ)*-general function* $J_{\overline{\varphi}^{(n,n)}}(x_0) : E_J \longrightarrow E_I$*, and one has*

$$\det J_\varphi(x_0) = \lim_{n\to+\infty} \det J_{\overline{\varphi}^{(n,n)}}(x_0).$$

Proof. Since φ is (m,σ)-general, from Remark 3.2, the linear function $J_\varphi(x_0)$ is (m,σ)-general; moreover, $\forall\, n \in \mathbf{N}$, $n \geq m$, from Proposition 2.4, there exists the linear function $J_{\overline{\varphi}^{(n,n)}}(x_0) : E_J \longrightarrow E_I$, and it is (m,σ)-general, from Remark 3.2; furthermore, we have $\mathcal{A}(J_\varphi(x_0)) = \mathcal{A}\left(J_{\overline{\varphi}^{(n,n)}}(x_0)\right)$.

If $\mathcal{A}(J_\varphi(x_0)) \neq \emptyset$, set $\overline{m} = \min \mathcal{A}(J_\varphi(x_0))$; $\forall\, n \geq \overline{m}$, we have

$$\det J_{\overline{\varphi}^{(n,n)}}(x_0) = \det J_{\varphi^{(n,n)}}(x_{0,j} : j \in J_n) \prod_{i \in I \setminus I_n} \varphi'_{i,\sigma(i)}\left(x_{0,\sigma(i)}\right); \tag{48}$$

if $|(I\setminus I_m)\setminus \mathcal{I}_\varphi| < +\infty$, set $i_0 = \max((I\setminus I_m)\setminus \mathcal{I}_\varphi)$ and $\widehat{m} = \max\{\overline{m}, |i_0|\}$; since $\prod_{i\in I\setminus I_{\widehat{m}}} \varphi'_{i,\sigma(i)}\left(x_{0,\sigma(i)}\right) \in \mathbf{R}^*$, we have $\lim_{n\to+\infty} \prod_{i\in I\setminus I_n} \varphi'_{i,\sigma(i)}\left(x_{0,\sigma(i)}\right) = 1$; then, from (48) and Theorem 3.6, we obtain

$$\lim_{n\to+\infty} \det J_{\overline{\varphi}^{(n,n)}}(x_0) = \lim_{n\to+\infty} \det J_{\varphi^{(n,n)}}(x_{0,j} : j \in J_n) = \det J_\varphi(x_0);$$

conversely, suppose that $|(I\setminus I_m)\setminus \mathcal{I}_\varphi| = +\infty$; for n sufficiently large, we have $\det J_{\varphi^{(n,n)}}(x_{0,j} : j \in J_n) = 0$, from which

$$\begin{aligned}
\det J_\varphi(x_0) &= \lim_{n\to+\infty} \det J_{\varphi^{(n,n)}}(x_{0,j} : j \in J_n) = 0 \\
&= \lim_{n\to+\infty} \det J_{\varphi^{(n,n)}}(x_{0,j} : j \in J_n) \prod_{i\in I\setminus I_n} \varphi'_{i,\sigma(i)}\left(x_{0,\sigma(i)}\right) \\
&= \lim_{n\to+\infty} \det J_{\overline{\varphi}^{(n,n)}}(x_0).
\end{aligned}$$

Moreover, if $\mathcal{A}(J_\varphi(x_0)) = \emptyset$, $\forall\, n \in \mathbf{N}$, $n \geq m$, we have $\mathcal{A}\left(J_{\overline{\varphi}^{(n,n)}}(x_0)\right) = \emptyset$, and so

$$\det J_\varphi(x_0) = 0 = \lim_{n\to+\infty} \det J_{\overline{\varphi}^{(n,n)}}(x_0).$$

$\square$

EXAMPLE 3.19: Consider the linear function $A = (a_{ij})_{i,j\in\mathbf{N}^*} : E_{\mathbf{N}^*} \longrightarrow E_{\mathbf{N}^*}$ given by

$$(Ax)_i = \begin{cases} \sum\limits_{j\in\mathbf{N}^*} 2^{-j}x_j & \text{if } i = 1 \\ x_1 + \sum\limits_{j\in\mathbf{N}^*} 2^{-j}x_j & \text{if } i = 2 \\ 2^{-i}x_1 + 2^{2^{-i}} & \text{if } i \in \mathbf{N}^*\backslash\{1,2\} \end{cases}, \forall\, x = (x_j : j \in \mathbf{N}^*) \in E_{\mathbf{N}^*}.$$

Then, A is a strongly (m,σ)-general function, where $I = J = \mathbf{N}^*$, $m = 2$, $I_m = J_m = \{1,2\}$, σ is the function given by $\sigma(i) = i$, $\forall\, i \in \mathbf{N}^*\backslash\{1,2\}$, and $\mathcal{A} = \mathbf{N}^*\backslash\{1\} \neq \emptyset$; moreover, we have $\lambda_i = 2^{2^{-i}}$, $\forall\, i \in \mathbf{N}^*\backslash\{1,2\}$.

In order to calculate $\det A$, observe that $A^{(\{2\},\mathbf{N}^*)} = u + v$, where $u = A^{(\{1\},\mathbf{N}^*)} \in E_{\mathbf{N}^*}$, and $v = (v_j : j \in \mathbf{N}^*) \in E_{\mathbf{N}^*}$, where $v_j = \delta_{j1}$, $\forall\, j \in \mathbf{N}^*$. Then, from Proposition 3.11, we have $\det A = \det U + \det V$, where $U = (u_{ij})_{i,j\in\mathbf{N}^*}$ and $V = (v_{ij})_{i,j\in\mathbf{N}^*}$ are the linear functions obtained by substituting the second row of A by u and v, respectively; moreover, since $U^{(\{1\},\mathbf{N}^*)} = U^{(\{2\},\mathbf{N}^*)}$, we have $\det U = 0$, from which

$$\det A = \det V = \lim_{n\longrightarrow+\infty} \det V^{(n,n)}. \tag{49}$$

Finally, $\forall\, n \in \mathbf{N}^*\backslash\{1,2\}$, we have

$$\begin{aligned} \det V^{(n,n)} &= (-1)^{n+1}2^{-n} \det V^{(n-1,\{2,\ldots,n\})} + 2^{2^{-n}} \det V^{(n-1,n-1)} \\ &= 2^{2^{-n}} \det V^{(n-1,n-1)}, \end{aligned} \tag{50}$$

since the second row of $V^{(n-1,\{2,\ldots,n\})}$ is zero, and so $\det V^{(n-1,\{2,\ldots,n\})} = 0$. Then, by recursion, from (50) we obtain

$$\det V^{(n,n)} = \det V^{(2,2)} \prod_{j=3}^{n} 2^{2^{-n}},$$

and so formula (49) implies

$$\det A = \lim_{n\longrightarrow+\infty} \det V^{(2,2)} \prod_{j=3}^{n} 2^{2^{-n}} = \det V^{(2,2)} 2^{\sum\limits_{j=3}^{+\infty} 2^{-n}} = -\frac{1}{4}\sqrt[4]{2}.$$

4. Problems for further study

A natural extension of this paper and of the paper [4] is the generalization of the change of variables' formula for the integration of the measurable real functions on $\left(\mathbf{R}^I, \mathcal{B}^{(I)}\right)$, by substituting the (m,σ)-standard functions for the (m,σ)-general functions.

Moreover, a natural application of this paper, in the probabilistic framework, is the development of the theory of the infinite-dimensional continuous random elements, defined in the paper [3]. In particular, we can prove the formula of the density of such random elements composed with the (m, σ)-general functions, with further properties. Consequently, it is possible to introduce many random elements that generalize the well known continuous random vectors in $\mathbf{R}^m$ (for example, the Beta random elements in E_I defined by the (m, σ)-general matrices), and to develop some theoretical results and some applications in the statistical inference. It is possible also to define a convolution between the laws of two independent and infinite-dimensional continuous random elements, as in the finite case.

Furthermore, we can generalize the paper [2] by considering the recursion $\{X_n\}_{n \in \mathbf{N}}$ on $[0, p)^{\mathbf{N}^*}$ defined by

$$X_{n+1} = AX_n + B_n \pmod p,$$

where $X_0 = x_0 \in E_I$, A is a bijective, linear, integer and (m, σ)-general function, $p \in \mathbf{R}^+$, and $\{B_n\}_{n \in \mathbf{N}}$ is a sequence of independent and identically distributed random elements on E_I. Our target is to prove that, with some assumptions on the law of B_n, the sequence $\{X_n\}_{n \in \mathbf{N}}$ converges with geometric rate to a random element with law $\bigotimes_{i \in \mathbf{N}^*} \left(\frac{1}{p} Leb\Big|_{\mathcal{B}([0,p))} \right)$. Moreover, we wish to quantify the rate of convergence in terms of A, p, m, and the law of B_n.

Finally, in the statistical mechanics, it is possible to describe the systems of smooth hard particles, by using the Boltzmann equation or the more general Master kinetic equation, described for example in the paper [9]. In order to study the evolution of these systems, we can consider the model of countable particles, such that their joint infinite-dimensional density can be determined by composing a particular random element with a (m, σ)-general function.

References

[1] L. Accardi, O. G. Smolyanov, and M. O. Smolyanova, *Change of variable formulas for infinite-dimensional distributions*, Mathematical notes **60** (1996), 212–215.

[2] C. Asci, *Convergence in total variation of an affine random recursion in* $[0, p)^k$ *to a uniform random vector*, Markov Processes and Related Fields **19** (2013), 125–140.

[3] C. Asci, *Integration over an infinite-dimensional Banach space and probabilistic applications*, International Journal of Analysis. doi: org/10.1155/2014/404186 (2014).

[4] C. Asci, *Differentiation theory over infinite-dimensional Banach spaces*, Journal of Mathematics. doi: org/10.1155/2016/2619087 (2016).

[5] R. Baker, *Lebesgue measure" on* $\mathbf{R}^\infty$, Proc. Amer. Math. Soc. **113** (1991), 1023–1029.
[6] S. Lang, *Analysis II*, Addison–Wesley, Reading, Massachusetts, 1969.
[7] S. Lang, *Linear algebra*, Springer, New York, 1987.
[8] R. Léandre, *Lebesgue measure in infinite dimension as an infinite-dimensional distribution*, Journal of Mathematical Sciences **159** (2009), 833–836.
[9] M. Tessarotto, C. Cremaschini, C. Asci, A. Soranzo, and G. Tironi, *Global validity of the Master kinetic equation for hard-sphere systems*, Eur. Phys. J. Plus. doi: org/10.1140/epjp/i2015-15169-2 **130** (2015).
[10] N. Tsilevich, A. Vershic, and M. Yor, *An infinite-dimensional analogue of the Lebesgue measure and distinguished properties of the Gamma process*, Journal of Functional Analysis **185** (2001), 274–296.
[11] J. Weidmann, *Linear operators in Hilbert spaces*, Springer, New York, 1980.

Author's address:

Claudio Asci
Dipartimento di Matematica e Geoscienze
Università degli Studi di Trieste
Via Valerio 12/1
34127 Trieste, Italy
E-mail: casci@units.it

Received August 01, 2017
Revised November 15, 2017
Accepted November 27, 2017

Rend. Istit. Mat. Univ. Trieste
Volume 50 (2018), 37–46
DOI: 10.13137/2464-8728/20578

On graded classical 2-absorbing submodules of graded modules over graded commutative rings

Khaldoun Al-Zoubi and Mariam Al-Azaizeh

Abstract. *Let G be a group with identity e. Let R be a G-graded commutative ring and M a graded R-module. In this paper, we will introduce the concept of graded classical 2-absorbing submodules of graded modules over a graded commutative ring as a generalization of graded classical prime submodules and investigate some basic properties of these classes of graded modules.*

Keywords: graded 2-absorbing submodule, graded classical prime submodule, graded classical 2-absorbing submodule.
MS Classification 2010: 13A02, 16W50.

1. Introduction and Preliminaries

Throughout this paper all rings are commutative with identity and all modules are unitary. Badawi in [8] introduced the concept of 2-absorbing ideals of commutative rings. We recall from [8] that a proper ideal I of R is called *a 2-absorbing ideal of* R if whenever $r, s, t \in R$ and $rst \in I$ implies $rs \in I$ or $rt \in I$ or $st \in I$. Later on, Anderson and Badawi in [7] generalized the concept of 2-absorbing ideals of commutative rings to the concept of n-absorbing ideals of commutative rings for every positive integer $n \geq 2$. We recall from [7] that a proper ideal I of R is called an n-absorbing ideal if whenever $x_1 \cdots x_{n+1} \in I$ for $x_1, \ldots, x_{n+1} \in R$, then there are n of the x_i's whose product is in I. In light of [8] and [7], many authors studied the concept of 2-absorbing submodules and n-absorbing submodules. Recently, H. Mostafanasab, U. Tekir and K.H. Oral in [12] studied classical 2-absorbing submodules of modules over commutative rings. Let M be an R-module. A proper submodule N of M is called classical 2-absorbing submodule, if whenever a, b, $c \in R$ and $m \in M$ with $abcm \in N$, then $abm \in N$ or $acm \in N$ $bcm \in N$.

The scope of this paper is devoted to the theory of graded modules over graded commutative rings. One use of rings and modules with gradings is in describing certain topics in algebraic geometry. Here, in particular, we are dealing with graded classical 2-absorbing submodules of graded modules over graded commutative rings. The notion of graded 2-absorbing ideals as a generalization of graded prime ideals was introduced and studied in [3, 13]. The notion of graded 2-absorbing ideals was extended to graded 2-absorbing submodules in [2, 11]. The notion of graded classical prime submodules as a generalization of graded prime submodules was introduced in [9] and studied in [1, 4, 5]. The purpose of this paper is to introduced the concept of graded classical 2-absorbing submodules as a generalization of graded classical prime submodules and give a number of its properties (see sec. 2).

First, we recall some basic properties of graded rings and modules which will be used in the sequel. We refer to [10, 14, 15, 16] for these basic properties and more information on graded rings and modules.

Let G be a group with identity e and R be a commutative ring with identity 1_R. Then R is a *G-graded ring* if there exist additive subgroups R_g of R such that $R = \bigoplus_{g\in G} R_g$ and $R_gR_h \subseteq R_{gh}$ for all $g, h \in G$. The elements of R_g are called to be *homogeneous* of degree g where the R_g's are additive subgroups of R indexed by the elements $g \in G$. If $x \in R$, then x can be written uniquely as $\sum_{g\in G} x_g$, where x_g is the component of x in R_g. Moreover, $h(R) = \bigcup_{g\in G} R_g$. Let I be an ideal of R. Then I is called a *graded ideal* of (R, G) if $I = \bigoplus_{g\in G}(I \bigcap R_g)$. Thus, if $x \in I$, then $x = \sum_{g\in G} x_g$ with $x_g \in I$. An ideal of a G-graded ring need not be G-graded.

Let R be a G-graded ring and M an R-module. We say that M is a *G-graded R-module* (or *graded R-module*) if there exists a family of subgroups $\{M_g\}_{g\in G}$ of M such that $M = \bigoplus\limits_{g\in G} M_g$ (as abelian groups) and $R_gM_h \subseteq M_{gh}$ for all $g, h \in G$. Here, R_gM_h denotes the additive subgroup of M consisting of all finite sums of elements r_gs_h with $r_g \in R_g$ and $s_h \in M_h$. Also, we write $h(M) = \bigcup\limits_{g\in G} M_g$ and the elements of $h(M)$ are called to be *homogeneous.* Let $M = \bigoplus\limits_{g\in G} M_g$ be a graded R-module and N a submodule of M. Then N is called a *graded submodule* of M if $N = \bigoplus\limits_{g\in G} N_g$ where $N_g = N \cap M_g$ for $g \in G$. In this case, N_g is called the *g-component* of N. Moreover, M/N becomes a G-graded R-module with g-component $(M/N)_g = (M_g + N)/N$ for $g \in G$.

Let R be a G-graded ring and $S \subseteq h(R)$ be a multiplicatively closed subset of R. Then the ring of fraction $S^{-1}R$ is a graded ring which is called the graded ring of fractions. Indeed, $S^{-1}R = \underset{g\in G}{\oplus}(S^{-1}R)_g$ where $(S^{-1}R)_g = \{r/s : r \in R, s \in S \text{ } and \text{ } g = (\deg s)^{-1}(\deg r)\}$. Let M be a graded module over a G-graded ring R and $S \subseteq h(R)$ be a multiplicatively closed subset of R. The module of fraction $S^{-1}M$ over a graded ring $S^{-1}R$ is a graded module which is called module of fractions, if $S^{-1}M = \underset{g\in G}{\oplus}(S^{-1}M)_g$ where $(S^{-1}M)_g = \{m/s : m \in M, s \in S \text{ } and \text{ } g = (\deg s)^{-1}(\deg m)\}$. We write $h(S^{-1}R) = \underset{g\in G}{\cup}(S^{-1}R)_g$ and $h(S^{-1}M) = \underset{g\in G}{\cup}(S^{-1}M)_g$. Consider the graded homomorphism $\eta : M \to$

$S^{-1}M$ defined by $\eta(m) = m/1$. For any graded submodule N of M, the submodule of $S^{-1}M$ generated by $\eta(N)$ is denoted by $S^{-1}N$. Similar to non graded case, one can prove that $S^{-1}N = \{\beta \in S^{-1}M : \beta = m/s \text{ for } m \in N \text{ and } s \in S\}$ and that $S^{-1}N \neq S^{-1}M$ if and only if $S \cap (N :_R M) = \phi$. If K is a graded submodule of $S^{-1}R$-module $S^{-1}M$, then $K \cap M$ will denote the graded submodule $\eta^{-1}(K)$ of M. Moreover, similar to the non graded case one can prove that $S^{-1}(K \cap M) = K$.

Let R be a G-graded ring and M a graded R-module.

A proper graded ideal P of R is said to be *a graded prime ideal* if whenever $rs \in P$, we have $r \in P$ or $s \in P$, where $r, s \in h(R)$ (see [18].) It is shown in [6, Lemma 2.1] that if N is a graded submodule of M, then $(N :_R M) = \{r \in R : rN \subseteq M\}$ is a graded ideal of R.

A proper graded submodule P of M is said to be *a graded prime submodule* if whenever $r \in h(R)$ and $m \in h(M)$ with $rm \in P$, then either $r \in (P :_R M)$ or $m \in P$ (see [6, 17].)

A proper graded ideal I of R is said to be *a graded 2-absorbing ideal of* R if whenever $r, s, t \in h(R)$ with $rst \in I$, then $rs \in I$ or $rt \in I$ or $st \in I$ (see [3, 13].)

A proper graded submodule N of M is called *a graded 2-absorbing submodule of* M if whenever $r, s \in h(R)$ and $m \in h(M)$ with $rsm \in N$, then either $rs \in (N :_R M)$ or $rm \in N$ or $sm \in N$ (see [2].)

A proper graded submodule N of M is called *a graded classical prime* submodule if whenever $r, s \in h(R)$ and $m \in h(M)$ with $rsm \in N$, then either $rm \in N$ or $sm \in N$ (see [4, 9].)

2. Results

Definition 2.1. *Let R be a G-graded ring, M a graded R-module, C a graded submodule of M and let $g \in G$.*

(i) *We say that C_g is a classical g-2-absorbing submodule of R_e-module M_g if $C_g \neq M_g$; and whenever $r, s, t \in R_e$ and $m \in M_g$ with $rstm \in C_g$, then either $rsm \in C_g$ or $rtm \in C_g$ or $stm \in C_g$.*

(ii) *We say that C is a graded classical 2-absorbing submodule of M if $C \neq M$; and whenever $r, s, t \in h(R)$ and $m \in h(M)$ with $rsm \in C$, then either $rsm \in C$ or $rtm \in C$ or $stm \in C$.*

Theorem 2.2. *Let R be a G-graded ring, M a graded R-module and C a graded submodule of M. If C is a graded classical 2-absorbing submodule of M, then C_g is a classical g-2-absorbing R_e-submodule of M_g for every $g \in G$.*

Proof. Suppose that C is a graded classical 2-absorbing submodule of M. For $g \in G$ assume that $rstm \in C_g \subseteq C$ where $r, s, t \in R_e$ and $m \in M_g$. Since C

is a graded classical 2-absorbing submodule of M, we have either $rsm \in C$ or $rtm \in C$ or $stm \in C$. Since $M_g \subseteq M$ and $C_g = C \cap M_g$, we conclude that either $rsm \in C_g$ or $rtm \in C_g$ or $stm \in C_g$. So C_g is classical g-2-absorbing R_e-submodule of M_g. □

Theorem 2.3. *Let R be a G-graded ring, M a graded R-module and C a proper graded submodule of M. Then the following statements hold:*

(i) *If C is a graded 2-absorbing submodule of M, then C is a graded classical 2-absorbing submodule of M.*

(ii) *C is a graded classical prime submodule of M if and only if C is a graded 2-absorbing submodule of M and $(C :_R M)$ is a graded prime ideal of R.*

Proof. (i) Assume that C is a graded 2-absorbing submodule of M. Let r, s, $t \in h(R)$ and $m \in h(M)$ such that $rstm \in C, rtm \notin C$ and $stm \notin C$. Since C is a graded 2-absorbing submodule of M, we conclude that $rs \in (C :_R M)$ and hence $rsm \in C$. Thus C is a graded classical 2-absorbing submodule of M.

(ii) Assume that C is a graded classical prime submodule of M. It is clear that C is a graded 2-absorbing submodule of M. Also by [4, Lemma 3.1.], $(C :_R M)$ is a graded prime ideal of R. Conversely, assume that C is a graded 2-absorbing submodule of M and $(C :_R M)$ is a graded prime ideal of R. Let r, $s \in h(R)$ and $m \in h(M)$ such that $rsm \in C$, $rm \notin C$ and $sm \notin C$. Since C is a graded 2-absorbing submodule of M, $rs \in (C :_R M)$. It follows that either $r \in (C :_R M)$ or $s \in (C :_R M)$ and hence $rm \in C$ or $sm \in C$, which is a contradiction. Thus C is a graded classical prime submodule of M. □

The following example shows that the converse of theorem 2.3(i) is not true.

Example 2.4: Let $G = (\mathbb{Z}, +)$ and $R = (\mathbb{Z}, +, .)$. Define

$R_g = \left\{ \begin{array}{ll} \mathbb{Z} & \text{if } g = 0 \\ 0 & \text{otherwise} \end{array} \right\}$. Then R is a G-graded ring. Let $M = \mathbb{Z}_2 \times \mathbb{Z}_3 \times \mathbb{Q}$. Then M is a G-graded R-module with

$$M_g = \left\{ \begin{array}{ll} \{0\} \times \mathbb{Z}_3 \times \mathbb{Q} & \text{if } g = 0 \\ \mathbb{Z}_2 \times \{0\} \times \mathbb{Q} & \text{if } g = 1 \\ \mathbb{Z}_2 \times \mathbb{Z}_3 \times \{0\} & \text{if } g = 2 \\ \{0\} \times \{0\} \times \{0\} & \text{otherwise} \end{array} \right\}.$$

Now consider a graded submodule $C = \{(0, 0, 0)\}$. One can easily see that C is a graded classical 2-absorbing submodule of M. Since $2.3.(1, 1, 0) = (0, 0, 0)$, but $3.(1, 1, 0) \notin C$, $2.(1, 1, 0) \notin C$ and $2.3.(1, 1, 1) \notin C$, we get C is not a graded 2-absorbing submodule. Also, part (ii) of theorem 2.3(ii) shows that C is note a graded classical prime submodule. Hence the two concepts of graded classical prime submodules and of graded classical 2-absorbing submodules are different in general.

Recall that a graded zero-divisor on a graded R-module M is an element $r \in h(R)$ for which there exists $m \in h(M)$ such that $m \neq 0$ but $rm = 0$. The set of all graded zero-divisors on M is denoted by $G\text{-}Zdv_R(M)$ (see [2].)

The following result studies the behavior of graded 2-absorbing submodules under localization.

THEOREM 2.5. *Let R be a G-graded ring, M a graded R-module and $S \subseteq h(R)$ a multiplication closed subset of R. Then the following hold:*

(i) *If C is a graded classical 2-absorbing submodule of M such that $(C :_R M) \cap S = \phi$, then $S^{-1}C$ is a graded classical 2-absorbing submodule of $S^{-1}M$.*

(ii) *If $S^{-1}C$ is a graded classical 2-absorbing submodule of $S^{-1}M$ and $S \cap G\text{-}Zdv_R(M/C) = \phi$, then C is a graded classical 2-absorbing submodule of M.*

Proof. (i) Let C be a graded classical 2-absorbing submodule of M and $(C :_R M) \cap S = \phi$. Suppose that $\frac{r_1}{s_1}\frac{r_2}{s_2}\frac{r_3}{s_3}\frac{m}{s_4} \in S^{-1}C$ for some $\frac{r_1}{s_1}, \frac{r_2}{s_2}, \frac{r_3}{s_3} \in h(S^{-1}R)$ and for some $\frac{m}{s_4} \in h(\ S^{-1}M)$. Hence there exists $k \in S$ such that $r_1r_2r_3(km) \in C$. Since C is a graded classical 2-absorbing submodule of M, we conclude that either $r_1r_2(km) \in C$ or $r_1r_3(km) \in C$ or $r_2r_3(km) \in C$. Thus $\frac{r_1r_2(km)}{s_1s_2s_4k} = \frac{r_1}{s_1}\frac{r_2}{s_2}\frac{m}{s_4} \in S^{-1}C$ or $\frac{r_1r_3(km)}{s_1s_3s_4k} = \frac{r_1}{s_1}\frac{r_3}{s_3}\frac{m}{s_4} \in S^{-1}C$ or $\frac{r_2r_3(km)}{s_2s_3s_4k} = \frac{r_2}{s_2}\frac{r3}{s_3}\frac{m}{s_4} \in S^{-1}C$. Therefore $S^{-1}C$ is a graded classical 2-absorbing submodule of $S^{-1}M$.

(ii) Assume that $S^{-1}C$ is a graded classical 2-absorbing submodule of $S^{-1}M$ and $S \cap G\text{-}Zdv_R(M/C) = \phi$. Let $r_1r_2r_3m \in C$ for some $r_1, r_2, r_3 \in h(R)$ and for some $m \in h(M)$. Then $\frac{r_1r_2r_3m}{1} = \frac{r_1}{1}\frac{r_2}{1}\frac{r_3}{1}\frac{m}{1} \in S^{-1}C$. Since $S^{-1}C$ is a graded classical 2-absorbing submodule of $S^{-1}M$, we conclude that either $\frac{r_1}{1}\frac{r_2}{1}\frac{m}{1} = \frac{r_1r_2m}{1} \in S^{-1}C$ or $\frac{r_1}{1}\frac{r_3}{1}\frac{m}{1} = \frac{r_1r_3m}{1} \in S^{-1}C$ or $\frac{r_2}{1}\frac{r_3}{1}\frac{m}{1} = \frac{r_2r_3m}{1} \in S^{-1}C$. If $\frac{r_1r_2m}{1} \in S^{-1}C$, then there exists $s \in S$ such that $sr_1r_2m \in C$ and since $S \cap G\text{-}Zdv_R(M/C) = \phi$, we have $r_1r_2m \in C$. With a same argument, we can show that if $\frac{r_1r_3m}{1} \in S^{-1}C$, then $r_1r_3m \in C$ and also we can show if $\frac{r_2r_3m}{1} \in S^{-1}C$, then $r_2r_3m \in C$. Therefore C is a graded classical 2-absorbing submodule of M. □

LEMMA 2.6. *Let R be a G-graded ring, M a graded R-module and C a graded classical 2-absorbing submodule of M. Let $I = \bigoplus_{g\in G} I_g$ be a graded ideal of R. Then for every $r, s \in h(R)$, $m \in h(M)$ and $g \in G$ with $rsI_gm \subseteq C$, either $rsm \in C$ or $rI_gm \subseteq C$ or $sI_gm \subseteq C$.*

Proof. Let $r, s \in h(R)$, $m \in h(M)$ and $g \in G$ such that $rsI_gm \subseteq C$, $rsm \notin C$, $rI_gm \nsubseteq C$ and $sI_gm \nsubseteq C$. Then there exist $i_{1g}, i_{2g} \in I_g$ such that $ri_{1g}m \notin C$ and $si_{2g}m \notin C$. Since C is a graded classical 2-absorbing submodule, $rsi_{1g}m \in$

C, $rsm \notin C$ and $ri_{1g}m \notin C$, we have $si_{1g}m \in C$. Also $rsi_{2g}m \in C$ implies that $ri_{2g}m \in C$, since C is a graded classical 2-absorbing submodule. Since $rs(i_{1g} + i_{2g})m \in C$, we conclude that $r(i_{1g} + i_{2g})m \in C$ or $s(i_{1g} + i_{2g})m \in C$ or $rsm \in C$ and hence either $rsm \in C$ or $ri_{1g}m \in C$ or $si_{2g}m \in C$, which is a contradiction. □

THEOREM 2.7. *Let R be a G-graded ring, M a graded R-module and C a graded classical 2-absorbing submodule of M. Let $I = \bigoplus_{g\in G} I_g$ and $J = \bigoplus_{g\in G} J_g$ be a graded ideals of R. Then for every $r \in h(R)$, $m \in h(M)$ and $g, h \in G$ with $rI_gJ_hm \subseteq C$, either $rI_gm \subseteq C$ or $rJ_hm \subseteq C$ or $I_gJ_hm \subseteq C$.*

Proof. Let $r \in h(R)$, $m \in h(M)$ and $g, h \in G$ such that $rI_gJ_hm \subseteq C$, $rI_gm \nsubseteq C$ and $rJ_hm \nsubseteq C$. We have to show that $I_gJ_hm \subseteq C$. Assume that $i_g \in I_g$ and $j_h \in J_h$. By assumption there exist $i'_g \in I_g$ and $j'_h \in J_h$ such that $ri'_gm \notin C$ and $rj'_hm \notin C$. Since $ri'_gJ_hm \subseteq C$, $ri'_gm \notin C$ and $rJ_hm \nsubseteq C$, by Lemma 2.6, we have $i'_gJ_hm \subseteq C$. Also since $rj'_hI_g\ m \subseteq C$, $rj'_hm \notin C$ and $rI_gm \nsubseteq C$, by Lemma 2.6, we have $j'_hI_g\ m \subseteq C$. By $(i_g + i'_g) \in I_g$ and $(j_h + j'_h) \in J_h$ it follows that $r(i_g + i'_g)(j_h + j'_h)m \in C$. Since C is a graded classical 2-absorbing submodule, either $r(i_g+i'_g)m \in C$ or $r(j_h+j'_h)m \in C$ or $(i_g+i'_g)(j_h+j'_h)m \in C$. If $r(i_g+i'_g)m = ri_gm+ri'_gm \in C$, then $ri_gm \notin C$ which implies that $i_gj_hm \in C$ by Lemma 2.6. Similarly, by $r(j_h + j'_h)m \in C$, we conclude that $i_gj_hm \in C$. If $(i_g + i'_g)(j_h + j'_h)m \in C$, then $i_gj_hm + i_gj'_hm + i'_gj_hm + i'_gj'_hm \in C$ and so $i_gj_hm \in C$. Thus $I_gJ_hm \subseteq C$. □

THEOREM 2.8. *Let R be a G-graded ring, M a graded R-module and C a proper graded submodule of M. Let $I = \bigoplus_{g\in G} I_g$, $J = \bigoplus_{g\in G} J_g$ and $K = \bigoplus_{g\in G} K_g$ be a graded ideals of R. Then the following statement are equivalent:*

(i) *C is a graded classical 2-absorbing submodule of M;*

(ii) *For every $g, h, \lambda \in G$ and $m \in h(M)$ with $I_gJ_hK_\lambda m \subseteq C$, either $I_gJ_hm \subseteq C$ or $I_gK_\lambda m \subseteq C$ or $J_hK_\lambda m \subseteq C$*

Proof. $(i) \Rightarrow (ii)$ Assume that C is a graded classical 2-absorbing submodule of M. Let $g, h, \lambda \in G$ and $m \in h(M)$ such that $I_gJ_hK_\lambda m \subseteq C$ and $I_gJ_hm \nsubseteq C$. Then by Theorem 2.7, for all $r_\lambda \in K_\lambda$ either $I_gr_\lambda m \subseteq C$ or $J_hr_\lambda m \subseteq C$. If $I_gr_\lambda m \subseteq C$, for all $r_\lambda \in K_\lambda$ we are done. Similarly if $J_hr_\lambda m \subseteq C$, for all $r_\lambda \in K_\lambda$ we are done. Suppose that $r_\lambda, r'_\lambda \in K_\lambda$ are such that $I_gr_\lambda m \nsubseteq C$ and $J_hr'_\lambda m \nsubseteq C$. It follows that $I_gr'_\lambda m \subseteq C$ and $J_hr_\lambda m \subseteq C$. Since $I_gJ_h(r_\lambda + r'_\lambda)m \subseteq C$, by Theorem 2.7, we have either $I_g(r_\lambda+r'_\lambda)m \subseteq C$ or $J_h(r_\lambda+r'_\lambda)m \subseteq C$. By $I_g(r_\lambda + r'_\lambda)m \subseteq C$ it follows that $I_gr_\lambda m \subseteq C$ which is a contradiction. Similarly by $J_h(r_\lambda + r'_\lambda)m \subseteq C$ we get a contradiction. Therefore $I_gK_\lambda m \subseteq C$ or $J_hK_\lambda m \subseteq C$.

$(ii) \Rightarrow (i)$Assume that (ii) holds. Let $r_g, s_h, t_\lambda \in h(R)$ and $m \in h(M)$ such that $r_gs_ht_\lambda m \in C$. Let $I = r_gR$, $J = s_hR$ and $K = t_\lambda R$ be a graded ideals of R generated by r_g, s_h and t_λ, respectively. Then $I_g J_h K_\lambda m \subseteq C$. By our assumption we obtain $I_g J_h m \subseteq C$ or $I_g K_\lambda m \subseteq C$ or $J_h K_\lambda m \subseteq C$. Hence $r_g s_h m \in C$ or $r_g t_\lambda m \in C$ or $s_h t_\lambda m \in C$. Therefore C is a graded classical 2-absorbing submodule of M. □

Let M and M' be two graded R-modules. A homomorphism of graded R-modules $\varphi : M \to M'$ is a homomorphism of R-modules verifying $\varphi(M_g) \subseteq M'_g$ for every $g \in G$.

Theorem 2.9. *Let R be a G-graded ring and M, M' be two graded R-modules and $\varphi : M \to M'$ be an epimorphism of graded modules.*

(i) *If C is a graded classical 2-absorbing submodule of M containing $Ker\varphi$, then $\varphi(C)$ is a graded classical 2-absorbing submodule submodule of M'.*

(ii) *If C' is a graded classical 2-absorbing submodule of M', then $\varphi^{-1}(C')$ is a graded classical 2-absorbing submodule of M.*

Proof. (i) Suppose that C is a graded classical 2-absorbing submodule of M and let $r, s, t \in h(R)$ and $m' \in h(M')$ such that $rstm' \in \varphi(C)$, $rsm' \notin \varphi(C)$ and $rtm' \notin \varphi(C)$. Since $rstm' \in \varphi(C)$, there exists $c \in C \cap h(M)$ such that $\varphi(c) = rstm'$. Since $m' \in h(M')$ and φ is an epimorphism, there exists $m \in h(M)$ such that $\varphi(m) = m'$. Then $\varphi(c) = rst\varphi(m)$ and so $\varphi(c - rstm) = 0$. Hence $c - rstm \in Ker\varphi \subseteq C$ and so $rstm \in C$. Since C is a graded classical 2-absorbing submodule of M, $rsm \notin C$ and $rtm \notin C$, we have $stm \in C$. Hence $stm' \in \varphi(C)$. Thus $\varphi(C)$ is a graded classical 2-absorbing submodule of M'.

(ii) Suppose that C' is a graded classical 2-absorbing submodule of M' and let $r, s, t \in h(R)$ and $m \in h(M)$ such that $rstm \in \varphi^{-1}(C')$, $rsm \notin \varphi^{-1}(C')$ and $rtm \notin \varphi^{-1}(C')$. Since φ is an epimorphism, $\varphi(rstm) = rst\varphi(m) \in C'$. Since C' is a graded classical 2-absorbing submodule of M', $rs\varphi(m) = \varphi(rsm) \notin C'$and $rt\varphi(m) = \varphi(rtm) \notin C'$, we have $st\varphi(m) = \varphi(stm) \in C'$ and hence $stm \in \varphi^{-1}(C')$. Thus $\varphi^{-1}(C')$ is a graded classical 2-absorbing submodule of M. □

As an immediate consequence of Theorem 2.9 we have the following corollary.

Corollary 2.10. *Let R be a G-graded ring, M a graded R-module and $K \subseteq C$ a graded submodules of M. Then C is a graded classical 2-absorbing submodule of M if and only if C/K is a graded classical 2-absorbing submodule of M/K.*

Lemma 2.11. *Let R be a G-graded ring, M a graded R-module and C a graded submodule of M. If C is an intersection of two graded classical prime submodules of M, then C is a graded classical 2-absorbing submodule of M.*

Proof. Suppose that $C = C_1 \cap C_2$, where C_1 and C_2 are graded classical prime submodules of M. Let r, s, $t \in h(R)$ and $m \in h(M)$ with $rstm \in C$. Since C_1 is a graded classical prime submodules of M, we have either $rm \in C_1$ or $sm \in C_1$ or $tm \in C_1$. Since C_2 is a graded classical prime submodules of M, we have either $rm \in C_2$ or $sm \in C_2$ or $tm \in C_2$. It follows that $rsm \in C_1 \cap C_2$ or $rtm \in C_1 \cap C_2$ or $stm \in C_1 \cap C_2$. Thus C is a a graded classical 2-absorbing submodule of M. □

Let R_i be a graded commutative ring with identity and M_i be a graded R_i-module, for $i = 1, 2$. Let $R = R_1 \times R_2$. Then $M = M_1 \times M_2$ is a graded R-module and each graded submodule of M is of the form $C = C_1 \times C_2$ for some graded submodules C_1 of M_1 and C_2 of M_2.

THEOREM 2.12. *Let $R = R_1 \times R_2$ be a graded ring and $M = M_1 \times M_2$ be a graded R-module where M_1 is a graded R_1-module and M_2 is a graded R_2-module. Let C_1 and C_2 be a proper graded submodules of M_1 and M_2, respectively.*

(i) *C_1 is a graded classical 2-absorbing submodule of M_1 if and only if $C = C_1 \times C_2$ is a graded classical 2-absorbing submodule of M.*

(ii) *C_2 is a graded classical 2-absorbing submodule of M_2 if and only if $C = M_1 \times C_2$ is a graded classical 2-absorbing submodule of M.*

(iii) *$C = C_1 \times C_2$ is a graded classical 2-absorbing submodule of M if and only if C_1 and C_2 are graded classical prime submodules of M_1 and M_2, respectively.*

Proof. (i) Suppose that $C = C_1 \times M_2$ is a graded classical 2-absorbing submodule of M. From our hypothesis, C_1 is proper, So $C_1 \neq M_1$. Set $M' = \frac{M}{\{0\} \times M_2}$. Hence $C' = \frac{C}{\{0\} \times M_2}$ is a graded classical 2-absorbing submodule of M by Corollary 2.10. Also observe that $M' \cong M_1$ and $C' \cong C_1$. Thus C_1 is a graded classical 2-absorbing submodule of M_1. Conversely, if C_1 is a graded classical 2-absorbing submodule of M_1, then it is clear that $C = C_1 \times M_2$ is a graded classical 2-absorbing submodule of M.

(ii) It can be easily verified similar to (i).

(iii) Assume that $C = C_1 \times C_2$ is a graded classical 2-absorbing submodule of M. We show that C_1 is a graded classical prime submodules of M_1. Since $C_2 \neq M_2$, there exists $m_2 \in M_2 \backslash C_2$. Let $rsm_1 \in C_1$ for r, $s \in h(R_1)$ and $m_1 \in h(M_1)$. Then $(r, 1)(s, 1)(1, 0)(m_1, m_2) = (rsm_1, 0) \in C = C_1 \times C_2$. Since $C = C_1 \times C_2$ is a graded classical 2-absorbing submodule of M and $m_2 \notin C_2$, either $(r, 1)(1, 0)(m_1, m_2) = (rm_1, 0) \in C = C_1 \times C_2$ or $(s, 1)(1, 0)(m_1, m_2) = (sm_1, 0) \in C = C_1 \times C_2$. Hence either $rm_1 \in C_1$ or $sm_1 \in C_1$ which shows that C_1 is a graded classical prime submodule of M_1. Similarly, one can show that C_2 is a graded classical prime submodule of M_2. Conversely, assume that

$C_1$and C_2 are graded classical prime submodules of M_1 and M_2, respectively. One can easily see that $(C_1 \times M_2)$ and $(M_1 \times C_2)$ are graded classical prime submodules of M. Hence $(C_1 \times M_2) \cap (M_1 \times C_2) = C_1 \times C_2 = C$ is a graded classical 2-absorbing submodule of M by Lemma 2.11.

□

References

[1] R. Abu-Dawwas and K. Al-Zoubi, *On graded weakly classical prime submodules*, Iran. J. Math. Sci. Inform. **12** (2017), no. 1, 153–161.

[2] K. Al-Zoubi and R. Abu-Dawwas, *On graded 2-absorbing and weakly graded 2-absorbing submodules*, J. Math. Sci. Adv. Appl. **28** (2014), 45–60.

[3] K. Al-Zoubi, R. Abu-Dawwas and S. Ceken, *On graded 2-absorbing and graded weakly 2-absorbing ideals*, submitted.

[4] K. Al-Zoubi, M. Jaradat and R. Abu-Dawwa, *On graded classical prime and graded prime submodules*, Bull. Iranian Math. Soc. **41** (2015), no. 1, 217–225.

[5] K. Al-Zoubi and M. Al-Dolat, *On graded classical primary submodules*, Adv. Pure Appl. Math. **7** (2016), no. 2, 93–96.

[6] S. E. Atani, *On graded prime submodules*, Chiang Mai J. Sci. **33** (2006), no. 1, 3–7.

[7] D. F. Anderson and A. Badawi, *On n-absorbing ideals of commutative rings*, Comm. Algebra **39** (2011), 1646–1672.

[8] A. Badawi, *On 2-absorbing ideals of commutative rings*, Bull. Aust. Math. Soc. **75** (2007), 417–429.

[9] A. Y. Darani and S. Motmaen, *Zariski topology on the spectrum of graded classical prime submodules*, Appl. Gen. Topol. **14** (2013), no. 2, 159–169.

[10] R. Hazrat, *Graded rings and graded Grothendieck groups*, Cambridge University Press, Cambridge, 2016.

[11] H. Mohammad and A. Ashour, *On graded n-absorbing submodules*, Le Matematiche **70** (2015), no. 2, 243–254.

[12] H. Mostafanasab, U. Tekir and K.H. Oral, *Classical 2-absorbing submodules of modules over commutative rings*, Eur. J. Pure Appl. Math. **8** (2015), no. 3, 417–430.

[13] S. R. Naghani and H. F. Moghimi, *On graded 2-absorbing and graded weakly 2-absorbing ideals of a commutative ring*, Cankaya Univ. J. Sci. Eng. **13** (2016), no. 2, 011–017.

[14] C. Nastasescu and V.F. Oystaeyen, *Graded and filtered rings and modules*, Lecture Notes in Math. 758, Berlin-New York, Springer, 1982.

[15] C. Nastasescu and F. Van Oystaeyen, *Graded ring theory*, Mathematical Library 28, North-Holland, Amsterdam, 1982.

[16] C. Nastasescu and V.F. Oystaeyen, *Methods of graded rings*, Lecture Notes in Math. 1836, Springer, Berlin-Heidelberg, 2004.

[17] K. H. Oral, U. Tekir and A. G. Agargun, *On graded prime and primary submodules*, Turkish J. Math. **35** (2011), 159–167.

[18] M. Refai and K. Al-Zoubi, *On graded primary ideals*, Turkish J. Math. **28** (2004), no. 3, 217–229.

Authors' addresses:

Khaldoun Al-Zoubi
Department of Mathematics and Statistics
Jordan University of Science and Technology
P.O.Box 3030
Irbid 22110, Jordan
E-mail: kfzoubi@just.edu.jo

Mariam Al-Azaizeh
Department of Mathematics and Statistics
Jordan University of Science and Technology
P.O.Box 3030
Irbid 22110, Jordan
E-mail: maalazaizeh15@sci.just.edu.jo

Received August 26, 2017
Revised November 20, 2017
Accepted January 12, 2018

Rend. Istit. Mat. Univ. Trieste
Volume 50 (2018), 47–63
DOI: 10.13137/2464-8728/22435

Some additive decompositions of semisimple matrices

Alberto Dolcetti and Donato Pertici

Abstract. *Under suitable hypotheses on the ground field and on the semisimple matrix M, we discuss some additive decompositions of M and of its image through a convergent power series.*

Keywords: (Normalized) $\mathbb{K}$-trace and $\mathbb{K}$-decomposition of an element of $\overline{\mathbb{K}}$, conjugacy and involution mapping, (normalized fine) $\mathbb{K}$-trace decomposition of a matrix, ordered fields, valued fields, ordered quadratically closed fields, real closed fields.
MS Classification 2010: 15A16, 12J10, 12J15, 15A18, 15A21.

1. Introduction

In this paper we carry on the study of some additive decompositions of matrices, started in [2] and in [1] with applications to the image of a matrix through a power series (see for instance also [3]). We work in a fixed algebraic closure $\overline{\mathbb{K}}$ of a fixed field $\mathbb{K}$ of characteristic 0.

In §2 we define a projection, $\tau_{\mathbb{K}} : \overline{\mathbb{K}} \to \overline{\mathbb{K}}$, whose image is $\mathbb{K}$ and which allows to decompose $\overline{\mathbb{K}}$ as direct sum $\mathbb{K} \oplus Ker(\tau_{\mathbb{K}})$ (see Remark 2.1). As a consequence, we get the $\mathbb{K}$-*trace decomposition* of a semisimple matrix $M \in M_n(\mathbb{K})$, i.e. we write (in a unique way) $M = T+F$, where $T, F \in M_n(\mathbb{K})$ are mutually commuting, T diagonalizable over $\mathbb{K}$ and F semisimple, with all eigenvalues in $Ker(\tau_{\mathbb{K}})$ (see Proposition 2.10). Finally, we obtain the *fine* $\mathbb{K}$-*trace decomposition* of any semisimple matrix $M \in M_n(\mathbb{K})$ (see Proposition 2.11 and Remark 2.13), which decomposes each summand of the $\mathbb{K}$-trace decomposition in simpler summands.

In §3, starting from the fine $\mathbb{K}$-trace decomposition of a semisimple matrix M, we get a formula for the image $f(M)$ through a power series under the further conditions that $\mathbb{K}$ is a *valued field* and that M is $\mathbb{K}$-*quadratic*, i.e. its eigenvalues have degree at most 2 over $\mathbb{K}$ (see Proposition 3.5 and in particular Examples 3.6).

In §4, we *normalize* the fine $\mathbb{K}$-trace decomposition of a semisimple $\mathbb{K}$-quadratic matrix M, when the field $\mathbb{K}$ is *ordered quadratically closed* and we write its image through a power series as above (see Proposition 4.10). When $\mathbb{K}$ is *real closed* too, this formula holds for every semisimple matrix in the domain of convergence of the series and it can be viewed as a generalization of the classical Rodrigues' formula for the exponential of a real skew symmetric matrix (see Examples 4.11).

2. Fine $\mathbb{K}$-trace decomposition

In this paper $\mathbb{K}$ is a fixed field of characteristic 0 and $\overline{\mathbb{K}}$ one of its algebraic closures.

Remark 2.1 ($\mathbb{K}$-decomposition): Let $\mathbb{L} \subseteq \overline{\mathbb{K}}$ be any finite extension of $\mathbb{K}$ and $\lambda \in \mathbb{L}$ be of degree d with minimal polynomial over $\mathbb{K}$:

$$m_\lambda(X) = X^d + a_{d-1}X^{d-1} + \cdots + a_1X + a_0\,.$$

The multiplication by λ is a $\mathbb{K}$-linear mapping from $\mathbb{L}$ into $\mathbb{L}$, whose characteristic polynomial is $m_\lambda(X)^{[\mathbb{L}:\mathbb{K}(\lambda)]}$ and whose trace is:

$$tr_{\mathbb{L}/\mathbb{K}}(\lambda) = -[\mathbb{L}:\mathbb{K}(\lambda)]a_{d-1} = -\frac{[\mathbb{L}:\mathbb{K}]}{[\mathbb{K}(\lambda):\mathbb{K}]}a_{d-1} = -\frac{[\mathbb{L}:\mathbb{K}]}{d}a_{d-1}$$

(see for instance [6, Ch. VI, Proposition 5.6]).

Hence the expression: $\dfrac{tr_{\mathbb{L}/\mathbb{K}}(\lambda)}{[\mathbb{L}:\mathbb{K}]} = -\dfrac{a_{d-1}}{d}$ depends only on λ and $\mathbb{K}$ and not on the finite extension $\mathbb{L} \subset \overline{\mathbb{K}}$ of $\mathbb{K}$, containing λ.

Therefore, for each $\lambda \in \overline{\mathbb{K}}$, we call $\tau_{\mathbb{K}}(\lambda) := -\dfrac{a_{d-1}}{d}$ the *normalized* $\mathbb{K}$-*trace of* λ.

It is easy to check that $\tau_{\mathbb{K}}$ is a $\mathbb{K}$-linear mapping from $\overline{\mathbb{K}}$ onto $\mathbb{K} \subseteq \overline{\mathbb{K}}$ with $\tau_{\mathbb{K}}^2 = \tau_{\mathbb{K}}$ (i.e. $\tau_{\mathbb{K}}$ is a *projection* as a $\mathbb{K}$-linear endomorphism of $\overline{\mathbb{K}}$) and therefore we get a canonical decomposition as $\mathbb{K}$-vector spaces: $\overline{\mathbb{K}} = \mathbb{K} \oplus Ker(\tau_{\mathbb{K}})$, (see for instance [4, p. 211]), i.e. every element $\lambda \in \overline{\mathbb{K}}$ has a unique decomposition $\lambda = \tau_{\mathbb{K}}(\lambda) + \varphi_{\mathbb{K}}(\lambda)$ with $\tau_{\mathbb{K}}(\lambda) \in \mathbb{K}$ and $\varphi_{\mathbb{K}}(\lambda) \in Ker(\tau_{\mathbb{K}})$.

We call this decomposition, $\tau_{\mathbb{K}}(\lambda)$ and $\varphi_{\mathbb{K}}(\lambda)$ respectively, the $\mathbb{K}$-*decomposition*, the $\mathbb{K}$-*part* and the $\mathbb{K}$-*trace-free part* of λ. We will write $\tau(\lambda)$ and $\varphi(\lambda)$ in absence of ambiguity about the field $\mathbb{K}$.

Remark 2.2 ($\mathbb{K}$-trace-free polynomial): Recall that two elements $\lambda, \lambda' \in \overline{\mathbb{K}}$ are said to be *conjugate over* $\mathbb{K}$, if they have the same minimal polynomial over $\mathbb{K}$ or, equivalently, if they are in the same orbit under $Aut(\overline{\mathbb{K}}/\mathbb{K})$: the group of all automorphisms of the field $\overline{\mathbb{K}}$ fixing each element of $\mathbb{K}$. Hence $\lambda, \lambda' \in \overline{\mathbb{K}}$ are conjugate over $\mathbb{K}$ if and only if $\tau(\lambda) = \tau(\lambda')$ and $\varphi(\lambda)$, $\varphi(\lambda')$ are conjugate over $\mathbb{K}$.

For every $\lambda \in \overline{\mathbb{K}}$ we denote $\nu_{\mathbb{K}}(\lambda) = \nu(\lambda)$ the *normalized norm of* λ *over* $\mathbb{K}$ as $\nu(\lambda) = (-1)^d a_0 = \lambda_1\lambda_2\ldots\lambda_d$, where $d = deg_{\mathbb{K}}(\lambda)$, a_0 is the constant term of the minimal polynomial of λ over $\mathbb{K}$ and $\{\lambda = \lambda_1, \lambda_2, \ldots, \lambda_d\}$ is the conjugacy class of λ over $\mathbb{K}$.

If $m_\lambda(X)$ is the minimal polynomial of λ over $\mathbb{K}$, then we call $\mathbb{K}$-*trace-free polynomial of* λ to be the polynomial $\widetilde{m}_\lambda(X) = m_\lambda(X + \tau(\lambda)) - m_\lambda(\tau(\lambda))$.

Note that $m_\lambda(X + \tau(\lambda))$ is the minimal polynomial of $\varphi(\lambda)$ over $\mathbb{K}$ and so $m_\lambda(\tau(\lambda)) = (-1)^d \nu(\varphi(\lambda))$, thus $\widetilde{m}_\lambda(X) = m_\lambda(X + \tau(\lambda)) + (-1)^{d+1}\nu(\varphi(\lambda))$. Moreover $\widetilde{m}_\lambda(X)$ is monic of degree d, its coefficient of the term of degree $d-1$ and its constant term are both zero. In particular, if $d = 2$, then $\widetilde{m}_\lambda(X) = X^2$ for every λ.

Remark 2.3 ($\mathbb{K}$-linear involution): The mapping $\lambda \mapsto \overline{\lambda} = \overline{\tau(\lambda) + \varphi(\lambda)} = \tau(\lambda) - \varphi(\lambda)$ is called the $\mathbb{K}$-*linear involution of* $\overline{\mathbb{K}}$. The $\mathbb{K}$-linear involution is an automorphism of $\overline{\mathbb{K}}$ as $\mathbb{K}$-vector space, but in general not as field; $\mathbb{K}$ is the set of fixed points of the $\mathbb{K}$-linear involution.

Lemma 2.4. *a) If $\lambda = \tau(\lambda) + \varphi(\lambda) \in \overline{\mathbb{K}}$ has degree 2 over $\mathbb{K}$, then $\varphi(\lambda)^2 \in \mathbb{K}$ and the unique conjugate of λ, distinct from λ, is $\overline{\lambda}$.*

b) If $\mathbb{L} \subseteq \overline{\mathbb{K}}$ is an extension of $\mathbb{K}$ of degree 2, then the $\mathbb{K}$-linear involution restricted to $\mathbb{L}$ is an element of $Aut(\mathbb{L}/\mathbb{K})$.

c) If $\beta, \beta' \in Ker(\tau)$ have degree 2 over $\mathbb{K}$, then $\beta\beta' \in \mathbb{K}$ if and only if β and β' are linearly dependent over $\mathbb{K}$, otherwise $\beta\beta' \in Ker(\tau)$ and its degree over $\mathbb{K}$ is 2.

d) If $\lambda \in \overline{\mathbb{K}}$ has degree 2 over $\mathbb{K}$, then $\lambda \in Ker(\tau)$ if and only if $\lambda^2 \in \mathbb{K}$.

Proof. a) Since $\varphi(\lambda)$ has degree 2 over $\mathbb{K}$ and its normalized $\mathbb{K}$-trace is zero, its minimal polynomial has the form $X^2 - \varphi(\lambda)^2$, hence $\varphi(\lambda)^2 \in \mathbb{K}$ and $\varphi(\lambda)$ and $-\varphi(\lambda)$ are conjugate over $\mathbb{K}$ and also λ and $\overline{\lambda}$ are conjugate over $\mathbb{K}$ (remember Remark 2.2).

b) Choose an element $\lambda \in \mathbb{L}$ of degree 2 over $\mathbb{K}$, having normalized $\mathbb{K}$-trace equal to zero. Hence the elements of $\mathbb{L}$ are of the form $k_1 + k_2\lambda$ with $k_1, k_2 \in \mathbb{K}$. We conclude by standard computations, because $\lambda^2 \in \mathbb{K}$ by (a).

c) If β and β' are as in (c), then, from (a), β^2, β'^2 are both in $\mathbb{K}$. Hence $\beta\beta'$ is root of $X^2 - \beta^2\beta'^2 \in \mathbb{K}[X]$ and so the degree of $\beta\beta'$ is at most 2. The degree of $\beta\beta'$ is 1 if and only if $\beta\beta' = t \in \mathbb{K}$, i. e. if and only if $\beta = \dfrac{t}{\beta'} = \dfrac{t}{\beta'^2}\beta'$ and so if and only if β, β' are linearly dependent over $\mathbb{K}$, because $\dfrac{t}{\beta'^2} \in \mathbb{K}$. Otherwise the degree of $\beta\beta'$ is 2; thus $X^2 - \beta^2\beta'^2$ is its minimal polynomial and so: $\tau(\beta\beta') = 0$.

Finally, one implication of (d) follows directly from (a), since $\tau(\lambda) = 0$. For the other implication it suffices to note that, if λ has degree 2 over $\mathbb{K}$ and $\lambda^2 \in \mathbb{K}$, then the minimal polynomial of λ over $\mathbb{K}$ is $X^2 - \lambda^2$ and so: $\tau(\lambda) = 0$. $\square$

Remark 2.5 (minimal polinomial): From now on M is a fixed *semisimple* matrix in the set $M_n(\mathbb{K})$ of the square matrices of order $n \geq 2$ with entries in $\mathbb{K}$ and with minimal polynomial $m(X) = (X-\gamma_1)\cdots(X-\gamma_s)\, m_1(X)\cdots m_t(X)$, where $\gamma_1, \cdots, \gamma_s$ are mutually distinct elements of $\mathbb{K}$ and $m_1(X), \cdots, m_t(X)$ are mutually distinct irreducible monic polynomials in $\mathbb{K}[X]$ of degrees $d_1, \cdots, d_t \geq 2$ respectively. We denote by $\lambda_{h1}, \cdots, \lambda_{hd_h}$ the d_h distinct roots of the factor $m_h(X)$ and by $\mathbb{F} \subseteq \overline{\mathbb{K}}$ the *splitting field* of $m(X)$. By conciseness we put $\lambda_h = \lambda_{h1}$ for every $h = 1, \cdots, t$.

Definition 2.6. *With the notations of Remark 2.5 we say that the semisimple matrix $M \in M_n(\mathbb{K})$ is* $\mathbb{K}$-quadratic *if every factor $m_h(X)$ has degree 2 (or if $m(X) = (X-\gamma_1)\cdots(X-\gamma_s)$).*

Remark 2.7 (Frobenius decomposition): The semisimple matrix M has a unique decomposition:

$$M = \sum_{i=1}^{s} \gamma_i \mathcal{A}_i + \sum_{h=1}^{t} \sum_{j=1}^{d_h} \lambda_{hj} C_{hj} \,,$$

where $\{\gamma_1\} \cup \cdots \{\gamma_s\} \cup_{h=1}^{t} \{\lambda_{h1}, \cdots, \lambda_{hd_h}\}$ is the set of all distinct eigenvalues of M (arranged in conjugacy classes) and the matrices $\mathcal{A}_i$'s, C_{hj}'s are *idempotent* and satisfying the conditions: $\mathcal{A}_i\mathcal{A}_{i'} = 0$ (if $i \neq i'$), $\mathcal{A}_i C_{hj} = C_{hj}\mathcal{A}_i = 0$ (for every i, j, h), $C_{hj}C_{h'j'} = 0$ (if $(h, j) \neq (h', j')$), $\sum_{i=1}^{s} \mathcal{A}_i + \sum_{h=1}^{t}\sum_{j=1}^{d_h} C_{hj} = I_n$ (the identity matrix of order n).

The above decomposition is called *Frobenius decomposition of* M and the matrices $\mathcal{A}_i$'s and C_{hj}'s, called *Frobenius covariants* of M, are uniquely determined by M (and by the previous conditions) and are polynomial expressions of M of degree strictly less than $deg(m(X))$; finally the matrices $\mathcal{A}_i$'s have coefficients in $\mathbb{K}$ and the matrices C_{hj}'s in $\mathbb{F}$ (see [1, § 1]).

Definition 2.8. *A* $\mathbb{K}$-trace decomposition *of M is an additive decomposition: $M = T + F$ where $T, F \in M_n(\mathbb{K})$ are mutually commuting, T is diagonalizable over $\mathbb{K}$ and F is semisimple with eigenvalues in $Ker(\tau_{\mathbb{K}})$. We say that T and F respectively are a* $\mathbb{K}$-part *and a* $\mathbb{K}$-trace-free part *of M.*

Remark 2.9: If A is a matrix in $M_n(\mathbb{K})$ such that all eigenvalues of A are in $Ker(\tau)$, then its usual *trace* is zero.

Indeed if λ is an eigenvalue of A of degree d over $\mathbb{K}$, then every conjugate of λ over $\mathbb{K}$ is again an eigenvalue of A, moreover the sum of the eigenvalues of the whole conjugacy class of λ over $\mathbb{K}$ is $d{\cdot}\tau(\lambda) = 0$ and so the trace of A is zero.

Therefore, if the matrix $M \in M_n(\mathbb{K})$ has a $\mathbb{K}$-trace decomposition, then the trace of its $\mathbb{K}$-trace-free part is zero.

Proposition 2.10. *The semisimple matrix $M \in M_n(\mathbb{K})$ has a unique $\mathbb{K}$-trace decomposition: $M = T(M) + F(M)$, where, with the notations of Remark 2.5 and Remark 2.7,*

$$T(M) = \sum_{i=1}^{s} \gamma_i \mathcal{A}_i + \sum_{h=1}^{t} \tau(\lambda_h) \sum_{j=1}^{d_h} C_{hj} \quad \text{and} \quad F(M) = \sum_{h=1}^{t} \sum_{j=1}^{d_h} \varphi(\lambda_{hj}) C_{hj}\,.$$

In particular $T(M)$, $F(M)$ are polynomial expressions of M.

Proof. By Remark 2.2, $\tau(\lambda_{hj}) = \tau(\lambda_h)$ for every j and, for every h, the set $\{\varphi(\lambda_{h1}), \cdots, \varphi(\lambda_{hd_h})\}$ is a conjugacy class over $\mathbb{K}$.

Now let $M = \sum_{i=1}^{s} \gamma_i \mathcal{A}_i + \sum_{h=1}^{t} \sum_{j=1}^{d_h} \lambda_{hj} C_{hj}$ be the Frobenius decomposition of M. By decomposing each λ_{hj} as $\tau(\lambda_h) + \varphi(\lambda_{hj})$, we get: $M = T(M) + F(M)$, with $T(M)$ and $F(M)$ as in the statement and therefore polynomial expressions of M. Arguing as in the proof of [1, Theorem 1.6], $T(M)$ and $F(M)$ are matrices with coefficients in $\mathbb{K}$, i.e. in the fixed field of $Aut(\mathbb{F}/\mathbb{K})$. Standard computations show that $T(M)$ and $F(M)$ are respectively diagonalizable over $\mathbb{K}$ and over $\mathbb{F}$ with eigenvalues $\{\gamma_i, \tau(\lambda_h)\}$ and $\{\varphi(\lambda_{hj})\}$ (see also Remark 2.7 and [1, Proposition 1.9]). This allows to conclude about the existence of a $\mathbb{K}$-trace decomposition in terms of polynomial expressions of M.

Now let $M = T' + F'$ be any other $\mathbb{K}$-trace decomposition of M. This implies $T(M) - T' = F' - F(M)$. Now T', F' commute with M and so with $T(M), F(M)$; moreover the four matrices are semisimple, hence, by simultaneous diagonalizability, every eigenvalue σ' of F' can be written as $\sigma' = \delta - \delta' + \sigma$ with δ, δ', σ eigenvalues of $T(M)$, T', $F(M)$ respectively. From the uniqueness of the $\mathbb{K}$-decomposition $\delta - \delta' = 0$ and $\sigma = \sigma'$. Therefore $T' = T(M)$ and $F' = F(M)$. □

Proposition 2.11. *Let $M \in M_n(\mathbb{K})$ be semisimple with eigenvalues: $\gamma_1, \cdots, \gamma_s$ distinct and belonging to $\mathbb{K}$ and the remaining (not in $\mathbb{K}$) $\{\lambda_h = \lambda_{h1}, \cdots, \lambda_{hd_h}\}$, $h = 1, \cdots, t$, arranged in distinct conjugacy classes. We have the decomposition*

$$M = \sum_{i=1}^{s} \gamma_i \mathcal{A}_i + \sum_{h=1}^{t} \frac{(-1)^{d_h+1}\,\tau(\lambda_h)}{\nu(\varphi(\lambda_h))} \widetilde{m}_h(\mathcal{B}_h) + \sum_{h=1}^{t} \mathcal{B}_h \tag{1}$$

with $\mathcal{B}_h = \mathcal{B}_h(M) = \sum_{j=1}^{d_h} \varphi(\lambda_{hj})\, C_{hj}$ $(h = 1, \cdots, t)$ matrix in $M_n(\mathbb{K})$ and $\widetilde{m}_h(X)$ $(h = 1, \cdots, t)$ the $\mathbb{K}$-trace free polynomial of λ_h.

Proof. By Proposition 2.10, we have:

$$M = \sum_{i=1}^{s} \gamma_i \mathcal{A}_i + \sum_{h=1}^{t} \tau(\lambda_h) \sum_{j=1}^{d_h} C_{hj} + \sum_{h=1}^{t} \sum_{j=1}^{d_h} \varphi(\lambda_{hj}) C_{hj} =$$

$$= \sum_{i=1}^{s} \gamma_i \mathcal{A}_i + \sum_{h=1}^{t} \tau(\lambda_h) \sum_{j=1}^{d_h} C_{hj} + \sum_{h=1}^{t} \mathcal{B}_h$$

As in [1, Proposition 1.5], it is easy to check that $\sigma(\mathcal{B}_h) = \mathcal{B}_h$ for every $\sigma \in Aut(\mathbb{F}/\mathbb{K})$, hence $\mathcal{B}_h \in M_n(\mathbb{K})$. Thus it suffices to prove that, for every $h = 1, \cdots, t$: $\widetilde{m}_h(\mathcal{B}_h) = (-1)^{d_h+1} \nu(\varphi(\lambda_h)) \sum_{j=1}^{d_h} C_{hj}$.

Since $\widetilde{m}_h(X)$ has constant term zero, from the properties of the matrices C_{hj}'s, we obtain $\widetilde{m}_h(\mathcal{B}_h) = \sum_{j=1}^{d_h} \widetilde{m}_h(\varphi(\lambda_{hj}))C_{hj}$ and so, by Remark 2.2, we can conclude that $\widetilde{m}_h(\mathcal{B}_h) = (-1)^{d_h+1} \nu(\varphi(\lambda_h)) \sum_{j=1}^{d_h} C_{hj}$. □

Remark 2.12: The matrices $\mathcal{A}_i$'s and $\mathcal{B}_h$'s in Proposition 2.11 are polynomial functions of M satisfying the following properties:

1) $\mathcal{A}_i \mathcal{A}_{i'} = \delta_{ii'} A_i$, for every i, i';
2) $\mathcal{A}_i \mathcal{B}_h = \mathcal{B}_h \mathcal{A}_i = 0$, for every i, h;
3) $\mathcal{B}_h \mathcal{B}_{h'} = 0$, provided $h \neq h'$;
4) $\mathcal{B}_h \, \widetilde{m}_h(\mathcal{B}_h) = (-1)^{d_h+1} \nu(\varphi(\lambda_h)) \, \mathcal{B}_h$, for every h.

Some of the previous properties have been already noted in Remark 2.7 and the others are easy to get by standard computations.

Remark 2.13 (fine $\mathbb{K}$-trace decomposition): It is easy to note that in Proposition 2.11, formula (1), every $\mathcal{B}_h$ is $\mathbb{K}$-trace free, while the $\mathbb{K}$-part and the $\mathbb{K}$-trace free part of M are respectively:

$$T(M) = \sum_{i=1}^{s} \gamma_i \mathcal{A}_i + \sum_{h=1}^{t} \frac{(-1)^{d_h+1} \, \tau(\lambda_h)}{\nu(\varphi(\lambda_h))} \widetilde{m}_h(\mathcal{B}_h)$$

with eigenvalues γ_i and $\tau(\lambda_h)$;

$$F(M) = \sum_{h=1}^{t} \mathcal{B}_h$$

with eigenvalues $\varphi(\lambda_{hj})$ and possibly 0.

Therefore we call the decomposition (1) in Proposition 2.11, the *fine* $\mathbb{K}$-*trace decomposition* of the semisimple matrix $M \in M_n(\mathbb{K})$.

Remark 2.14: By Lemma 2.4-(a) and Remark 2.2, if the matrix $M \in M_n(\mathbb{K})$ is semisimple and $\mathbb{K}$-quadratic, then the fine $\mathbb{K}$-trace decomposition of M becomes:

$$M = \sum_{i=1}^{s} \gamma_i \mathcal{A}_i + \sum_{h=1}^{t} \frac{\tau(\lambda_h)}{\varphi(\lambda_h)^2} \mathcal{B}_h^2 + \sum_{h=1}^{t} \mathcal{B}_h \,, \tag{2}$$

while the property (4) in Remark 2.12 becomes

4') $\mathcal{B}_h^3 = \varphi(\lambda_h)^2 \, \mathcal{B}_h$, for every h.

Moreover, from the properties of the Frobenius covariants of M and from Lemma 2.4-(a), we get:

$$I_n = \sum_{i=1}^{s} \mathcal{A}_i + \sum_{h=1}^{t} \frac{\mathcal{B}_h^2}{\varphi(\lambda_h)^2} \,.$$

3. A formula for power series of matrices over a valued field.

Remark 3.1: In this section we assume that $\mathbb{K}$ (of characteristic 0) is endowed with an *absolute value* $|\cdot|$. We call such a pair $(\mathbb{K}, |\cdot|)$ a *valued field*. We refer for instance to [5, Ch. 9], to [10, Ch. III], to [6, Ch. XII] and to [7, Ch. 23] for more information.

Let $(\mathbb{K}, |\cdot|)$ be a valued field. The absolute value over $\mathbb{K}$ extends in a unique way to its *completion* $\mathbb{K}^c$; this one extends in a unique way to an absolute value over a fixed algebraic closure $\overline{\mathbb{K}^c}$ of $\mathbb{K}^c$ and finally the last one extends in a unique way to the completion $(\overline{\mathbb{K}^c})^c$ (see for instance [7] Theorem 2 p. 48, Ostrowski's Theorem p. 55 and Theorem 4' p. 60). We denote all extensions always by the same symbol $|\cdot|$.

Note that the field $\{\alpha \in \overline{\mathbb{K}^c} \;/\; \alpha \text{ is algebraic over } \mathbb{K}\}$ is the unique algebraic closure of $\mathbb{K}$ contained in $\overline{\mathbb{K}^c}$ and therefore it can be identified with $\overline{\mathbb{K}}$.

By restriction, we get an absolute value over $\overline{\mathbb{K}}$, extending the absolute value of $\mathbb{K}$.

Lemma 3.2. *Let $(\mathbb{K}, |\cdot|)$ be a valued field.*

a) If $\mathbb{K}$ is algebraically closed, then its completion $\mathbb{K}^c$ (with respect to $|\cdot|$) is algebraically closed too.

b) $(\overline{\mathbb{K}^c})^c$ is algebraically closed and complete with respect to the unique extension to it of $|\cdot|$.

Proof. The proof of (a) follows easily from Ostrowski's Theorem in archimedean case (see [7, p. 55]), while for the non-archimedean case we refer to [7, Appendix 24.15, p. 316]. Part (b) follows directly from (a). □

DEFINITION 3.3. *Let $f(X) = \sum_{m=0}^{+\infty} a_m X^m$ be a series with coefficients in the valued field $(\mathbb{K}, |\cdot|)$ and $R_f \in \mathbb{R} \cup \{+\infty\}$ be the* radius of convergence *of the associated real series $\sum_{m=0}^{+\infty} |a_m| X^m$.*

We denote by $\Lambda_{f,\mathbb{K}}$ the subset of the matrices $A \in M_n(\mathbb{K})$ such that all of their eigenvalues $\lambda = \tau(\lambda) + \varphi(\lambda)$ (with their $\mathbb{K}$-decompositions) satisfy:

i) $|\tau(\lambda)| + |\varphi(\lambda)| < R_f$, if the absolute value of $\mathbb{K}$ is archimedean,

ii) $max(|\tau(\lambda)|, |\varphi(\lambda)|) < R_f$, if the absolute value of $\mathbb{K}$ is non-archimedean.

For every eigenvalue λ of a matrix $A \in \Lambda_{f,\mathbb{K}}$ with $\mathbb{K}$-decomposition $\lambda = \tau(\lambda) + \varphi(\lambda)$, denoted by $\lfloor x \rfloor$ the integer part *of the real number x, we introduce the formal series:*

$$\mathcal{T}f(\lambda) = \sum_{m=0}^{+\infty} a_m \sum_{h=0}^{\lfloor m/2 \rfloor} \binom{m}{2h} \tau(\lambda)^{m-2h} \varphi(\lambda)^{2h},$$
$$\mathcal{F}f(\lambda) = \sum_{m=1}^{+\infty} a_m \sum_{h=1}^{\lfloor (m+1)/2 \rfloor} \binom{m}{2h-1} \tau(\lambda)^{m-2h+1} \varphi(\lambda)^{2h-1}.$$

REMARK 3.4: a) If $A \in \Lambda_{f,\mathbb{K}}$, then $f(A) \in M_n(\mathbb{K}^c)$ (see [1, Remark-Definition 3.1(c)]).

b) If λ is an eigenvalue of a matrix in $\Lambda_{f,\mathbb{K}}$, then $f(\lambda), \mathcal{T}f(\lambda), \mathcal{F}f(\lambda)$ converge in $(\overline{\mathbb{K}})^c \subseteq (\overline{\mathbb{K}^c})^c$ and $f(\lambda) = \mathcal{T}f(\lambda) + \mathcal{F}f(\lambda)$.

The previous assertions follow from the definitions, by standard computations.

c) Assume that λ is an eigenvalue of degree 2 over $\mathbb{K}$ of a matrix in $\Lambda_{f,\mathbb{K}}$. Then: $\mathcal{T}f(\lambda) \in \mathbb{K}^c$, $\mathcal{F}f(\lambda), f(\lambda) \in \mathbb{K}^c(\lambda)$. Moreover, if $\lambda \notin \mathbb{K}^c$, then $f(\lambda) = \mathcal{T}f(\lambda) + \mathcal{F}f(\lambda)$ is the $\mathbb{K}^c$-decomposition of $f(\lambda)$.

Indeed if $\lambda \notin \mathbb{K}^c$ has degree 2 over $\mathbb{K}$ and $\lambda = \tau(\lambda) + \varphi(\lambda)$ is its $\mathbb{K}$-decomposition, then $\varphi(\lambda) \notin \mathbb{K}^c$ and $\varphi(\lambda)^2 \in \mathbb{K}$. Hence, looking at the partial sums and at their limits, we get that $\mathcal{T}f(\lambda) \in \mathbb{K}^c$, while $\mathcal{F}f(\lambda)$ is a multiple of $\varphi(\lambda)$ with coefficient in $\mathbb{K}^c$ and so it belongs to $Ker(\tau_{\mathbb{K}^c})$. We can conclude by uniqueness of $\mathbb{K}^c$-decomposition.

PROPOSITION 3.5. *Let $f(X) = \sum_{m=0}^{+\infty} a_m X^m$ be a series with coefficients in the valued field $(\mathbb{K}, |\cdot|)$ and $M \in \Lambda_{f,\mathbb{K}}$ be a semisimple $\mathbb{K}$-quadratic matrix with* *fine $\mathbb{K}$-trace decomposition:*

$$M = \sum_{i=1}^{s} \gamma_i \mathcal{A}_i + \sum_{h=1}^{t} \frac{\tau_{\mathbb{K}}(\lambda_h)}{\varphi_{\mathbb{K}}(\lambda_h)^2} \mathcal{B}_h^2 + \sum_{h=1}^{t} \mathcal{B}_h$$

as in Remark 2.14. Then

a) $f(M) = \sum_{i=1}^{s} f(\gamma_i)\mathcal{A}_i + \sum_{h=1}^{t} \left[\frac{\mathcal{T}f(\lambda_h)}{\varphi_{\mathbb{K}}(\lambda_h)^2} \mathcal{B}_h^2 + \frac{\mathcal{F}f(\lambda_h)}{\varphi_{\mathbb{K}}(\lambda_h)} \mathcal{B}_h \right]$,

with $f(\gamma_i)$, $\frac{\mathcal{T}f(\lambda_h)}{\varphi_{\mathbb{K}}(\lambda_h)^2}$, $\frac{\mathcal{F}f(\lambda_h)}{\varphi_{\mathbb{K}}(\lambda_h)}$ $\in \mathbb{K}^c$ *for every* i, h;

b) furthermore, if no eigenvalue λ_h of degree 2 over $\mathbb{K}$ is in $\mathbb{K}^c$,

$$f(M) = \sum_{i=1}^{s} f(\gamma_i)\mathcal{A}_i + \sum_{h=1}^{t} \left[\frac{\tau_{\mathbb{K}^c}(f(\lambda_h))}{\varphi_{\mathbb{K}}(\lambda_h)^2} \mathcal{B}_h^2 + \frac{\varphi_{\mathbb{K}^c}(f(\lambda_h))}{\varphi_{\mathbb{K}}(\lambda_h)} \mathcal{B}_h \right];$$

c) in general $f(M)$ is semisimple, $\mathbb{K}^c$-quadratic and its $\mathbb{K}^c$-trace decomposition is

$$f(M) = T(f(M)) + F(f(M)),$$

$$\text{where} \qquad T(f(M)) = \sum_{i=1}^{s} f(\gamma_i)\mathcal{A}_i + \sum_{h=1}^{t} \mathcal{T}f(\lambda_h) \frac{\mathcal{B}_h^2}{\varphi_{\mathbb{K}}(\lambda_h)^2},$$

whose (possibly repeated) eigenvalues are $f(\gamma_i)$ and $\mathcal{T}f(\lambda_h)$ for every i, h and

$$F(f(M)) = \sum_{h=1}^{t} \frac{\mathcal{F}f(\lambda_h)}{\varphi_{\mathbb{K}}(\lambda_h)} \mathcal{B}_h,$$

whose (possibly repeated) eigenvalues are $\pm\mathcal{F}f(\lambda_h)$ for every h and possibly 0.

Proof. Parts (b) and (c) follow from part (a) via Remark 3.4 and ordinary computations.

For (a), we denote $\alpha_j = \tau_{\mathbb{K}}(\lambda_j)$, $n_j = -\varphi_{\mathbb{K}}(\lambda_j)^2$, so that we can write the $\mathbb{K}$-decomposition of λ_j as $\lambda_j = \alpha_j + \sqrt{-n_j}$. From $\mathcal{B}_j^3 = -n_j \mathcal{B}_j$ we get:

$$\mathcal{B}_j^{2k} = (-n_j)^{k-1} \mathcal{B}_j^2 \text{ and } \mathcal{B}_j^{2k-1} = (-n_j)^{k-1} \mathcal{B}_j \text{ for every } k \geq 1.$$

Therefore, for every $m \geq 1$, standard computations allow to get:

$$\left[\frac{\tau_{\mathbb{K}}(\lambda_j)}{\varphi_{\mathbb{K}}(\lambda_j)^2}\mathcal{B}_j^2+\mathcal{B}_j\right]^m=\left[\sum_{h=0}^{\lfloor m/2\rfloor}\frac{\binom{m}{2h}\alpha_j^{m-2h}(\sqrt{-n_j})^{2h}}{(-n_j)}\right]\mathcal{B}_j^2$$
$$+\left[\sum_{h=1}^{\lfloor (m+1)/2\rfloor}\frac{\binom{m}{2h-1}\alpha_j^{m-2h+1}(\sqrt{-n_j})^{2h-1}}{\sqrt{-n_j}}\right]\mathcal{B}_j.$$

We have:

$$f(M)=a_oI_n+\sum_{m=1}^{+\infty}a_m\left[\sum_{i=1}^{s}\gamma_i\mathcal{A}_i+\sum_{j=1}^{t}(\frac{\alpha_j}{(-n_j)}\mathcal{B}_j^2+\mathcal{B}_j)\right]^m,$$

thus, remembering the properties of the various matrices on the right:

$$f(M)=a_oI_n+\sum_{m=1}^{+\infty}a_m\sum_{i=1}^{s}\gamma_i^m\mathcal{A}_i+\sum_{m=1}^{+\infty}a_m\sum_{j=1}^{t}\left[\frac{\alpha_j}{(-n_j)}\mathcal{B}_j^2+\mathcal{B}_j\right]^m$$
$$=a_0\left[I_n-\sum_{i=1}^{s}\mathcal{A}_i+\sum_{j=1}^{t}\frac{\mathcal{B}_j^2}{n_j}\right]+\sum_{i=1}^{s}f(\gamma_i)\mathcal{A}_i$$
$$+\sum_{j=1}^{t}\left[\frac{\sum_{m=0}^{+\infty}a_m\sum_{h=0}^{\lfloor m/2\rfloor}\binom{m}{2h}\alpha_j^{m-2h}\sqrt{-n_j}^{2h}}{(-n_j)}\right]\mathcal{B}_j^2$$
$$+\sum_{j=1}^{t}\left[\frac{\sum_{m=1}^{+\infty}a_m\sum_{h=1}^{\lfloor (m+1)/2\rfloor}\binom{m}{2h-1}\alpha_j^{m-2h+1}\sqrt{-n_j}^{2h-1}}{\sqrt{-n_j}}\right]\mathcal{B}_j.$$

Now, remembering Remark 2.14 and the definitions of the various matrices, we get the expression of $f(M)$ in the statement. Note that the expressions of $\dfrac{\mathcal{T}f(\lambda_j)}{\varphi_{\mathbb{K}}(\lambda_j)^2}$ and of $\dfrac{\mathcal{F}f(\lambda_j)}{\varphi_{\mathbb{K}}(\lambda_j)}$ are invariant under exchanging λ_j with its conjugate $\overline{\lambda}_j$. □

Examples 3.6: Assume that $(\mathbb{K},|\cdot|)$ is a valued field. Then the restriction to fundamental field $\mathbb{Q}$ of $|\cdot|$ is equivalent either to the usual euclidean absolute value (archimedean case), or to the trivial absolute value, or to a p-adic absolute value for some prime number p (see for instance [7, Ch. 23, Theorem 1, p. 44]). Hence, if the absolute value is non-archimedean, we say that the valued field $(\mathbb{K},|\cdot|)$ has *trivial fundamental restriction* or *p-adic fundamental restriction* respectively. In all cases we can define as power series, as in ordinary real case, the exponential function, the sine, the cosine, the hyperbolic sine and the hyperbolic cosine. These series have the same radius of convergence: $R = +\infty$ if the absolute value is archimedean, $R = 1$ if the absolute value has trivial fundamental restriction, and $R = \left(\frac{1}{p}\right)^{\frac{1}{p-1}}$ if the absolute value has p-adic fundamental restriction (see for instance [9, pp. 70–72]).

We put $\Lambda_{\mathbb{K}} = \Lambda_{exp,\mathbb{K}}$ (remember Definition 3.3). If $M \in \Lambda_{\mathbb{K}}$ and $\lambda = \tau_{\mathbb{K}}(\lambda) + \varphi_{\mathbb{K}}(\lambda)$ is an eigenvalue of M with its $\mathbb{K}$-decomposition, then, for $f(\lambda) = \exp(\lambda)$, we have:

$$\mathcal{T}f(\lambda) = \exp(\tau_{\mathbb{K}}(\lambda))\cosh(\varphi_{\mathbb{K}}(\lambda)) \text{ and } \mathcal{F}f(\lambda) = \exp(\tau_{\mathbb{K}}(\lambda))\sinh(\varphi_{\mathbb{K}}(\lambda))\,.$$

Now if $M \in \Lambda_{\mathbb{K}}$ is a semisimple $\mathbb{K}$-quadratic matrix, then from Proposition 3.5 we get:

$$\begin{aligned}\exp(M) = \sum_{i=1}^{s} \exp(\gamma_i)\mathcal{A}_i + \sum_{j=1}^{t} \frac{\exp(\tau_{\mathbb{K}}(\lambda_j))\cosh(\varphi_{\mathbb{K}}(\lambda_j))}{\varphi_{\mathbb{K}}(\lambda_j)^2}\mathcal{B}_j^2 \\ + \sum_{j=1}^{t} \frac{\exp(\tau_{\mathbb{K}}(\lambda_j))\sinh(\varphi_{\mathbb{K}}(\lambda_j))}{\varphi_{\mathbb{K}}(\lambda_j)}\mathcal{B}_j.\end{aligned}$$

Analogously we can get the formulas for other power series; for instance if $M \in \Lambda_{\mathbb{K}}$ is semisimple and $\mathbb{K}$-quadratic, then

$$\begin{aligned}\cos(M) = \sum_{i=1}^{s} \cos(\gamma_i)\mathcal{A}_i + \sum_{j=1}^{t} \frac{\cos(\tau_{\mathbb{K}}(\lambda_j))\cos(\varphi_{\mathbb{K}}(\lambda_j))}{\varphi_{\mathbb{K}}(\lambda_j)^2}\mathcal{B}_j^2 \\ - \sum_{j=1}^{t} \frac{\sin(\tau_{\mathbb{K}}(\lambda_j))\sin(\varphi_{\mathbb{K}}(\lambda_j))}{\varphi_{\mathbb{K}}(\lambda_j)}\mathcal{B}_j.\end{aligned}$$

4. Matrices over an ordered quadratically closed field.

REMARK 4.1: In this section we assume that $\mathbb{K}$ is an *ordered quadratically closed field*, i.e. $\mathbb{K}$ is an ordered field such that all of its positive elements have square root in $\mathbb{K}$ (for this notion we follow [6, Ch. XI, p. 462] rather than other definitions of quadratically closed field in literature).

For every $a \in \mathbb{K}$, $a > 0$, we denote by $\sqrt{a}$ its unique positive square root in $\mathbb{K}$. Moreover we fix a square root of -1 in $\overline{\mathbb{K}} \setminus \mathbb{K}$, denoted by $\sqrt{-1}$.

REMARK 4.2: It is known that an ordered quadratically closed field has characteristic 0 and it has a unique order as field (see for instance [6, Ch. XI, p. 462]).

Definition 4.3. *The field* $\mathbb{K}$ *is said to be a* real closed field*, if it can be endowed with a structure of ordered field such that its positive elements have a square root in* $\mathbb{K}$ *and every polynomial of odd degree of* $\mathbb{K}[X]$ *has a root in* $\mathbb{K}$.

Remark 4.4: It follows directly from the definitions that every real closed field is an ordered quadratically closed field. It is known that, for every ordered field $\mathbb{K}$, there exists an algebraic extension, contained in $\overline{\mathbb{K}}$, which is real closed and whose order extends the order of $\mathbb{K}$ and, moreover, that any two such extensions are $\mathbb{K}$-isomorphic (see for instance [5, Theorem 11.4] or [8, Theorem 15.9]). We call any such extension $\mathbb{L}$ a *real closure of the ordered field* $\mathbb{K}$ *in* $\overline{\mathbb{K}}$. Note that $\overline{\mathbb{K}}$ is the algebraic closure of $\mathbb{L}$ too.

For more information, further characterizations and properties of real closed fields we refer for instance to [6, Ch. XI.2] and to [8, Ch. 15]. In particular it is known that $\mathbb{K}$ is a real closed field if and only if $\sqrt{-1} \notin \mathbb{K}$ and $\mathbb{K}(\sqrt{-1})$ is algebraically closed (see for instance [8, characterization (1), p. 221]).

In Proposition 4.12 we point out other simple characterizations of real closed fields.

Proposition 4.5. *Assume that* $\mathbb{K}$ *is an ordered quadratically closed field and choose one of its real closures,* $\mathbb{L}$ *in* $\overline{\mathbb{K}}$.

a) For every $z \in \overline{\mathbb{K}}$ *there exist* $x, y \in \mathbb{L}$ *such that* $z = x + \sqrt{-1}\,y$*; such elements are uniquely determined by* $\mathbb{L}$ *and by* $\sqrt{-1}$,.

   We denote $x = \mathbf{Re}(z)$ *and* $y = \mathbf{Im}(z)$*: the* real *and the* imaginary *part of* z.

b) For every $z \in \overline{\mathbb{K}}$ *of degree* 2 *over* $\mathbb{K}$, $\mathbf{Re}(z)$ *and* $\mathbf{Im}(z)$ *are both elements of* $\mathbb{K}$ *and moreover* $\tau_{\mathbb{K}}(z) = \mathbf{Re}(z)$ *and* $\varphi_{\mathbb{K}}(z) = \sqrt{-1}\,\mathbf{Im}(z)$*; hence, in this case,* $\mathbf{Re}(z)$ *and* $\mathbf{Im}(z)$ *are independent of* $\mathbb{L}$.

Proof. Part (a) follows from Remark 4.4 since $\overline{\mathbb{K}} = \mathbb{L}(\sqrt{-1})$.

Let $z \in \overline{\mathbb{K}}$ as in (b) and write $z = \tau_{\mathbb{K}}(z) + \varphi_{\mathbb{K}}(z)$, where the minimal polynomial of $\varphi_{\mathbb{K}}(z)$ is $g(X) = X^2 - \varphi_{\mathbb{K}}(z)^2$ (remember Lemma 2.4-(d)); thus $-\varphi_{\mathbb{K}}(z)^2 > 0$, being $g(X)$ irreducible; so $\pm\sqrt{-\varphi_{\mathbb{K}}(z)^2}$ are both elements of the ordered quadratically closed field $\mathbb{K}$. Now $\varphi_{\mathbb{K}}(z) = \sqrt{-1}\,[\pm\sqrt{-\varphi_{\mathbb{K}}(z)^2}]$ and we conclude (b) by uniqueness of the decomposition in (a). □

Lemma 4.6. *Let* $(\mathbb{K}, |.|)$ *be a valued field and* $\lambda \in \overline{\mathbb{K}} \subseteq \overline{\mathbb{K}^c}$.
Then $(\mathbb{K}(\lambda))^c = \mathbb{K}^c(\lambda)$.

Proof. The element λ is algebraic over $\mathbb{K}^c$ too. Hence, by [6, Ch. XII, Proposition 2.5], we get that $\mathbb{K}^c(\lambda)$ is complete. Since it contains $\mathbb{K}(\lambda)$, it contains also its completion and this gives the first inclusion.

Now let $\vartheta \in \mathbb{K}^c(\lambda)$ and denote by l the degree of λ over $\mathbb{K}^c$. Therefore $\vartheta = \sum_{i=0}^{l-1} h_i\lambda^i$ with $h_0, \cdots, h_{l-1} \in \mathbb{K}^c$. Since $\mathbb{K}$ is dense in $\mathbb{K}^c$, there exist some sequences in $\mathbb{K}$, $\{k_m^{(i)}\}_{m\in\mathbb{N}}$, $0 \leq i \leq l-1$, such that each $k_m^{(i)}$ converges to h_i. Since the topology over $\mathbb{K}^c(\lambda)$ is the product topology (see the proof of [6, Ch. XII, Proposition 2.2]), we have that: $k_m^{(0)} + k_m^{(1)}\lambda + \cdots + k_m^{(l-1)}\lambda^{l-1}$ is a sequence in $\mathbb{K}(\lambda)$, which converges to ϑ. Thus $\vartheta \in (\mathbb{K}(\lambda))^c$. □

PROPOSITION 4.7. *Let* $(\mathbb{K}, |.|)$ *be a real closed valued field and denote by* $\mathbb{K}^c$ *its completion.*

If $\sqrt{-1} \in \mathbb{K}^c$*, then* $\mathbb{K}^c$ *is algebraically closed.*

If $\sqrt{-1} \notin \mathbb{K}^c$*, then* $\mathbb{K}^c$ *is real closed.*

Proof. By Lemma 4.6, we have: $(\overline{\mathbb{K}})^c = (\mathbb{K}(\sqrt{-1}))^c = \mathbb{K}^c(\sqrt{-1})$. Since the completion of an algebraically closed field is algebraically closed too (remember Lemma 3.2-(a)), $\mathbb{K}^c(\sqrt{-1})$ is algebraically closed. Hence, if $\sqrt{-1} \in \mathbb{K}^c$, then $\mathbb{K}^c$ is algebraically closed. Otherwise $\mathbb{K}^c$ is real closed by the characterization recalled in Remark 4.4. □

COROLLARY 4.8. *If* $(\mathbb{K}, |.|)$ *is a real closed valued field with* p*-adic fundamental restriction for some prime* p*, then its completion* $\mathbb{K}^c$ *is algebraically closed.*

Proof. Let us consider the sequence $\{x_n\}_{n\geq 1}$, $x_n = \sqrt{p^n - 1}$. Since $\mathbb{K}$ is real closed (hence ordered and quadratically closed), $x_n \in \mathbb{K}$. Now $x_n^2 + 1 = p^n \to 0$ in $(\mathbb{K}, |.|)$.

If there exists a subsequence $\{x_{n_k}\}$ such that $(x_{n_k} + \sqrt{-1}) \to 0$, then $\pm\sqrt{-1} \in \mathbb{K}^c$.

Otherwise there exists a real number $\delta > 0$ such that $|x_n + \sqrt{-1}| \geq \delta > 0$ for every n. In this case $|x_n - \sqrt{-1}| = \dfrac{|x_n^2+1|}{|x_n + \sqrt{-1}|} \leq \dfrac{|x_n^2+1|}{\delta} \to 0$. So $x_n \to \sqrt{-1}$ and again $\sqrt{-1} \in \mathbb{K}^c$. Hence, by the previous Proposition, $\mathbb{K}^c$ is algebraically closed. □

REMARK 4.9 (normalized fine $\mathbb{K}$-trace decomposition): Let $\mathbb{K}$ be an ordered quadratically closed field and $M \in M_n(\mathbb{K})$ be a semisimple $\mathbb{K}$-quadratic matrix. As remarked in Proposition 4.5 (after choosing $\sqrt{-1} \in \overline{\mathbb{K}}$), for every eigenvalue λ of M, the decomposition $\lambda = \mathbf{Re}(\lambda) + \sqrt{-1}\,\mathbf{Im}(\lambda)$ is independent of the choice of a real closure of $\mathbb{K}$ into $\overline{\mathbb{K}}$, being $\mathbf{Re}(\lambda) = \tau_{\mathbb{K}}(\lambda)$ and $\sqrt{-1}\,\mathbf{Im}(\lambda) = \varphi_{\mathbb{K}}(\lambda)$. Remembering Remark 2.12, we choose $\lambda_1, \ldots, \lambda_t$, the eigenvalues of M not in $\mathbb{K}$, so that $\mathbf{Im}(\lambda_j) > 0$ for every j, and we define the matrices:

$$\mathbf{A}_i = \mathcal{A}_i \text{ for every } i = 1, \cdots, s,,$$

$$\text{and} \qquad \mathbf{B}_j = \frac{\mathcal{B}_j}{\mathbf{Im}(\lambda_j)} \text{ for every } j = 1, \cdots, t\,.$$

They are in $M_n(\mathbb{K}) \setminus \{0\}$ and are polynomial expressions of M. Moreover

$$\begin{aligned}
&\mathbf{A}_i\mathbf{A}_j = \delta_{ij}\mathbf{A}_i \text{ for every } i, j\,;\\
&\mathbf{A}_i\mathbf{B}_j = \mathbf{B}_j\mathbf{A}_i = 0 \text{ for every } i, j\,;\\
&\mathbf{B}_i\mathbf{B}_j = 0 \text{ for every } i \neq j \text{ and}\\
&\mathbf{B}_j^3 = -\mathbf{B}_j \text{ for every } j\,.
\end{aligned}$$

Then by Remark 2.14 we get

$$M = \sum_{i=1}^{s} \gamma_i\, \mathbf{A}_i - \sum_{j=1}^{t} \mathbf{Re}(\lambda_j)\, \mathbf{B}_j^2 + \sum_{j=1}^{t} \mathbf{Im}(\lambda_j)\, \mathbf{B}_j \quad (\text{with } \mathbf{Im}(\lambda_j) > 0). \tag{3}$$

We call the above decomposition *normalized fine* $\mathbb{K}$*-trace decomposition of* M.

PROPOSITION 4.10. *Assume that* $(\mathbb{K}, |\cdot|)$ *is an ordered quadratically closed valued field, that* $f(X)$ *is a power series with coefficients in* $\mathbb{K}$ *and that* $M \in \Lambda_{f,\mathbb{K}}$ *is a semisimple* $\mathbb{K}$*-quadratic matrix with normalized fine* $\mathbb{K}$*-trace decomposition:*

$$M = \sum_{i=1}^{s} \gamma_i\, \mathbf{A}_i - \sum_{j=1}^{t} \mathbf{Re}(\lambda_j)\, \mathbf{B}_j^2 + \sum_{j=1}^{t} \mathbf{Im}(\lambda_j)\, \mathbf{B}_j, \quad \textit{with } \mathbf{Im}(\lambda_j) > 0.$$

Then

a) $f(M) = \sum_{i=1}^{s} f(\gamma_i)\mathbf{A}_i - \sum_{j=1}^{t} \mathcal{T}f(\lambda_j)\, \mathbf{B}_j^2 + \sum_{j=1}^{t} \mathcal{G}f(\lambda_j)\mathbf{B}_j$,

where $f(\gamma_i)$*,* $\mathcal{T}f(\lambda_j)$ *and* $\mathcal{G}f(\lambda_j) := \dfrac{\mathcal{F}f(\lambda_j)}{\sqrt{-1}}$ *belong to* $\mathbb{K}^c$ *for every* i, j*;*

b) if $\mathbb{K}^c$ *is ordered quadratically closed too , then*

$$f(M) = \sum_{i=1}^{s} f(\gamma_i)\mathbf{A}_i - \sum_{j=1}^{t} \mathbf{Re}(f(\lambda_j))\, \mathbf{B}_j^2 + \sum_{j=1}^{t} \mathbf{Im}(f(\lambda_j))\mathbf{B}_j,$$

where $f(\gamma_i)$*,* $\mathbf{Re}(f(\lambda_j))$*,* $\mathbf{Im}(f(\lambda_j)) \in \mathbb{K}^c$ *for every* i, j*.*

Proof. Part (a) follows from Proposition 3.5-(a). Indeed it suffices to remark that

$$\mathcal{G}f(\lambda_j)\mathbf{B}_j = \frac{\mathcal{F}f(\lambda_j)}{\sqrt{-1}} \frac{\mathcal{B}_j}{\mathbf{Im}(\lambda_j)} = \frac{\mathcal{F}f(\lambda_j)}{\varphi_{\mathbb{K}}(\lambda_j)}\, \mathcal{B}_j\,.$$

From the expression of $\mathcal{F}f(\lambda_j)$ it follows that $\mathcal{G}f(\lambda_j) \in \mathbb{K}^c$.

Part (b) follows from part (a), from Remark 3.4-(c) and from Proposition 4.5-(b), since, for every j, $f(\lambda_j)$ has degree at most 2 over $\mathbb{K}^c$ and $\lambda_j \notin \mathbb{K}^c$ (because $\varphi_{\mathbb{K}}(\lambda_j)^2 < 0$ and $\mathbb{K}^c$ is an ordered field). □

EXAMPLES 4.11: Assume that $(\mathbb{K}, |\cdot|)$ is an ordered quadratically closed valued field. If $M \in \Lambda_{\mathbb{K}}$ is semisimple and $\mathbb{K}$-quadratic, then, by Proposition 4.10 and remarking that $\mathcal{T}\exp(\lambda_j) = \exp(\mathbf{Re}(\lambda_j))\cos(\mathbf{Im}(\lambda_j))$ and $\mathcal{G}\exp(\lambda_j) = \exp(\mathbf{Re}(\lambda_j))\sin(\mathbf{Im}(\lambda_j))$, we have:

$$\exp(M) = \sum_{i=1}^{s} \exp(\gamma_i)\mathbf{A}_i - \sum_{j=1}^{t} \exp(\mathbf{Re}(\lambda_j))\cos(\mathbf{Im}(\lambda_j))\mathbf{B}_j^2 + \sum_{j=1}^{t} \exp(\mathbf{Re}(\lambda_j))\sin(\mathbf{Im}(\lambda_j))\mathbf{B}_j,$$

and, analogously,

$$\cos(M) = \sum_{i=1}^{s} \cos(\gamma_i)\mathbf{A}_i - \sum_{j=1}^{t} \cos(\mathbf{Re}(\lambda_j))\cosh(\mathbf{Im}(\lambda_j))\mathbf{B}_j^2 - \sum_{j=1}^{t} \sin(\mathbf{Re}(\lambda_j))\sinh(\mathbf{Im}(\lambda_j))\mathbf{B}_j,$$

where the λ_j's are the eigenvalues of M not in $\mathbb{K}$, having positive imaginary part.

The previous formulas point out the analogous formulas in Examples 3.6. Moreover the expression of $\exp(M)$ extends the classical Rodrigues' formula for the exponential of a real skew symmetric matrix (see for instance [3, Theorem 2.2]).

PROPOSITION 4.12. *If char*$(\mathbb{K}) = 0$*, the following assertions are equivalent:*

a) $\mathbb{K}$ *is real closed;*

b) $\mathbb{K}$ *is not algebraically closed and the irreducible polynomials of* $\mathbb{K}[X]$ *have degree at most* 2*;*

c) the $\mathbb{K}$*-linear involution of* $\overline{\mathbb{K}}$ *is an element of Aut*$(\overline{\mathbb{K}}/\mathbb{K})$ *different from the identity;*

d) Ker$(\tau_{\mathbb{K}})$ *is the* $\mathbb{K}$*-vector space generated by* $\sqrt{-1}$*;*

e) $\mathbb{K}$ *is not algebraically closed and every semisimple matrix with entries in* $\mathbb{K}$ *is* $\mathbb{K}$*-quadratic.*

Proof. The implication (a) $\Rightarrow$ (b) follows from Remark 4.4. For the converse it suffices to prove that $\overline{\mathbb{K}} = \mathbb{K}(\sqrt{-1})$, since $\mathbb{K}$ is not algebraically closed. Let $t \in \overline{\mathbb{K}} \setminus \mathbb{K}$, thus it has degree 2. We decompose $t = \alpha + \beta$ as sum of an element $\alpha \in \mathbb{K}$ and of an element $\beta \in Ker(\tau_{\mathbb{K}}) \setminus \{0\}$. By Lemma 2.4-(a) the conjugate of t is $\bar{t} = \alpha - \beta$ and $\beta^2 \in \mathbb{K}$. Now we consider the polynomial of $\mathbb{K}[X]$: $q(X) = X^4 - \beta^2 = (X - \sqrt{\beta})(X + \sqrt{\beta})(X - \sqrt{-\beta})(X + \sqrt{-\beta})$ with its factorization in $\overline{\mathbb{K}}[X]$ (note that its roots are not in $\mathbb{K}$). Since $q(X)$ has degree 4, it is reducible over $\mathbb{K}$, so it is product of two irreducible polynomials of $\mathbb{K}[X]$. Since $\beta \notin \mathbb{K}$, one of the two factors must have the form $(X - \sqrt{\beta})(X \pm \sqrt{-\beta})$ and therefore $\sqrt{-\beta^2} \in \mathbb{K} \setminus \{0\}$. Hence $\beta = \pm\sqrt{-\beta^2}\sqrt{-1} \in \mathbb{K}(\sqrt{-1})$. This implies that $t = \alpha + \beta \in \mathbb{K}(\sqrt{-1})$, therefore $\overline{\mathbb{K}} \setminus \mathbb{K} \subseteq \mathbb{K}(\sqrt{-1})$ and so $\overline{\mathbb{K}} = \mathbb{K}(\sqrt{-1})$.

By Lemma 2.4-(b), (a) implies (c). Now assume (c), so that $\mathbb{K}$ is not algebraically closed. Let $\lambda = \alpha + \beta \in \overline{\mathbb{K}}$ with its $\mathbb{K}$-decomposition. In particular $\mathbb{K}(\lambda) = \mathbb{K}(\beta)$. From (c) we have: $\overline{\beta^2} = \overline{\beta}^2 = (-\beta)^2 = \beta^2$. Hence, by Remark 2.3, we get that $\beta^2 \in \mathbb{K}$ and so both β and λ have degree at most 2 over $\mathbb{K}$. This gives (b).

Next we prove the equivalence between (a) and (d). Assume first (d). By Remark 2.1, $\overline{\mathbb{K}} = \mathbb{K} \oplus Ker(\tau_{\mathbb{K}}) = \mathbb{K} \oplus Span(\sqrt{-1})$, thus $\sqrt{-1} \notin \mathbb{K}$ and $\mathbb{K}(\sqrt{-1}) = \overline{\mathbb{K}}$ is algebraically closed. For the converse, every element in $\overline{\mathbb{K}} \setminus \mathbb{K}$ is algebraic of order 2 over $\mathbb{K}$. Since $\mathbb{K}$ is ordered quadratically closed too, by Lemma 2.4-(d), $\beta \in Ker(\tau_{\mathbb{K}})$ if and only if $\beta = \pm\sqrt{t}$ with $t \in \mathbb{K}$ and $t \leq 0$, i.e. if and only if $\beta = \pm\sqrt{-t}\,\sqrt{-1}$ with $\sqrt{-t} \in \mathbb{K}$ and this allows to conclude.

Now (b) implies (e) by obvious reasons. For the converse, we note that every monic irreducible polynomial $q(X) \in \mathbb{K}[X]$ with $deg(q(X)) \geq 2$ is the minimal polynomial of its *companion matrix*, which is therefore semisimple and so $\mathbb{K}$-quadratic, by (e). Since $q(X)$ is irreducible, we get that $deg(q(X)) = 2$. $\square$

REMARK 4.13: Assume that $(\mathbb{K}, |\cdot|)$ is a real closed valued field and that $f(X)$ is a power series with coefficients in $\mathbb{K}$, then (see Proposition 4.12-(e)) the formula of Proposition 4.10-(a) (and possibly of Proposition 4.10-(b)) holds for every semisimple matrix $M \in \Lambda_{f,\mathbb{K}}$. In particular, the same formulas of Examples 4.11 hold for every semisimple matrix $M \in \Lambda_{\mathbb{K}}$.

Acknowledgements

This research was partially supported by GNSAGA-INdAM (Italy).

References

[1] A. Dolcetti and D. Pertici, *Some remarks on the Jordan-Chevalley decomposition*, São Paulo J. Math. Sci. **11(2)** (2017), 385–404.

[2] A. Dolcetti and D. Pertici, *Skew symmetric logarithms and geodesics on* $O_n(\mathbb{R})$, Adv. Geom. **18(4)** (2018), 495–507.

[3] J. Gallier and D. Xu, *Computing exponentials of skew-symmetric matrices and logarithms of orthogonal matrices*, Internat. J. Robotics Automation **17(4)** (2002), 10–20.

[4] K. Hoffman and R. Kunze, *Linear Algebra*, second ed., Prentice-Hall Inc., Englewood Cliffs, 1971.
[5] N. Jacobson, *Basic algebra II*, second ed., W.H. Freeman and Company, New York, 1989.
[6] S. Lang, *Algebra*, revised third ed., GTM 211, Springer, New York, 2002.
[7] F. Lorenz, *Algebra. Volume II: fields with structure, algebras and advanced topics*, Universitext, Springer, New York, 2008.
[8] A. R. Rajwade, *Squares*, Lecture Note Series 171, Cambridge University Press, Cambridge, 1993.
[9] W. H. Schikhof, *Ultrametric calculus. an introduction to p-adic analysis*, Cambridge Stud. Adv. Math. 4, Cambridge University Press, Cambridge, 1984.
[10] S. Warner, *Topological fields*, Mathematics Studies 157, North Holland, Amsterdam, 1989.

Authors' addresses:

Alberto Dolcetti
Università degli Studi di Firenze
Dipartimento di Matematica e Informatica "Ulisse Dini"
Viale Morgagni 67/a
50134 Firenze, Italy
E-mail: alberto.dolcetti@unifi.it

Donato Pertici
Università degli Studi di Firenze
Dipartimento di Matematica e Informatica "Ulisse Dini"
Viale Morgagni 67/a
50134 Firenze, Italy
E-mail: donato.pertici@unifi.it

Received July 5, 2018
Accepted July 16, 2018

Rend. Istit. Mat. Univ. Trieste
Volume 50 (2018), 65–79
DOI: 10.13137/2464-8728/21598

Veronesean almost binomial almost complete intersections

Thomas Kahle and André Wagner

Abstract. *The second Veronese ideal I_n contains a natural complete intersection J_n of the same height, generated by the principal 2-minors of a symmetric $(n \times n)$-matrix. We determine subintersections of the primary decomposition of J_n where one intersectand is omitted. If I_n is omitted, the result is a direct link in the sense of complete intersection liaison. These subintersections also yield interesting insights into binomial ideals and multigraded algebra. For example, if n is even, I_n is a Gorenstein ideal and the intersection of the remaining primary components of J_n equals $J_n + \langle f \rangle$ for an explicit polynomial f constructed from the fibers of the Veronese grading map.*

Keywords: Veronese, complete intersection, binomial ideal, multigrading.
MS Classification 2010: 05E40, 13A02, 13H10, 14M25, 52B20.

1. Introduction

Ideals generated by minors of matrices are a mainstay of commutative algebra. Here we are concerned with ideals generated by 2-minors of symmetric matrices. Ideals generated by arbitrary minors of symmetric matrices have been studied by Kutz [18] who proved, in the context of invariant theory, that the quotient rings are Cohen–Macaulay. Results of Goto show that the quotient ring is normal with divisor class group $\mathbb{Z}_2$ and Gorenstein if the format of the symmetric matrix has the same parity as the size of the minors [11, 12]. Conca extended these results to more general sets of minors of symmetric matrices [4] and determined Gröbner bases and multiplicity [5].

Here we are concerned only with the binomial ideal I_n generated by the 2-minors of a symmetric $(n \times n)$-matrix. This ideal cuts out the second Veronese variety and was studied classically, for example by Gröbner [15]. It contains a complete intersection J_n, generated by the principal 2-minors (Definition 2.2). Both ideals are of height $\binom{n}{2}$. Coming from liaison theory one may ask for the ideal $K_n = J_n : I_n$ on the other side of the complete intersection link via J_n. In this paper we determine K_n.

EXAMPLE 1.1: Consider the ideal $J_3 = \langle ad-b^2, af-c^2, df-e^2\rangle \subseteq \mathbb{Q}[a,b,c,d,e,f]$ generated by the principal 2-minors of the symmetric matrix $\begin{pmatrix} a & b & c \\ b & d & e \\ c & e & f \end{pmatrix}$. The ideal J_3 is a complete intersection because it has an initial ideal with this property. Using, for example, MACAULAY2 [14], one finds the prime decomposition $J_3 = I_3 \cap K_3$ where

$$I_3 = J_3 + \langle ae - bc, cd - be, ce - bf\rangle$$

is the second Veronese ideal, generated by all 2-minors, and

$$K_3 = J_3 + \langle ae + bc, cd + be, ce + bf\rangle$$

is the image of I_3 under the automorphism of $\mathbb{Q}[a,\dots,f]$ that maps b, c, and e to their negatives and the remaining indeterminates to themselves. As predicted by Theorem 2.11, the generator $ae+bc$ is the sum of monomials whose exponents are the lattice points of the fiber

$$\left\{ u \in \mathbb{N}^6 : \begin{pmatrix} 2 & 1 & 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 2 & 1 & 0 \\ 0 & 0 & 1 & 0 & 1 & 2 \end{pmatrix} \cdot u = \begin{pmatrix} 2 \\ 1 \\ 1 \end{pmatrix} \right\}$$

of the $\mathbb{Z}$-linear map V_3 that defines the fine grading of $\mathbb{Q}[a,\dots,f]/I_3$. We call this the *generating function* of the fiber. For $n \geq 4$ the extra generators are not binomials anymore and K_n is an intersection of ideals obtained from I_n by twisting automorphisms (Definition 2.7). In Example 2.12, for $n = 4$, we find $K_4 = J_4 + \langle p\rangle$ for one quartic polynomial p with eight terms.

Results on Gorenstein biliaison of ideals of minors of symmetric matrices have been obtained by Gorla [9, 10] but here we study direct complete intersection links. Our methods rely on the combinatorics of binomial ideals and since K_n is not binomial and we do not know of a natural binomial complete intersection contained in K_n, we cannot explore the linkage class more with the present method. Instead we are motivated by general questions about binomial ideals and their intersections. For example, [17, Problem 17.1] asks, when the intersection of binomial ideals is binomial. From the primary (in fact, prime) decomposition of J_n we remove I_n and intersect the remaining binomial prime ideals. The result is not binomial. If n is even, $K_n = J_n + \langle p\rangle$ for one additional polynomial p. In the terminology of [1], K_n is thus an *almost complete intersection*. It is also *almost binomial*, as it is principal modulo its binomial part—the binomial ideal spanned over $\Bbbk$ by all binomials in the ideal [16, Definition 2.1]. If n is odd, then there are n additional polynomials (Theorem 2.11). While these numbers can be predicted from general liaison theory, our explicit formulas reveal interesting structures at the boundary of binomiality and are thus a first step towards [17, Problem 17.1]

We determine K_n with methods from combinatorial commutative algebra, multigradings in particular (see [19, Chapters 7 and 8]). The principal observation that drives the proofs in Section 2 is that the Veronese-graded Hilbert function of the quotient $\Bbbk[\mathbf{x}]/J_n$ becomes eventually constant (Remark 2.14). The eventual value of the Hilbert function bounds the number of terms that a graded polynomial can have. The extra generators of K_n are the lowest degree polynomials that realize the bound. We envision that this structure could be explored independently and brought to unification with the theory of toral modules from [7]. Our results also have possible extensions to higher Veronese ideals as we outline in Section 3.

Denote by $c_n := \binom{n}{2}$ the entries of the second diagonal in Pascal's triangle. Throughout, let $[n] := \{1, \ldots, n\}$ be the set of the first n integers. The second Veronese ideal lives in the polynomial ring $\Bbbk[\mathbb{N}^{c_{n+1}}]$ in c_{n+1} indeterminates over a field $\Bbbk$. For polynomial rings and quotients modulo binomial ideals we use monoid algebra notation (see, for instance, [17, Definition 2.15]). We make no a-priori assumptions on $\Bbbk$ regarding its characteristic or algebraic closure, although care is necessary in characteristic two. The variables of $\Bbbk[\mathbb{N}^{c_{n+1}}]$ are denoted x_{ij}, for $i, j \in [n]$ with the implicit convention that $x_{ij} = x_{ji}$. For brevity we avoid a comma between i and j. We usually think about upper triangular matrices, that is $i \leq j$. The Veronese ideal I_n is the toric ideal of the *Veronese multigrading* $\mathbb{N}V_n$, defined by the $(n \times c_{n+1})$-matrix V_n with entries

$$(V_n)_{i,jk} := \begin{cases} 2 & \text{if } i = j = k, \\ 1 & \text{if } i = j, \text{ or } i = k, \text{ but not both,} \\ 0 & \text{otherwise.} \end{cases}$$

That is, the columns of V_n are the non-negative integer vectors of length n and weight two. For $\mathbf{b} \in \mathbb{N}V_n$, the *fiber* is $V_n^{-1}[\mathbf{b}] = \{u \in \mathbb{N}^{c_{n+1}} : V_n u = \mathbf{b}\}$. Computing the V_n-degree of a monomial is easy: just count how often each row or column index appears in the monomial. For example, $\deg(x_{12}x_{nn}) = (1, 1, 0, \ldots, 0, 2)$. We do not distinguish row and column vectors notationally, in particular we write columns as rows when convenient. Gröbner bases for a large class of toric ideals including I_n have been determined by Sturmfels [20, Theorem 14.2]. The *Veronese lattice* $L_n \subseteq \mathbb{Z}^{c_{n+1}}$ is the kernel of V_n. The rank of L_n is c_n since the rank of V_n is n and $c_{n+1} - n = c_n$. Lemma 2.1 gives a lattice basis. With $\{e_{ij}, i \leq j \in [n]\}$ a standard basis of $\mathbb{Z}^{c_{n+1}}$, we use the following notation

$$[ij|kl] := e_{ik} + e_{jl} - e_{il} - e_{jk} \in \mathbb{Z}^{c_{n+1}}.$$

Then $[ij|kl]$ is the exponent vector of the minor $x_{ik}x_{jl} - x_{il}x_{jk}$.

Example 1.2: The Veronese lattice $L_3 \subseteq \mathbb{Z}^6$ is of rank $3 = \binom{4}{2} - 3$ and minimally generated by the following elements

$$[13|13] = \begin{pmatrix} 1 & 0 & -2 \\ & 0 & 0 \\ & & 1 \end{pmatrix}, \; [13|23] = \begin{pmatrix} 0 & 1 & -1 \\ & 0 & -1 \\ & & 1 \end{pmatrix}, \; [23|23] = \begin{pmatrix} 0 & 0 & 0 \\ & 1 & -2 \\ & & 1 \end{pmatrix}.$$

1.1. Acknowledgement

The authors are grateful to Aldo Conca for posing the question of determining the link of the Veronese variety through its complete intersection of principal minors and valuable comments on an early version of the manuscript. During the initial stages of this work, Thomas Kahle was supported by the research focus dynamical systems of the state Saxony-Anhalt. André Wagner's research is carried out in the framework of Matheon supported by Einstein Foundation Berlin.

2. Decomposing and Recomposing

Lemma 2.1. *The set*

$$\mathcal{B} = \{[in|jn] : i, j \in [n-1]\}$$

is a lattice basis of the Veronese lattice L_n.

Proof. Write the elements of $\mathcal{B}$ as the columns of a $(c_{n+1} \times c_n)$-matrix B. Deleting the rows corresponding to indices (i, n) for $i \in [n]$ yields the identity matrix I_{c_n}. Thus $\mathcal{B}$ spans a lattice of the correct rank and that lattice is saturated. Indeed, the Smith normal form of B must equal the identity matrix I_{c_n} concatenated with a zero matrix. Thus the quotient by the lattice spanned by $\mathcal{B}$ is free. □

The Veronese ideal contains a codimension c_n complete intersection J_n generated by the principal 2-minors.

Definition 2.2. *The* principal minor ideal J_n *is generated by all principal* 2*-minors* $x_{ii}x_{jj} - x_{ij}^2$ *of a generic symmetric matrix. The* principal minor lattice L'_n *is the lattice generated by the corresponding exponent vectors* $[ij|ij]$, $i, j \in [n]$.

It can be seen that the principal minor lattice is minimally generated by $[ij|ij]$. It is an unsaturated lattice meaning that it cannot be written as the kernel of an integer matrix, or equivalently, that the quotient $\mathbb{Z}^{c_{n+1}}/L'_n$ has torsion. Since there are no non-trivial coefficients on the binomials in J_n, Proposition 2.5 below says that it is a lattice ideal with lattice L'_n. Its torsion subgroup is given in the following proposition.

Proposition 2.3. *The principal minor lattice is minimally generated by*

$$\mathcal{B}' = \{2[in|jn] : i \neq j \in [n-1]\} \cup \{[in|in] : i \in [n-1]\}.$$

Furthermore the group L_n/L'_n *is (isomorphic to)* $(\mathbb{Z}/2\mathbb{Z})^{c_{n-1}}$.

Proof. It holds that $2[in|jn] = [in|in] + [jn|jn] - [ij|ij]$ and the map which includes the span of the elements $[ij|ij]$ into L'_n is unimodular. A presentation of the group can be read off the Smith normal form of the matrix whose columns are a lattice basis. Since $\mathcal{B}'$ is a basis of L'_n, the columns and rows can be arranged so that the diagonal matrix $\operatorname{diag}(2,\ldots,2,1,\ldots,1)$ with c_{n-1} entries 2 is the top $(c_n \times c_n)$-matrix of the Smith normal form. Any entry below a two is divisible by two and thus row operations can be used to zero out the the bottom part of the matrix. This yields the Smith normal form. □

Example 2.4: For $n = 3$, the basis $\mathcal{B}'$ is given in matrix notation as

$$\begin{pmatrix} 0 & 2 & -2 \\ & 0 & -2 \\ & & 2 \end{pmatrix}, \quad \begin{pmatrix} 1 & 0 & -2 \\ & 0 & 0 \\ & & 1 \end{pmatrix}, \quad \begin{pmatrix} 0 & 0 & 0 \\ & 1 & -2 \\ & & 1 \end{pmatrix}.$$

The advantage of $\mathcal{B}'$ over the basis in Definition 2.2 is that the transition matrix from $\mathcal{B}$ to $\mathcal{B}'$ is diagonal. This makes it easy to understand the quotient of the Veronese lattice modulo the principal minor lattice.

If $\operatorname{char}(\mathbb{k}) = 2$, then J_n is primary over I_n. In all other characteristics one can see that the Veronese ideal I_n is a minimal prime and in fact a primary component of J_n. These statements follow from [8] and are summarized in Proposition 2.10 below. Towards this observation, the next proposition says that J_n is a mesoprime ideal, that is, it equals the kernel of a monomial $\mathbb{k}$-algebra homorphism from $\mathbb{k}[\mathbb{N}^{c_{n+1}}]$ to a twisted group algebra [17, Definition 10.4]. The adjective twisted implements the general coefficients on the binomials in [17]. Here all coefficients are equal to $1_{\mathbb{k}}$. The ideal J_n is a lattice ideal as a kernel of a monomial homomorphism onto an ordinary group algebra.

Proposition 2.5. *J_n is a mesoprime binomial ideal and its associated lattice is L'_n.*

Proof. We show that $J_n = \langle x^{u^+} - x^{u^-} : u \in L'_n \rangle$, since the quotient by this ideal is the group algebra $\mathbb{k}[\mathbb{Z}^{c_{n+1}}/L'_n]$. By the correspondence between non-negative lattice walks and binomial ideals [6, Theorem 1.1] we prove that for any $u = u^+ - u^- \in L'_n$, the parts $u^+, u^- \in \mathbb{N}^{c_{n+1}}$ can be connected using *moves* $[ij|ij]$ without leaving $\mathbb{N}^{c_{n+1}}$.

The vectors u^+, u^- can be represented by upper triangular non-negative integer matrices. From Definition 2.2 it is obvious that all off-diagonal entries of $u^+ - u^-$ are divisible by two. Since

$$\langle x^{u^+} - x^{u^-} : u \in L'_n \rangle : x_{ij} = \langle x^{u^+} - x^{u^-} : u \in L'_n \rangle$$

for any variable x_{ij}, we can assume that u^+ and u^- have disjoint supports and thus individually have off-diagonal entries divisible by two. Consequently the moves $[ij|ij]$ allow to reduce all off-diagonal entries to zero, while increasing the diagonal entries. As visible from its basis, the lattice L'_n contains no nonzero diagonal matrices and thus u^+ and u^- have been connected to the same diagonal matrix. $\square$

Remark 2.6: From Proposition 2.3 it follows immediately that the group algebra $\Bbbk[\mathbb{Z}^{c_{n+1}}]/J_n\Bbbk[\mathbb{Z}^{c_{n+1}}]$ is isomorphic to $\Bbbk[\mathbb{Z}^n \oplus (\mathbb{Z}/2\mathbb{Z})^{c_{n-1}}]$. In particular $\Bbbk[\mathbb{N}^{c_{n+1}}]/J_n$ is finely graded by the monoid $\mathbb{N}V_n \oplus (\mathbb{Z}/2\mathbb{Z})^{c_{n-1}}$.

Definition 2.7. *A $\mathbb{Z}_2$-*twisting *is a ring automorphism of a (Laurent) polynomial ring that maps the indeterminates either to themselves or to their negatives.*

A *fundamental parallelepiped* of the lattice L'_n is the quotient $\mathbb{Q}^{c_{n+1}}/L'_n$, embedded as a half-open polytope in $\mathbb{Q}^{c_{n+1}}$. The lattice points in it play an important role in the following developments. The most succinct way to encode them is using their generating function, a Laurent polynomial in the ring $\Bbbk[\mathbb{Z}^{c_{n+1}}]$. Its explicit form depends on the chosen coordinates. The next lemma is immediate from the definition of $\mathcal{B}'$.

Lemma 2.8. *Let $M = \{[in|jn] : i \neq j \in [n-1]\}$. The generating function of the fundamental parallelepiped of $\mathcal{B}'$ is*

$$p_n = \prod_{m\in M} (x^m + 1) = \prod_{m\in M} (x^{m^+} + x^{m^-})$$

It is useful for the further development to pick the second representation of p_n in Lemma 2.8 as a representative of p_n in polynomial ring $\Bbbk[\mathbb{N}^{c_{n+1}}]$. Its image in the quotient $\Bbbk[\mathbb{N}^{c_{n+1}}]/J_n$ also has a natural representation. The terms of p_n can be identified with upper triangular integer matrices which arise as sums of positive and negative parts of elements of M. A positive part of $[in|jn] \in M$ has entries 1 at positions (i,j) and (n,n) while a negative part has two entries 1 in the last column, but not at (n,n). Modulo the moves $\mathcal{B}'$, any exponent matrix of a monomial of p_n can be reduced to have only entries 0 or 1 in its off-diagonal positions.

Remark 2.9: A simple count yields that p_n has V_n-degree $(n-2,\dots,n-2,2c_{n-1})$. In the natural representation of monomials of p_n as integer matrices with entries $0/1$ off the diagonal, there is a lower bound for the value of the (n,n) entry. To achieve the lowest value, one would fill the last column with entries 1 using negative parts of elements of M, and then use positive parts (which increase (n,n)). For example, if n is even, there is one term of p_n whose last column arises from the negative parts of $[1n|2n], [3n|4n], \dots, [(n-3)n|(n-2)n]$ and then positive parts of the remaining elements of M. If n is odd, then there is one term of p_n, whose n-th column is $(1, \ldots, 1, \sigma_{n-1})$ for some value σ_{n-1}. In fact, since $|M| = c_{n-1}$, the lowest possible value of the (n, n) entry is $\sigma_{n-1} = c_{n-1} - \lfloor \frac{n-1}{2} \rfloor$.

The primary decomposition of J_n is given by [8, Theorem 2.1 and Corollary 2.2].

PROPOSITION 2.10. *If* $\operatorname{char}(\mathbb{k}) = 2$*, the* J_n *is primary over* I_n*. In all other characteristics, there exist* $\mathbb{Z}_2$*-twistings* ϕ_i *for* $i = 1, \ldots, 2^{c_{n-1}}$ *such that the complete intersection* J_n *has prime decomposition*

$$J_n = \bigcap_i \phi_i(I_n). \tag{1}$$

THEOREM 2.11. *If* n *is odd, intersecting all but one of the components in* (1) *yields*

$$\bigcap_{i \neq l} \phi_i(I_n) = J_n + \langle \phi_l(p^+_{n,i}) : i \in [n] \rangle,$$

where $p^+_{n,i} \in \mathbb{k}[\mathbb{N}^{c_{n+1}}]$ *are homogeneous polynomials of degree* $\frac{(n-1)^2}{2}$ *that are given as generating functions of the fibers* $V_n^{-1}[(n-2, \ldots, n-2) + e_i]$*. If* n *is even, then the same holds for a single polynomial* p^+_n *of degree* $\frac{n(n-2)}{2}$*, given as the generating function of* $V_n^{-1}[(n-2, \ldots, n-2)]$*.*

The proof of Theorem 2.11 occupies the remainder of the section after the following example.

EXAMPLE 2.12: The complete intersection J_4 is a lattice ideal for the lattice L'_4. In the basis $\mathcal{B}'$, it is generated by the six elements

$$\{2[i4|j4] : i < j \in [3]\} \cup \{[i4|i4] : i \in [3]\}.$$

Three of the six elements correspond to principal minors

$$x_{11}x_{44} - x_{14}^2, x_{22}x_{44} - x_{24}^2, x_{33}x_{44} - x_{34}^2.$$

The other elements give the binomials

$$x_{12}^2x_{44}^2 - x_{14}^2x_{24}^2, x_{13}^2x_{44}^2 - x_{14}^2x_{34}^2, x_{23}^2x_{44}^2 - x_{24}^2x_{34}^2.$$

These six binomials do not generate J_4, but J_4 equals the saturation with respect to the product of the variables [19, Lemma 7.6]. The $2^3 = 8$ minimal prime components of J_n are obtained by all possible twist combinations of the monomials $\pm x_{14}x_{24}$, $\pm x_{14}x_{34}$, $\pm x_{24}x_{34}$. Consider the mysterious polynomial

$$p_4 = (x_{12}x_{44} + x_{14}x_{24})(x_{13}x_{44} + x_{14}x_{34})(x_{23}x_{44} + x_{24}x_{34}),$$

which is the generating function of the fundamental parallelepiped of L_4' in the basis $\mathcal{B}'$ and of V_4-degree $(2,2,2,6)$. *In the Laurent polynomial ring* $\Bbbk[\mathbb{Z}^{10}]$, the desired ideal $J_4 : I_4$ equals $J_4 + \langle p_4 \rangle$. To do the computation in the polynomial ring, we need to saturate with respect to $\prod_{ij} x_{ij}$. If n is even, this saturation generates one polynomial, if n is odd, it generates n polynomials. Here, where $n = 4$, the ideal $J_4 : I_4$ is generated by J_4 and the single polynomial

$$\begin{aligned} p_4^+ = &\, x_{11}x_{22}x_{33}x_{44} + x_{11}x_{23}x_{24}x_{34} + x_{13}x_{14}x_{22}x_{34} + x_{12}x_{14}x_{24}x_{33} \\ &+ x_{13}x_{14}x_{23}x_{24} + x_{12}x_{14}x_{23}x_{34} + x_{12}x_{13}x_{24}x_{34} + x_{12}x_{13}x_{23}x_{44}. \end{aligned}$$

Modulo the binomials in J_4, the polynomial p_4^+ equals p_4/x_{44}^2 (Lemma 2.18).

As a first step towards the proof of Theorem 2.11 we compute the monoid Q under which $\Bbbk[\mathbb{N}^{c_{n+1}}]/J_n$ is finely graded, meaning that its Q-graded Hilbert function takes values only zero or one. That is, we make Remark 2.6 explicit.

LEMMA 2.13. *Fix* $\mathbf{b} \in \operatorname{cone}(V_n)$ *for some* n. *The equivalence classes of lattice points in the fiber* $V_n^{-1}[\mathbf{b}]$, *modulo the moves* $\mathcal{B}'$, *are in bijection with set of symmetric 0/1 matrices* $u \in \{0,1\}^{n\times n}$ *of the following form*

- $u_{ii} = 0$, *for all* $i \in [n]$
- $u_{in} = 0$, *for all* $i \in [n]$
- $\mathbf{b} - V_n u \in \mathbb{N}^n$.

Proof. Each equivalence class of upper triangular matrices has a representative whose off-diagonal entries are all either zero or one. The bijection maps such an equivalence class to the c_{n-1} entries that are off-diagonal and off the last column. To prove that this is a bijection it suffices to construct the inverse map. To this end, let u satisfy the properties in the statement. In each row $i = 1, \dots, n$, there are two values unspecified: the diagonal entry and the entry in the last column. Given $\mathbf{b}_i$, using the representative modulo $\mathcal{B}'$ whose last column entries are either 0 or 1 fixes the diagonal entry by linearity. Therefore the map is a bijection. □

REMARK 2.14: If $\mathbf{b}_i \geq (n-2)$ for all $i \in [n]$, then any 0/1 upper triangular $(n-2)$-matrix is a possible choice for the off-diagonal off-last column entries of u in Lemma 2.13. An upper triangular $(n-2)$-matrix has c_{n-1} entries. Thus all those fibers have equivalence classes modulo $\mathcal{B}'$ that are in bijection with $\{0,1\}^{c_{n-1}}$. In particular, each of those fibers, has the same number of equivalence classes.

REMARK 2.15: Remark 2.14 implies that in the V_n-grading, $\Bbbk[\mathbb{N}^{c_{n+1}}]/J_n$ is *toral* as in [7, Definition 4.3]: its V_n-graded Hilbert function is globally bounded by $2^{c_{n-1}}$.

If n is odd, then $(n-2,\dots,n-2) \notin \mathbb{N}V_n$. Therefore the minimal (with respect to addition in the semigroup $\operatorname{cone}(V_n)$) fibers that satisfy Remark 2.14 are $(n-1,n-2,\dots,n-2),\dots,(n-2,\dots,n-2,n-1)$. If n is even, there is only one minimal fiber.

For the proof of Theorem 2.11 it is convenient to work in the quotient ring $\Bbbk[\mathbb{N}^{c_{n+1}}]/J_n$. Since $I_n \supseteq J_n$ and I_n is finely graded by $\mathbb{N}V_n$, each equivalence class is contained in a single fiber $V_n^{-1}[\mathbf{b}]$ and each fiber breaks into equivalence classes. The following definition sums the monomials corresponding to these classes for specific fibers.

Definition 2.16. *The* minimal saturated fibers *are the minimal fibers that satisfy Remark 2.14. The generating function of the equivalence class in a minimal saturated fiber is denoted by $p^+_{n,i}$. That is*

$$p^+_{n,i} = \sum_{\mathbf{a}\in V_n^{-1}[\mathbf{b}_i]/L'_n} \mathbf{x}^{\mathbf{a}} \in \Bbbk[\mathbb{N}^{c_{n+1}}]/J_n.$$

where $\mathbf{b}_i := (n-2,\dots,n-2)+e_i$ if n is odd and $\mathbf{b}_i = (n-2,\dots,n-2)$ if n is even.

If n is even, Definition 2.16 postulates only one polynomial which is simply denoted p_n^+ when convenient. Sometimes, however, it can be convenient to keep the indices.

Remark 2.17: The construction of a generating function of equivalence classes of elements of the fiber in Definition 2.16 can be carried out for any fiber of V_n. For the fiber $V_n^{-1}[(n-2,\dots,n-2,2c_{n-1})]$ we get the polynomial p_n from Lemma 2.8.

The quantity $\sigma_{n-1} = c_{n-1} - \lfloor \frac{n-1}{2} \rfloor$ (that is $c_{n-1} - \frac{n-2}{2} = \frac{(n-2)^2}{2}$ for even n, and $c_{n-1} - \frac{n-1}{2} = \frac{(n-1)(n-3)}{2}$ for odd n) appeared in Remark 2.9 and shows up again in the next lemma: it almost gives the saturation exponent when passing from the Laurent polynomial ring to the polynomial ring.

Lemma 2.18. *As elements of $\Bbbk[\mathbb{N}^{c_{n+1}}]/J_n$, if n is even then, $x_{nn}^{\sigma_{n-1}} p^+_{n,i} = p_n$, and if n is odd, then, $x_{nn}^{\sigma_{n-1}+1} p^+_{n,i} = x_{in}p_n$.*

Proof. If n is even, the product $x_{nn}^{\sigma_{n-1}} p^+_{n,i}$ has V_n-degree $(n-2,\dots,n-2,2c_{n-1})$. If n is odd, the degree of $x_{nn}^{\sigma_{n-1}+1} p^+_{n,i}$ equals $(n-2,\dots,n-2,2c_{n-1}+1)+e_i$. Now these products equal p_n if n is even and $x_{in}p_n$ if n is odd by Remarks 2.14 and 2.17. □

Lemma 2.19. *If n is odd, then for any triple of distinct indices $i,j,k \in [n]$, in $\Bbbk[\mathbb{N}^{c_{n+1}}]/J_n$ we have $x_{ij}p^+_{n,k} = x_{jk}p^+_{n,i}$.*

Proof. Since in a group algebra all monomials are invertible, Proposition 2.5 implies in particular that the variables are nonzerodivisors on $\Bbbk[\mathbb{N}^{c_{n+1}}]/J_n$. The multidegree of $p^+_{n,k}$ satisfies the conditions of Remark 2.14, thus there are bijections between the monomials of $x_{ij}p^+_{n,k}$ and $x_{jk}p^+_{n,i}$. Since all relations in J_n are equalities of monomials, multiplication with a variable does not touch coefficients. □

The following lemma captures an essential feature of our situation. Since the V_n-graded Hilbert function of $\Bbbk[\mathbb{N}^{c_{n+1}}]$ is globally bounded, there is a notion of *longest homogeneous polynomial* as one that uses all monomials in a given V_n-degree. For any multidegree $\mathbf{b}$ that satisfied $\mathbf{b}_i \geq (n-2)$, by Remark 2.14, if a longest polynomial of multidegree $\mathbf{b}$ is multiplied by a term, then it remains a longest polynomial.

LEMMA 2.20. *The V_n-graded Hilbert functions of the $\Bbbk[\mathbb{N}^{c_{n+1}}]/J_n$-modules, $\langle p_n\rangle$ and $\langle p^+_{n,i}\rangle$, $i = 1,\dots,n$ take only zero and one as their values.*

Proof. We only prove the statement for $\langle p_n\rangle$ since the same argument applies also to $\langle p^+_{n,i}\rangle$. The claim is equivalent to the statement that any $f \in \langle p_n\rangle$ is a term (that is, a monomial times a scalar) times p_n. Let $f = gp_n$ with a V_n-homogeneous g. Let $t_1,\dots,t_s$ be the terms of g. Since p_n is the sum of all monomials of degree $\deg(p_n)$, and multiplication by a term does not produce any cancellation, the number of terms of t_ip_n equals that of p_n. By Remark 2.14, the monomials in degree $\deg(t_ip_n)$ are in bijection with the monomials in degree $\deg(p_n)$, and therefore all t_ip_n are scalar multiples of the generating function of the fiber for $\deg(t_ip_n)$ and this generating function is equal to mp_n for any monomial m of multidegree $\deg(gp_n) - \deg(p_n)$. □

LEMMA 2.21. *For any $i \in [n]$, $\langle p^+_{n,i}\rangle : \left(\prod_{ij} x_{ij}\right)^\infty = \langle p^+_{n,k} : k \in [n]\rangle$.*

Proof. If n is odd, the containment of $p^+_{n,k}$ in the left hand side follows immediately from Lemma 2.19. If n is even, it is trivial. For the other containment, let f be a V_n-homogeneous polynomial that satisfies $mf \in \langle p^+_{n,i}\rangle$ for some monomial m. We want $f \in \langle p^+_{n,k} : k \in [n]\rangle$. By Lemma 2.20, $mf = tp^+_{n,i}$ for some term t. Since mf has the same number of terms as f and also the same number of terms as $tp^+_{n,i}$, this number must be $2^{c_{n-1}}$. By Remark 2.14, the only V_n-homogeneous polynomials with $2^{c_{n-1}}$ terms are monomial multiples of the $p^+_{n,k}$ for $k \in [n]$. □

PROPOSITION 2.22. $(J_n + \langle p_n\rangle) : \left(\prod_{ij} x_{ij}\right)^\infty = J_n + \langle p^+_{n,j}, j \in [n]\rangle$.

Proof. Throughout we work in the quotient ring $S := \Bbbk[\mathbb{N}^{c_{n+1}}]/J_n$ and want to show

$$\langle p_n \rangle : \left(\prod_{ij} x_{ij}\right)^\infty = \langle p^+_{n,j}, j \in [n] \rangle.$$

Lemma 2.18 gives the inclusion $\supseteq$, since it shows that, modulo J_n, a monomial multiple of $p^+_{n,i}$ is equal to either p_n or $x_{in}p_n$ and thus lies in $\langle p_n \rangle$. For the other containment let

$$f \in \langle p_n \rangle : \left(\prod_{ij} x_{ij}\right)^\infty,$$

that is $mf \in \langle p_n \rangle$ for some monomial m in S. This implies $mf = gp_n$ for some polynomial $g \in S$. By Lemma 2.18, $x_{in}mf = g'p^+_{n,i}$ for some $g' \in S$. So, $x_{in}mf \in \langle p^+_{n,i} \rangle$ and thus $f \in \langle p^+_{n,i} \rangle : x_{in}m$. Lemma 2.21 shows that $f \in \langle p_{n,k} : k \in [n] \rangle$. □

Having identified the minimal saturated fibers, the longest polynomials, and computed the saturation with respect to the variables x_{ij}, we are now ready to prove Theorem 2.11.

Proof of Theorem 2.11. After a potential renumbering, assume ϕ_1 is the identity. It suffices to prove the theorem for the omission of the Veronese ideal $i = 1$ from the intersection. The remaining cases follow by application of ϕ_l to the ambient ring.

Consider the extensions $J_n\Bbbk[\mathbb{Z}^{c_{n+1}}]$ and $I_n\Bbbk[\mathbb{Z}^{c_{n+1}}]$ to the Laurent polynomial ring. By the general Theorem 2.23

$$\bigcap_{i \neq 1} \phi_i(I_n\Bbbk[\mathbb{Z}^{c_{n+1}}]) = J_n\Bbbk[\mathbb{Z}^{c_{n+1}}] + \langle p_n \rangle.$$

Pulling back to the polynomial ring, we have

$$\bigcap_{i \neq 1} \phi_i(I_n) = (J_n + \langle p_n \rangle) : (\prod_{x_{ij}} x_{ij})^\infty.$$

Contingent on Theorem 2.23, the result now follows from Proposition 2.22. □

We have reduced the proof of Theorem 2.11 to a general result on intersection in the Laurent polynomial ring. It is a variation of [8, Theorem 2.1]. According to [8, Section 2], any binomial ideal in the Laurent polynomial ring $\Bbbk[\mathbb{Z}^n]$ is defined by its lattice $L \subseteq \mathbb{Z}^n$ of exponents and a partial character $\rho : L \to \Bbbk^*$. Such an ideal is denoted $I(\rho)$ where the lattice L is part of the definition of ρ. Let now L be a saturated lattice, $\rho : L \to \Bbbk^*$ a partial character, and $1 : L \to \Bbbk^*$ the trivial character that maps all of L to $1 \in \Bbbk$. The ideal $I(\rho)$

can be constructed by appropriately twisting the ideal $I(1)$. Specifically, if $\Bbbk$ is algebraically closed, there exists an automorphism ϕ_ρ of $\Bbbk[\mathbb{Z}^n]$ that maps each variable to a scalar multiple of the same variable and such that $\phi_\rho(I(1)) = I(\rho)$. Suitable coefficients $a_1, \ldots, a_n$ of the variables that define such an automorphism can be computed by solving the equations $a^{-m_i} = \rho(m_i)$ for any lattice basis $m_1, \ldots, m_r$ of L. These equations are solvable over an algebraically closed field and the resulting automorphisms generalize the $\mathbb{Z}_2$-twistings from Definition 2.7.

Theorem 2.23. *Let $\Bbbk$ be a field such that* $\operatorname{char}(\Bbbk)$ *is either zero or does not divide the order of the torsion part of $\mathbb{Z}^n/L$ and $I(\rho) \subseteq \Bbbk[\mathbb{Z}^n]$ be the binomial ideal for some partial character $\rho : L \to \Bbbk^*$. Let $I(\rho) = I(\rho'_1) \cap \ldots \cap I(\rho'_k)$ be a primary decomposition of $I(\rho)$ over the algebraic closure $\overline{\Bbbk}$ of $\Bbbk$. Omitting one component $I(\rho'_{i^*})$ yields*

$$\bigcap_{i \neq i^*} I(\rho'_i) = I(\rho) + \phi_{\rho'_{i^*}}(p_L)$$

where p_L is the generating function of a fundamental parallelepiped of the lattice L.

Proof. A linear change of coordinates in $\mathbb{Z}^n$ corresponds to a multiplicative change of coordinates in $\Bbbk[\mathbb{Z}^n]$. Since the inclusion of $L \subseteq \mathbb{Z}^n$ can be diagonalized using the Smith normal form, one can reduce to the case that $I(\rho)$ is generated by binomials $x_i^{q_i} - c_i$ for some coefficients $c_i \in \Bbbk$. This case follows by multiplication of the results in the univariate case. In the univariate case, the factors of $x^n - c$ are the n-th roots $\zeta_1, \ldots, \zeta_n$ of c. Then ρ is defined by $n \mapsto c$ and ρ'_i by $1 \mapsto \zeta_i$. One has

$$\prod_{i \neq i^*} (x - \zeta_i) = \phi_{i^*}((x^n - 1)/(x - 1)),$$

where ϕ_{i^*} is the automorphism of $\Bbbk[\mathbb{Z}]$ defined by $x \mapsto \zeta_{i^*}^{-1} x$. □

The assumption on $\operatorname{char}(\Bbbk)$ in Theorem 2.23 can be relaxed at the cost of a case distinction similar to that in [8, Theorem 2.1].

The explicit form of p_L depends on a choice of lattice basis. Because the notions lattice basis ideal and lattice ideal are not the same in the polynomial ring (they are in the Laurent polynomial ring), one needs to pull back using colon ideals to get a result in the polynomial ring. Even if in the Laurent polynomial ring the subintersection in Theorem 2.23 is principal modulo $I(\rho)$, it need not be principal in the polynomial ring (as visible in Theorem 2.11). It would be very nice to find more effective methods for binomial subintersections in the polynomial ring, but at the moment the following remark is all we have.

Remark 2.24: Under the field assumptions in Theorem 2.23, let $I \subseteq \Bbbk[\mathbb{N}^n]$ be a lattice ideal in a polynomial ring with indeterminates $x_1, \ldots, x_n$. There exists a partial character $\rho : L \to \Bbbk^*$ such that $I = I(\rho) \cap \Bbbk[\mathbb{N}^n]$. The intersection of all but one minimal primary components of I is

$$(I(\rho) + \phi_\rho(p_L)) \cap \Bbbk[\mathbb{N}^n] = (I + \phi_\rho(p)m) : \left(\prod_{i=1}^{n} x_i\right)^\infty .$$

where p_L is the generating function of a fundamental parallelepiped of L, and m is any monomial such that $\phi_\rho(p_L)m \in \Bbbk[\mathbb{N}^n]$.

3. Extensions

The broadest possible generalization of the results in Section 2 may start from an arbitrary toric ideal $I \subseteq \Bbbk[\mathbb{N}^n]$, corresponding to a grading matrix $V \in \mathbb{N}^{d\times n}$, and a subideal $J \subseteq I$, for example a lattice basis ideal. One can then ask when the quotient $\Bbbk[\mathbb{N}^n]/J$ is toral in the grading V. The techniques in Section 2 depend heavily on this property and the very controllable stabilization of the Hilbert function. One can get the feeling that this happens if $J \subseteq I$ is a lattice ideal for some lattice that is of finite index in the saturated lattice $\ker_{\mathbb{Z}}(V)$. However, such a J cannot always be found: by a result of Cattani, Curran, and Dickenstein, there exist toric ideals that do not contain a binomial complete intersection of the same dimension [3].

A more direct generalization of the results of Section 2 was suggested to us by Aldo Conca. The d-th Veronese grading $V_{d,n}$ has as its columns all vectors of length n and weight d. The corresponding toric ideal is the d-th Veronese ideal $I_{d,n} \subseteq S = \Bbbk[\mathbb{N}^N]$ and it contains a natural complete intersection $J_{d,n}$ defined as follows. The set of columns of $V_{d,n}$ includes the multiples of the unit vectors $D := \{de_i, i = 1, \ldots, n\}$. For any column $v \notin D$, let $f_v = x_v^d - \prod_i x_{de_i}^{v_i}$. Then $J = \langle f_v : v \notin D\rangle \subseteq I_{d,n}$ is a complete intersection with $\operatorname{codim}(J_{d,n}) = \operatorname{codim}(I_{d,n})$. It is natural to conjecture that a statement similar to Proposition 2.5 is true. In this case, however, the group L/L' (cf. Proposition 2.3) has higher torsion. This implies that the binomial primary decomposition of J exists only if $\Bbbk$ has corresponding roots of unity. By results of Goto and Watanabe [13, Chapter 3] on the canonical module (cf. [2, Exercise 3.6.21]) the ring S/I is Gorenstein if and only if $d|n$, so that $J : I$ is equal to $J + (p)$ for some polynomial p exactly in this situation.

In Section 2, the notation can be kept in check because there is a nice representation of monomials as upper triangular matrices (Proposition 2.5, Lemma 2.13, etc.). To manage the generalization, it will be an important task to find a similarly nice representation. It is entirely possible that something akin to the string notation of [20, Section 14] does the job. Additionally, experimentation with Macaulay2—which has informed the authors of this paper—will be hard. For example, for $d = 3, n = 4$, the group L/L' from Proposition 2.3 is isomorphic to $(\mathbb{Z}/3\mathbb{Z})^{13}$ which means that a prime decomposition of $J_{3,4}$ has 1594323 components. Computing subintersections of it is out of reach. It may be possible to compute a colon ideal like $(J_{3,4} : I_{3,4})$ directly, but off-the-shelf methods failed for us.

References

[1] R. Berger, *Differentialmoduln eindimensionaler lokaler Ringe*, Mathematische Zeitschrift **81** (1963), no. 4, 326–354.

[2] W. Bruns and J. Herzog, *Cohen–Macaulay rings*, Cambridge University Press, 1998.

[3] E. Cattani, R. Curran, and A. Dickenstein, *Complete intersections in toric ideals*, Proceedings of the American Mathematical Society **135** (2007), no. 2, 329–335.

[4] A. Conca, *Divisor class group and canonical class of determinantal rings defined by ideals of minors of a symmetric matrix*, Archiv der Mathematik **63** (1994), no. 3, 216–224.

[5] A. Conca, *Gröbner bases of ideals of minors of a symmetric matrix*, Journal of Algebra **166** (1994), no. 2, 406–421.

[6] P. Diaconis, D. Eisenbud, and B. Sturmfels, *Lattice walks and primary decomposition*, Mathematical Essays in Honor of Gian-Carlo Rota (B. Sagan and R. Stanley, eds.), Progress in Mathematics, vol. 161, Birkhauser, Boston, 1998, pp. 173–193.

[7] A. Dickenstein, L. Matusevich, and E. Miller, *Combinatorics of binomial primary decomposition*, Mathematische Zeitschrift **264** (2010), no. 4, 745–763.

[8] D. Eisenbud and B. Sturmfels, *Binomial ideals*, Duke Mathematical Journal **84** (1996), no. 1, 1–45.

[9] E. Gorla, *The G-biliaison class of symmetric determinantal schemes*, Journal of Algebra **310** (2007), no. 2, 880–902.

[10] E. Gorla, *Symmetric ladders and G-biliaison*, Liaison, Schottky Problem and Invariant Theory, Springer, 2010, pp. 49–62.

[11] S. Goto, *The divisor class group of a certain Krull domain*, Journal of Mathematics of Kyoto University **17** (1977), 47–50.

[12] S. Goto, *On the Gorensteinness of determinantal loci*, Journal of Mathematics of Kyoto University **19** (1979), no. 2, 371–374.

[13] S. Goto and K. Watanabe, *On graded rings I*, Journal of the Mathematical Society of Japan **30** (1978), no. 2, 179–213.

[14] D. R. Grayson and M. E. Stillman, *Macaulay2, a software system for research in algebraic geometry*, Available at `http://www.math.uiuc.edu/Macaulay2/`.

[15] W. Gröbner, *Über Veronesesche Varitäten und deren Projektionen*, Archiv der Mathematik **16** (1965), no. 1, 257–264.

[16] A. Jensen, T. Kahle, and L. Katthän, *Finding binomials in polynomial ideals*, Research in the Mathematical Sciences **4** (2017), no. 16, 10p.

[17] T. Kahle and E. Miller, *Decompositions of commutative monoid congruences and binomial ideals*, Algebra and Number Theory **8** (2014), no. 6, 1297–1364.

[18] R. E. Kutz, *Cohen–Macaulay rings and ideal theory in rings of invariants of algebraic groups*, Transactions of the American Mathematical Society **194** (1974), 115–129.

[19] E. Miller and B. Sturmfels, *Combinatorial commutative algebra*, GTM, vol. 227, Springer, Berlin, 2005.

[20] B. Sturmfels, *Gröbner bases and convex polytopes*, University Lecture Series, vol. 8, American Mathematical Society, Providence, RI, 1996.

Authors' addresses:

Thomas Kahle
Fakultät für Mathematik
Otto-von-Guericke Universität Magdeburg
Universitätsplatz 2
39106 Magdeburg, Germany
E-mail: thomas.kahle@ovgu.de

André Wagner
Institut für Mathematik
Sekretariat MA 6-2
Technischen Universität Berlin
Strasse des 17. Juni 136
10623 Berlin, Germany
E-mail: wagner@math.tu-berlin.de

Received January 30, 2018
Revised April 16, 2018
Accepted April 23, 2018

Rend. Istit. Mat. Univ. Trieste
Volume 50 (2018), 81–100
DOI: 10.13137/2464-8728/21599

Cones and matrix invariance: a short survey

Michela Brundu and Marino Zennaro

Abstract. *In this survey we collect and revisit some notions and results regarding the theory of cones and matrices admitting an invariant cone. The aim is to provide a self-contained treatment to form a convenient background to further researches. In doing this, we introduce some new intermediate concepts and propose several new proofs.*

Keywords: Cone, duality, matrix, invariant set, leading eigenvalue, leading eigenvector.
MS Classification 2010: 15A18, 15A48, 52A30, 52B55.

1. Introduction

In the framework of Linear Algebra, the description of the eigenvalues of an endomorphism of a vector space is one of the most classical problems.

A sufficient condition for the existence of a *leading* eigenvalue equal to the spectral radius was determined in 1907, in the real and finite dimensional case, by the mile-stone Theorem of Perron [11, 12], giving an affirmative answer as far as a positive matrix (associated to the endomorphism) is concerned.

In 1912, Frobenius [5] extended this result to irreducible nonnegative matrices. From then, the so called *Perron–Frobenius Theory* played a very important role within matrix theory, leading to several applications in Probability, Dynamical Systems, Economics, etc.

In the subsequent decades, this theory admitted a wide development, together with several generalizations which, in turn, have been applied to other branches of Mathematics and to applied sciences such as Physics, Social Sciences, Biology, etc.

The observation that a real positive $d \times d$ matrix corresponds to an endomorphism of $\mathbb{R}^d$ mapping the positive orthant into itself has naturally led to investigate endomorphisms admitting an *invariant cone* (the natural generalization of the orthant). In this context we mention, in particular, the generalization of the Perron–Frobenius Theorem due to Birkhoff [1] and the work by Vandergraft [17], where necessary and sufficient conditions on a matrix to have an invariant cone are given.

In this survey we collect some known notions and revisit several results regarding the theory of cones and matrices admitting an invariant cone.

The aim is to provide a convenient background to our papers [3, 2].

In doing this, on the one hand we introduce some new intermediate concepts. On the other hand, in order to provide a self-contained treatment, we fill in some gaps and, hence, we propose several new proofs.

2. Notation

We refer to $\mathbb{R}^d$ as a real vector space endowed with the Euclidean scalar product, denoted by $x^T y$ for any $x, y \in \mathbb{R}^d$. The metric and topological structures of this Euclidean space are induced by this pairing.

In this framework, if U is a nonempty subset of $\mathbb{R}^d$, we denote by $\mathrm{cl}(U)$ its *closure*, by $\mathrm{conv}(U)$ its *convex hull*, by $\mathrm{int}(U)$ its *interior* and by ∂U its *boundary* as a subset of $\mathbb{R}^d$. We also denote by $\mathrm{span}(U)$ the smallest vector subspace containing U. Finally, we set

$$\mathbb{R}_+ U := \left\{ \alpha x \;\middle|\; \alpha \geq 0 \text{ and } x \in U \right\}$$

and

$$U^\perp := \left\{ h \in \mathbb{R}^d \;\middle|\; h^T x = 0 \text{ for all } x \in U \right\}$$

denotes the *orthogonal set* of U.

In particular, if H is a (vector) hyperplane of $\mathbb{R}^d$ (i.e., a linear subspace of $\mathbb{R}^d$ of dimension $d-1$), then $H = \{h\}^\perp$ for a suitable vector $h \in \mathbb{R}^d \setminus \{0\}$, unique up to a scalar.

The hyperplane H splits $\mathbb{R}^d$ into two parts, say the *positive* and the *negative* semi-space

$$S_+^h := \{x \in \mathbb{R}^d \mid h^T x \geq 0\} \qquad \text{and} \qquad S_-^h := \{x \in \mathbb{R}^d \mid h^T x \leq 0\},$$

respectively. Clearly,

$$\mathrm{int}\left(S_+^h\right) = \{x \in \mathbb{R}^d \mid h^T x > 0\} \qquad \text{and} \qquad \mathrm{int}\left(S_-^h\right) = \{x \in \mathbb{R}^d \mid h^T x < 0\},$$

$$\mathbb{R}^d = \mathrm{int}\left(S_+^h\right) \cup H \cup \mathrm{int}\left(S_-^h\right) \qquad \text{and} \qquad \partial S_+^h = \partial S_-^h = H.$$

3. Cones and duality

The notion of *proper cone* is standard enough in the literature (see, e.g., Tam [16], Schneider and Tam [14] and Rodman, Seyalioglu and Spitkovsky [13]). The more general notion of *cone* is, instead, not universally shared: accordingly to the various authors, it involves a variable subset (or even all, see Schneider and Vidyasagar [15]) of the requirements for proper cones.

In this survey we shall deal with proper cones, as defined in the standard way, and with cones that verify a particular subset of the possible properties. We shall also find it useful to consider a weaker instance of our definition of cone, that we refer to as *quasi-cone.*

DEFINITION 3.1. *Let K be a nonempty closed and convex set of $\mathbb{R}^d$ and consider the following conditions:*

c1) $\mathbb{R}_+ K \subseteq K$ *(i.e., K is* positively homogeneous*);*

c2) $K \cap -K = \{0\}$ *(i.e., K is* pointed *or* salient*);*

c3) $\operatorname{span}(K) = \mathbb{R}^d$ *(i.e., K is* full *or* solid*).*

We say that K is a quasi-cone *if it satisfies (c1). If, in addition, it satisfies (c2), we say that K is a* cone*. Finally, if it satisfies all the above properties, we say that K is a* proper cone*.*

If a quasi-cone K is not solid, we also say that it is a degenerate quasi-cone*.*

The most known example of proper cone is the *positive orthant*

$$\mathbb{R}^d_+ = \left\{x \in \mathbb{R}^d \;\middle|\; x_i \geq 0,\ i = 1, \ldots, d\right\}.$$

In this section we recall some of the basic properties of quasi-cones. Most is well known and we refer the reader, e.g., to Fenchel [4], Schneider and Vidyasagar [15] and Tam [16].

The following invariants of a quasi-cone measure, in some sense, how far it is from being either pointed or full, respectively.

DEFINITION 3.2. *For any quasi-cone K, we denote by $L(K)$ the largest vector subspace included in K, called the* lineality space *of K, and by $l(K)$ the dimension of $L(K)$.*

Moreover, we denote by $d(K)$ the dimension of $\operatorname{span}(K)$*, called the* (linear) dimension *of K.*

REMARK 3.3: If K is a quasi-cone, it is clear that:

(i) $L(K) = K \cap -K$;

(ii) K is pointed if and only if $l(K) = 0$;

(iii) K is solid if and only if $d(K) = d$ or, equivalently, if and only if $\operatorname{int}(K) \neq \emptyset$.

If K is degenerate, then it is solid in the linear space $\mathrm{span}(K) \cong \mathbb{R}^{d(K)}$. So we can give the following definition.

DEFINITION 3.4. *If K is a quasi-cone, its interior as a subset of* $\mathrm{span}(K)$ *is called the* relative interior *of K and is denoted by* $\mathrm{int}_{rel}(K)$.

Note that, if K is a quasi-cone, then $l(K) \leq d(K)$ and the equality holds if and only if one of the following equivalent conditions is satisfied:

(i) $L(K) = K$;

(ii) $K = \mathrm{span}(K)$;

(iii) K is a linear subspace;

(iv) $\mathrm{int}_{rel}(K) = K$.

The next notion is well known.

DEFINITION 3.5. *Given a hyperplane H, we say that a nonempty positively homogeneous set $U \subset \mathbb{R}^d$ is* supported by H *(or, briefly, H-*supported *) if*

$$U \subseteq S_+^h \quad \text{or} \quad U \subseteq S_-^h.$$

Moreover, we say that U is strictly supported by H *(or, briefly,* strictly H-supported *) if*

$$U \setminus \{0\} \subseteq \mathrm{int}\left(S_+^h\right) \quad \text{or} \quad U \setminus \{0\} \subseteq \mathrm{int}\left(S_-^h\right).$$

REMARK 3.6: Let K be a cone and H be a hyperplane. Then K is strictly H-supported if and only if $K \cap H = \{0\}$.

PROPOSITION 3.7. *If $K \neq \mathrm{span}(K)$ is a quasi-cone of $\mathbb{R}^d$, then there exists a hyperplane H which supports K and*

$$H \cap \mathrm{int}_{rel}(K) = \emptyset.$$

Proof. First assume that K is solid. In this case, there exists a hyperplane H which supports K. (see Fenchel [4] (Corollary 1)).

If there exists $v \in H \cap \mathrm{int}(K)$, then we can consider a d-dimensional ball U_v, centered in v and contained in $\mathrm{int}(K)$. Clearly, U_v meets both $\mathrm{int}(S_+^h)$ and $\mathrm{int}(S_-^h)$, against the fact that K is H-supported .

Otherwise, if K is degenerate, let $S := \mathrm{span}(K)$, $s := d(K)$ its dimension and let T be a $(d-s)$-dimensional subspace such that $S \oplus T = \mathbb{R}^d$. Clearly,

K is solid in S and, so, from the previous case, we obtain the existence of a hyperplane V of S which supports K and $V \cap \mathrm{int}_{rel}(K) = \emptyset$. Now set $H := V \oplus T$, so that K is clearly H-supported and

$$H \cap \mathrm{int}_{rel}(K) = H \cap S \cap \mathrm{int}_{rel}(K) = V \cap \mathrm{int}_{rel}(K) = \emptyset,$$

as required. □

DEFINITION 3.8. *Given a nonempty set $U \subset \mathbb{R}^d$, the intersection of all the quasi-cones containing U (i.e., the smallest quasi-cone containing U) is called the* quasi-cone generated by U *and we denote it by* $\mathrm{qcone}(U)$.

Note that, while $\mathrm{qcone}(U)$ is defined for any set U, the smallest cone containing U may well not exist. Anyway, if it does exist, then it coincides with $\mathrm{qcone}(U)$.

DEFINITION 3.9. *Consider a nonempty set $U \subset \mathbb{R}^d$ and assume that* $\mathrm{qcone}(U)$ *is a cone. Then we denote it by* $\mathrm{cone}(U)$ *and call it the* cone generated by U.

The quasi-cone generated by U can be represented explicitly in formula by the aid of the following properties, whose proofs are straightforward.

PROPOSITION 3.10. *Let $U \subset \mathbb{R}^d$ be a nonempty set. Then*

(i) $\mathrm{conv}\,(\mathbb{R}_+ U) = \mathbb{R}_+ \mathrm{conv}\,(U)$;

(ii) $\mathrm{cl}\,(\mathbb{R}_+ U) \supseteq \mathbb{R}_+ \mathrm{cl}\,(U)$ *and, consequently,* $\mathrm{cl}\,(\mathbb{R}_+ U) = \mathbb{R}_+ \mathrm{cl}\,(\mathbb{R}_+ U)$;

(iii) $\mathrm{cl}\,(\mathrm{conv}\,(U)) \supseteq \mathrm{conv}\,(\mathrm{cl}\,(U))$ *and, consequently,* $\mathrm{cl}\,(\mathrm{conv}\,(U)) = \mathrm{conv}\,(\mathrm{cl}\,(\mathrm{conv}(U)))$.

COROLLARY 3.11. *For any nonempty set $U \subset \mathbb{R}^d$, we have*

$$\mathrm{qcone}(U) = \mathrm{cl}\,(\mathrm{conv}\,(\mathbb{R}_+ U)) = \mathrm{cl}\,(\mathbb{R}_+ \mathrm{conv}\,(U))\,. \tag{1}$$

Proof. The second equality in (1) is obtained just by taking the closure of both sides of (i) in Proposition 3.10.

Concerning the first equality, note that $\mathrm{cl}\,(\mathrm{conv}\,(\mathbb{R}_+ U))$ contains U, is convex (by (iii) in Proposition 3.10) and positively homogeneous (by (i) and (ii) in Proposition 3.10). Thus, by Definitions 3.1 and 3.8, we obtain $\mathrm{qcone}(U) \subseteq \mathrm{cl}\,(\mathrm{conv}\,(\mathbb{R}_+ U))$.

Conversely, since $\mathrm{qcone}(U)$ is positively homogeneous, $\mathrm{qcone}(U) \supseteq \mathbb{R}_+ U$. Moreover, it is convex and, hence, $\mathrm{qcone}(U) \supseteq \mathrm{conv}\,(\mathbb{R}_+ U)$. The fact that $\mathrm{qcone}(U)$ is also closed completes the proof. □

PROPOSITION 3.12. *A nonempty set* $U \subset \mathbb{R}^d$ *is contained in a closed semispace* S_+^h *if and only if* $\mathrm{qcone}(U) \neq \mathbb{R}^d$.

Proof. It is clear that $U \subseteq S_+^h$ if and only if $\mathrm{qcone}(U) \subseteq S_+^h$. On the other hand, by Proposition 3.7, this condition is equivalent to $\mathrm{qcone}(U) \neq \mathbb{R}^d$. □

The notion of *duality* is essential in the study of cones. Now we summarize a few basic definitions and properties.

DEFINITION 3.13. *Let* U *be a nonempty set of* $\mathbb{R}^d$. *Then*

$$U^* := \{h \in \mathbb{R}^d \mid h^T x \geq 0 \quad \forall x \in U\}$$

is called the dual set *of* U. *By convention, we also define* $\emptyset^* := \mathbb{R}^d$.

REMARK 3.14: If U is a subset of $\mathbb{R}^d$, then it is clear that $U \subseteq S_+^h$ if and only if $h \in U^* \setminus \{0\}$.

The proofs of the following relationships are straightforward.

PROPOSITION 3.15. *Let* U *and* V *be nonempty sets of* $\mathbb{R}^d$. *Then we have:*

(i) $U \subseteq U^{**}$;

(ii) $U \subseteq V$ *implies* $U^* \supseteq V^*$;

(iii) $(U \cup V)^* = U^* \cap V^*$;

(iv) $(U \cap V)^* \supseteq U^* \cup V^*$.

REMARK 3.16: Note that $\{0\}^* = \mathbb{R}^d$, $(\mathbb{R}^d)^* = \{0\}$ and, if $x \in \mathbb{R}^d \setminus \{0\}$, then

$$\{x\}^* = \{h \in \mathbb{R}^d \mid h^T x \geq 0\} = S_+^x$$

is the positive semi-space determined by x. Consequently, if U is a nonempty subset of $\mathbb{R}^d$, then

$$U^* = \bigcap_{x \in U} S_+^x \,.$$

Hence, U^* is closed, convex and positively homogeneous, i.e., U^* is a quasi-cone.

The above observation shows that the notion of dual of a set is deeply related to that of quasi-cone, as is evident also from the following fact.

PROPOSITION 3.17. *Let* U *be a subset of* $\mathbb{R}^d$ *and* U^* *be its dual set. Then*

$$U^* = (\mathrm{qcone}(U))^*.$$

Proof. Since for any $V \subseteq \mathbb{R}^d$ we easily have $V^* = (\mathrm{cl}(V))^*$, $V^* = (\mathrm{conv}(V))^*$ and $V^* = (\mathbb{R}_+ V)^*$, the claim follows immediately from (1). $\square$

DEFINITION 3.18. *If K is a quasi-cone of $\mathbb{R}^d$, the set*

$$K^* = \{h \in \mathbb{R}^d \mid h^T x \geq 0 \quad \forall x \in K\}$$

is called the dual quasi-cone *of K.*

As we saw in Proposition 3.15, $*$ is not completely a "geometric duality" on the subsets of $\mathbb{R}^d$. Namely, even if it is compatible with the union and contravariant with respect to the inclusion, a generic subset is not reflexive. Besides the category of vector subspaces of $\mathbb{R}^d$, that of quasi-cones fulfils the reflexivity, too. To this purpose, we recall that, for any quasi-cone K, we have

$$K^{**} = K \tag{2}$$

(see [4], Corollary to Theorem 3). Consequently, using the general implication in Proposition 3.15-(ii), we obtain

$$K^{(1)} \subseteq K^{(2)} \quad \Longleftrightarrow \quad (K^{(1)})^* \supseteq (K^{(2)})^* \tag{3}$$

for any pair $K^{(1)}$ and $K^{(2)}$ of quasi-cones.

REMARK 3.19: Let $K \neq \mathbb{R}^d$ be a quasi-cone. Then, thanks to Proposition 3.7, it is supported by some hyperplane H. As observed in Remark 3.14, this fact is equivalent to $K^* \neq \{0\}$.

The following key-fact can be found in Fenchel [4] (Theorem 5 and its Corollary).

PROPOSITION 3.20. *Let K be a quasi-cone of $\mathbb{R}^d$. Then*

$$d(K) + l(K^*) = d \qquad \text{and} \qquad d(K^*) + l(K) = d. \tag{4}$$

Remark 3.3 and Proposition 3.20 immediate yield the next consequence.

COROLLARY 3.21. *Let K be a quasi-cone. Then K is pointed if and only if K^* is solid and, dually, K^* is pointed if and only if K is solid. In particular, K is a proper cone if and only if K^* is a proper cone.*

Moreover, $K = \mathrm{span}(K)$ if and only if $K^* = \mathrm{span}(K^*)$.

This observation allows us to describe the lineality space of a quasi-cone in terms of its dual quasi-cone.

LEMMA 3.22. *Let K be a quasi-cone. Then*

$$L(K) = (K^*)^{\perp}. \tag{5}$$

Proof. Let us first show that $L(K) \subseteq (K^*)^{\perp}$. To this purpose, let $h \in K^*$. Since $L(K) \subseteq K$, for each $z \in L(K)$ we have $h^T z \geq 0$. Since $L(K)$ is a vector space, it also contains $-z$ and, hence, $h^T(-z) \geq 0$. Therefore, $h^T z = 0$ for each $z \in L(K)$ and, so, $L(K) \subseteq \{h\}^{\perp}$.

To prove the equality, it is enough to observe that $(K^*)^{\perp} = (\mathrm{span}(K^*))^{\perp}$. Hence, $\dim((K^*)^{\perp}) = d - d(K^*) = l(K)$, where the second equality follows from (4). □

PROPOSITION 3.23. *If $K \neq \mathrm{span}(K)$ is a quasi-cone, then*

$$L(K) \cap \mathrm{int}_{rel}(K) = \emptyset.$$

Proof. On one hand, by Proposition 3.7, there exists a hyperplane H supporting K such that $H \cap \mathrm{int}_{rel}(K) = \emptyset$. On the other hand, by Lemma 3.22 and Remark 3.19, we have that $L(K) \subseteq H$. □

LEMMA 3.24 ([4], Theorem 12). *If K is a quasi-cone and $h \in K^* \setminus \{0\}$, then*

$$h \in \mathrm{int}_{rel}(K^*) \iff K \cap \{h\}^{\perp} = L(K). \tag{6}$$

Note that, if $K = \mathrm{span}(K)$, then it is clear that $K^* = K^{\perp}$ and Lemma 3.24 just says that $K \cap \{h\}^{\perp} = K$ for each $h \in K^* \setminus \{0\}$.

Now we are in a position to prove a stronger version of Proposition 3.7.

PROPOSITION 3.25. *Let K be a quasi-cone. Then it is a cone if and only if it is strictly supported by some hyperplane H.*

Proof. Assume that K is a cone. So, by Corollary 3.21, its dual K^* is solid. Then just take $h \in \mathrm{int}(K^*)$ and set $H = \{h\}^{\perp}$. By Lemma 3.24, we have $K \cap H = \{0\}$ and, hence, by Remark 3.6, K is strictly H-supported .

Conversely, if K is strictly H-supported for some H, then $K \cap H = \{0\}$. Thus, $K \cap -K = \{0\}$ and, by Remark 3.3, K is pointed. □

The above discussion allows us to show the inclusions opposite to (*ii*) and (*iii*) of Proposition 3.10 hold in some particular cases.

Lemma 3.26. *Let X be a bounded subset of $\mathbb{R}^d$. Then* $\operatorname{conv}(\operatorname{cl}(X))$ *is closed and, hence,*

$$\operatorname{cl}(\operatorname{conv}(X)) = \operatorname{conv}(\operatorname{cl}(X)). \tag{7}$$

In addition, if $0 \notin \operatorname{cl}(X)$*, then also* $\mathbb{R}_+\operatorname{cl}(X)$ *is closed and, hence,*

$$\operatorname{cl}(\mathbb{R}_+ X) = \mathbb{R}_+\operatorname{cl}(X). \tag{8}$$

Proof. The first claim is well known. Hence, since $\operatorname{conv}(X) \subseteq \operatorname{conv}(\operatorname{cl}(X))$, we have that $\operatorname{cl}(\operatorname{conv}(X)) \subseteq \operatorname{conv}(\operatorname{cl}(X))$. Therefore, equality (7) follows from Proposition 3.10-(iii).

Now let $Y := \operatorname{cl}(X)$ and let $x \in \partial(\mathbb{R}_+Y) \setminus \{0\}$. Then there exists a sequence $(x_n)_n \subset \mathbb{R}_+Y$ converging to x and, so, there exists $M > 0$ such that definitively

$$\| x_n \| \leq M.$$

On the other hand, we can write

$$x_n = \lambda_n a_n$$

where $\lambda_n \in \mathbb{R}_+$ and $a_n \in Y$ for all n.

Since Y is compact, the sequence $(a_n)_n$ (or a suitable subsequence) converges to a point, say a, of Y. Necessarily, $a \neq 0$ because $0 \notin Y$. Thus, there exists $\mu > 0$ such that definitively

$$\| a_n \| \geq \mu > 0.$$

Since $\| x_n \| = |\lambda_n| \ \| a_n \|$, we then obtain definitively

$$\lambda_n \leq M/\mu.$$

Therefore, the sequence $(\lambda_n)_n$ (or a suitable subsequence) converges to a certain $\lambda \in \mathbb{R}_+$.

Finally, we obtain that (a suitable subsequence of) $(x_n)_n$ converges to λa. This implies that $x = \lambda a \in \mathbb{R}_+Y$. So $\mathbb{R}_+Y$ is closed. By using Proposition 3.10-(ii), similarly as before (8) follows. □

Proposition 3.27. *Let $X \subset \mathbb{R}^d$ be positively homogeneous and such that* $\operatorname{cl}(X)$ *is strictly supported by some hyperplane H. Then*

$$\operatorname{cl}(\operatorname{conv}(X)) = \operatorname{conv}(\operatorname{cl}(X)).$$

Proof. Denote by S the unit d-sphere of $\mathbb{R}^d$ and consider the compact set $\mathrm{cl}(X) \cap S$. Therefore, by Lemma 3.26, $\mathrm{conv}(\mathrm{cl}(X) \cap S)$ is closed.

Moreover, observe that $0 \notin \mathrm{conv}(\mathrm{cl}(X) \cap S)$ since $\mathrm{cl}(X)$ is strictly H-supported by assumption. Thus, by the second part of Lemma 3.26, we obtain that $\mathbb{R}_+\mathrm{conv}(\mathrm{cl}(X) \cap S)$ is closed.

On the other hand, $\mathrm{cl}(X)$ is positively homogeneous. Therefore, as is easy to see, $\mathbb{R}_+(\mathrm{cl}(X) \cap S) = \mathrm{cl}(X)$. Hence,

$$\mathrm{conv}(\mathbb{R}_+(\mathrm{cl}(X) \cap S)) = \mathrm{conv}(\mathrm{cl}(X))$$

and, so, Proposition 3.10-(i) yields

$$\mathbb{R}_+\mathrm{conv}(\mathrm{cl}(X) \cap S) = \mathrm{conv}(\mathrm{cl}(X)).$$

Therefore, $\mathrm{conv}(\mathrm{cl}(X))$ is closed and, using Proposition 3.10-(iii), like in the first part of Lemma 3.26 we get the thesis. □

COROLLARY 3.28. *Consider a nonempty set $U \subset \mathbb{R}^d$ and assume that* $\mathrm{qcone}(U)$ *is a cone. Then*

$$\mathrm{cone}(U) = \mathrm{conv}\left(\mathrm{cl}\left(\mathbb{R}_+U\right)\right) = \mathrm{cl}\left(\mathrm{conv}\left(\mathbb{R}_+U\right)\right) = \mathrm{cl}\left(\mathbb{R}_+\mathrm{conv}\left(U\right)\right). \tag{9}$$

Proof. Note first that

$$\mathrm{cl}\left(\mathbb{R}_+U\right) \subseteq \mathrm{cl}\left(\mathrm{conv}\left(\mathbb{R}_+U\right)\right) = \mathrm{cone}(U),$$

where the equality follows from Corollary 3.11. Therefore, $\mathrm{cl}\left(\mathbb{R}_+U\right)$ is strictly supported by some hyperplane H by Proposition 3.25.

Consequently, $\mathbb{R}_+U$ satisfies the assumptions on the set X of Proposition 3.27 which, in turn, gives the second equality in (9).

Finally, (1) gives the third equality. □

A more detailed study of the notion of dual of a quasi-cone leads us to the forthcoming Proposition 3.30.

LEMMA 3.29. *If $K \neq \mathrm{span}(K)$ is a quasi-cone and $h \in \mathbb{R}^d \setminus \{0\}$, then the following conditions are equivalent:*

(i) $h \in K^*$ and $K \cap \{h\}^\perp = L(K)$;

(ii) $h^T x > 0$ for all $x \in K \setminus L(K)$.

Proof. $(i) \Rightarrow (ii)$ Since $h \in K^*$, then $h^T x \geq 0$ for all $x \in K$. Now, if $x \in K \setminus L(K)$, then (i) implies that $x \notin \{h\}^\perp$, i.e., $h^T x \neq 0$.

$(ii) \Rightarrow (i)$ By Proposition 3.23 we have that $K \setminus L(K) \supseteq \mathrm{int}_{rel}(K)$ and, hence, the assumption implies that $h^T x > 0$ for all $x \in \mathrm{int}_{rel}(K)$. Therefore, the continuity of the scalar product proves that $h^T x \geq 0$ for all $x \in K$, i.e., $h \in K^*$. In turn, this fact implies that $K \cap \{h\}^\perp \supseteq L(K)$ holds (see (5)). So we are left to show that $K \cap \{h\}^\perp \subseteq L(K)$. If $x \in K$ and $h^T x = 0$, then necessarily $x \notin K \setminus L(K)$ by assumption, and this proves the requested inclusion. ☐

PROPOSITION 3.30. *Let K be a quasi-cone of $\mathbb{R}^d$. Then we have:*

(i) $$\mathrm{int}_{rel}(K^*) = \{h \in \mathbb{R}^d \mid h^T x > 0 \quad \forall x \in K \setminus L(K)\}$$
and, if K is a cone, then
$$\mathrm{int}(K^*) = \{h \in \mathbb{R}^d \mid h^T x > 0 \quad \forall x \in K \setminus \{0\}\}.$$

(ii) $$K^* \setminus L(K^*) = \{h \in \mathbb{R}^d \mid h^T x > 0 \quad \forall x \in \mathrm{int}_{rel}(K)\}$$
and, if K is solid, then
$$K^* \setminus \{0\} = \{h \in \mathbb{R}^d \mid h^T x > 0 \quad \forall x \in \mathrm{int}(K)\}.$$

Proof. (i) The first equality follows immediately from Lemmas 3.24 and 3.29. In particular, if K is a cone, then $L(K) = 0$ and the second equality is also proved.
(ii) It is clear that (i) implies
$$K \setminus L(K) \subseteq \{x \in \mathbb{R}^d \mid h^T x > 0 \quad \forall h \in \mathrm{int}_{rel}(K^*)\}.$$
Conversely, let $x \in \mathbb{R}^d$ be such that $h^T x > 0$ for all $h \in \mathrm{int}_{rel}(K^*)$. Then $x \notin \{h\}^\perp$ and, hence, $x \notin L(K)$ by (5). Moreover, still by the continuity of the scalar product, we also get $h^T x \geq 0$ for all $h \in K^*$. This means that $x \in K^{**} = K$. In this way we have shown that
$$K \setminus L(K) = \{x \in \mathbb{R}^d \mid h^T x > 0 \quad \forall h \in \mathrm{int}_{rel}(K^*)\}.$$
Exchanging the role of K and K^* and applying the reflexivity of the quasi-cones (see (2)), we obtain the requested equality.

Finally, if K is solid, then $L(K^*) = \{0\}$. ☐

A straightforward consequence of the above proposition follows.

COROLLARY 3.31. *If $K^{(1)}$ and $K^{(2)}$ are quasi-cones, then*
$$K^{(1)} \setminus L(K^{(1)}) \subseteq \mathrm{int}_{rel}(K^{(2)}) \quad \Longrightarrow \quad \mathrm{int}_{rel}((K^{(1)})^*) \supseteq (K^{(2)})^* \setminus L((K^{(2)})^*).$$

The last part of this section is devoted to some properties concerning the quasi-cone generated by a finite union of quasi-cones.

LEMMA 3.32. *Let $K^{(1)}, \dots, K^{(r)}$ be quasi-cones of $\mathbb{R}^d$ and $U := \bigcup_{i=1}^r K^{(i)}$. Then*

$$(\mathrm{qcone}\,(U))^* = U^* = \bigcap_{i=1}^r \left(K^{(i)}\right)^* .$$

Moreover, the above set, which is a quasi-cone, is $\neq \{0\}$ if and only if U is supported by some hyperplane H.

Proof. The first equality follows from Proposition 3.17 and the second from Proposition 3.15-(iii). Moreover,

$$(\mathrm{qcone}\,(U))^* \neq \{0\} \quad \Longleftrightarrow \quad \mathrm{qcone}\,(U) \neq \mathbb{R}^d,$$

and this is equivalent to U being H-supported (see Proposition 3.12). □

DEFINITION 3.33. *Let $K^{(1)}, \dots, K^{(r)}$ be quasi-cones. Their* sum *is defined as*

$$K^{(1)} + \dots + K^{(r)} := \{x_1 + \dots + x_r \mid x_i \in K^{(i)},\ i = 1, \dots, r\}.$$

LEMMA 3.34. *Let $K^{(1)}, \dots, K^{(r)}$ be quasi-cones. Then*

$$K^{(1)} + \dots + K^{(r)} = \mathrm{conv}(K^{(1)} \cup \dots \cup K^{(r)}) \tag{10}$$

and

$$\mathrm{cl}(K^{(1)} + \dots + K^{(r)}) = \mathrm{qcone}(K^{(1)} \cup \ldots \cup K^{(r)}). \tag{11}$$

Proof. Equality (10) proved in Kusraev and Kutateladze [9], 1.1.8.

Equality (11) immediately follows from (10). In fact, since the quasi-cones $K^{(i)}$ are positively homogeneous, equality (1) implies that $\mathrm{qcone}(K^{(1)} \cup \dots \cup K^{(r)}) = \mathrm{cl}(\mathrm{conv}(K^{(1)} \cup \dots \cup K^{(r)}))$. □

We recall that the notion of separatedness of two closed convex subsets of $\mathbb{R}^d$ has to be slightly modified (e.g., following Klee [7]) to adapt it to the case of cones.

DEFINITION 3.35. *Two cones $K^{(1)}$ and $K^{(2)}$ of $\mathbb{R}^d$ are said to be* separated *if there exists a hyperplane $H = \{h\}^\perp$ such that*

$$K^{(1)} \setminus \{0\} \subseteq \mathrm{int}\left(S^h_+\right) \quad \text{and} \quad K^{(2)} \setminus \{0\} \subseteq \mathrm{int}\left(S^h_-\right).$$

Moreover, we say that such an H is a separating hyperplane *for $K^{(1)}$ and $K^{(2)}$.*

Let us mention two well know results, the first of which is the "cone version" of a "separation-type" theorem, obtained directly from Klee [7], Theorem 2.7 (see also Holmes [6]).

THEOREM 3.36. *Two cones $K^{(1)}$ and $K^{(2)}$ of $\mathbb{R}^d$ are separated if and only if $K^{(1)} \cap K^{(2)} = \{0\}$.*

In other words, $K^{(1)} \cap -K^{(2)} = \{0\}$ if and only if $K^{(1)} \cup K^{(2)}$ is strictly supported by some hyperplane H. So the next statement immediately comes from Klee [7], Proposition 2.1.

PROPOSITION 3.37. *Let $K^{(1)}$ and $K^{(2)}$ be two cones of $\mathbb{R}^d$. If $K^{(1)} \cup K^{(2)}$ is strictly supported by some hyperplane H, then $K^{(1)} + K^{(2)}$ is closed.*

Let $U \subset \mathbb{R}^d$. Clearly, if $K = \text{qcone}(U)$ is strictly H-supported , then $U \setminus \{0\} \subseteq \text{int}(S^h_+)$. The converse is false as long as U is a generic set. For instance, let $U \subset \mathbb{R}^2$ be the unit open ball centered in the point $(0,1)$. Clearly, $U = U \setminus \{0\}$ is contained in $\text{int}(S^h_+)$, where $h = (0,1)$, but, at the same time, $\text{qcone}(U) = S^h_+$.

Nevertheless, the converse is true whenever U is a finite union of cones.

PROPOSITION 3.38. *Let $K^{(1)}, \ldots, K^{(r)}$ be cones of $\mathbb{R}^d$, H a hyperplane and*

$$K := \text{qcone}(K^{(1)} \cup \ldots \cup K^{(r)}).$$

Then the following statements are equivalent:

(i) *K is strictly H-supported ;*

(ii) *$K^{(1)} + \cdots + K^{(r)}$ is strictly H-supported and, hence, closed;*

(iii) *$K^{(1)} \cup \ldots \cup K^{(r)}$ is strictly H-supported .*

In this case, $K = K^{(1)} + \cdots + K^{(r)}$ is a cone, too.

Proof. With reference to (ii), we begin by observing that, if $K^{(1)} + \cdots + K^{(r)}$ is strictly H-supported , then it is closed. In fact, this can be easily proved by induction on r using Proposition 3.37.
$(i) \Rightarrow (ii)$ By (11).
$(ii) \Rightarrow (iii)$ By (10).
$(iii) \Rightarrow (ii)$ From the assumption, there exists h such that $\{h\}^\perp = H$ and $h^T z > 0$ for all $z \in K^{(1)} \cup \ldots \cup K^{(r)}$, $z \neq 0$. Hence, $h^T(z_1 + \cdots + z_r) > 0$ for all $z_i \in K^{(i)}$, $i = 1, \ldots, r$, such that $z_1 + \cdots + z_r \neq 0$.
$(ii) \Rightarrow (i)$ Since $K^{(1)} + \cdots + K^{(r)}$ is closed, then it coincides with K by (11) and, hence, K is strictly H-supported as well. □

4. Matrices with invariant cones

Let $\mathbb{F}$ denote either the real field $\mathbb{R}$ or the complex field $\mathbb{C}$. Throughout this paper we denote by $\mathbb{F}^{d\times d}$ the space of the $d \times d$ matrices on $\mathbb{F}$.

If $A \in \mathbb{F}^{d\times d}$, we identify it with the corresponding endomorphism

$$f_A : \mathbb{F}^d \to \mathbb{F}^d$$

defined by $f_A(x) = Ax$. Hence, the kernel and the image of f_A will be simply denoted by $\ker(A)$ and $\mathrm{range}(A)$, respectively, and, if U is a subset of $\mathbb{F}^d$, its image will be denoted by $A(U)$.

Nevertheless, the preimage of a subset V of $\mathbb{F}^d$ will be explicitly denoted by $f_A^{-1}(V)$.

Definition 4.1. *A subset U of $\mathbb{R}^d$ is said to be* invariant under the action of the matrix A *on* $\mathbb{R}^d$ *(in short,* invariant for A*) if $A(U) \subseteq U$.*

Assumption 4.1. *In order to avoid trivial cases, from now on we assume that A is a nonzero matrix.*

If $\lambda \in \mathbb{F}$ and a nonzero vector $v \in \mathbb{F}^d$ are such that $Av = \lambda v$, then they are called *eigenvalue* and *eigenvector* of A, respectively.

The set $V_\lambda(A)$, or simply V_λ, consisting of such eigenvectors and of the zero vector, is a linear subspace called the *eigenspace* corresponding to λ. Obviously, V_λ is invariant under the action of A.

Denoting by $\mu_a(\lambda)$ the *algebraic multiplicity* of λ (as root of the characteristic polynomial $\det(A-\lambda I)$) and by $\mu_g(\lambda)$ the *geometric multiplicity* of λ (i.e., $\dim_{\mathbb{F}}(V_\lambda)$), it is also well known that $\mu_g(\lambda) \le \mu_a(\lambda)$. If the equality holds, then λ is called *nondefective*. Otherwise, it is called *defective*.

Definition 4.2. *Let λ be an eigenvalue of A and $k = \mu_a(\lambda)$. Then the linear space*

$$W_\lambda(A) := \ker((A-\lambda I)^k) \subseteq \mathbb{F}^d$$

is called generalized eigenspace *corresponding to λ and each of its nonzero elements which does not belong to V_λ is called* generalized eigenvector.

If no misunderstanding occurs, we shall simply write W_λ instead of $W_\lambda(A)$.

It is clear that W_λ is a linear subspace invariant for A and it is well known that $\dim_{\mathbb{F}}(W_\lambda) = \mu_a(\lambda)$ (see, e.g., Lax [10], Theorem 11). Therefore, $V_\lambda = W_\lambda$ if and only if λ is nondefective.

In this paper we shall deal with real matrices only. Clearly, if A is a real matrix, we can take $\mathbb{F} = \mathbb{R}$ or $\mathbb{F} = \mathbb{C}$.

If $\lambda \in \mathbb{R}$, then W_λ is a linear subspace of $\mathbb{R}^d$ and $\dim_{\mathbb{R}}(W_\lambda) = \mu_a(\lambda)$.

Otherwise, if $\lambda \in \mathbb{C} \setminus \mathbb{R}$, take $\mathbb{F} = \mathbb{C}$ and consider $W_\lambda \subseteq \mathbb{C}^d$. Since the conjugate of λ is an eigenvalue as well, set $U_\mathbb{C}(\lambda, \bar{\lambda}) := W_\lambda \oplus W_{\bar{\lambda}} \subseteq \mathbb{C}^d$. With $k := \mu_a(\lambda) = \dim_\mathbb{C}(W_\lambda)$, it is clear that $\dim_\mathbb{C}(U_\mathbb{C}(\lambda, \bar{\lambda})) = 2k$. Setting also $U_\mathbb{R}(\lambda, \bar{\lambda}) := U_\mathbb{C}(\lambda, \bar{\lambda}) \cap \mathbb{R}^d$, it turns out that $\dim_\mathbb{R}(U_\mathbb{R}(\lambda, \bar{\lambda})) = 2k$ and that this linear space is spanned by the real and the imaginary parts of the vectors of W_λ. Clearly, $U_\mathbb{R}(\lambda, \bar{\lambda})$ is invariant for A.

Therefore, if $\lambda_1, \ldots, \lambda_r \in \mathbb{R}$ and $\mu_1, \bar{\mu}_1, \ldots, \mu_s, \bar{\mu}_s \in \mathbb{C} \setminus \mathbb{R}$ are the distinct roots of the characteristic polynomial, then

$$\mathbb{R}^d = \bigoplus_{i=1}^{r} W_{\lambda_i} \oplus \bigoplus_{i=1}^{s} U_\mathbb{R}(\mu_i, \bar{\mu}_i). \tag{12}$$

Finally, recall that the set $\sigma(A)$ of the (real or complex) eigenvalues is called the *spectrum* of A and the nonnegative real number

$$\rho(A) := \max_{\lambda \in \sigma(A)} |\lambda|$$

is called the *spectral radius* of A.

It is well known that either $\rho(A) > 0$ or $A^d = 0$.

The eigenvalues whose modulus is $\rho(A)$ are called *leading eigenvalues* and the corresponding eigenvectors are called *leading eigenvectors*. (For the convenience of the reader, we recall that, in the literature, these objects are also known as *principal eigenvalues* and *principal eigenvectors*).

The remaining eigenvalues and eigenvectors are called *secondary eigenvalues* and *secondary eigenvectors*, respectively.

REMARK 4.3: If the matrix A admits a real leading eigenvalue λ_1, we can write

$$\mathbb{R}^d = W_A \oplus H_A,$$

where

$$W_A := W_{\lambda_1} \qquad \text{and} \qquad H_A := \bigoplus_{i=2}^{r} W_{\lambda_i} \oplus \bigoplus_{i=1}^{s} U_\mathbb{R}(\mu_i, \bar{\mu}_i). \tag{13}$$

Observe that both W_A and H_A are linear subspaces invariant for A.

PROPOSITION 4.4. *Let A be a matrix admitting a real leading eigenvalue $\lambda_1 > 0$ and let $x \in \mathbb{R}^d$. Then*

$$Ax \in H_A \implies x \in H_A.$$

Proof. Using (13), we can write $x = v + u$ for suitable $v \in W_A$ and $u \in H_A$ and thus $Ax = Av + Au$. Clearly, $Ax \in H_A$ by assumption and $Au \in H_A$ since H_A is invariant for A. Therefore, $Av \in W_A \cap H_A = \{0\}$ and, hence, $v \in \ker(A) = W_0$. But $W_A \cap W_0 = \{0\}$ since $\lambda_1 > 0$. Therefore, $v = 0$ and the proof is complete. $\square$

It is clear that, if $\lambda > 0$ is a real eigenvalue and $\dim(V_\lambda) = 1$, both the two half-lines which constitute V_λ are invariant for A. Therefore, it makes sense to extend the search of invariant sets from linear subspaces to cones.

In the case of cones the notion of invariance is the general one (see Definition 4.1), but it is useful to recall the following refinement.

DEFINITION 4.5. *We say that a quasi-cone K is* strictly invariant under the action of the matrix A *on* $\mathbb{R}^d$ *(in short,* strictly invariant for A*) if*

$$A(K \setminus L(K)) \subseteq \mathrm{int}_{rel}(K).$$

In particular, if K is a cone, the above inclusion reads $A(K \setminus \{0\}) \subseteq \mathrm{int}_{rel}(K)$.

For example, the positive orthant $\mathbb{R}^d_+$ is invariant for a real matrix with nonnegative entries, whereas it is strictly invariant for a matrix with all strictly positive entries.

We recall that A and the *transpose matrix* A^T have the same eigenvalues with the same multiplicities. More precisely, for any eigenvalue $\lambda \in \mathbb{C}$ it holds that $\dim(V_\lambda(A)) = \dim(V_\lambda(A^T))$ and $\dim(W_\lambda(A)) = \dim(W_\lambda(A^T))$.

The following result is well known in the case of proper cones.

PROPOSITION 4.6. *A quasi-cone K is invariant (respectively, strictly invariant) for a matrix A if and only if the dual quasi-cone K^* is invariant (respectively, strictly invariant) for the transpose matrix A^T.*

We recall the following well-known Perron-Frobenius theorems, which may be found, for instance, in Vandergraft [17].

THEOREM 4.7. *Let a proper cone K be invariant for a nonzero matrix A. Then the following facts hold:*

(i) the spectral radius $\rho(A)$ is an eigenvalue of A;

(ii) the cone K contains an eigenvector v corresponding to $\rho(A)$.

THEOREM 4.8. *Let a proper cone K be strictly invariant for a nonzero matrix A. Then the following facts hold:*

(i) the spectral radius $\rho(A)$ is a simple positive eigenvalue of A and $|\lambda| < \rho(A)$ for any other eigenvalue λ of A;

(ii) $\mathrm{int}(K)$ *contains the unique leading eigenvector v (corresponding to $\rho(A)$);*

(iii) the secondary eigenvectors and generalized eigenvectors of A do not belong to K.

Under the hypotheses of Theorem 4.7, in the next Theorem 4.10 we prove a stronger version of the analogous counterpart of Theorem 4.8-(iii). Moreover, following the same line, in Theorem 4.12 we then easily obtain a stronger version of Theorem 4.8-(iii) itself.

LEMMA 4.9. *Let A be a matrix having a real leading eigenvalue $\rho(A)$. Then $W_{A^T} = (H_A)^{\perp}$.*

Proof. Set $B := (A-\rho(A)I)^k$ and recall that $W_A = \ker(B)$ (see Definition 4.2). Moreover, H_A is invariant for B since it is invariant for A.

From Remark 4.3 we then obtain that $\text{range}(B) = B(H_A) = H_A$, where the second equality holds since the matrix B is nonsingular on H_A.

Recalling that $\text{range}(B) = (\ker(B^T))^{\perp}$, we get $H_A = (\ker(B^T))^{\perp}$ and, finally, the equality $W_{A^T} = \ker(B^T)$ concludes the proof. ☐

Note that, if A is a matrix having an invariant proper cone K, then $\lambda_1 = \rho(A)$ is a real leading eigenvalue by Theorem 4.7. So, keeping the notation of Remark 4.3, we have the following result.

THEOREM 4.10. *Let A be a matrix having an invariant proper cone K. Then*

$$\text{int}(K) \cap H_A = \emptyset.$$

Proof. Let us consider $y \in \text{int}(K) \cap H_A$. Then, by Proposition 3.30-(i) applied to K^*, we have that $y^T w > 0$ for all $w \in K^* \setminus \{0\}$.

Let us observe that $\rho(A) = \rho(A^T)$ and that K^* is a proper cone invariant for A^T (see Proposition 4.6). Therefore, by Theorem 4.7, there exists a leading eigenvector $\bar{w}$ of A^T which belongs to K^*, i.e., $\bar{w} \in W_{A^T} \cap K^*$.

Since, by Lemma 4.9, $W_{A^T} = (H_A)^{\perp}$, we have $y^T \bar{w} = 0$, which gives a contradiction. ☐

COROLLARY 4.11. *If $\rho(A) > 0$, in the assumptions of the previous theorem, we have*

$$\text{int}(K) \cap \ker(A) = \emptyset.$$

Proof. If 0 is not an eigenvalue, the equality trivially holds. Otherwise, 0 is a secondary eigenvalue and, so, $W_0 \subseteq H_A$. On the other hand, $\ker(A) = V_0 \subseteq W_0$ and, thus, Theorem 4.10 concludes the proof. ☐

THEOREM 4.12. *If K is a strictly invariant proper cone for a matrix A, then*

$$K \cap H_A = \{0\}.$$

Proof. Let us consider $y \in K \cap H_A$. Then, by Proposition 3.30-(ii) applied to K^*, we have that $y^T w > 0$ for all $w \in \operatorname{int}(K^*)$.

The result is easily obtained by reasoning as in the proof of Theorem 4.10 by using Theorem 4.8 in place of Theorem 4.7. □

The previous result may be also found, for example, in Krasnosel'skiĭ, Lifshits and Sobolev [8] with a different proof.

The analogue of Corollary 4.11 clearly holds.

COROLLARY 4.13. *In the assumptions of the previous theorem we also have*

$$K \cap \ker(A) = \{0\}.$$

We conclude this survey by considering a particular class of matrices, which turns out to be the only one we can meet in the strictly invariant case.

DEFINITION 4.14. *A matrix A is said to be* asymptotically rank-one *if the following conditions hold:*

(i) $\rho(A) > 0$*;*

(ii) exactly one between $\rho(A)$ and $-\rho(A)$ is an eigenvalue of A and, moreover, it is a simple eigenvalue;

(iii) $|\lambda| < \rho(A)$ *for any other eigenvalue λ of A.*

The unique leading eigenvalue of A will be denoted by λ_A.

REMARK 4.15: A matrix A is asymptotically rank-one if and only if A^T is so.

The term "asymptotically rank-one" is inspired by the following known fact.

PROPOSITION 4.16. *If A is an asymptotically rank-one matrix, then there exists*

$$\hat{A}^\infty := \lim_{k\to\infty} A^k/\lambda_A^k$$

and such a limit is the rank-one matrix

$$\hat{A}^\infty = (v_A^T h_A)^{-1} v_A h_A^T,$$

where v_A and h_A are the (unique) leading eigenvectors of A and A^T, respectively.

Proof. We need to observe that the Jordan canonical form $\hat{J}$ of the normalized matrix $\hat{A} := A/\lambda_A$ may be assumed to be block diagonal. More precisely, the first block is 1×1 and consists in the maximum simple eigenvalue $\lambda_{\hat{A}} = 1$. The second one is a $(d-1) \times (d-1)$-block, upper bidiagonal, whose diagonal entries are the secondary eigenvalues of $\hat{A}$, all with modulus < 1, and the upper diagonal entries are equal to 1 or to 0. Therefore, when we take the kth power of $\hat{J}$, the first block remains unchanged, while the second clearly goes to zero. Hence, we obtain the rank-one limit matrix $\hat{J}^\infty$ with only one nonzero entry equal to 1 in the left upper corner.

Finally, the form of the limit $\hat{A}^\infty$ is easily determined by taking into account that it has the leading eigenvector v_A related to the eigenvalue 1 and that, analogously, its transpose $(\hat{A}^\infty)^T$ has the leading eigenvector h_A. □

The following characterization rephrases Theorem 4.4 in Vandergraft [17].

THEOREM 4.17. *A matrix A is asymptotically rank-one if and only if A or $-A$ admits a strictly invariant proper cone.*

Acknowledgements

The research was supported by funds from the University of Trieste (Grant FRA 2015) and from INdAM-GNCS.

The second author is a member of the INdAM Research group GNCS.

References

[1] G. BIRKHOFF, *Linear transformations with invariant cones*, Amer. Math. Month. **74** (1967), 274–276.

[2] M. BRUNDU AND M. ZENNARO, *Invariant multicones for families of matrices*, Ann. Mat. Pur. Appl. (2018), https://doi.org/10.1007/s10231–018–0790–4.

[3] M. BRUNDU AND M. ZENNARO, *Multicones, duality and matrix invariance*, J. Convex Anal. **26** (2019).

[4] W. FENCHEL, *Convex Cones, Sets and Functions*, Princeton University Notes, Princeton, 1953.

[5] G.F. FROBENIUS, *Über Matrizen aus nicht negativen Elementen*, Sitzungsber. Kön. Preuss. Akad. Wiss. Berlin (1912), 456–477.

[6] R.B. HOLMES, *Geometric Functional Analysis and its Applications*, Graduate Texts in Mathematics, no. 24, Springer-Verlag, New York - Heidelberg - Berlin, 1975.

[7] V.L. KLEE, *Separation properties of convex cones*, Proc. Amer. Math. Soc. **6-2** (1955), 313–318.

[8] M.A. KRASNOSEL'SKIĬ, E.A. LIFSHITS, AND A.V. SOBOLEV, *Positive Linear Systems: the Method of Positive Operators*, Sigma Series in Applied Mathematics, no. 5, Heldermann Verlag, Berlin, 1989.

[9] A.G. Kusraev and S.S. Kutateladze, *Subdifferentials: Theory and Applications*, Springer Science Business Media, 2012.

[10] P.D. Lax, *Linear Algebra and its Applications*, John Wiley and Sons, New York, 2007.

[11] O. Perron, *Grundlagen für eine Theorie des Jacobischen Kettenbruchalgorithmus*, Math. Ann. **63** (1907), 1–76.

[12] O. Perron, *Zur Theorie der Über Matrizen*, Math. Ann. **64** (1907), 248–263.

[13] L. Rodman, H. Seyalioglu, and I.M. Spitkovsky, *On common invariant cones for families of matrices*, Linear Algebra Appl. **432** (2010), 911–926.

[14] H. Schneider and B.S. Tam, *Matrices leaving a cone invariant*, in Handbook of Linear Algebra, L. Hogben Editor, Chapman and Hall, 2007.

[15] H. Schneider and M. Vidyasagar, *Cross-positive matrices*, SIAM J. Numer. Anal. **7** (1970), 508–519.

[16] B.S. Tam, *A cone-theoretic approach to the spectral theory of positive linear operators: the finite dimensional case*, Taiwanese J. Math. **5** (2001), 207–277.

[17] J.S. Vandergraft, *Spectral properties of matrices which have invariant cones*, SIAM J. Appl. Math. **16** (1968), 1208–1222.

Authors' addresses:

Michela Brundu
Università degli Studi di Trieste
Dipartimento di Matematica e Geoscienze
Via Valerio 12/1
34100 Trieste, Italy
E-mail: brundu@units.it

Marino Zennaro
Università degli Studi di Trieste
Dipartimento di Matematica e Geoscienze
Via Valerio 12/1
34100 Trieste, Italy
E-mail: zennaro@units.it

Received March 27, 2018
Accepted May 19, 2018

Rend. Istit. Mat. Univ. Trieste
Volume 50 (2018), 101–123
DOI: 10.13137/2464-8728/21596

Chevalley-Weil formula for hypersurfaces in $\mathbf{P}^n$-bundles over curves and Mordell–Weil ranks in function field towers

Remke Kloosterman

Abstract. *Let X be a complex hypersurface in a $\mathbf{P}^n$-bundle over a curve C. Let $C' \to C$ be a Galois cover with group G. In this paper we describe the $\mathbf{C}[G]$-structure of $H^{p,q}(X \times_C C')$ provided that $X \times_C C'$ is either smooth or $n = 3$ and $X \times_C C'$ has at most ADE singularities. As an application we obtain a geometric proof for an upper bound by Pál for the Mordell–Weil rank of an elliptic surface obtained by a Galois base change of another elliptic surface.*

Keywords: Elliptic surfaces, Mordell-Weil rank under base change.
MS Classification 2010: 14J27.

1. Introduction

Let k be a field of characteristic zero, C/k a smooth, geometrically integral curve, and let $f : C' \to C$ be a (ramified) Galois cover with Galois group G. Let $E/k(C)$ be a non-isotrivial elliptic curve, i.e., with $j(E) \in k(C) \setminus k$ and let $\pi : X \to C$ be the associated relatively minimal elliptic surface with section. Let $R \subset C$ be the set of points over which f is ramified and let s be the number of points in R. Let e be the Euler characteristic of $C \setminus R$, i.e., $e = 2 - 2g(C) - s$.

Assume that the discriminant of π does not vanish at any point in R. Let c_E and d_E be the degree of the conductor of $E/k(C)$ and the degree of the minimal discriminant of E, respectively. Pál showed in [12] using equivariant Grothendieck–Ogg–Shafarevich theory that

$$\operatorname{rank} E(k(C')) \leq \epsilon(G,k)\left(c_E - \frac{d_E}{6} - e\right) \tag{1}$$

where $\epsilon(G,k)$ is the Ellenberg constant of (G,k), for a definition see [3]. This constant depends only on the group G and the field k, but not on E. In this paper we will give an alternative proof for this bound.

As noted in [12] it suffices to prove that $E(k(C')) \otimes_{\mathbf{Z}} \mathbf{C}$ is a quotient of a free $k[G]$-module of rank $c_E - d_E/6 - e$, and by the Lefschetz principle it suffices to prove this slightly stronger statement only in the case $k = \mathbf{C}$.

Let $X' = \widetilde{X \times_C C'}$ be the elliptic surface associated with $E/\mathbf{C}(C')$. Our starting point is that the following ingredients would lead to a proof for the fact that $E(\mathbf{C}(C'))$ is a quotient of $\mathbf{C}[G]^{\oplus c_E + d_E/6 - e}$.

1. $E(\mathbf{C}(C')) \otimes \mathbf{C}$ is a quotient of $H^{1,1}(X', \mathbf{C})$.
2. Let μ be the total Milnor number of X. Then the kernel of the natural map $H^{1,1}(X', \mathbf{C}) \to E(\mathbf{C}(C')) \otimes \mathbf{C}$ contains $\mathbf{C}^2 \oplus \mathbf{C}[G]^{\mu}$.
3. $H^0(K_{C'})^{\oplus 2}$ is a quotient of $\mathbf{C}[G]^{-e}$.
4. $\mu = d_E - c_E$.
5. The $\mathbf{C}[G]$-structure of $H^{1,1}(X', \mathbf{C})$ is $\mathbf{C}[G]^{\oplus \frac{5}{6} d_E} \oplus H^0(K_{C'})^{\oplus 2}$.

The first point is part of the standard proofs for the Shioda–Tate formula for the Mordell–Weil rank of an elliptic surface and the Lefschetz $(1,1)$-theorem. The second point follows similarly, but here we need to use our assumptions on the ramification of f. The third point is straightforward (Lemma 3.3), the fourth point is not difficult (Corollary 4.15). Hence the crucial point is to determine the $\mathbf{C}[G]$-structure of $H^{1,1}(X', \mathbf{C})$.

If C' is rational and all singular fibers of X' are irreducible then the $\mathbf{C}[G]$-structure of $H^{1,1}(X')$ can be determined as follows: Since C' is rational we have that X' is birational to a surface $W' \subset \mathbf{P}(2k, 3k, 1, 1)$ of degree $6k$, for some k. The surface W' is not unique, but if we take k minimal then is it unique. The surface W' is called the Weierstrass model of X'. From our assumptions that all fibers of X' are irreducible it follows that all singularities of W' along the singular locus of $\mathbf{P}(2k, 3k, 1, 1)$. Moreover, in this case W' is quasismooth: its affine quasi-cone is smooth away from the vertex.

From the fact that W' is quasismooth it follows that the co-kernel of the injective map $H^{1,1}(W')_{\text{prim}} \to H^{1,1}(X')$ is two-dimensional, and G acts trivially on this co-kernel. Steenbrink [15] presented a method to find an explicit basis for $H^{1,1}(W')_{\text{prim}}$ in terms of the Jacobian ideal of W', extending Griffiths' method for hypersurfaces in $\mathbf{P}^n$. A straightforward calculation then yields the $\mathbf{C}[G]$-structure of $H^{1,1}(W')$.

If C' is rational, but X' has reducible fibers then there are two possible ways to generalize this result. The first approach uses a deformation argument to show that X' is the limit for $t = 0$ of a family X'_t of elliptic surfaces admitting a G-action, such that all for $t \neq 0$ the elliptic fibration on X'_t has only irreducible fibers. The second approach uses a result of Steenbrink [16] where he extends his method to describe $H^{p,q}(W')_{\text{prim}}$ to the case where, very roughly, the sheaves of Du Bois differentials and of Barlet differentials on W'

coincide (this condition holds for Weierstrass models of elliptic surfaces, the precise condition on W' is formulated in [16]).

This paper grew out of an attempt to generalize the latter approach to the case where $g(C') > 0$. However, in this case some additional technical complications occur. Let $\pi : X \to C$ be an elliptic surface, and let $S \subset X$ be the image of the zero section. Let $N_{S/X}$ be the normal bundle of S. Then one can find a Weierstrass model W of X in $\mathbf{P}(\mathcal{E})$ where $\mathcal{E} = \mathcal{O} \oplus \mathcal{L}^{-2} \oplus \mathcal{L}^{-3}$, with $\mathcal{L} = (\pi_* N_{S/X})^*$. Similarly the Weierstrass model of the base changed elliptic surface is a surface W' in $\mathbf{P}(f^*\mathcal{E}) =: \mathbf{P}$. The Griffiths–Steenbrink approach yields two injective maps

$$\frac{H^0(K_{\mathbf{P}}(2W'))}{H^0(K_{\mathbf{P}}(W') \oplus dH^0(\Omega^2(2W'))} \hookrightarrow H^{1,1}(W') \hookrightarrow H^{1,1}(X').$$

Using our assumptions on f we can easily describe the $\mathbf{C}[G]$-action on the left hand side. The cokernel of the second map is isomorphic to $\mathbf{C}[G]^{\mu}$. The dimension of the cokernel of the first map is $2 + h^1(f^*\mathcal{L})$. The 2 corresponds to two copies of the trivial representation, however, it is not that easy to describe the $\mathbf{C}[G]$-action on the vector space of dimension $h^1(f^*\mathcal{L})$. From this it follows that the Griffiths–Steenbrink approach works as long as $h^1(f^*\mathcal{L})$ vanishes. This happens only if the degree of the ramification divisor $C' \to C$ is small compared to $\deg(f)$ and $\deg(\mathcal{L})$.

To avoid this restriction on $h^1(\mathcal{L})$ we work with equivariant Euler characteristic: Let $K(\mathbf{C}[G])$ be the Grothendieck group of all finitely generated $\mathbf{C}[G]$-modules. For a coherent sheaf $\mathcal{F}$ on a scheme with a G-action one defines

$$\chi_G(\mathcal{F}) = \sum_i (-1)^i [H^i(X, \mathcal{F})].$$

We use the ideas behind the Griffiths–Steenbrink approach to prove that the class of $H^{1,1}(W')$ in $K(\mathbf{C}[G])$ equals

$$2[\mathbf{C}] - \chi_G(\Omega^2_{\mathbf{P}}(W')) + \chi_G(K_{\mathbf{P}}(2W')) - \chi_G(H^0(\mathcal{T})) - \chi_G(K_{\mathbf{P}}(W')).$$

Here $\mathcal{T}$ is a skyscraper sheaf supported on the singular locus of W', such that its stalk is isomorphic to the Tjurina algebra of the singularity, and $\Omega^{2,\mathrm{cl}}_{\mathbf{P}}$ is the sheaf of closed 2-forms. The remaining Euler characteristics can be calculated by fairly standard techniques and thereby yielding a proof of the point (5) mentioned above.

One can easily describe $H^{1,1}(X')$ (as $\mathbf{C}[G]$-module) in terms of the regular representation $\mathbf{C}[G]$ and $H^{1,1}(W')$. The $\mathbf{C}[G]$-structure on the other $H^{p,q}(X')$ can be determined by standard techniques. In the sequel we show:

Proposition 1.1. *Let $\pi : X \to C$ be an elliptic surface and set $\mathcal{L} = (\pi_* N_{S/X})^*$. Let $f : C' \to C$ be a ramified Galois cover with group G and let $X' \to C'$ be* *the smooth minimal elliptic surface birational to $X \times_C C'$. Suppose that over each branch point of f the fiber of π is smooth or semistable. Then we have the following identities in $K(\mathbf{C}[G])$:*

$$\begin{array}{rcl} [H^{0,1}(X',\mathbf{C})] & = & [H^{1,0}(X',\mathbf{C})] = [H^0(C',K_{C'})]; \\ [H^{2,0}(X',\mathbf{C})] & = & [H^0(C',K_{C'})] - [\mathbf{C}] + \deg(\mathcal{L})[\mathbf{C}[G]]; \\ [H^{1,1}(X',\mathbf{C})] & = & 2[H^0(C',K_{C'})] + 10\deg(\mathcal{L})[\mathbf{C}[G]]. \end{array}$$

Since X' is smooth we can use Poincaré duality to describe the $\mathbf{C}[G]$-structure of $H^{p,q}(X')$ for all other p,q. As argued above, this Proposition is sufficient to prove the bound (1), see Corollary 4.15.

We would like to make one remark concerning this bound of Pál: From the Shioda–Tate formula it follows that

$$\operatorname{rank} E(k(C')) \leq \#G\left(c_E - \frac{d_E}{6}d_E\right) + 2g(C') - 2.$$

If each of the elements of G is defined over k, then the Ellenberg constant equals the number of elements of G. Hence Pál's bound reads

$$E(k(C')) \leq \#G\left(c_E - \frac{d_E}{6}\right) + \#G\,(2g(C) - 2 + s)$$

in this case. From Riemman-Hurwitz it follows that $2g(C')-2$ is at most $\#G(2g(C)-2+s)$ (and equality holds if and only if $s=0$). Hence the bound (1) is weaker than the bound from the Shioda-Tate formula in this case. However, if the absolute Galois group of k acts highly non-trivially on G then the Ellenberg constant is small and therefore this bound is very useful.

Our approach to determine the $\mathbf{C}[G]$-structure of $H^{p,q}$ works for a larger class of varieties. To formulate this result we need to introduce a skyscraper sheaf $\mathcal{T}$, which can be defined for a hypersurface X' with isolated singularities in a smooth ambient space, its support is the singular locus of X' and the stalk at a point $x \in X'$ is the Tjurina algebra of X' at x.

THEOREM 1.2. *Let C be a smooth projective curve and $\mathcal{E}$ a rank r vector bundle over C, which is a direct sum of line bundles. Let $X \subset \mathbf{P}(\mathcal{E})$ be a hypersurface. Let $f : C' \to C$ be a Galois cover and let $X' = X \times_C C'$. Assume that either X' is smooth or $r = 3$ and X' is a surface with at most ADE singularities.*

Moreover, assume $H^i(X') \cong H^i(\mathbf{P}(f^*\mathcal{E}))$ for $i \leq r-2$.

Then we have the following identity in $K(\mathbf{C}[G])$

$$[H^{p,q}(X')] = a[\mathbf{C}[G]] + b\chi_G(\mathcal{O}_C) + c[\mathbf{C}] + d[H^0(\mathcal{T})]$$

for some integers a,b,c,d, which can be determined explicitly and depend on p, q, the degrees of the direct summands of $\mathcal{E}$ and the fiber degree of X.

There are many other results on the behaviour of the Mordell–Weil rank under base change. Most of these results assume that the fibers over the critical values are very singular, e.g., the results by Fastenberg [4, 5, 6] and by Heijne [8]. Bounds in the case where the fibers over the critical values are smooth and where the base change map is étale, are obtained by Silverman [14]. Ellenberg proved a slightly weaker bound in a much more general setting, namely he showed that

$$\operatorname{rank} E(k(C')) \leq \epsilon(G,k)(c_E - 2e)$$

without imposing any condition on G, and assuming only that 6 is invertible in k.

The $\mathbf{C}[G]$-structure of the cohomology of a ramified cover $X \to Y$ has been studied in general, but we could not find any result that was sufficiently precise to prove (1). The first result in this direction was by Chevalley–Weil [1] in the curve case. There are several results by Nakajima in the higher-dimensional case [10].

In Section 2 we discuss the construction of Weierstrass models associated with elliptic surfaces. In Section 3 we prove Theorem 1.2. In Section 4 we determine the constants a, b, c, d for the case of Weierstrass models of elliptic surfaces and give a proof for (1).

2. Weierstrass models and Projective bundles

In this section let C be a smooth projective curve and $\mathcal{L}$ a line bundle on a curve C, of positive degree. We recall the construction of Weierstrass models of elliptic surfaces with fundamental line bundle $\mathcal{L}$. Most of the results of this section are also present in [9, Chapter II and III], but we included them for the reader's convenience.

Let $\mathcal{E} = \mathcal{O} \oplus \mathcal{L}^{-2} \oplus \mathcal{L}^{-3}$, let $\mathbf{P}(\mathcal{E})$ be the associated projective bundle, parametrizing one-dimensional quotients of $\mathcal{E}$. Let $\varphi : \mathbf{P} \to C'$ be the projection map. Then $\varphi_*(\mathcal{O}_{\mathbf{P}}(1)) = \mathcal{E}$. Let

$$\begin{aligned} X &= (0,1,0) \in H^0(\varphi^*\mathcal{L}^2(1)) = H^0(\mathcal{L}^2) \oplus H^0(\mathcal{O}_C) \oplus H^0(\mathcal{L}^{-1}), \\ Y &= (0,0,1) \in H^0(\varphi^*\mathcal{L}^3(1)) = H^0(\mathcal{L}^3) \oplus H^0(\mathcal{L}) \oplus H^0(\mathcal{O}_C), \\ Z &= (1,0,0) \in H^0(\mathcal{O}_{\mathbf{P}}(1)) = H^0(\mathcal{O}) \oplus H^0(\mathcal{L}^{-2}) \oplus H^0(\mathcal{L}^{-3}) \end{aligned}$$

be the standard coordinates.

Definition 2.1. *A (minimal) Weierstrass model W is an element*

$$F := -Y^2Z - a_1XYZ - a_3YZ^2 + X^3 + a_2X^2Z + a_4XZ^2 + a_6Z^3$$

in $|\mathcal{L}^6 \otimes \mathcal{O}_{\mathbf{P}(\mathcal{E})}(3)|$, such that $V(F) \subset \mathbf{P}(\mathcal{E})$ has at most isolated ADE singularities.

Remark 2.2: The restriction of φ to a Weierstrass model W is a morphism with only irreducible fibers, and the generic fiber is a genus one curve. For a fixed Weierstrass model W denote with X its minimal resolution of singularities and with $\pi : X \to C$ the induced fibration.

Lemma 2.3. *The minimal resolution of singularities of a Weierstrass model is an elliptic surface $\pi : X \to C$. The section $\sigma_0 : C \to W$, which maps a point p to the point $[0:1:0]$ in the fiber over p, extends to a section $C \to X$.*

Proof. The first statement is straightforward. From the shape of the polynomial F it follows that W_{sing} is contained in $V(Y)$. Recall that $\sigma_0(C) = V(X,Z)$. Hence $\sigma_0(C)$ does not intersect W_{sing} and we can extend $\sigma_0 : C \to X$. $\square$

Remark 2.4: Conversely, every elliptic surface over C admits a minimal Weierstrass model for a proper choice of line bundle $\mathcal{L}$, namely $\mathcal{L}$ is the inverse of the push forward of the normal bundle of the zero section. The line bundle $\mathcal{L}$ is of non-negative degree. If the degree of $\mathcal{L}$ is zero then the fibration has no singular fibers and after a finite étale base change the elliptic surface is a product. See [9, Section III.3].

Remark 2.5: Since we work in characteristic zero we may, after applying an automorphism of $\mathbf{P}(\mathcal{E})/C$ if necessary, assume that a_1, a_2 and a_3 vanish. In the sequel we work with a short Weierstrass equation

$$-Y^2Z + X^3 + AXZ^2 + BZ^3$$

with $A \in H^0(\mathcal{L}^4)$ and $B \in H^0(\mathcal{L}^6)$.

This is the equation of a minimal Weierstrass model if and only if for each point $p \in C$ we have either $v_p(A) \le 3$ or $v_p(B) \le 5$.

Lemma 2.6. *The Weierstrass model W is smooth if and only if all singular fibers of π are of type I_1 and II.*

Proof. The Weierstrass model W is smooth if and only if $X \cong W$. Since all fibers of $W \to C$ are irreducible, this is equivalent to the fact that all singular fibers of π are irreducible. Hence these fibers are of type I_1 or II. $\square$

Lemma 2.7. *Let W be a Weierstrass model with associated line bundle $\mathcal{L}$. Let $f : C' \to C$ be a finite morphism of curves. Suppose that over the branch points of f the fiber of π is either smooth or semi-stable.*

Then $W' := W \times_C C'$ is a Weierstrass model (with associated line bundle $f^*(\mathcal{L})$).

Proof. Consider the induced map $\mathbf{P}(f^*(\mathcal{E})) \to \mathbf{P}$. Then W' is the zero set of

$$-Y^2Z + X^3 + f^*(A)XZ^2 + f^*(B)Z^3.$$

If W' is not a Weierstrass model then there is a point $p \in C'$ such that $v_p(f^*(A)) \geq 4$ and $v_p(f^*(B)) \geq 6$.

Since W is Weierstrass model we have $v_q(A) \leq 3$ or $v_q(B) \leq 5$ for all $q \in C$. Let e_p be the ramification index of p then $v_p(f^*A) = e_p v_q(A)$ and $v_p(f^*B) = e_p v_p(B)$ for $q = f(p)$. Hence if $v_p(f^*A) \geq 4$ and $v_p(f^*B) \geq 6$ then $e_p > 1$, i.e. f is ramified at p. However, in this case the fiber of $f(p)$ is either smooth or multiplicative. This implies that at least one of $A(q)$ or $B(q)$ is nonzero. Hence at least one $v_p(f^*A)$ or $v_p(f^*B)$ vanishes and therefore W' is a minimal Weierstrass model. □

Since W has only ADE singularities we have that the cohomology of W and X are closely related:

PROPOSITION 2.8. *Let W be a Weierstrass model and $\pi : X \to C$ the elliptic fibration on the minimal resolution of singularities of W. Let μ be the total number of fiber-components of π which do not intersect the image of the zero-section. Then μ equals the total Milnor number of the singularities of X.*

Moreover, the natural mixed Hodge structure on $H^i(W)$ is pure for all i and we have $h^{p,q}(X) = h^{p,q}(W)$ for $(p,q) \neq (1,1)$ and $h^{1,1}(X) = h^{1,1}(W) + \mu$.

Proof. All fibers of $W \to C$ are irreducible by construction. Hence the number of fiber components not intersecting the image of the zerosection equals the number of irreducible components of the exceptional divisor $X \to W$.

The resolutions of ADE surfaces singularities are well-known, and the number of irreducible components of the exceptional divisor equals the Milnor number, proving the first claim.

The intersection graph of the exceptional divisor of a resolution of an ADE singularity is also well-known and from this it follows that the exceptional divisors are simply connected complex curves. Hence if we have s singular points with total Milnor number μ and E is the total exceptional divisor then $H^0(E) = \mathbf{C}^s$ and $H^2(E) = \mathbf{C}(-1)^\mu$ and all other cohomology groups vanish.

Let $\Sigma = W_{\text{sing}}$. From [13, Corollary-Definition 5.37] it follows that we have a long exact sequence of MHS

$$\cdots \to H^i(W) \to H^i(X) \oplus H^i(\Sigma) \to H^i(E) \to H^{i+1}(W) \to \dots \tag{2}$$

Note that $h^i(\Sigma) = 0$ for $i \neq 0$. Moreover, the map $H^0(\Sigma) \to H^0(E)$ is clearly an isomorphism, combining this with the fact that $H^i(E) = 0$ for $i \neq 0, 2$ we obtain that $H^i(X) \cong H^i(W)$ for $i \neq 2, 3$.

To prove the proposition it suffices to show that the map $H^2(E) \to H^3(W)$ is zero. As $H^2(E) = \mathbf{C}(-1)^\mu$ has a pure Hodge structure of weight 2 it suffices to show that all the nontrival Hodge weights of $H^3(W)$ are at least 3. If W is smooth then this is trivially true, so suppose that W is singular.

Consider the long exact sequence of the pair (W, W_{smooth}). Since W has only ADE singularities and the dimension of W is even it follows that $H^i_\Sigma(W) =$

0 for $i \neq 4$, and $H^4_\Sigma(W) = \mathbf{C}(-2)^s$. The long exact sequence of the pair (W', W'_{smooth}) now yields isomorphisms $H^i(W) \cong H^i(W_{\text{smooth}})$ for $i \neq 3, 4$ and an exact sequence

$$0 \to H^3(W) \to H^3(W_{\text{smooth}}) \to \mathbf{C}(-2)^{\#\Sigma} \to H^4(W') \to 0 = H^4(W_{\text{smooth}}).$$

Since W_{smooth} is smooth we have that the Hodge weights of $H^3(W_{\text{smooth}})$ are at least 3, and hence the same statement holds true for $H^3(W)$. □

LEMMA 2.9. *Consider the inclusion $i : W \to \mathbf{P}$. Then $i^* : H^k(\mathbf{P}) \to H^k(W)$ is an isomorphism for $k = 0, 1, 3$, is injective for $k = 2$ and is surjective for $k = 4$.*

Proof. For $k = 0$ the statement is trivial. The case $k = 1$ can be shown as follows: Consider $\sigma_0 : C \to W$ and $i \circ \sigma_0 : C \to \mathbf{P}$. Combining these morphisms with $\pi : W \to C$, respectively $\varphi : \mathbf{P} \to C$, yield the identity on C. This implies that $\pi^* \circ \sigma_0^*$ and $\varphi^* \circ (i \circ \sigma_0)^*$ are isomorphisms and that $\sigma_0^* : H^k(C) \to H^k(W)$ is injective.

From [9, Lemma IV.1.1] it follows that $h^1(C) = h^1(X)$ and by the previous proposition we have $h^1(W) = h^1(X)$. In particular σ_0^* and $(i\sigma_0)^*$ are isomorphisms and therefore i^* is an isomorphism.

For $k = 2$ note that $H^2(\mathbf{P})$ is generated by the first Chern classes of a fiber of φ and $\mathcal{O}_{\mathbf{P}}(1)$. Their images in $H^2(X)$ are clearly independent, hence the composition $H^2(\mathbf{P}) \to H^2(W) \to H^2(X)$ is injective. For $k = 4$ note that the selfintersection of $c_1(\mathcal{O}_{\mathbf{P}}(1)) \in H^4(\mathbf{P})$ is mapped to a nonzero element in the one-dimensional vector space $H^4(X)$. Hence $H^4(\mathbf{P}) \to H^4(W) \to H^4(X)$ is surjective. Since $H^4(W) \cong H^4(X)$ this case follows also.

The case $k = 3$ is slightly more complicated. By successively blowing up points in $\mathbf{P}$ we find a variety $\tilde{\mathbf{P}}$ such that the strict transform of W is isomorphic with X. Now let H be an ample class of $\tilde{\mathbf{P}}$ and H_X its restriction to X. From the hard Lefschetz theorem it follows that the cupproduct with the class of $H|_X$ induces an isomorphism $H^1(X) \to H^3(X)$. Since $i^* : H^1(\mathbf{P}) \to H^1(W)$ is an isomorphism it follows that $H^1(\tilde{\mathbf{P}}) \to H^1(X)$ is an isomorphism. Therefore we find a morphsim $H^1(\tilde{\mathbf{P}}) \to H^3(X)$. We can factor this morphism also as first taking the cupproduct with H, and then applying i. Hence $i^* : H^3(\tilde{\mathbf{P}}) \to H^3(X)$ is surjective. Since we blow up only smooth points in $\mathbf{P}$ we find $H^3(\tilde{\mathbf{P}}) = H^3(\mathbf{P})$ and we showed before that $H^3(X) = H^3(W)$. Hence $H^3(\mathbf{P}) \to H^3(X)$ is surjective, and is an isomorphism because both vector spaces are of the same dimension. □

3. The $\mathbf{C}[G]$-structure of $H^{p,q}(X')$

Let $\mathcal{E}$ be a rank $n+1$ vector bundle on a smooth curve C. Let $X \subset \mathbf{P}(\mathcal{E})$ be a hypersurface such that either X is smooth or X is a surface with ADE singularities.

Let $f : C' \to C$ be a Galois cover with group G, such that $X' := X \times_C C'$ is smooth or X' is a surface with ADE singularities.

We now want to describe the $\mathbf{C}[G]$-module structure of $H^{p,q}(X')$. For this we prove first four easy lemmas concerning identities between representations.

Definition 3.1. *For a scheme Z with a G-action and a sheaf $\mathcal{F}$, denote with $\chi_G(\mathcal{F})$ the* equivariant Euler characteristic

$$\sum_i (-1)^i [H^i(Z, \mathcal{F})]$$

in $K(\mathbf{C}[G])$, the Grothendieck group of all finitely generated $\mathbf{C}[G]$-modules.

In the sequel we use the following lemma, which can be proven by "the usual devissage argument" and Serre duality:

Lemma 3.2 ([11, Lemma 5.6]). *Let $f : C' \to C$ be a ramified Galois cover with group G. If $\mathcal{M}$ is a line bundle on C, then*

$$\chi_G(f^*\mathcal{M}) = \deg(\mathcal{M})\mathbf{C}[G] + \chi_G(\mathcal{O}_{C'})$$

and

$$\chi_G(f^*\mathcal{M} \otimes K_{C'}) = \deg(\mathcal{M})\mathbf{C}[G] - \chi_G(\mathcal{O}_{C'}).$$

Let R be the set over which f is ramified. If R is non-empty then let Z be the zero-dimensional scheme on C' such that

$$0 \to K_{C'} \to f^*K_C(R) \to \mathcal{O}_Z \to 0 \tag{3}$$

is exact. Let s be the number of points in R.

Lemma 3.3. *Let $f : C' \to C$ be a Galois cover of curves, with group G. If f is unramified then*

$$[H^0(K_{C'})] = [H^0(f^*K_C)] = [\mathbf{C}] + (g(C) - 1)[\mathbf{C}[G]].$$

If f is ramified then

$$2[H^0(K_{C'})] + [H^0(\mathcal{O}_Z)] = 2[\mathbf{C}] + (2g(C) - 2 + s)[\mathbf{C}[G]].$$

Proof. If f is ramified then the degree of $f^*K_C(R)$ is strictly larger than $2g(C') - 2$, hence its first cohomology group vanishes and we obtain from Lemma 3.2 that

$$[H^0(f^*K_C(R))] = [\mathbf{C}] - [H^0(K_{C'})] + (2g(C) - 2 + s)[\mathbf{C}[G]].$$

From the exact sequence (3) we obtain that

$$[H^0(K_{C'})] - \mathbf{C} = [H^0(K_{C'})] - [H^1(K_{C'})] = [H^0(f^*K_C(R))] - [H^0(\mathcal{O}_Z)].$$

Combining this yields

$$2[H^0(K_{C'})] + [H^0(\mathcal{O}_Z)] = 2[\mathbf{C}] + (2g(C) - 2 + s)[\mathbf{C}[G]].$$

If f is unramified then $f^*K_C = K'_C$. Lemma 3.2 implies now

$$\chi_G(K_{C'}) = \deg(K_C)[\mathbf{C}[G]] + \chi_G(\mathcal{O}_{C'}).$$

From $\chi_G(\mathcal{O}_{C'}) = -\chi_G(K_{C'})$ we obtain

$$2\chi_G(K_{C'}) = (2g(C) - 2)[\mathbf{C}[G]].$$

The result now follows from $\chi_G(K_{C'}) = [H^0(K_{C'})] - [\mathbf{C}]$. □

LEMMA 3.4. *Let $f : C' \to C$ be a Galois cover of curves, with group G. Then $H^0(K_{C'})^{\oplus 2}$ is a quotient of $\mathbf{C}^{\oplus 2} \oplus \mathbf{C}[G]^{\oplus 2g(C)-2+s}$.*

Proof. This follows directly from the previous lemma. □

REMARK 3.5: The Chevalley–Weil formula gives a precise description of the $\mathbf{C}[G]$-structure of $H^0(K_{C'})$, see [1].

We will now go back to our hypersurface $X' \subset \mathbf{P}(f^*(\mathcal{E}))$. Denote with $\varphi : \mathbf{P}(f^*\mathcal{E}) \to C'$ and $\varphi_0 : \mathbf{P}(\mathcal{E}) \to C$ the natural projection maps.

We will now prove a structure theorem for the $\mathbf{C}[G]$-module $H^{p,q}(X')$.

PROPOSITION 3.6. *Suppose that $\mathcal{E}$ is a direct sum of line bundles. Let $X \subset \mathbf{P}(\mathcal{E})$ be a hypersurface, and $X' = X \times_C C'$. Then for $i > 0, k \geq 0$ we have that $\chi_G(\Omega^i_{\mathbf{P}(f^*\mathcal{E})}(kX'))$ is a direct sum of copies of $\mathbf{C}[G]$ and $\chi_G(\mathcal{O}_{C'})$.*

Proof. Let $\varphi : \mathbf{P}(f^*(\mathcal{E})) \to C'$ be the natural projection map. Consider the short exact sequence

$$0 \to \varphi^*K_{C'} \to \Omega^1_{\mathbf{P}(f^*\mathcal{E})} \to \Omega^1_\varphi \to 0.$$

On $\Omega^t_{\mathbf{P}(f^*\mathcal{E})}$ there is a filtration such that $\mathrm{Gr}^p = \wedge^p \varphi^*(K_{C'}) \otimes \Omega^{t-p}_\varphi$ [7, Exer. II.5.16]. From $\wedge^p \varphi^* K_{C'} = 0$ for $p > 1$ it follows that at most two of the Gr^ps are nonzero and they fit in the exact sequence

$$0 \to \varphi^*(K_{C'}) \otimes \Omega^{t-1}_\varphi \to \Omega^t_{\mathbf{P}(f^*(\mathcal{E}))} \to \Omega^t_\varphi \to 0. \tag{4}$$

Similarly, consider the Euler sequence

$$0 \to \Omega^1_\varphi \to (\varphi^* f^* \mathcal{E})(-1) \to \mathcal{O}_{\mathbf{P}(f^*\mathcal{E})} \to 0.$$

By using the filtration constructed in [7, Exer. II.5.16] again we obtain the following exact sequence

$$0 \to \Omega^t_\varphi \to \wedge^t(\varphi^* f^* \mathcal{E})(-1) \to \Omega^{t-1}_\varphi \to 0. \tag{5}$$

Let $\mathcal{L} \in \operatorname{Pic}(C)$ and $d > 0$ be such that $\mathcal{O}_{\mathbf{P}(f^*(\mathcal{E}))}(kX') = (\varphi^* f^*(\mathcal{L}))(d)$. A straightforward exercise using the exact sequence (4) tensored with $\mathcal{O}(kX')$, the exact sequence (5) tensored with $\mathcal{O}(kX')$ respectively with $\mathcal{O}(kX') \otimes \varphi^*(K_{C'})$ and induction on t yields that $\chi_G(\Omega^i_{\mathbf{P}(f^*\mathcal{E})}(\varphi^* f^* \mathcal{L})(d))$ equals

$$\sum_{i=0}^{t}(-1)^{t-i}\chi_G((\Lambda_i \otimes \varphi^* f^* \mathcal{L})(d)) + \sum_{i=0}^{t-1}(-1)^{t-i}\chi_G((\Lambda_i \otimes \varphi^*(f^*\mathcal{L} \otimes K_{C'}))(d))$$

with

$$\Lambda_t := \wedge^t(\varphi^* f^* \mathcal{E})(-1).$$

Using that $R^i\varphi_*(\mathcal{O}(k)) = 0$ for $i > 0, k \geq -1$ (see [17]) and the projection formula again we obtain that $\chi_G(\mathcal{F}) = \chi_G(\varphi_*\mathcal{F})$ where $\mathcal{F}$ is one of

$$(\wedge^t(\varphi^* f^* \mathcal{E})(d-1)) \otimes \varphi^*(f^*(\mathcal{L})), (\wedge^t(\varphi^* f^* \mathcal{E})(d-1)) \otimes \varphi^*(K_{C'} \otimes f^*(\mathcal{L})). \tag{6}$$

Since $\mathcal{E}$ is a sum of line bundles, we obtain that

$$(\wedge^t f^* \mathcal{E})$$

is a direct sum of line bundles pulled back from C. Similarly we obtain that

$$R^i\varphi_*\mathcal{O}(k) = \operatorname{Sym}^k(f^*\mathcal{E})$$

is a direct sum of line bundles pulled back from C and by using the projection formula we have that $\varphi_*\mathcal{F}$ is the direct sum of line bundles pulled back from C, for $\mathcal{F}$ as in (6).

We can therefore calculate the relevant equivariant Euler characteristic by Lemma 3.2, and we obtain that $\chi_G(\varphi_*(\mathcal{F}))$ is a sum of copies $\chi_G(K_{C'})$ and $\mathbf{C}[G]$ for $\mathcal{F}$ as in (6). The multiplicity of $\mathbf{C}[G]$ depends on the sum of degrees of the direct summands and the multiplicity of $\chi_G(K_{C'})$ on the rank of $\mathcal{F}$. Hence the multiplicity of $\chi_G(K_{C'})$ and $\mathbf{C}[G]$ in $\chi_G(\Omega^i(kX'))$ depend only on i, k, the fiberdegree of X' and the degrees of the direct summand of $\mathcal{E}$. □

Remark 3.7: Note that the proof of the theorem also yields a method to determine the number of copies of $\mathbf{C}[G]$, respectively, $\chi_G(\mathcal{O})$ which occur. In the next section we make this precise for the case $\mathcal{E} = \mathcal{O} \oplus \mathcal{L}^{-2} \oplus \mathcal{L}^{-3}$, $X \in |(\varphi^* f^* \mathcal{L}^6)(3)|$ and $(i, k) = (2, 1), (3, 1), (3, 2)$.

Proposition 3.8. *Let $n \geq 2$. Let $X \subset \mathbf{P}$ be a n-dimensional smooth hypersurface. Assume that for $i : X \subset \mathbf{P}$ we have that $i^* : H^k(\mathbf{P}, \mathbf{C}) \to H^k(X, \mathbf{C})$ is an isomorphism for $k < n$ and that for $k = n$ this map is injective. Let $U = \mathbf{P} \setminus X$. Then $H^i(U) = 0$ for $i \neq 0, 1, 2, n+1$ Moreover, we have isomorphisms $H^0(U) \cong \mathbf{C}$, $H^1(U) \cong H^1(C)$, $H^2(X) \cong \mathbf{C}(-1)$ and $H^n(U)(1) \cong \operatorname{coker} H^{n-1}(\mathbf{P}) \to H^{n-1}(X)$.*

Proof. Consider the Gysin exact sequence for cohomology with compact support

$$\cdots \to H^k_c(U) \to H^k_c(\mathbf{P}) \to H^k_c(X) \to H^{k+1}_c(U) \to \ldots$$

Our assumption on i^* now yields $H^k_c(U) = 0$ for $k \leq n$.

Let $\mathcal{M}$ be an ample line bundle on $\mathbf{P}$, and $\mathcal{M}'$ be its restriction to X. Then by the hard Lefschetz theorem we get that the k-fold cupproduct with $c_1(\mathcal{M}')$ yields an isomorphism $H^k(X, \mathbf{C}) \to H^{n-k}(X, \mathbf{C})$. For $0 < k \leq n$ we obtain an isomorphism

$$H^k(\mathbf{P}, C) \to H^k(X, \mathbf{C}) \to H^{n-k}(X, \mathbf{C}).$$

We can factor this isomorphism as first taking the k-fold cupproduct with $c_1(\mathcal{M})$ and then applying i^*. In particular the map $H^{n-k}(\mathbf{P}) \to H^{n-k}(X)$ is surjective. The Betti numbers of $\mathbf{P}$ are well-known, namely $h^0(\mathbf{P})$ and $h^{2n+2}(\mathbf{P})$ equal 1,

$$h^{2k}(\mathbf{P}) = 2 \text{ for } k = 1, \ldots, n \text{ and } h^{2k+1} = h^1(C) \text{ for } k = 0, \ldots, n.$$

These facts yield that $H^i(\mathbf{P}) \cong H^i(X)$ for $i = 0, \ldots, n-1$ and $i = n+1, \ldots, 2n-1$. Hence $H^i_c(U) = 0$ for $i \neq n+1, 2n, 2n+1, 2n+2$. Moreover we have two exact sequences

$$0 \to H^n(\mathbf{P}) \to H^n(X) \to H^{n+1}_c(U) \to 0$$

and

$$0 \to H^{2n}_c(U) \to H^{2n}(\mathbf{P}) \to H^{2n}(X) \to 0$$

and isomoprhisms $H^i_c(U) \cong H^i_c(\mathbf{P})$ for $i = 2n+1, 2n+2$.

Applying Poincaré duality now gives the result. □

Denote with $\Omega^{p,\mathrm{cl}}_{\mathbf{P}}$ or $\Omega^{p,\mathrm{cl}}$ the sheaf of closed p-forms on $\mathbf{P}$. Recall that for a hypersurface $X \subset \mathbf{P}$ we have $\Omega^{p,\mathrm{cl}}(X) = \Omega^{p,\mathrm{cl}}(\log X)$.

Proposition 3.9. *Let $X \subset \mathbf{P}$ be a n-dimensional smooth hypersurface. Suppose $n \geq 2$. Let $G \subset \operatorname{Aut}(\mathbf{P}, X)$ be a subgroup. Assume that for $i : X \subset \mathbf{P}$ we have that $i^* : H^k(\mathbf{P}, \mathbf{C}) \to H^k(X, \mathbf{C})$ is an isomorphism for $k < n$ and that for $k = n$ this map is injective.*

Then for $p \geq 1$ we have $(-1)^{n-p}([H^{p,n-p}(X)] - [H^{p,n-p}(\mathbf{P})])$ equals

$$\sum_{k=1}^{n-p+1} (-1)^k \chi_G(\Omega^{p+k}(kX)) + \sum_{k=1}^{n-p}(-1)^k \chi_G(\Omega^{p+1+k}(kX)).$$

and for $p = 0$ we find that

$$[H^0(K_{C'})] - [\mathbf{C}] + (-1)^n [H^{0,n}(X)]$$

equals

$$\sum_{k=1}^{n+1}(-1)^k \chi_G(\Omega^k(kX)) + \sum_{k=1}^{n}(-1)^k \chi_G(\Omega^{k+1}(kX)).$$

Proof. Let U be the complement of X in $\mathbf{P}$. From the previous proposition it follows that

$$[H^{p,n-p}(X)] - [H^{p,n-p}(\mathbf{P})] = [\mathrm{Gr}_F^{p+1} H^{n+1}(U)].$$

Hence we will focus on determining the $\mathbf{C}[G]$ structure of $\mathrm{Gr}_F^{p+1} H^{n+1}(U)$.

From Deligne's construction of the Hodge filtration on the cohomology of U we get

$$F^p H^k(U, \mathbf{C}) = \mathrm{Im}(\mathbf{H}^k(\Omega^{\geq p}_{\mathbf{P}(\mathcal{E})}(\log X)) \to \mathbf{H}^k(\Omega^{\bullet}_{\mathbf{P}(\mathcal{E})}(\log X))).$$

The map is injective by the degeneracy of the Fröhlicher spectral sequence at E_1. Recall that $\Omega^{p,\mathrm{cl}}(X)$ is the kernel of $d : \Omega^p(X) \to \Omega^{p+1}(2X)$. For $p \geq 1$ we have that the filtered de Rham complex is a resolution of $\Omega^{p,\mathrm{cl}}(X)$. Combining these fact we obtain for $p \geq 1$ that

$$F^p H^{p+q}(U, \mathbf{C}) = H^q(X, \Omega^{p,\mathrm{cl}}(X)).$$

For $p > 1$ we have $\mathrm{Gr}_F^p H^{p+q}(U, \mathbf{C}) = 0$ except possibly for $q = n + 1 - p$. In particular, $H^q(\Omega^{p,\mathrm{cl}}(X)) = 0$ for $q \neq n + 1 - p$, $p \geq 2$. Hence for $p \geq 2$ we obtain that $\chi_G(\Omega^{p,\mathrm{cl}}(X))$ equals

$$(-1)^{n+1-p}[H^{n+1-p}(X, \Omega^{p,c}(X))] = (-1)^{n+1-p} F^p H^{n+1}(U, \mathbf{C}).$$

The exact sequence

$$0 \to \Omega^{p,\mathrm{cl}}(tX) \to \Omega^p(tX) \to \Omega^{p+1,\mathrm{cl}}((t+1)(X)) \to 0$$

then yields

$$\chi_G(\Omega^{p,\mathrm{cl}}(tX)) = \sum_{k=0}^{n+1-p} (-1)^k \chi_G(\Omega^{p+k}((t+k)X)).$$

From this we obtain that for $p \geq 1$ we have that

$$\mathrm{Gr}_F^p \operatorname{coker}(H^n(\mathbf{P}) \to H^n(X)) = \mathrm{Gr}_F^{p+1} H^{n+1}(U)$$

equals $(-1)^{n-p}$ times

$$\sum_{k=1}^{n-p+1} (-1)^k \chi_G(\Omega^{p+k}(kX)) + \sum_{k=1}^{n-p} (-1)^k \chi_G(\Omega^{p+1+k}(kX))).$$

For $p = 0$ we find

$$\begin{aligned} \chi_G(\Omega^{1,\mathrm{cl}}(X)) &= [F^1H^1(U,\mathbf{C})] - [F^1H^2(U,\mathbf{C})] + (-1)^n[F^1H^{n+1}(U,\mathbf{C})] \\ &= [H^0(\Omega^{1,\mathrm{cl}}(X))] - [H^1(\Omega^{1,\mathrm{cl}})] + (-1)^n[H^n(\Omega^{1,\mathrm{cl}}(X))]. \end{aligned}$$

From Proposition 3.8 it follows that

$$[F^1H^1(U,\mathbf{C})] = [H^0(K_{C'}))] \text{ and } [F^1H^2(U,\mathbf{C})] = [\mathbf{C}]$$

holds. As above we find that

$$[H^0(K_{C'})] - [\mathbf{C}] + (-1)^n[\mathrm{Gr}_F^0 \operatorname{coker}(H^n(\mathbf{P}) \to H^n(X)]$$

equals

$$\sum_{k=1}^{n+1}(-1)^k \chi_G(\Omega^k(kX)) + \sum_{k=1}^{n}(-1)^k \chi_G(\Omega^{k+1}(kX))).$$

□

Let P be a smooth compact Kähler manifold. Steenbrink [16] extended Deligne's approach to the class of hypersurfaces $X \subset P$, such that the sheaf of Du Bois differentials of X and the sheaf of Barlet differentials of X coincide. This happens only for few classes of singularities. The only known singular varieties for which this property holds are surfaces. Streenbrink [16] gave three classes of examples, one of which are surfaces with ADE singularities [16, Section 3].

To explain Steenbrink's results, let $X \subset P$ be a hypersurface, with at most isolated singularities. Let $\mathcal{T}$ be the skyscraper sheaf supported on the singular locus, such that at each point p the stalk $\mathcal{T}_p$ is the Tjurina algebra of the singularity (X, p).

The following proposition summarizes Steenbrink's method in the case of a three-dimensional ambient space P: Note that if X is a surface with at most ADE singularities then the mixed Hodge structure on $H^i(X)$ is pure of weight i. Hence it makes sense to define $H^{p,q}(X) := \mathrm{Gr}_F^p H^{p+q}(X)$.

PROPOSITION 3.10. *Let P be a smooth compact three-dimensional Kähler manifold, and let $X \subset P$ be a surface with at most ADE singularities. For all $G \subset \mathrm{Aut}(P, X)$ we have $[H^{0,2}(X)] = [H^0(K_P(X))]$ and that $[H^{1,1}(X)]$ equals*

$$[H^{2,0}(P)] + [H^{2,2}(P)] + [H^{1,0}(X)] + [H^{1,2}(X)] - [H^{2,1}(P)] - [H^{2,3}(P)] \\ - \chi_G(\Omega^2_P(X)) + \chi_G(K_P(2X)) - \chi_G(K_P(X)) - \chi_G(\mathcal{T})$$

in $K(\mathbf{C}[G])$.

Proof. Since ADE singularities are rational we get that

$$H^{0,2}(X) = H^0(K_P(X))$$

(see, e.g., [16, Introduction]).

The second equality follows from [16]:

Let $\Omega^2_X(\log X)$ be the kernel of $\Omega^2(X) \xrightarrow{d} K_P(2X)/K_P(X)$. Since X has ADE singularities we have that the cokernel of d is $\mathcal{T}$ [16, Section 2]. Define $\omega^1_X = \Omega^2_P(\log X)/\Omega^2_P$ to be the sheaf of Barlet 1-forms on X.

Consider now the filtered de Rham complex $\tilde{\Omega}^\bullet_X$ on X, as introduced by Du Bois [2].

Since X has ADE singularities it follows from [16, Section 4] that $\mathrm{Gr}^1_F \tilde{\Omega}^\bullet_X$ is concentrated in degree one, and in this degree it is isomorphic to $\tilde{\Omega}^1_X$. Moreover, in the same section Steenbrink shows that for a surface with ADE singularities we have $\tilde{\Omega}^1_X \cong \omega^1_X$. This implies $H^i(\omega^1_X) = \mathrm{Gr}^1_F H^{1+i}(X)$ and hence

$$\chi_G(\omega^1_X) = [H^{1,0}(X)] - [H^{1,1}(X)] + [H^{1,2}(X)].$$

The definition of ω^1_X yields the equality

$$\chi_G(\omega^1_X) = \chi_G(\Omega^2_P(\log X)) - \chi_G(\Omega^2_P).$$

Since P is a smooth threefold we find that

$$\chi_G(\Omega^2_P) = [H^{2,0}(P)] - [H^{2,1}(P)] + [H^{2,2}(P)] - [H^{2,3}(P)].$$

Using the definition of $\Omega^2_P(\log X)$ we find

$$\chi_G(\Omega^2_P(\log X)) = \chi_G(\Omega^2_P(X)) - \chi_G(K_P(2X)) + \chi_G(K_P(X)) + \chi_G(\mathcal{T}).$$

□

REMARK 3.11: If $H^i(X) \cong H^i(P)$ holds for $i = 1$ and $i = 3$ then

$$[H^{1,0}(X)] + [H^{1,2}(X)] = [H^{2,1}(P)] + [H^{2,3}(P)].$$

If, moreover, $H^{2,0}(P) = 0$ we have further simplifications in the formula from Proposition 3.10.

In case $P = \mathbf{P}(\mathcal{O} \oplus f^*\mathcal{L}^{-2} \oplus f^*\mathcal{L}^{-3})$ and X a Weierstrass model all these cancellations happen, and, moreover, $[H^{2,2}(P)] = 2[\mathbf{C}]$ in $K(\mathbf{C}[G])$.

Corollary 3.12. *Let $\mathcal{E}$ be a direct sum of at least three line bundles on a smooth projective curve C. Let $X \subset \mathbf{P}(\mathcal{E})$ be a hypersurface. Let $f : C' \to C$ be a Galois cover. Let $X' = X \times_C C' \subset \mathbf{P}(f^*\mathcal{E})$ be the base-changed hypersurface. Assume that the natural map $H^i(\mathbf{P}(f^*(\mathcal{E}))) \to H^i(X')$ is an isomorphism for $0 \leq i < \dim X'$ and for $i = \dim X'$ this map is injective.*

If X' is smooth then for each $p, q \in \mathbf{Z}$ there exist integers a, b, c, depending on p, q, the degrees of the direct summands of $\mathcal{E}$ and the fiber degree of X, such that $[H^{p,q}(X')] = a[\mathbf{C}] + b\chi_G(\mathcal{O}) + c[\mathbf{C}[G]]$.

If X' is surface with at most ADE singularities for each $p, q \in \mathbf{Z}$ there exist integers a, b, c, depending on p, q, the degrees of the direct summand of $\mathcal{E}$ and the fiber degree of X, such that $[H^{p,q}(X')] = a\mathbf{C} + b\chi_G(\mathcal{O}) + c[\mathbf{C}[G]] + \delta[H^0(\mathcal{T})]$, where $\delta = 0$ for $(p, q) \neq (1, 1)$ and $\delta = 1$ for $(p, q) = (1, 1)$.

Corollary 3.13. *Let $\mathcal{E}$ be a direct sum of three line bundles. Let $W \subset \mathbf{P}(\mathcal{E})$ be a surface. Let $C' \to C$ be a Galois base change such that $W' := W \times_C C'$ is a surface with at most ADE singularities and such that $H^1(W') \cong H^1(\mathbf{P})$. Let X' be the desingularization of W'. Then $[H^{1,1}(W')]$ equals*

$$2[\mathbf{C}] - \chi_G(\Omega^2(W')) + \chi_G(K_{\mathbf{P}(f^*\mathcal{E})}(2W')) - \chi_G(K_{\mathbf{P}(f^*\mathcal{E})}(W')) - \chi_G(\mathcal{T})$$

and

$$[H^{1,1}(X')] = 2[\mathbf{C}] - \chi_G(\Omega^2(W')) + \chi_G(K_{\mathbf{P}(f^*\mathcal{E})}(2W')) - \chi_G(K_{\mathbf{P}(f^*\mathcal{E})}(W')).$$

Proof. The formula for $[H^{1,1}(W')]$ follows directly from Proposition 3.10. The quotient $H^{1.1}(X')/H^{1,1}(W')$ is generated by the irreducible components of the resolution $X' \to W'$ and one easily checks that the representation induced by G-action on these irreducible components equlas $\mathcal{T}$. $\square$

Remark 3.14: Note that $[H^{1,1}(X')]$ depends only on the linear equivalence class of W', and not on the singularities of W'. If $|W|$ is base point free then there is a different approach to obtain this statement. In this case W' is the limit of a family of smooth surfaces, all of which are pulled back from $\mathbf{P}(\mathcal{E})$, and W' has at most ADE singularities. In particular there is a simultaneous resolution of singularities of this family. The central fiber of this resolution is X', and this implies the $\mathbf{C}[G]$-structure of $H^{p,q}(X')$ is the same as the one on the general member of this family.

4. The $\mathbf{C}[G]$-structure of the cohomology of Weierstrass models

We want to apply the results of the previous section to the special case of Weierstrass models. In the first part of the section we only assume that $\mathcal{E}$ is a direct sum of three line bundles. Let $C, C', X, X', \mathbf{P}_0, \mathbf{P}, \varphi, \varphi_0$ be as in the previous section. Assume that $\dim X = 2$.

We want to determine the $\mathbf{C}[G]$-structure of $H^{1,1}(X)$ and of $H^{2,0}(X)$. By Corollary 3.13 it suffices to determine the $\mathbf{C}[G]$-structure of

$$\chi_G(\Omega^2_{\mathbf{P}}(X)), \chi_G(K_{\mathbf{P}}(X)) \text{ and } \chi_G(K_{\mathbf{P}}(2X))$$

and the $\mathbf{C}[G]$-structure on $H^0(\mathcal{T})$.

We will determine the structure on $H^0(\mathcal{T})$ below. A strategy to calculate the three equivariant Euler characteristics is given in the proof of Proposition 3.6. The main ingredients are

1. $\Omega^3_{\mathbf{P}} \cong \varphi^* \det(f^*\mathcal{E} \otimes K_{C'})(-3)$ (adjunction).
2. $\Omega^2_{\varphi} \cong \varphi^*(\det(f^*\mathcal{E}))(-3)$.
3. $0 \to \Omega^1_{\varphi} \to \varphi^* f^* \mathcal{E}(-1) \to \mathcal{O}_{\mathbf{P}} \to 0$ (Euler sequence).
4. $0 \to \Omega^1_{\varphi} \otimes \varphi^* K_{C'} \to \Omega^2_{\mathbf{P}} \to \Omega^2_{\varphi} \to 0$.

The points (2)-(4) easily yield

Lemma 4.1. *Let $X \subset \mathbf{P}(\mathcal{E})$ be a hypersurface in $|(\varphi^* f^* \mathcal{L})(d)|$, fixed under G. Then $\chi_G(\Omega^2(X))$ equals*

$$\chi_G(\varphi^* f^*(\mathcal{L} \otimes \det \mathcal{E})(d-3)) + \chi_G(\varphi^* f^*(\mathcal{L} \otimes \mathcal{E})(d-1)) - \chi_G(\varphi^*(f^*\mathcal{L} \otimes K_{C'})(d)).$$

It turns out that if $\mathcal{E}$ is a direct sum of line bundles then we can express all of the above equivariant Euler characteristics in terms of equivariant Euler characteristics of sheaves of the form $(\varphi^* f^* \mathcal{F})(k)$ and $\varphi^*(f^*\mathcal{F} \otimes K_{C'})(k)$, where $\mathcal{F}$ is a direct sum of line bundles on C. The following lemmas are helpful in calculating χ_G of such sheaves.

Lemma 4.2. *Suppose $\mathcal{E} = \mathcal{O}_{C'} \oplus \mathcal{L} \oplus \mathcal{M}$, with $\deg(\mathcal{L}), \deg(\mathcal{M}) \leq 0$. Then $\varphi_* \mathcal{O}_{\mathbf{P}(\mathcal{E})}(t)$ is the pullback under f^* of a direct sum of $\binom{k+2}{2}$ line bundles, such that the sum of the degrees equals*

$$\frac{1}{6} t(t+1)(t+2)(\deg(\mathcal{L}) + \deg(\mathcal{M})).$$

Proof. Since $\mathcal{E} = \mathcal{O}_C \oplus \mathcal{L} \oplus \mathcal{M}$ we can pick canonical sections X, Y, Z in $H^0(\varphi^* \mathcal{L}^{-1}(1))$, $H^0(\varphi^* \mathcal{M}^{-1}(1))$ and $H^0(\mathcal{O}_{\mathbf{P}}(1))$ (cf. Section 2). Note that

$$\varphi_* \mathcal{O}(t) = \oplus_{0 \leq i+j \leq t} (f^* \mathcal{L}^i \otimes f^* \mathcal{M}^j) X^i Y^j Z^{t-i-j}.$$

Hence the sum of the degrees equals

$$\sum_{0 \leq i+j \leq t} (\deg(\mathcal{L}) i + \deg(\mathcal{M}) j) = \frac{1}{6} t(t+1)(t+2)(\deg(\mathcal{L}) + \deg(\mathcal{M})).$$

□

Lemma 4.3. *Suppose* $\mathcal{E} = \mathcal{O}_{C'} \oplus f^*\mathcal{L} \oplus f^*\mathcal{M}$, *with* $\deg(\mathcal{L}), \deg(\mathcal{M}) \leq 0$. *Let* $\mathcal{N}$ *be a line bundle on* C. *Let* $t \geq 0$ *be an integer. Set*

$$d = \binom{t+2}{3}(\deg(\mathcal{L}) + \deg(\mathcal{M})) + \binom{t+2}{2}\deg(\mathcal{N}).$$

Then

$$\chi_G((\varphi^* f^*\mathcal{N})(t)) = d\mathbf{C}[G] + \frac{t+2}{2}\chi_G(\mathcal{O}_{C'})$$

and

$$\chi_G(\varphi^*(K_{C'} \otimes f^*\mathcal{N})(t)) = d\mathbf{C}[G] - \frac{t+2}{2}\chi_G(\mathcal{O}_{C'}).$$

Proof. Since $R^i\varphi_*\mathcal{O}(t) = 0$ for $i > 0$ we find that

$$H^k(X, (\varphi^* f^*\mathcal{N})(t)) = H^k(X, \varphi_*((\varphi^* f^*\mathcal{N})(t))).$$

Combining this with the projection formula yields

$$\chi_G((\varphi^* f^*\mathcal{N})(t)) = \chi_G((f^*\mathcal{N}) \otimes \varphi_*\mathcal{O}(t)).$$

Since $\varphi_*\mathcal{O}(t)$ is a direct sum of line bundles pulled back from C, the same holds for $f^*\mathcal{N} \otimes \varphi_*\mathcal{O}(t)$. The sum of the degree of the line bundles on C equals d. It follows now from Lemma 3.2 that

$$\chi_G((f^*\mathcal{N}) \otimes \varphi_*\mathcal{O}(t)) = d\mathbf{C}[G] + \frac{t+2}{2}\chi_G(\mathcal{O}_{C'}).$$

The Euler characteristic $\chi_G(\varphi^*(K_{C'} \otimes f^*\mathcal{N})(t))$ can be calculated similarly, by using Serre duality on C'. □

From here on we assume that $\mathcal{E} = \mathcal{O} \oplus f^*\mathcal{L}^{-2} \oplus f^*\mathcal{L}^{-3}$ and that $W \in |\varphi_0^*\mathcal{L}^6(3)|$ and hence that $X = W' \in |\varphi^* f^*\mathcal{L}^6(3)|$.

We will now repeatedly apply Lemma 4.3 to determine all the relevant Euler characteristics:

Lemma 4.4. *In* $K(\mathbf{C}[G])$ *we have*

$$\chi_G(K_{\mathbf{P}}(W')) = \deg(\mathcal{L})[\mathbf{C}[G]] - \chi_G(\mathcal{O}_{C'})$$

and

$$\chi_G(K_{\mathbf{P}}(2W')) = 20\deg(\mathcal{L})[\mathbf{C}[G]] - 10\chi_G(\mathcal{O}_{C'}).$$

Proof. Note that

$$K_{\mathbf{P}} = \varphi^*(\det(\mathcal{E}) \otimes K_{C'}(-3)) = \varphi^*(f^*\mathcal{L}^{-5} \otimes K_{C'})(-3).$$

Hence $K_{\mathbf{P}}(W') = \varphi^* f^*(\mathcal{L} \otimes K_{C'})$. From Lemma 4.3 it now follows that $\chi_G(K_{\mathbf{P}}(W')) = \deg(\mathcal{L})[\mathbf{C}[G]] - \chi_G(\mathcal{O}_{C'})$.

Similarly $K_{\mathbf{P}}(W') = \varphi^* f^*(\mathcal{L}^7 \otimes K_{C'})(3)$. From Lemma 4.3 it follows now that

$$\chi_G(K_{\mathbf{P}}(2W')) = 20 \deg(\mathcal{L})[\mathbf{C}[G]] - 10\chi_G(\mathcal{O}_{C'}).$$

□

LEMMA 4.5. *In* $K(\mathbf{C}[G])$ *we have*

$$\chi_G(\Omega^2_\varphi(W')) = \deg(\mathcal{L})[\mathbf{C}[G]] + \chi_G(\mathcal{O}_{C'}).$$

Proof. Note that $\Omega^2_\varphi(W') = (\varphi^* f^* \mathcal{L}^{-5})(-3) \otimes \mathcal{L}^6(3) = \varphi^* f^*(\mathcal{L})$. Lemma 4.3 now yields

$$\chi_G(\Omega^2_\varphi(W')) = \deg(\mathcal{L})[\mathbf{C}[G]] + \chi_G(\mathcal{O}_{C'}).$$

□

LEMMA 4.6. *In* $K(\mathbf{C}[G])$ *we have*

$$\chi_G(\varphi^*(K_{C'}(W'))) = 10 \deg(\mathcal{L})[\mathbf{C}[G]] - 10\chi_G(\mathcal{O}_{C'}).$$

Proof. Using $\varphi^*(K_{C'})(W') = \varphi^*(K_{C'} \otimes f^*\mathcal{L}^6)(3)$ we obtain from Lemma 4.3

$$\chi_G(\varphi^*(K_{C'}(W'))) = 10 \deg(\mathcal{L})[\mathbf{C}[G]] - 10\chi_G(\mathcal{O}_{C'}).$$

□

LEMMA 4.7. *In* $K(\mathbf{C}[G])$ *we have*

$$\chi_G(\varphi^*(\mathcal{E} \otimes K_{C'})(W')(-1)) = 18 \deg(\mathcal{L})[\mathbf{C}[G]] - 18\chi_G(\mathcal{O}_{C'}).$$

Proof. Note that $\varphi^*(\mathcal{E} \otimes K_{C'})(W')(-1) = \varphi^*(\mathcal{E} \otimes K_{C'} \otimes f^*\mathcal{L}^6)(2)$. Hence

$$\varphi^*(\mathcal{E} \otimes K_{C'} \otimes f^*\mathcal{L}^6)(2) = \varphi^*((f^*\mathcal{L}^6 \oplus f^*\mathcal{L}^4 \oplus f^*\mathcal{L}^3) \otimes K_{C'})(2).$$

From Lemma 4.3 it follows that its Euler characteristic equals

$$18 \deg(\mathcal{L})[\mathbf{C}[G]] - 18\chi_G(\mathcal{O}_{C'}).$$

□

LEMMA 4.8. *In* $K(\mathbf{C}[G])$ *we have*

$$\chi_G(\Omega^2(W')) = 9 \deg(\mathcal{L})[\mathbf{C}[G]] - 7\chi_G(\mathcal{O}_{C'}).$$

Proof. From

$$0 \to \Omega^1_\varphi \otimes \varphi^* K_{C'}(W') \to \Omega^2(W') \to \Omega^2_\varphi(W') \to 0$$

and

$$0 \to \Omega^1_\varphi \otimes \varphi^* K_{C'}(W') \to \mathcal{E} \otimes \varphi^* K_C(W')(-1) \to \varphi^* K_C(W') \to 0.$$

It follows that $\chi_G(\Omega^2(W'))$ equals

$$\begin{aligned} \chi_G(\Omega^2_\varphi(W')) + \chi_G(\mathcal{E} \otimes \varphi^* K_C(W')(-1)) - \chi_G(\varphi^* K_C(W')) \\ = 9 \deg(\mathcal{L})[\mathbf{C}[G]] - 7\chi_G(\mathcal{O}_{C'}). \end{aligned}$$

□

Collecting everything we find:

PROPOSITION 4.9. *We have the following identities in $K(\mathbf{C}[G])$:*

$$[H^{2,0}(W')] = [H^{2,0}(X')] = \deg(\mathcal{L})\mathbf{C}[G] + [H^0(K_{C'})] - [\mathbf{C}],$$

$$[H^{1,1}(W')] = 10 \deg(\mathcal{L})[\mathbf{C}[G]] + 2[H^0(K_{C'})] - [H^0(\mathcal{T})]$$

and

$$[H^{1,1}(X')] = 10 \deg(\mathcal{L})[\mathbf{C}[G]] + 2[H^0(K_{C'})]$$

REMARK 4.10: A different proof for the formula for $H^{2,0}(X')$ can be found in [12, Theorem 2.5].

The $\mathbf{C}[G]$ action on $H^0(\mathcal{T})$ is hard to describe in general. However, if we make some assumption on the ramification locus then it simplifies a lot:

LEMMA 4.11. *Suppose the ramification locus of $W' \to W$ does not intersect W'_{sing}. Then*

$$[H^0(\mathcal{T})] = \mu[\mathbf{C}[G]]$$

where μ is the total Milnor number of W.

Proof. Let $\mathcal{T}_W$ and $\mathcal{T}_{W'}$ be the sheaves on W, resp. on W', such that at each point p the stalk is isomorphic to the Tjurina algebra at p. The length of $\mathcal{T}_W$ is the total Tjurina number of W, which equals the total Milnor number of W.

Since $\mathcal{T}_{W'}$ is supported outside the ramification locus, we find that $\mathcal{T}_{W'}$ is the pull back of $\mathcal{T}_W$ and it consists of $\#G$ copies of $\mathcal{T}_W$. In particular the G action on $H^0(\mathcal{T}_{W'})$ consists of μ copies of the regular representation. □

To obtain Pál's upper bound for the Mordell–Weil rank we need the following result, which directly implies the Shioda–Tate formula for the Mordell–Weil rank of an elliptic surface.

Proposition 4.12. *We have a short exact sequence of* $\mathbf{C}[G]$*-modules*

$$0 \to \mathbf{C}^2 \oplus H^0(\mathcal{T}) \to \mathrm{NS}(X') \to E(\mathbf{C}(C')) \to 0.$$

Proof. Let $T \subset \mathrm{NS}(X')$ be the trivial sub-lattice, the lattice generated by the class of a fiber, the image of the zero-section and the classes of irreducible components of reducible fibers. Shioda and Tate both showed that $E(\mathbf{C}(C'))$ is isomorphic to $\mathrm{NS}(X')/T$ as abelian groups.

The group G acts on T, $\mathrm{NS}(X')$ and $E(\mathbf{C}')$, and from the construction of this map it follows directly that this isomorphism is G-equivariant. Moreover the fiber components which do not intersect the zero-section are precisely the exceptional divisors of $X' \to W'$, i.e., they span a subspace isomorphic to $H^0(\mathcal{T})$. Since G maps a fiber to a fiber, and fixes the zero section, we find

$$0 \to \mathbf{C}^2 \oplus H^0(\mathcal{T}) \to \mathrm{NS}(X') \to E(\mathbf{C}(C')) \to 0$$

is exact. □

Theorem 4.13. *Let $X \to C$ be an elliptic surface and let $f : C' \to C$ be a Galois cover such that the fibers of π over the branch points of f are smooth. Let E be the general fiber of π. Let μ be the number of fiber-components not intersecting the zero-section, which equals the total Milnor number of W.*

Then $E(\mathbf{C}(C')) \otimes_{\mathbf{Z}} \mathbf{C}$ is a quotient of a $\mathbf{C}[G]$-module M such that

$$[M] = (10\deg(\mathcal{L}) - \mu)[\mathbf{C}[G]] + 2[H^0(K_{C'})] - 2[\mathbf{C}].$$

Proof. From Proposition 4.12 it follows $E(\mathbf{C}(C'))$ equals $\mathrm{NS}(X')/T(X')$. Now $\mathrm{NS}(X') \otimes_{\mathbf{Z}} \mathbf{C}$ (as $\mathbf{C}[G]$-module) is a quotient of $H^{1,1}(X')$. Hence $E(k(C')) \otimes_{\mathbf{Z}} \mathbf{C}$ is a quotient of $H^{1,1}(X')/T(X')$.

Note that the Weierstrass model of W' is the pullback of the Weierstrass model of W. In particular the minimal discriminant of $X' \to C'$ is the pullback of the minimal discriminant of $X \to C$. Our assumption on the singular fibers of $X \to C$ imply that the singular fibers are outside the ramification locus of $X' \to X$. If $q \in W'_{\mathrm{sing}}$ then q is a point on a singular fiber, hence q is outside the ramification locus of $W' \to W$. Hence we may apply Lemma 4.11 and obtain that $[T(X')] = \mu[\mathbf{C}[G]] + 2[\mathbf{C}]$.

From the previous section it follows that $[H^{1,1}(X')] = 10\deg\mathcal{L}[\mathbf{C}[G]] + 2[H^0(K_{C'})]$, which yields the theorem. □

Remark 4.14: If we allow the fibers over the branch points of f to be semi-stable then the $\mathbf{C}[G]$-structure of T is harder to describe. E.g., suppose we have a I_1 fiber over a branch point, with ramification index 2 and $G = \mathbf{Z}/2\mathbf{Z}$. Then $X' \to C'$ has a I_2 fiber and this contributes a one dimensional vector space to T, on which G acts via a non-trivial character.

COROLLARY 4.15. *Let $X \to C$ be an elliptic surface over a field k of characteristic zero. Let $C' \to C$ be a Galois cover such that the fibers of π over the branch points of f are smooth. Let E be the general fiber of π. Then*

$$\operatorname{rank} E(k(C')) \leq \epsilon(G,k)\left(c_E + \frac{d_E}{6} + 2g - 2 + s\right).$$

Proof. As explained in [12, Section 1] we may assume that $k = \mathbf{C}$. Moreover, in the same section it is shown that it suffices to prove that $E(\mathbf{C}(C')) \otimes_{\mathbf{Z}} \mathbf{C}$ is a quotient of $\mathbf{C}[G]^{c_E + \frac{d_E}{6} + 2g - 2 + s}$.

From the Tate algorithm it follows that the number of fiber components in a singular fiber equals $v_p(\Delta) - 1$ if the reduction is multiplicative and $v_p(\Delta) - 2$ if the reduction is additive. Denote with a the number of additive fibers and with m the number of multiplicative fibers. Hence $\mu = d_E - m - 2a$. Now $c_E = m + 2a$ and $d_E = 12 \deg(\mathcal{L})$. It follows from the previous theorem that $E(k(C')) \otimes_{\mathbf{Z}} \mathbf{C}$ is a quotient of the $\mathbf{C}[G]$-module M, with

$$[M] = \left(c_E + \frac{d_E}{6}\right)[\mathbf{C}[G]] + 2[H^0(K'_C)] - 2[\mathbf{C}].$$

If $C' \to C$ is unramified that $H^0(K'_C) = \mathbf{C}[G]^{g(C)}$. If $C' \to C$ is ramified then $H^0(\mathcal{O}_Z)$ is a quotient of $\mathbf{C}[G]^s$, where s is the number of critical values and we find $2H^0(K'_C)$ is a quotient of $\mathbf{C}^{\oplus 2} \oplus \mathbf{C}[G]^{\oplus 2g-2+s}$

In both cases $E(\mathbf{C}(C')) \otimes_{\mathbf{Z}} \mathbf{C}$ is a quotient of $\mathbf{C}[G]^{\oplus c_E + \frac{d_E}{6} + 2g - 2 + s}$. $\square$

Acknowledgements

The author would like to thank Gavril Farkas for pointing out the existence of [1], Mike Roth for various discussions on the best way to calculate the cohomology of $\Omega^i_{\mathbf{P}(\mathcal{E})}(kX)$ for a hypersurface $X \subset \mathbf{P}(\mathcal{E})$ and Orsola Tommasi for various remarks on a previous version of this paper. The author would like to thank the referee for several useful comments to improve the exposition.

The author acknowledges support from INdAM (GNSAGA).

REFERENCES

[1] C. CHEVALLEY AND A. WEIL, *Über das Verhalten der Integrale 1. Gattung bei Automorphismen des Funktionenkörpers.*, Abh. Math. Semin. Univ. Hambg. **10** (1934), 358–361 (German).

[2] PH. DU BOIS, *Complexe de de Rham filtré d'une variété singulière*, Bull. Soc. Math. France **109** (1981), 41–81.

[3] J. S. ELLENBERG, *Selmer groups and Mordell-Weil groups of elliptic curves over towers of function fields*, Compos. Math. **142** (2006), 1215–1230.

[4] L. A. Fastenberg, *Mordell-Weil groups in procyclic extensions of a function field*, Duke Math. J. **89** (1997), 217–224.

[5] L. A. Fastenberg, *Computing Mordell-Weil ranks of cyclic covers of elliptic surfaces*, Proc. Amer. Math. Soc. **129** (2001), 1877–1883.

[6] L. A. Fastenberg, *Cyclic covers of rational elliptic surfaces*, Rocky Mountain J. Math. **39** (2009), 1895–1903.

[7] R. Hartshorne, *Algebraic geometry*, Graduate Texts in Mathematics, vol. 52, Springer, New York-Heidelberg, 1977.

[8] B. Heijne, *The maximal rank of elliptic Delsarte surfaces*, Math. Comp. **81** (2012), no. 278, 1111–1130.

[9] R. Miranda, *The basic theory of elliptic surfaces*, Dottorato di Ricerca in Matematica, ETS Editrice, Pisa, 1989, Available at `www.researchgate.net/publication/247789879_The_Basic_Theory_of_Elliptic_Surfaces`.

[10] S. Nakajima, *Galois module structure of cohomology groups for tamely ramified coverings of algebraic varieties*, J. Number Theory **22** (1986), 115–123.

[11] A. Pál, *The Manin constant of elliptic curves over function fields*, Algebra Number Theory **4** (2010), 509–545.

[12] A. Pál, *Hodge theory and the Mordell-Weil rank of elliptic curves over extensions of function fields*, J. Number Theory **137** (2014), 166–178.

[13] C. A. M. Peters and J. H. M. Steenbrink, *Mixed Hodge structures*, Ergebnisse der Mathematik und ihrer Grenzgebiete. 3. Folge., vol. 52, Springer, Berlin, 2008.

[14] J. H. Silverman, *The rank of elliptic surfaces in unramified abelian towers*, J. Reine Angew. Math. **577** (2004), 153–169.

[15] J. H. M. Steenbrink, *Intersection form for quasi-homogeneous singularities*, Compos. Math. **34** (1977), 211–223.

[16] J. H. M. Steenbrink, *Adjunction conditions for one-forms on surfaces in projective three-space*, Singularities and computer algebra, London Math. Soc. Lecture Note Ser., vol. 324, Cambridge Univ. Press, Cambridge, 2006, pp. 301–314.

[17] J. L. Verdier, *Le théorème de Le Potier*, Différents aspects de la positivité (Sém. Géom. Anal., École Norm. Sup., Paris, 1972–1973), Soc. Math. France, Paris, 1974, pp. 68–78. Astérisque, No. 17.

Author's address:

Remke Kloosterman
Dipartimento di Matematica "Tullio Levi-Civita"
Università degli Studi di Padova
Via Trieste 63
35121 Padova, Italy
E-mail: `klooster@math.unipd.it`

Received December 17, 2017
Revised June 22, 2018
Accepted June 25, 2018

Rend. Istit. Mat. Univ. Trieste
Volume 50 (2018), 125–131
DOI: 10.13137/2464-8728/22436

A note on finite group-actions on surfaces containing a hyperelliptic involution

Bruno P. Zimmermann

Abstract. *By topological methods using the language of orbifolds, we give a short and efficient classification of the finite diffeomorphism groups of closed orientable surfaces of genus $g \geq 2$ which contain a hyperelliptic involution; in particular, for each $g \geq 2$ we determine the maximal possible order of such a group.*

Keywords: hyperelliptic Riemann surface, hyperelliptic involution, finite diffeomorphism group.
MS Classification 2010: 57M60, 57S17, 30F10.

1. Introduction

Every finite group occurs as the isometry group of a closed hyperbolic 3-manifold [7]; on the other hand, the class of isometry groups ofhyperbolic, hyperelliptic 3-manifolds (i.e., hyperbolic 3-manifolds which are 2-fold branched coverings of S^3, branched along a knot or link) is quite restricted but a complete classification turns out to be difficult (see [9]). More generally one can ask: what are the finite groups which act on a closed 3-manifold and contain a hyperelliptic involution, i.e. an involution with quotient space S^3? Due to classical results for hyperelliptic Riemann surfaces, the situation is much better understood in dimension 2; motivated by the 3-dimensional case, in the present note we discuss the situation for surfaces. All surfaces in the present paper will be orientable, and all finite group-actions orientation-preserving.

Let F_g be a closed orientable surface of genus $g \geq 2$; we call a finite group G of diffeomorphisms of F_g *hyperelliptic* if G contains a hyperelliptic involution, i.e. an involution with quotient space S^2. The quotient F_g/G is a 2-orbifold (a closed surface with a finite number of branch points), and such a 2-orbifold can be given the structure of a hyperbolic 2-orbifold by uniformizing it by a Fuchsian group (see [12, Chapter 6]). Lifting the hyperbolic structure to F_g, this becomes a hyperbolic surface such that the group G acts by isometries. In particular, G acts as a group of conformal automorphisms of the underlying Riemann surface F_g; if G contains a hyperelliptic involution,

F_g is a hyperelliptic Riemann surface. A hyperelliptic Riemann surface has a unique hyperelliptic involution, with $2g+2$ fixed points, which is central in its automorphism group (see [4, Section III.7] for basic facts about hyperelliptic Riemann surfaces, and [10, Chapter 13] for the language of orbifolds). In particular, a hyperelliptic involution h in a finite group of diffeomorphisms G of F_g is unique and central, and the factor group $\bar{G} = G/\langle h \rangle$ acts on the quotient-orbifold $F_g/\langle h \rangle \cong \mathcal{S}^2(2^{2g+2})$, which denotes the 2-sphere with $2g+2$ hyperelliptic branch points of order 2, and G permutes the set $\mathcal{B}$ of the $2g+2$ hyperelliptic branch points. Note that any two hyperelliptic involutions of a surface F_g are conjugate by a diffeomorphism (since they have the same quotient $\mathcal{S}^2(2^{2g+2})$) and, if distinct, generate an infinite dihedral group of diffeomorphisms.

Conversely, if $\bar{G}$ is a finite group acting on the orbifold $\mathcal{S}^2(2^{2g+2})$ (in particular, mapping the set $\mathcal{B}$ of hyperelliptic branch points to itself), then the set of all lifts of elements of $\bar{G}$ to F_g defines a group G with $G/\langle h \rangle \cong \bar{G}$ and $F_g/G = \mathcal{S}^2(2^{2g+2})/\bar{G}$. The finite groups $\bar{G}$ which admit an orientation-preserving action on the 2-sphere S^2 are cyclic $\mathbb{Z}_n$ with quotient-orbifold $\mathcal{S}^2(n,n)$, dihedral $\mathbb{D}_{2n}$ of order $2n$ with quotient $\mathcal{S}^2(2,2,n)$, tetrahedral $\mathbb{A}_4$ of order 12 with quotient $\mathcal{S}^2(2,3,3)$, octahedral $\mathbb{S}_4$ of order 24 with quotient $\mathcal{S}^2(2,3,4)$, or dodecahedral A_5 of order 60 with quotient $\mathcal{S}^2(2,3,5)$.

In the following theorem we classify large hyperelliptic group-actions; however, the methods apply easily also to arbitrary actions, see Remark 2.3.

THEOREM 1.1. *Let G be a finite group of diffeomorphisms of a closed orientable surface F_g of genus $g \geq 2$ containing a hyperelliptic involution; suppose that $|G| \geq 4g$ and that G is maximal, i.e. not contained in a strictly larger finite group of diffeomorphisms of F_g. Then G belongs to one of the following cases (see 2.1 for the definitions of the groups $A_{8(g+1)}$ and B_{8g}):*

$$\begin{array}{lll}
G = A_{8(g+1)}, & \bar{G} \cong \mathbb{D}_{4(g+1)}, & F_g/G = \mathcal{S}^2(2,4,2g+2); \\
G = B_{8g}, & \bar{G} \cong \mathbb{D}_{4g}, & F_g/G = \mathcal{S}^2(2,4,4g); \\
G \cong \mathbb{Z}_{4g+2}, & \bar{G} \cong \mathbb{Z}_{2g+1}, & F_g/G = \mathcal{S}^2(2,2g+1,4g+2); \\
|G| = 120, & \bar{G} \cong \mathbb{A}_5, & g = 5,9,14,15,20,24,29,30; \\
|G| = 48, & \bar{G} \cong \mathcal{S}_4, & g = 2,3,5,6,8,9,11,12; \\
|G| = 24, & \bar{G} \cong \mathbb{A}_4, & g = 4.
\end{array}$$

In each of the cases, up to conjugation by diffeomorphisms of F_g there is a unique group G for each genus g (see Sections 2.3 and 2.4 for the quotient orbifolds in the last three cases).

COROLLARY 1.2. *Let $m_h(g)$ denote the maximal order of a hyperelliptic group of diffeomorphisms of a closed orientable surface of genus $g \geq 2$; then $m_h(g) = 8(g+1)$, with the exceptions $m_h(2) = m_h(3) = 48$ and $m_h(5) = m_h(9) = 120$.*

The maximal order $m(g)$ of a finite group of diffeomorphisms of closed surface of genus $g \geq 2$ is not known in general; there is the classical Hurwitz bound $m(g) \leq 84(g-1)$ [6] which is both optimal and non-optimal for infinitely many values of g. Considering hyperelliptic groups as in Theorem 1.1 one has $m(g) \geq 8(g+1)$, and Accola and Maclachlan proved that $m(g) = 8(g+1)$ for infinitely many values of g, see Remark 2.2 in Section 2.

The group $G \cong \mathbb{Z}_{4g+2}$ in Theorem 1.1 realizes the unique action of a cyclic group of maximal possible order $4g+2$ on a surface of genus $g \geq 2$, see Remark 2.1.

2. Proof of Theorem 1.1

2.1. Dihedral groups

Let $\bar{G} = \mathbb{D}_{2n}$ be a dihedral group of order $2n$ acting on the orbifold $\mathcal{S}^2(2^{2g+2})$. The action of $\mathbb{D}_{2n}$ on the 2-sphere has one orbit $\mathcal{O}_2$ consisting of the two fixed points of the cyclic subgroup $\mathbb{Z}_n$ of $\mathbb{D}_{2n}$, two orbits $\mathcal{O}_n$ and $\mathcal{O}'_n$ each of n elements whose union is the set of $2n$ fixed point of the n reflections in the dihedral group $\mathbb{D}_{2n}$, and regular orbits $\mathcal{O}_{2n}$ with $2n$ elements. We consider various choices for the set $\mathcal{B}$ of $2g+2$ hyperelliptic branch points in $\mathcal{S}^2(2^{2g+2})$.

i) $\mathcal{B} = \mathcal{O}_n, \quad n = 2g+2, \quad \mathcal{S}^2(2^{2g+2})/\bar{G} = \mathcal{S}^2(2,4,2g+2).$

We define $A_{8(g+1)}$ as the group G of order $8(g+1)$ of all lifts of elements of $\bar{G}$ to the 2-fold branched covering F_g of $\mathcal{S}^2(2^{2g+2})$. It is easy to find a presentation of $A_{8(g+1)}$: considering the central extension $1 \to \mathbb{Z}_2 = \langle h \rangle \to A_{8(g+1)} \to \mathbb{D}_{4(g+1)} \to 1$ and the presentation $\mathbb{D}_{4(g+1)} = <\bar{x}, \bar{y} \mid \bar{x}^2 = \bar{y}^2 = (\bar{x}\bar{y})^{2(g+1)} = 1>$, one obtains the presentation $A_{8(g+1)} = <x, y, h \mid h^2 = 1,\ [x,h] = [y,h] = 1,\ y^2 = h,\ x^2 = y^4 = (xy)^{2(g+1)} = 1>$.

ii) $\mathcal{B} = \mathcal{O}_n \cup \mathcal{O}_2, \quad n = 2g$ even, $\quad \mathcal{S}^2(2^{2g+2})/\bar{G} = \mathcal{S}^2(2,4,4g).$

The lift G of $\bar{G}$ to F_g defines a group B_{8g} of order $8g$, with presentation $B_{8g} = <x, y, h \mid h^2 = 1,\ [x,h] = [y,h] = 1,\ y^2 = (xy)^{2g} = h,\ x^2 = y^4 = (xy)^{4g} = 1>$.

iii) $\mathcal{B} = \mathcal{O}_n \cup \mathcal{O}'_n, \quad n = g+1, \quad \mathcal{S}^2(2^{2g+2})/\bar{G} = \mathcal{S}^2(4,4,g+1).$

This orbifold has an involution with quotient $\mathcal{S}^2(2,4,2(g+1))$ which lifts to $\mathcal{S}^2(2^{2g+2})$, hence G is a subgroup of index 2 in $A_{8(g+1)}$.

iv) $\mathcal{B} = \mathcal{O}_{2n}, \quad n = g+1, \quad \mathcal{S}^2(2^{2g+2})/\bar{G} = \mathcal{S}^2(2,2,2,g+1).$

This orbifold has an involution with quotient $\mathcal{S}^2(4, 2, 2(g+1))$ which lifts to $\mathcal{S}^2(2^{2g+2})$, and G is a subgroup of index 2 in $A_{8(g+1)}$.

v) $\mathcal{B} = \mathcal{O}_n \cup \mathcal{O}'_n \cup \mathcal{O}_2, \quad n = g, \quad \mathcal{S}^2(2^{2g+2})/\bar{G} = \mathcal{S}^2(4, 4, 2g).$

Again there is an involution, with quotient $\mathcal{S}^2(2, 4, 4g)$, hence G is a subgroup of index 2 of B_{8g}.

vi) $\mathcal{B} = \mathcal{O}_{2n} \cup \mathcal{O}_2, \quad n = g, \quad \mathcal{S}^2(2^{2g+2})/\bar{G} = \mathcal{S}^2(2, 2, 2, 2g).$

Dividing out a further involution one obtains $\mathcal{S}^2(4, 2, 4g)$, and G is a subgroup of index 2 in B_{8g}.

Note that for any other choice of $\mathcal{B}$ one obtains groups G of order less than $4g$.

Remark 2.1: Incidentally, by results of Accola [1] and Maclachlan [8], for infinitely many values of g the groups $A_{8(g+1)}$ in i) realize the maximal possible order of a group acting on a surface of genus $g \geq 2$. Moreover, the group $A_{8(g+1)}$ has an abelian subgroup $\mathbb{Z}_{2g+2} \times \mathbb{Z}_2$ of index two which realizes the maximal possible order of an abelian group acting on a surface of genus $g \geq 2$ (see [12, 4.14.27]).

2.2. Cyclic groups

Next we consider the case of a cyclic group $\bar{G} = \mathbb{Z}_n$. There are two orbits with exactly one point, the fixed points of $\mathbb{Z}_n$, all other orbits have n points (regular orbits).

If $\mathcal{B}$ consists of a regular orbit and exactly one of the two fixed points of $\mathbb{Z}_n$, with $n+1 = 2g+2$, $n = 2g+1$ odd and $\mathcal{S}^2(2^{2g+2})/\bar{G} = \mathcal{S}^2(2, 2g+1, 2(2g+1))$, then the 2-fold branched covering of $\mathcal{S}^2(2^{2g+2})$ is a surface of genus g on which a cyclic group $G \cong \mathbb{Z}_{4g+2}$ acts.

If $\mathcal{B}$ consists of one regular orbit, then $n = 2g+2$, $\mathcal{S}^2(2^{2g+2})/\bar{G} = \mathcal{S}^2(2, 2g+2, 2g+2)$ which is a 2-fold branched covering of $\mathcal{S}^2(2, 4, 2g+2)$, hence $G \cong \mathbb{Z}_{2g+2} \times \mathbb{Z}_2$ is a subgroup of index 2 in $A_{8(g+1)}$.

If $\mathcal{B}$ consists of a regular orbit and the two fixed points of $\mathbb{Z}_n$, then $n + 2 = 2g+2$, $n = 2g$, $\mathcal{S}^2(2^{2g+2})/\bar{G} = \mathcal{S}^2(2, 4g, 4g)$ which is a 2-fold cover of $\mathcal{S}^2(2, 4, 4g)$, and $G \cong \mathbb{Z}_{4g}$ is a subgroup of index 2 in B_{8g}.

Remark 2.2: By a result of Wiman [11], $4g+2$ is the maximal possible order of a cyclic group-action on a surface of genus $g \geq 2$, and the action of $G \cong \mathbb{Z}_{4g+2}$ above is the unique action of a cyclic group realizing this maximal order (see [5], [12, 4.14.27]). The group $G \cong \mathbb{Z}_{2g+2} \times \mathbb{Z}_2$ instead realizes the maximum order of an abelian group-action on a surface of genus $g \geq 2$, see Remark 2.1.

2.3. Dodecahedral groups

Now let $\bar{G} = \mathbb{A}_5$ be the dodecahedral group acting on S^2. The orbits of the action are $\mathcal{O}_{12}$ consisting of the 12 fixed points of the 6 subgroups $\mathbb{Z}_5$ of $\mathbb{A}_5$ (the centers of the 12 faces of a regular dodecahedron), $\mathcal{O}_{20}$ consisting of the twenty fixed points of the 10 subgroups $\mathbb{Z}_3$ (the 20 vertices of a regular dodecahedron), $\mathcal{O}_{30}$ consisting of the 30 fixed points of the 15 involutions (the centers of the 30 edges of a regular dodecahedron), and regular orbits $\mathcal{O}_{60}$. The list of the different choices of $\mathcal{B}$, the genera g and the corresponding quotient-orbifolds are as follows:

$$\begin{array}{lll}
\mathcal{B} = \mathcal{O}_{12}: & g = 5, & \mathcal{S}^2(2^{12})/\bar{G} \cong \mathcal{S}^2(2,3,10); \\
\mathcal{B} = \mathcal{O}_{20}: & g = 9, & \mathcal{S}^2(2^{20})/\bar{G} \cong \mathcal{S}^2(2,6,5); \\
\mathcal{B} = \mathcal{O}_{30}: & g = 14, & \mathcal{S}^2(2^{30})/\bar{G} \cong \mathcal{S}^2(4,3,5); \\
\mathcal{B} = \mathcal{O}_{60}: & g = 29, & \mathcal{S}^2(2^{60})/\bar{G} \cong \mathcal{S}^2(2,2,3,5); \\
\mathcal{B} = \mathcal{O}_{12} \cup \mathcal{O}_{20}: & g = 15, & \mathcal{S}^2(2^{32})/\bar{G} \cong \mathcal{S}^2(2,6,10); \\
\mathcal{B} = \mathcal{O}_{12} \cup \mathcal{O}_{30}: & g = 20, & \mathcal{S}^2(2^{42})/\bar{G} \cong \mathcal{S}^2(4,3,10); \\
\mathcal{B} = \mathcal{O}_{20} \cup \mathcal{O}_{30}: & g = 24, & \mathcal{S}^2(2^{50})/\bar{G} \cong \mathcal{S}^2(4,6,5); \\
\mathcal{B} = \mathcal{O}_{12} \cup \mathcal{O}_{20} \cup \mathcal{O}_{30}: & g = 30, & \mathcal{S}^2(2^{62})/\bar{G} \cong \mathcal{S}^2(4,6,10).
\end{array}$$

These are exactly the genera in the Theorem for the case $\bar{G} \cong A_5$. For $g = 5$, 9, 15 and 29, the group G is isomorphic to $\mathbb{A}_5 \times \mathbb{Z}_2$, for $g = 14$, 20, 24 and 30 to the binary dodecahedral group $\mathbb{A}_5^*$ (these are the two central extensions of $\mathbb{A}_5$).

Remark 2.3: For each of the finite groups $\bar{G}$ acting on S^2 one can easily produce a complete list of the possible quotient orbifolds $F_g/G = \mathcal{S}^2(2^{2g+2})/\bar{G}$ (i.e., without the restriction $|G| \geq 4g$ in the Theorem). For the case of $\bar{G} = \mathbb{A}_5$, the possible quotient-orbifolds are as follows (where $m \geq 0$ denotes the number of regular orbits $\mathcal{O}_{60}$ in the set $\mathcal{B}$ of hyperelliptic branch points).

$$\begin{array}{llll}
\mathcal{S}^2(2^m,2,3,5), & \mathcal{S}^2(2^m,2,3,10), & \mathcal{S}^2(2^m,2,6,5), & \mathcal{S}^2(2^m,4,3,5), \\
\mathcal{S}^2(2^m,2,6,10), & \mathcal{S}^2(2^m,4,3,10), & \mathcal{S}^2(2^m,4,6,5), & \mathcal{S}^2(2^m,4,6,10).
\end{array}$$

Each of these orbifolds defines a unique action of $\bar{G} \cong \mathbb{A}_5$ on an orbifold $\mathcal{S}^2(2^{2g+2})$ and of $G \cong \mathbb{A}_5 \times \mathbb{Z}_2$ or $\mathbb{A}_5^*$ on a surface F_g, and this gives the complete classification of the actions of the groups G of type $\bar{G} \cong \mathbb{A}_5$, up to conjugation.

2.4. Octahedral and tetrahedral groups

Finally, playing the same game for the groups $\mathbb{S}_4$ and $\mathbb{A}_4$, one produces similar lists also for these cases. This gives the list of genera for the groups G of type $\mathbb{S}_4$ in the Theorem; the groups G of type $\mathbb{A}_4$ are all subgroups of index 2 in groups G of type $\mathbb{S}_4$ except for $g = 4$ (with $\mathcal{B} = \mathcal{O}_4 \cup \mathcal{O}_6$ and quotient-orbifold $\mathcal{S}^2(3,4,6)$). The groups G are isomorphic to one of the two central extensions $\mathbb{A}_4 \times \mathbb{Z}_2$ and $\mathbb{A}_4^*$ of $\mathbb{A}_4$, or to one of four central extensions of $\mathbb{S}_4$.

Note added for the revised version. The referee provided the two additional references [2] and [3] in which similar results are obtained, in an algebraic language and by different algebraic methods. The main point of the present note is a short, topological approach to the classification: After the preliminary fact from complex analysis that a hyperelliptic involution of a Riemann surface is unique and central in its automorphism group, the methods in the present note are purely topological, offering a short and efficient topological approach to the classification.

References

[1] R. D. M. Accola, *On the number of automorphisms of a closed Riemann surface*, Trans. Amer. Math. Soc. **131** (1968), 398–408.

[2] R. Brandt and H. Stichtenoth, *Die Automorphismengruppen hyperelliptischer Kurven*, Manuscripta Math. **55** (1986), no. 1, 83–92.

[3] E. Bujalance, J. M. Gamboa, and G. Gromadzki, *The full automorphism groups of hyperelliptic Riemann surfaces*, Manuscripta Math. **79** (1993), no. 3-4, 267–282.

[4] H. M. Farkas and I. Kra, *Riemann surfaces*, 2nd ed., Springer-Verlag,, New York, 1992.

[5] W. J. Harvey, *Cyclic groups of automorphisms of a compact Riemann surface*, Quart. J. Math. Oxford Ser. (2) **17** (1966), 86–97.

[6] A. Hurwitz, *Über algebraische Gebilde mit eindeutigen Transformationen in sich*, Math. Ann. **41** (1892), no. 3, 403–442.

[7] S. Kojima, *Isometry transformations of hyperbolic* 3*-manifolds*, Topology Appl. **29** (1988), no. 3, 297–307.

[8] C. Maclachlan, *A bound for the number of automorphisms of a compact Riemann surface*, J. London Math. Soc. **44** (1969), 265–272.

[9] M. Mecchia, *Finite groups acting on hyperelliptic* 3*-manifolds*, arXiv:1805.06331, 2018.

[10] W. P. Thurston, *The Geometry and Topology of* 3*-manifolds*, Lecture Notes Princeton University, Princeton University, 1978, http://library.msri.org/books/gt3m/PDF/1.pdf.

[11] A. Wiman, *Über die hyperelliptischen Curven und diejenigen vom Geschlecht* $p = 3$*, welche eindeutige Transformationen auf sich zulassen*, Bihang Til Kongl. Svenska Veienskaps-Akademiens Hadlinga, Stockholm **21** (1895-86), 1–23.

[12] H. Zieschang, E. Vogt, and H.-D. Coldewey, *Surfaces and planar discontinuous groups*, Lecture Notes in Mathematics, vol. 835, Springer, Berlin, 1980.

Author's address:

Bruno P. Zimmermann
Università degli Studi di Trieste
Dipartimento di Matematica e Geoscienze
Via Valerio 12/1
34127 Trieste, Italy
E-mail: zimmer@units.it

Received June 6, 2018
Revised October 1, 2018
Accepted October 3, 2018

Rend. Istit. Mat. Univ. Trieste
Volume 50 (2018), 133–137
DOI: 10.13137/2464-8728/22432

A bijection between phylogenetic trees and plane oriented recursive trees

Helmut Prodinger

Abstract. *Phylogenetic trees are binary non-plane trees with labelled leaves, and plane oriented recursive trees are planar trees with an increasing labelling. Both families are enumerated by double factorials. A bijection is constructed, using the respective representations as 2-partitions and trapezoidal words.*

Keywords: Phylogenetic tree, plane oriented recursive tree, trapezoidal word, bijection.
MS Classification 2010: 05A19, 05A10.

1. Introduction

There are many occurrences of the *double factorials*

$$(2n-1)!! = 1 \cdot 3 \cdots (2n-1)$$

in the combinatorial literature. A nice survey is by Callan [2].

Two manifestations of them deal with trees, and it is our objective to establish a bijection between them.

The phylogenetic trees are binary non-plane trees with the leaves labelled by the numbers $1, 2, \dots, n+1$. Their number is given by $(2n-1)!!$. Stanley describes (codes) them by set partitions of $\{1, 2, \dots, 2n\}$ into n sets of 2 elements each. These are easily enumerated by the double factorials: Just note that they are counted by $\frac{(2n)!}{2^n n!}$, where we start with all permutations of $2n$ elements, but divide by $n!$, since the order of the blocks does not count, and by 2^n, since in each block the order of the 2 elements is immaterial. We call these set partitions 2-partitions. Stanley's coding will be reviewed in a later section.

Plane oriented recursive trees (PORTs) [3], also known as heap ordered trees, are planar trees, where the nodes are labelled by the integers $1, \dots, n+1$, and the labels are increasing towards the leaves. They are also enumerated by $(2n-1)!!$. They are also coded by simple objects called *trapezoidal words*, which are reviewed in the next section.

Our bijection will in fact be between the two codings, i.e., between 2-partitions and trapezoidal words.

2. Trapezoidal words and Plane oriented recursive trees

One of the easiest manifestations of double factorials is by words $x_1x_2\dots x_n$, where $1 \le x_i \le 2i-1$; they were called *trapezoidal words* by Riordan [4]; see also [2].

Plane oriented recursive trees (PORTs) are rooted planar trees, where the n nodes are labelled by the numbers $1,\dots,n$ in an increasing way from the root to the leaves.

If one has already such a PORT with n nodes, there are $2n-1$ positions where a new node labelled $n+1$ can be attached, whence the enumeration by double factorials: The number of PORTs with $n+1$ nodes is given by $(2n-1)!!$, and the trees are in obvious bijection with trapezoidal words of length n, simply by recording the position where one node after the other was inserted. PORTs were also called heap ordered trees, but we adopted the notation from [3].

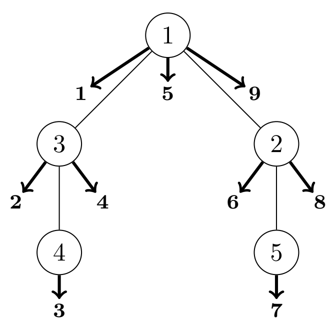


Figure 1: A PORT with 5 nodes and the 9 positions where node labelled 6 could be inserted.

3. 2-partitions and Phylogenetic trees

Phylogenetic trees are non-plane binary trees, with the leaves labelled by the numbers $1,\dots,n+1$ in an arbitrary way. Stanley [5] describes the procedure in Figure 2 to label the remaining nodes as well (except for the root): The numbers $n+2,\dots,2n$ are used as labels in this order as follows: among the pairs of children that are both labelled, but the parent isn't, find the smallest label of a child; it is the parent who gets the current label. The procedure can be seen from Figure 2. At the end, the labels of each pair of 2 children form a subset of 2 elements, leading to n such pairs and thus to a 2-partition. See also [1] for more results and references about phylogenetic trees.

Although Stanley leaves it to the reader to figure out why this works, we

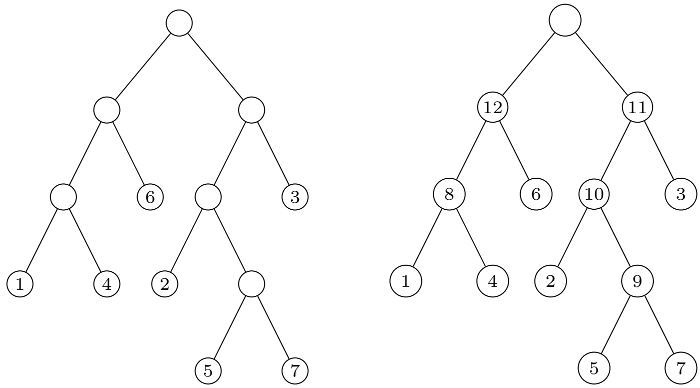


Figure 2: Left: A non-plane binary tree with leaves labelled by $1, \ldots, 7$. Right: The remaining nodes (except for the root) are now labelled by $8, \ldots, 12$. The two children of each node form the 2-partition: $\{1,4\}, \{2,9\}, \{5,7\}, \{6,8\}, \{3,10\}, \{11,12\}$.

sketch a possible answer by showing how a phylogenetic tree can be reconstructed from a 2-partition:

We use the consecutive labels $n+1, \ldots, 2n$ to work as a parent. For that, we look at the block, such that both entries are smaller than the current new label, and among them at the one in which the smallest number occurs. After that, the pair is discarded, and the process continues until all pairs have been processed. The final root may be thought of having the label ∞.

Thus, in the running example, the number 8 is the current new parent, and $\{1,4\}, \{5,7\}$ are such that both members are smaller than 8. The block $\{1,4\}$ is chosen. Then we move to number 9. Candidates are $\{5,7\}, \{6,8\}$; the block containing the number 5 is used, then discarded, and so on.

4. The bijection between 2-partitions and trapezoidal words

Our strategy of proof is to grow a 2-partition of $2n-2$ elements to one of $2n$ elements (hereby establishing once again the $(2n-1)!!$ formula), and describing how the corresponding trapezoidal word of length $n-1$ grows to one of length n. The correspondence is bijective at each step. Our argument is essentially by induction. It should be noted that the way a 2-partition (and a trazoidal word) grows towards a final partition/word is *unique*.

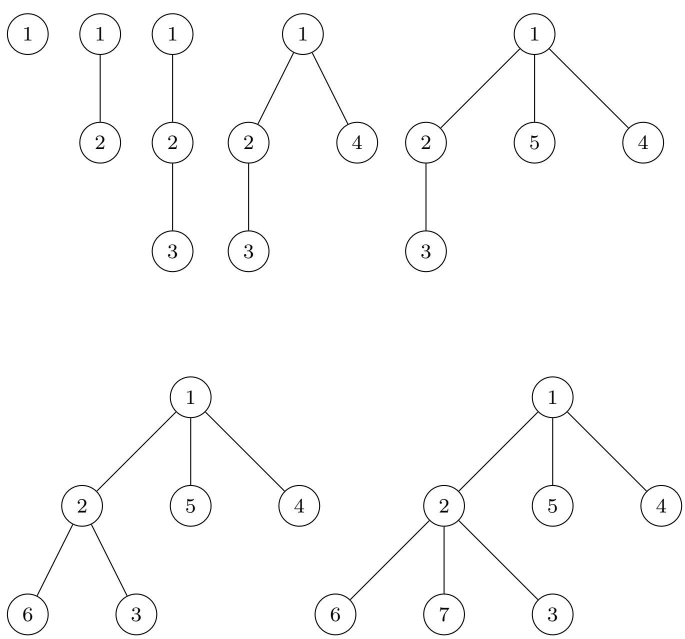
Two new elements $2n-1, 2n$ can form a class of their own, and this can happen in $(2n-3)!!$ ways. Otherwise, $2n$ matches with an element $1 \le b \le 2n-2$ (in $2n-2$ ways), and $2n-1$ matches with the former partner a of b. So, the set $\{a, b\}$ is replaced by the sets $\{a, 2n-1\}, \{b, 2n\}$. Such an operation is often called a *rotation*. Thus, we get altogether $(1+(2n-2)) \cdot (2n-3)!! = (2n-1)!!$ 2-partitions, as to be expected.

Then we augment the corresponding trapezoidal word $x_1 \dots x_{n-1}$, by x_n, defined a follows: if the second case happened and $2n$ matches with an element $1 \le b \le 2n-2$, then we set $x_n := b$, otherwise, if $2n-1, 2n$ form a class of their own we set $x_n := 2n-1$. It is easy to see that this operation is reversible.

As an example, let us see how the trapezoidal word $1, 2, 5, 5, 2, 4$ lets the 2-partition grow. The first 1 translates into the starting partition $\{1, 2\}$.

$$\begin{gathered}
\{1,2\} \xrightarrow{2} \{1,3\}\{2,4\} \xrightarrow{5} \{1,3\}\{2,4\}\{5,6\} \xrightarrow{5} \{1,3\}\{2,4\}\{6,7\}\{5,8\} \\
\xrightarrow{2} \{1,3\}\{2,10\}\{4,9\}\{6,7\}\{5,8\} \\
\xrightarrow{4} \{1,3\}\{2,10\}\{4,12\}\{9,11\}\{6,7\}\{5,8\}.
\end{gathered}$$

And here is how the PORT develops:



References

[1] M. Bóna, *On the number of vertices of each rank in phylogenetic trees and their generalizations*, Discrete Math. Theor. Comput. Sci. **18** (2016), Paper no. 7, 7 pp.

[2] D. Callan, *A combinatorial survey of identities for the double factorial*, arXiv:0906.1317v1 (2009).

[3] H.-K. Hwang, *Profiles of random trees: Plane-oriented recursive trees*, Random Structures and Algorithms **30** (2006), 380–413.

[4] J. Riordan, *The blossoming of Schröder's fourth problem*, Acta. Math. **137** (1976), 1–16.

[5] R. Stanley, *Enumerative Combinatorics, Vol. 2*, Cambridge University Press, 1999.

Author's address:

Helmut Prodinger
Department of Mathematical Sciences
Stellenbosch University
7602 Stellenbosch, South Africa
E-mail: hproding@sun.ac.za

Received November 6, 2017
Accepted October 8, 2018

Rend. Istit. Mat. Univ. Trieste
Volume 50 (2018), 139–147
DOI: 10.13137/2464-8728/22433

On Fröberg-Macaulay conjectures for algebras

Mats Boij and Aldo Conca

Abstract. *Macaulay's theorem and Fröberg's conjecture deal with the Hilbert function of homogeneous ideals in polynomial rings over a field K. In this short note we present some questions related to variants of Macaulay's theorem and Fröberg's conjecture for K-subalgebras of polynomial rings.*

Keywords: Hilbert functions, Macaulay theorem.
MS Classification 2010: 13D40, 14M25.

1. Introduction

Macaulay's theorem and Fröberg's conjecture deal with the Hilbert function of homogeneous *ideals* in polynomial rings S over a field K. In this short note we present some questions related to variants of Macaulay's theorem and Fröberg's conjecture for K-*subalgebras* of polynomial rings. In details, given a subspace V of forms of degree d we consider the K-subalgebra $K[V]$ of S generated by V. What can be said about Hilbert function of $K[V]$? The analogy with the ideal case suggests several questions. To state them we start by recalling Macaulay's theorem, Fröberg's conjecture and Gotzmann's persistence theorem for ideals. Then we presents the variants for K-subalgebras along with some partial results and examples.

2. Macaulay's theorem and Fröberg's conjecture for ideals

Let K be a field and $S = K[x_1, \dots, x_n]$ be the polynomial ring equipped with its standard grading, i.e., with $\deg x_i = 1$ for $i = 1, \dots, n$. Then $S = \bigoplus_{j=0}^{\infty} S_j$ where S_j denotes the K-vector space of homogeneous polynomials of degree j. Given positive integers d, u such that $u \leq \dim S_d$ let $G(u, S_d)$ be the Grassmannian of all u-dimensional K-subspaces of S_d. For a given subspace $V \in G(u, S_d)$, the homogeneous components of the ideal $I = (V)$ of S generated by V are the vector spaces S_jV, i.e., the vector spaces generated by the elements fg with $f \in S_j$ and $g \in V$.

Question 2.1. *What can be said about the dimension of S_jV in terms of $u = \dim V$?*

2.1. Lower bound

Macaulay's theorem on Hilbert functions, see [10], provides a lower bound for $\dim S_jV$ given $\dim V$. It asserts that there exists a subspace $L \in G(u, S_d)$ such that

$$\dim S_jL \leq \dim S_jV$$

for every j and every $V \in G(u, S_d)$. Furthermore $\dim S_jL$ can be expressed combinatorially in terms of d and u by means of the so-called *Macaulay expansion*, see [2, 13] for details. The vector space L can be taken to be generated by the largest u monomials of degree d with respect to the lexicographic order. Such an L is called the *lex-segment* (vector space) associated to the pair d and u and it is denoted by $\mathrm{Lex}(u, S_d)$.

2.2. Persistence

A vector space $L \in G(u, S_d)$ is called *Gotzmann* if it satisfies

$$\dim S_1L = \dim S_1 \mathrm{Lex}(u, S_d),$$

i.e., if

$$\dim S_1L \leq \dim S_1V,$$

for all $V \in G(u, S_d)$. Gotzmann's persistence theorem [8] asserts that if $L \in G(u, S_d)$ is Gotzmann then S_1L is Gotzmann as well. In particular if L is Gotzmann one has

$$\dim S_jL \leq \dim S_jV,$$

for all j and all $V \in G(u, S_d)$.

2.3. Upper bound

Clearly, the upper bound for $\dim S_jV$ is given by the $\dim S_jW$ for a "general" W in $G(u, S_d)$. More precisely, there exists a non-empty Zariski open subset U of $G(u, S_d)$ such that for every $V \in G(u, S_d)$, for every $j \in \mathbb{N}$ and every $W \in U$ one has

$$\dim S_jV \leq \dim S_jW.$$

Fröberg's conjecture predicts the values of the upper bound $\dim S_jW$. For a formal power series $f(z) = \sum_{i=0}^{\infty} f_iz^i \in \mathbb{Z}[\![z]\!]$ one denotes $[f(z)]_+$ the series

$\sum_{i=0}^{\infty} g_i z^i$, where $g_i = f_i$ if $f_j \geq 0$ for all $j \leq i$ and $g_i = 0$ otherwise. Given n, u and d one considers the formal power series:

$$\sum c_i z^i = \left[\frac{(1-z^d)^u}{(1-z)^n}\right]_+$$

and then Fröberg's conjecture asserts that $\dim S_j W = \dim S_{j+d} - c_{d+j}$ for all j. It is known to be true in these cases:

(1) $n \leq 3$ and any u, d, j, [1, 7],

(2) $u \leq n+1$ and any d, j, [12],

(3) $j = 1$ and any n, u, d, [9]

and it remains open in general. See [11] for some recent contributions.

3. Macaulay's theorem and Fröberg's conjecture for subalgebras

For any subspace $V \in G(u, S_d)$ we can consider the K-subalgebra $K[V] \subseteq S$ generated by V. Indeed, $K[V]$ is the coordinate ring of the closure of the image of the rational map $\mathbb{P}^{n-1} \dashrightarrow \mathbb{P}^{u-1}$ associated to V.

The homogeneous component of degree j of $K[V]$ is the vector space V^j, i.e., the K-subspace of S_{jd} generated by the elements of the form $f_1 \cdots f_j$ with $f_1, \ldots, f_j \in V$.

Question 3.1. *What can be said about the dimension of V^j? In other words, what can be said about the Hilbert function of the K-algebra $K[V]$?*

Definition 3.2. *For positive integers, n, d, u and j, define*

$$L(n, d, u, j) = \min\{\dim V^j : V \in G(u, S_d)\}$$

and

$$M(n, d, u, j) = \max\{\dim V^j : V \in G(u, S_d)\}.$$

3.1. Lower bound

Recall that a monomial vector space W is said to be *strongly stable* if $mx_i/x_j \in W$ for every monomial $m \in W$ and $i < j$ such that $x_j | m$. Intersections, sums and products of strongly stable vector spaces are strongly stable. Given monomials $m_1, \ldots, m_c \in S_d$ the smallest strongly stable vector space containing them is denoted by $\mathrm{St}(m_1, \ldots, m_c)$ and it is called the strongly stable vector space generated by $m_1, \ldots, m_c$.

PROPOSITION 3.3. *Given n, d, u and j there exists a strongly stable vector space $W \in G(u, S_d)$ such that*

$$L(n, d, u, j) = \dim W^j$$

independently of the field K.

Proof. Given a term order $<$ on S for every $V \in G(u, S_d)$ one has $\mathrm{in}(V)^j \subseteq \mathrm{in}(V^j)$ for every j. Hence one has $\dim V_0^j \leq \dim V^j$ where $V_0 = \mathrm{in}(V)$. Therefore the lower bound $L(n, d, u, j)$ is achieved by a monomial vector space. Comparing the vector space dimension of monomial algebras is a combinatorial problem and hence we may assume the base field has characteristic 0. Applying a general change of coordinates, we may put V in "generic coordinates" and hence $\mathrm{in}(V)$ is the generic initial vector space of V with respect to some term order. Being such it is Borel fixed. Since the base field has characteristic 0, we have that $\mathrm{in}(V)$ is strongly stable. Therefore the lower bound $L(n, d, u, j)$ is achieved by a strongly stable vector space. $\square$

EXAMPLE 3.4: For $n = 3, d = 4, u = 7$ there are 3 strongly stable vector spaces:

$$\begin{array}{lll} 1) \ \mathrm{St}\{xy^3, x^2z^2\} & = & \langle xy^3, x^2z^2, x^2yz, x^2y^2, x^3z, x^3y, x^4\rangle \\ & & \qquad\qquad\qquad \text{– the Lex Segment} \\ 2) \ \mathrm{St}\{xy^2z\} & = & \langle xy^2z, xy^3, x^2yz, x^2y^2, x^3z, x^3y, x^4\rangle \\ 3) \ \mathrm{St}\{y^4, x^2yz\} & = & \langle y^4, xy^3, x^2yz, x^2y^2, x^3z, x^3y, x^4\rangle \\ & & \qquad\qquad\qquad \text{– the RevLex Segment} \end{array}$$

In this case 2) and 3) turns out to give rational normal scrolls of type $(3, 2)$ and $(4, 1)$ respectively and they give the minimal possible Hilbert function in all values.

EXAMPLE 3.5: For $n = 3$, $d = 5$ and $u = 12$, there are five strongly stable subspaces of S_d:

$$\begin{aligned} W_1 = \mathrm{St}\{x^2z^2, xy^3z\} = \langle & x^5, x^4y, x^4z, x^3y^2, x^3yz, x^3z^2, x^2y^3, x^2y^2z, \\ & x^2yz^2, x^2z^3, xy^4, xy^3z\rangle, \\ W_2 = \mathrm{St}\{xy^2z^2\} = \langle & x^5, x^4y, x^4z, x^3y^2, x^3yz, x^3z^2, x^2y^3, x^2y^2z, \\ & x^2yz^2, xy^4, xy^3z, xy^2z^2\rangle, \\ W_3 = \mathrm{St}\{x^2z^3, y^5\} = \langle & x^5, x^4y, x^4z, x^3y^2, x^3yz, x^3z^2, x^2y^3, x^2y^2z, \\ & x^2yz^2, x^2z^3, xy^4, y^5\rangle, \end{aligned}$$

$$W_4 = \mathrm{St}\{x^2z^2y, xy^3z, y^5\} = \langle x^5, x^4y, x^4z, x^3y^2, x^3yz, x^3z^2, x^2y^3, x^2y^2z, x^2yz^2, xy^4, xy^3z, y^5\rangle,$$
$$W_5 = \mathrm{St}\{x^3z^2, y^4z\} = \langle x^5, x^4y, x^4z, x^3y^2, x^3yz, x^3z^2, x^2y^3, x^2y^2z, xy^4, xy^3z, y^5, y^4z\rangle.$$

In this case, neither the Lex segment, W_1, nor the RevLex segment, W_5, achieve the minimum Hilbert function. The Hilbert series are given by

$$\mathrm{HS}_{K[W_1]}(z) = \mathrm{HS}_{K[W_5]}(z) = \frac{1+9z+3z^2}{(1-z)^3},$$
$$\mathrm{HS}_{K[W_2]}(z) = \mathrm{HS}_{K[W_4]}(z) = \frac{1+9z+2z^2}{(1-z)^3},$$
$$\mathrm{HS}_{K[W_3]}(z) = \frac{1+9z+5z^2}{(1-z)^3}.$$

and the minimum turns out to be $L(3,5,12,j) = \dim W_2^j = \dim W_4^j = 6j^2 + 5j + 1$, for $j \geq 1$.

QUESTIONS 3.6. *(1) Does there exist a (strongly stable) subspace $W \in G(u, S_d)$ such that $L(n,d,u,j) = \dim W^j$ for every j?*

(2) Given n, d, u, j can one characterize combinatorially the strongly stable subspace(s) W with the property $L(n,d,u,j) = \dim W^j$?

(3) Persistence: Assume $W \in G(u, S_d)$ satisfies $L(n,d,u,2) = \dim W^2$. Does it satisfy also $L(n,d,u,j) = \dim W^j$ for all j?

REMARK 3.7: For $n = 2$ there exists only one strongly stable vector space in $G(u, S_d)$, i.e. $\langle x^d, x^{d-1}y, \ldots, x^{d-u+1}y^{u-1}\rangle$ (which is both the Lex and RevLev segment) and the questions in 3.6 have all straightforward answers.

REMARK 3.8: It is proved in [5] that Lex-segments, RevLex-segments and principal strongly stable vector spaces define normal and Koszul toric rings (in particular Cohen-Macaulay). Furthermore in [6] it is proved that a strongly stable vector space with two strongly stable generators define a Koszul toric ring. On the other hand, there are examples of strongly stable vector spaces with a non-Cohen-Macaulay and non-quadratic toric ring, see [3, Example 1.3].

3.2. Upper bound

As in the ideal case, the upper bound is achieved by a general subspace W, i.e., for W in a non-empty Zariski open subset of $G(u, S_d)$.

QUESTION 3.9. *What can be said about the value $M(n,d,u,j)$?*

Obviously,

$$M(n,d,u,j) \leq \min\left\{\dim S_{jd}, \binom{u-1+j}{u-1}\right\} \tag{1}$$

and the naive expectation is that equality holds in (1), i.e., if $f_1,\ldots,f_u$ are general forms of degree d, then the monomials of degree j in the f_i's are either linearly independent or they span S_{jd}. It turns out that in nature things are more complex than expected at first. First of all, if $u > n$ then equality in (1) would imply that for a generic W one would have $W^j = S_{jd}$ for large j. This fact can be stated in terms of projections of the d-th Veronese variety: the projection associated to W is an isomorphism. Recall that a generic linear projection of a smooth projective variety of dimension m from some projective space where its embedded, into a projective space of dimension c is an isomorphism if $c \geq 2m+1$. Hence we have that if $u \geq 2n$ then equality in (1) holds at least for large j. On the other hand, for $n+1 \leq u < 2n$ equality in (1) should not be expected unless one knows that the corresponding projection of the Veronese variety behaves in an unexpected way.

Summing up, the most natural question turns out to be:

QUESTION 3.10. *Assume that $u \geq 2n$. Is it true that*

$$M(n,d,u,j) = \min\left\{\dim S_{jd}, \binom{u-1+j}{u-1}\right\}$$

holds for all j?

The answer turns out to be negative as the following example shows:

PROPOSITION 3.11. *For any space W generated by eight quadrics in four variables the dimension of W^2 is at most 34 independently of the base field K. That is:*

$$M(4,2,8,2) \leq 34 < \min\left\{\dim S_4, \binom{7+2}{7}\right\} = 35.$$

REMARK 3.12: This assertion was proven in [4, Theorem 2.4] using a computer algebra calculation. Here we present a more conceptual argument.

Proof. Firstly we may assume that K has characteristic 0 and is algebraically closed. Secondly we may assume that W is generic. The 8-dimension space of quadrics W is apolar to a 2-dimension space of quadrics, call it V. A pair of generic quadrics can be put simultaneously in diagonal form, i.e., that V is generated by $x_1^2+x_2^2+x_3^2+x_4^2$ and $a_1x_1^2+a_2x_2^2+a_3x_3^2+a_4x_4^2$. See for example [14]. Indeed, it is sufficient that V contains a quadric of rank 4 since that can be put into the form $x_1^2+x_2^2+x_3^2+x_4^2$ and the other form can then be diagonalized preserving the first. As a consequence, after a change of coordinates W contains x_ix_j with $1 \le i < j \le 4$. Since $(x_1x_4)(x_2x_3) = (x_1x_2)(x_3x_4)$ and $(x_1x_3)(x_2x_4) = (x_1x_2)(x_3x_4)$ we have at least two independent relations among the 36 generators of W^2. Therefore $\dim W^2 \le 34$. □

More precisely one has:

Proposition 3.13. *One has $M(4,2,8,2) = 34$ independently of the base field K.*

Proof. We have already argued that $M(4,2,8,2) \le 34$.

Therefore it is enough to describe an 8-dimension space of quadrics W in 4 variables such that $\dim W^2 = 34$. We set

$$W_0 = \langle x_ix_j : 1 \le i < j \le 4\rangle$$

and

$$F = a_1x_1^2 + a_2x_2^2 + a_3x_3^2 + a_4x_4^2 \text{ and } G = b_1x_1^2 + b_2x_2^2 + b_3x_3^2 + b_4x_4^2 .$$

Then we set $W_1 = \langle F, G\rangle$ and then

$$W = W_0 + W_1.$$

We consider two conditions on the coefficients $a_1, a_2, \dots, b_4$:

Conditions 3.14. *(1) All the 2-minors of*

$$\begin{pmatrix} a_1 & a_2 & a_3 & a_4 \\ b_1 & b_2 & b_3 & b_4 \end{pmatrix}$$

are non-zero.

(2) The matrix

$$\begin{pmatrix} a_1^2 & a_2^2 & a_3^2 & a_4^2 \\ b_1^2 & b_2^2 & b_3^2 & b_4^2 \\ a_1b_1 & a_2b_2 & a_3b_3 & a_4b_4 \end{pmatrix}$$

has rank 3.

We observe that W_0^2 is generated by the 19 monomials of degree 4 and largest exponent ≤ 2. Then we note that if W_1 contains a quadric q supported on x_i^2, x_j^2 and x_h^2 with i, j, h distinct and $k \notin \{i, j, h\}$ then $x_kx_iq = x_kx_i^3$ mod (W_0^2) and similarly for j and h. This implies that if Condition (1) holds then $W_0^2 + W_0W_1$ is generated by the 31 monomials different from $x_1^4, \dots, x_4^4$. Assuming that Condition (1) holds, we have that the matrix representing F^2, G^2, FG in $S_4/W_0^2 + W_0W_1$ is exactly the one appearing in Condition (2). Then F^2, G^2, FG are linearly independent mod $W_0^2 + W_0W_1$ if and only if Condition (2) holds. Summing up, if Conditions (1) and (2) hold then $\dim W^2 = 34$.

Finally we observe that for $F = x_1^2+x_3^2+x_4^2$ and $G = x_2^2+\alpha x_3^2+x_4^2$ the conditions (1) and (2) are satisfied provided $\alpha \neq 0$ and $\alpha \neq 1$. Hence this (conceptual) argument works for any field but $\mathbb{Z}/2\mathbb{Z}$. Over $\mathbb{Z}/2\mathbb{Z}$ one can consider the space W generated by $x_1^2, x_2^2, x_3^2, x_1x_3, x_2x_4, x_3x_4, x_2x_3 + x_1x_4, x_1x_2 + x_4^2$ and check with the help of a computer algebra system that $\dim W^2 = 34$. $\square$

As far as we know the case discussed in Proposition 3.11 is the only known case where $u \geq 2n$ and the actual value of $M(n, d, u, j)$ is smaller than

$$\min\left\{\dim S_{jd}, \binom{u-1+j}{u-1}\right\}.$$

Acknowledgements

The first author was partially supported by VR2013-4545 and the second author was partially supported by INdAM-GNSAGA.

We thank Winfred Bruns and David Eisenbud for helpful discussions on the material of this paper.

References

[1] D.J. Anick, *Thin algebras of embedding dimension three*, J. Algebra, **100** (1986), no. 1, 235–259,

[2] W. Bruns and J. Herzog, *Cohen-Macaulay rings*, Cambridge Stud. Adv. Math., vol. 39, Cambridge University Press, Cambridge, 1993.

[3] W. Bruns and A. Conca, *Linear resolutions of powers and products*, Singularities and computer algebra, Springer, Cham, 2017, 47–69.

[4] L. Chiantini, F. Orecchia, and I. Ramella *Maximal rank and minimal generation of some parametric varieties*, J. Pure Appl. Algebra **186** (2004), no. 1, 21–31.

[5] E. De Negri, *Toric rings generated by special stable sets of monomials*, Math. Nachr. **203** (1999), 31–45.

[6] M. DiPasquale, C. Francisco, J. Mermin, J. Schweig, and G. Sosa, *The Rees algebra of a two-Borel ideal is Koszul*, Proc. Amer. Math. Soc., to apper.

[7] R. Fröberg, *An inequality for Hilbert series of graded algebras*, Math. Scand. **56** (1985), no. 2, 117–144.

[8] G. Gotzmann, *Eine Bedingung für die Flachheit und das Hilbertpolynom eines graduierten Ringes*, Math. Z. **158** (1978), no. 1, 61–70 (German).

[9] M. Hochster and D. Laksov, *The linear syzygies of generic forms*, Comm. Algebra **15** (1987), no. 1-2, 227–239.

[10] F. S. Macaulay, *Some properties of enumeration in the theory of modular systems*, Proc. London Math. Soc. (2) **26** (1927), 531–555.

[11] G. Nenashev, *A note on Fröberg's conjecture for forms of equal degrees*, C. R. Math. Acad. Sci. Paris **355** (2017), no. 3, 272–276.

[12] R. P. Stanley, *Weyl groups, the hard Lefschetz theorem, and the Sperner property*, SIAM J. Algebraic Discrete Methods **1** (1980), no.2, 168–184.

[13] B. Sury, *Macaulay expansion*, Amer. Math. Monthly **121** (2014), no.4, 359–360.

[14] M. J. Wonenburger, *Simultaneous diagonalization of symmetric bilinear forms*, J. Math. Mech. **15** (1966), 617–622.

Authors' addresses:

Mats Boij
KTH - Royal Institute of Technology
SE-100 44 Stockholm, Sweden
E-mail: `boij@kth.se`

Aldo Conca
Dipartimento di Matematica
Università di Genova
Via Dodecaneso 35
I-16146 Genova, Italy
E-mail: `conca@dima.unige.it`

Received October 26, 2018
Accepted November 5, 2018

Rend. Istit. Mat. Univ. Trieste
Volume 50 (2018), 149–159
DOI: 10.13137/2464-8728/22434

Tannakian categories, fundamental groups and Higgs bundles

Ugo Bruzzo

Abstract. *After recalling the basic notions concerning profinite and proalgebraic group completions and Tannakian categories, we review how the latter can be used to define generalizations of the notion of fundamental group of a space, such as the Nori and Langer fundamental groups, and the algebraic fundamental group introduced by Simpson. Then we discuss how one can define a Tannakian category whose objects are Higgs bundles on a complex projective variety that are "numerically flat" in a suitable sense, and show how the Higgs fundamental group is related to a conjecture about semistable Higgs bundles.*

Keywords: Fundamental groups, Tannakian categories, Higgs bundles, curve semistability.
MS Classification 2010: 14F05, 14F35, 14H60, 14J60, 18D35.

1. Introduction

Tannakian categories are abelian tensor categories that satisfy some additional properties and are equipped with a functor to the category of vector spaces. They all turn out to be equivalent to categories of representations of proalgebraic affine group schemes, so that there is natural duality between Tannakian categories and such group schemes. This "Tannaka duality" has been used to devise generalizations of the notion of fundamental group, with the purpose of better capturing the geometry of such geometric structures as schemes and algebraic varieties. A classical example is the Nori fundamental group [21, 22], and more recently, the S-fundamental group introduced by Langer [16, 17]. The latter is the Tannaka dual of the category of numerically flat vector bundles, i.e., vector bundles that are numerically effective together with their duals (this group was introduced in the case of curves also in [5]). C. Simpson considered the category of semi-harmonic bundles on a smooth projective variety over $\mathbb{C}$, i.e., semistable Higgs vector bundles with vanishing rational Chern classes [24, 25]. The resulting fundamental group scheme is a proalgebraic completion of the topological fundamental group. Since flat (Higgs) bundles are essentially finite, numerically flat, and semi-harmonic, and the topological fundamental group represents the category of flat bundles, there is a natural morphism from the usual fundamental group to each of these groups.

Notions of numerical effectiveness and numerical flatness for Higgs bundles were introduced in [6, 7] , motivated by the remark that the universal quotient bundles over the Grassmann bundles $\mathrm{Gr}_k(E)$ of a numerically effective vector bundle are numerically effective. Given a Higgs vector bundle $\mathfrak{E} = (E, \phi)$, we consider closed subschemes $\mathfrak{Gr}_k(\mathfrak{E}) \subset \mathrm{Gr}_k(E)$ that parameterize locally free Higgs quotients on $\mathfrak{E}$. Then $\mathfrak{E}$ is said to be H-numerically effective if the universal Higgs quotients on $\mathfrak{Gr}_k(\mathfrak{E})$ are H-numerically effective, according to a definition which is recursive on the rank. Finally, a Higgs bundle is said to be H-numerically flat if $\mathfrak{E}$ and its dual Higgs bundle $\mathfrak{E}^*$ are H-numerically effective. H-numerically flat Higgs bundles make up again a neutral Tannakian category; the corresponding group scheme is denoted $\pi_1^H(X, x)$ [4].

Numerically flat vector bundles equipped with the zero Higgs field are H-numerically flat, hence there is a faitfhfully flat morphism $\pi_1^H(X, x) \to \pi_1^S(X, x)$. The relation of $\pi_1^H(X, x)$ with Simpson's proalgebraic fundamental group $\pi_1^{\mathrm{alg}}(X, x)$ is more subtle: semi-harmonic bundles are H-numerically flat, so that there is faitfhfully flat morphism $\pi_1^H(X, x) \to \pi_1^{\mathrm{alg}}(X, x)$. The fact that the groups may be isomorphic is related with a conjecture about the so-called *curve semistable Higgs bundles* — i.e., Higgs bundles that are semistable after pullback to any smooth projective curve [7, 11, 18] (Conjecture 4.7 in the text). This conjecture states that if a Higgs bundle (E, ϕ) on a projective variety is semistable after pullback to any projective curve, then its rational characteristic class

$$\Delta(E) = c_2(E) - \frac{r-1}{2r} c_1(E)^2$$

vanishes (here $r = \mathrm{rk}\, E$).

2. Completions

Generalized fundamental groups are defined in terms of, or are related to, completions of discrete groups. In this section we briefly review the definitions of *profinite* and *proalgebraic* completion of a discrete group.

DEFINITION 2.1. *A profinite group is a topological group which is the inverse limit of an inverse system of discrete finite groups. The profinite completion $\hat{G}$ of a group G is the inverse limit of the system formed by the quotients groups G/N of G, where N are normal subgroups of G of finite index, ordered by inclusion.*

For instance, the profinite completion of $\mathbb{Z}$ is

$$\hat{\mathbb{Z}} = \prod_p \mathbb{Z}(p)\,,$$

where p runs over the prime numbers, and $\mathbb{Z}(p)$ is the ring of p-adic integers [19].

An interesting geometric example of a profinite completion is *Grothendieck's fundamental group* [14]. The idea for its introduction may be regarded as a generalization of the usual fundamental group, recalling that for X a topological space, $\pi_1(X)$ is the group of deck transformations of the universal covering of X. To get a suitable replacement for schemes, one substitutes covering spaces with étale covers. So, if X a connected and locally noetherian scheme over a field $\Bbbk$, let x be a geometric point in X, i.e., a morphism $\operatorname{Spec}\bar{\Bbbk} \to X$, where $\bar{\Bbbk}$ is a separable closure of $\Bbbk$. Let I be the set of pairs (p, y), where $p\colon Y \to X$ is a finite étale cover, and $y \in Y$ is a geometric point such that $p(y) = x$, partially ordered by the relation $(p, y) \geq (p', y')$ if there is a commutative diagram

$$\begin{tikzcd} Y \arrow[r, "f"] \arrow[dr, "p"'] & Y' \arrow[d, "p'"] \\ & X \end{tikzcd}$$

with with $y' = f(y)$. Then one sets

$$\pi_1^{\text{ét}}(X, x) = \varprojlim_{i \in I} \operatorname{Aut}_X(p_i, y_i).$$

If X is a scheme of finite type over $\mathbb{C}$, *the étale fundamental group* $\pi_1^{\text{ét}}(X, x)$ *is a profinite completion of the topological fundamental group* $\pi_1(X, x)$ [14].

In spite of the naturalness of its definition, the étale fundamental group, for a field of positive characteristic, fails to enjoy some quite reasonable properties; for instance, it is not a birational invariant, and is not necessarily zero for rational varieties [21, 22]. Nori's fundamental group solves some of these problems. It is defined in terms of Tannaka duality (see next Section) and involves the notion of proalgebraic completion of a discrete group [3].

A proalgebraic group over $\Bbbk$ is the inverse limite of a system of algebraic groups over $\Bbbk$.

Definition 2.2. *Let* Γ *be a discrete group. A proalgebraic completion of* Γ *over* $\Bbbk$ *is a proalgebraic group* $A(\Gamma)$ *over* $\Bbbk$ *with a homomorphism* $\rho : \Gamma \to A(\Gamma)$ *such that every morphism* $\Gamma \to H$*, where* H *is a proalgebraic group over* $\Bbbk$*, uniquely filters through* $A(\Gamma)$ *via* ρ

$$\begin{tikzcd} & \Gamma \arrow[dl, "\rho"'] \arrow[d] \\ A(\Gamma) \arrow[r, dotted] & H \end{tikzcd}$$

A proalgebraic completion for Γ is unique up to unique isomorphism. The image of ρ is Zariski dense in $A(\Gamma)$. A proalgebraic completion can be built via

Tannaka duality, as the group of tensor product preserving automorphisms of the forgetful functor from the category of finite dimensional Γ-modules to the category of finite dimensional $\Bbbk$-vector spaces.

3. Tannakian categories

In this section we recall the main notions and establish the basic notation about Tannakian categories. For a detailed introduction the reader may refer to [12].

A category $\mathfrak{C}$ is *additive* if

- the Hom classes are abelian groups and the composition of morphisms is bilinear;
- $\mathfrak{C}$ has finite direct sums and direct products;
- it has a zero object.

An additive category is *abelian* if

- every morphism has both a kernel and a cokernel (the notion of kernel and cokernel are defined in terms of suitable universal properties);
- every monomorphism is a kernel of some morphism, and every epimorphism is a cokernel of some morphism.

An additive category is $\Bbbk$-*linear* over a field $\Bbbk$ if the Hom groups are $\Bbbk$-vector spaces, and the composition of morphisms is $\Bbbk$-linear. A *tensor* category is an abelian category with a biproduct satisfying the usual properties of the tensor product (including the existence of a unit object 1 for the tensor product).

A tensor category is *rigid* if

- Hom and $\otimes$ satisfy the natural distributive property over finite families;
- all objects are reflexive, i.e., the natural maps to their double duals are isomorphisms (the dual $A^\vee$ of an object A of $\mathfrak{C}$ is the object $\mathrm{Hom}(A, 1)$).

Definition 3.1. *A* neutral Tannakian category *over a field* $\Bbbk$ *is a rigid Abelian* $\Bbbk$*-linear tensor category* $\mathfrak{T}$ *together with an exact faithful* $\Bbbk$*-linear tensor functor* $\omega\colon \mathfrak{T} \longrightarrow \mathbf{Vect}_{\Bbbk}$*, called the* fiber functor.

The archetypical Tannakian category is the category $\mathrm{Rep}(G)$ of representations (on vector spaces over $\Bbbk$) of an affine group scheme G over $\Bbbk$. The fiber functor is defined as the forgetful functor

$$\omega(\rho, V) = V \qquad \text{if} \qquad \rho\colon G \to \mathrm{Aut}(V).$$

Categories of representations of affine group schemes are much more than just examples: it turns out that every neutral Tannakian category is equivalent to one of them [12].

THEOREM 3.2 (Tannaka duality). *For every neutral Tannakian category* $(\mathfrak{T}, \omega)$ *there is a proalgebraic affine group scheme* G *such that* $\mathfrak{T} \simeq \mathrm{Rep}(G)$.

The group G is recovered as the group of automorphisms of the fiber functor that are compatible with the tensor product, $G = \mathrm{Aut}^{\otimes}(\omega)$. If $\mathfrak{T} \simeq \mathrm{Rep}(G)$, one also writes $G = \pi_1(\mathfrak{T})$.

EXAMPLES 3.3:

- The category $\mathbf{Vect}_{\Bbbk}$ of vector spaces over $\Bbbk$ with the identity as fiber functor is a neutral Tannakian category. Its corresponding affine group scheme is the trivial group $G = \{e\}$, i.e., $\pi_1(\mathbf{Vect}_{\Bbbk}) = \{e\}$.

- The category of modules over a commutative ring with unit R is an abelian tensor category. It may fail to be rigid as there are R-modules that are not reflexive.

- If $\mathfrak{g}$ is a semisimple Lie algebra over a field $\Bbbk$, the category $\mathrm{Rep}(\mathfrak{g})$ of representations of $\mathfrak{g}$, with the fiber functor given by the forgetful functor that only keeps the vector space structure of $\mathfrak{g}$, is a neutral Tannakian category, and $\pi_1(\mathrm{Rep}(\mathfrak{g})) = G$, where G is the unique connected simply connected Lie group whose Lie algebra is $\mathfrak{g}$.

- If X is a smooth projective variety over $\mathbb{C}$, the category of vector bundles on X with a flat connection (a.k.a. local systems), with a functor which to a bundle E associates its fiber at $x \in X$, is Tannakian, and is equivalent to the category $\mathrm{Rep}(\pi_1(X, x))$ of representations of the topological fundamental group of X. The dual group via Tannaka duality, i.e. the group $\pi_1(\mathrm{Rep}(\pi_1(X, x)))$, is the proalgebraic completion of $\pi_1(X, x)$.

4. Tannakian categories and fundamental groups

The basic idea for using Tannaka duality to define fundamental groups is to single out a class of geometric objects on a scheme X that make up a neutral Tannakian category, and take the associated group scheme. We briefly review two examples, Nori's and Langer's fundamental groups. Next we shall introduce the Higgs fundamental group and discuss its relation with Simpson's proalgebraic fundamental group; this will be related to a conjecture about semistable Higgs bundles on projective varieties.

Nori's fundamental group

The first example of such a fundamental group was provided by Nori [21, 22]. A vector bundle E over a scheme X is *essentialy finite* if there exists a principal bundle $\pi\colon P \to X$, with a finite structure group, such that π^*E is trivial. Essentially finite vector bundles make up a neutral Tannakian category, where the fiber functor maps E to the fiber over a fixed point $x \in X$ (some assumptions on the scheme X have to be made). The affine group scheme representing this Tannakian category is the *Nori fundamental group scheme* $\pi_1^N(X, x)$. It turns out that there is a faithfully flat (i.e., flat and surjective) morphism

$$\pi_1^N(X,x) \twoheadrightarrow \pi_1^{\text{ét}}(X,x)$$

which is an isomorphism when $\operatorname{char} \Bbbk = 0$.

A related notion, that of *F-fundamental group,* was introduced in [2], and some properties of it were studied in [1]. Another generalization was proposed in [23].

Langer's fundamental group

Let X be a smooth projective variety over an algebraically closed field. We can define intersections between divisors D and curves C in X by letting

$$C \cdot D = \deg f^* \mathcal{O}_X(D)$$

where $f\colon \tilde{C} \to C$ is a normalization of C. In the same way, we can define the intersection product between a line bundle and a curve. Then we have the usual notion of numerical effectiveness.

Definition 4.1. *L is numerically effective (nef) if $L \cdot C \geq 0$ for all irreducible curves C in X. A vector bundle E on X is numerically effective if its hyperplane line bundle $\mathcal{O}_{\mathbb{P}(E)}(1)$ on the projectivization $\mathbb{P}(E)$ is. E is numerically flat if both E and its dual bundle $E^\vee$ are nef.*

As proved by Langer [16, 17], numerically flat vector bundles make up a neutral Tannakian category, so that one can define a "fundamental group" $\pi_1^S(X,x)$ as its dual (this group was introduced in the case of curves also in [5]). Essentially finite vector bundles are numerically flat, so that there is a morphism

$$\pi_1^S(X,x) \twoheadrightarrow \pi_1^N(X,x)$$

which is again faithfully flat, and is an isomorphism when $\operatorname{char} \Bbbk = 0$. Some properties of this fundamental group, e.g. its birational invariance, were proved in [15].

Higgs fundamental group

We follow this pattern to introduce a fundamental group which "feels" the behavior of Higgs bundles on a projective variety. To do that we restrict to varieties over the complex numbers, and start by considering ordinary bundles.

So, let X be a smooth projective variety over $\mathbb{C}$, and E a vector bundle on X of rank r. We shall consider the characteristic class (the discriminant of E)

$$\Delta(E) = c_2(E) - \frac{r-1}{2r} c_1(E)^2 \in H^4(X,\mathbb{R}) .$$

Moreover, after equipping X with an ample line bundle L, and denoting by H it first Chern class (a polarization on X), we define the degree of E as

$$\deg E = c_1(E) \cdot H^{n-1}$$

where $n = \dim X$. If X is a smooth irreducible projective curve, it has a canonical polarization, given by the class of a closed point of X. Whenever X is such a curve, one implicitly assumes this choice of a polarization.

Definition 4.2. *E is semistable (with respect to the chosen polarization) if for every coherent subsheaf $F \subset E$, with $0 < \operatorname{rk} F < r$, one has*

$$\frac{\deg F}{\operatorname{rk} F} \le \frac{\deg E}{r} .$$

E is curve semistable if for all morphisms $f\colon C \to X$, where C is a smooth projective irreducible curve, the pullback bundle $f^*(E)$ is semistable.

The following theorem was proved in a slightly weaker form by Nakayama [20] and strengthened into its present form by Hernández Ruipérez and the author [9].

Theorem 4.3. *The following conditions are equivalent:*

- *E is curve semistable;*
- *E is semistable with respect to a polarization, and $\Delta(E) = 0$.*

The following corollary is not hard to prove [9].

Corollary 4.4. *E is numerically flat if and only if it is curve semistable and $c_1(E) = 0$.*

It is quite natural to ask if a result such as Theorem 4.3 also works for Higgs bundles. A Higgs sheaf is a pair (E,ϕ) where E is a coherent sheaf and

$$\phi\colon E \to E \otimes \Omega^1_X, \qquad \phi \wedge \phi = 0.$$

A Higgs bundle is a locally free Higgs sheaf. A notion of semistability is given as for ordinary vector bundles, but the inequality is required to hold only for ϕ-invariant subsheaves. There is a notion of nefness/numerical flatness for Higgs bundles [9, 7], which we briefly review here. If E is a vector bundle of rank r on X, and $s < r$ is a positive integer, we can consider the Grassmann bundle $\mathrm{Gr}_s(E)$ on X. Denote by $p_s : \mathrm{Gr}_s(E) \longrightarrow X$ the natural projection. There is a universal short exact sequence

$$0 \longrightarrow S_{r-s,E} \xrightarrow{\psi} p_s^* E \xrightarrow{\eta} Q_{s,E} \longrightarrow 0 \tag{1}$$

of vector bundles on $\mathrm{Gr}_s(E)$, with $S_{r-s,E}$ the universal subbundle of rank $r-s$ and $Q_{s,E}$ the universal quotient of rank s [13]. The Grassmannian $\mathrm{Gr}_s(E)$ parameterizes locally free rank s quotients of E, in the sense that if $f\colon Y \to X$ is a morphism, and G is a quotient bundle of $f^*(E)$, there is a morphism $g\colon Y \to \mathrm{Gr}_s(E)$ such that $G \simeq g^* Q_{s,E}$, and the diagram

$$\begin{tikzcd} & \mathrm{Gr}_s(E) \arrow[d, "p_s"] \\ Y \arrow[ur, "g"] \arrow[r, "f"'] & X \end{tikzcd}$$

commutes [13].

Given a Higgs bundle $\mathfrak{E} = (E, \phi)$, we define closed subschemes $\mathfrak{Gr}_s(\mathfrak{E}) \subset \mathrm{Gr}_s(E)$ parameterizing rank s locally free Higgs quotients, i.e., locally free quotients of E whose corresponding kernels are ϕ-invariant. The Grassmannian of locally free rank s Higgs quotients of $\mathfrak{E}$, denoted $\mathfrak{Gr}_s(\mathfrak{E})$, is the closed subscheme of $\mathrm{Gr}_s(E)$ defined by the vanishing of the composition of morphisms

$$(\eta \otimes \mathrm{Id}) \circ p_s^*(\phi) \circ \psi \colon S_{r-s,E} \longrightarrow Q_{s,E} \otimes p_s^* \Omega^1_X. \tag{2}$$

Let $\rho_s := p_s|_{\mathfrak{Gr}_s(\mathfrak{E})} : \mathfrak{Gr}_s(\mathfrak{E}) \longrightarrow X$ be the induced projection. The restriction of (1) to $\mathfrak{Gr}_s(\mathfrak{E})$ yields a universal exact sequence

$$0 \longrightarrow \mathfrak{S}_{r-s,\mathfrak{E}} \xrightarrow{\psi} \rho_s^* \mathfrak{E} \xrightarrow{\eta} \mathfrak{Q}_{s,\mathfrak{E}} \longrightarrow 0, \tag{3}$$

where $\mathfrak{Q}_{s,\mathfrak{E}} := Q_s|_{\mathfrak{Gr}_s(\mathfrak{E})}$ is endowed with the quotient Higgs field induced by the Higgs field $\rho_s^*\phi$. A morphism of $\Bbbk$-varieties $f : T \to X$ factors through $\mathfrak{Gr}_s(\mathfrak{E})$ if and only if the pullback $f^*(E)$ admits a Higgs quotient of rank s. The pullback of the above universal sequence on $\mathfrak{Gr}_s(E)$ gives a quotient of $f^*(E)$.

Definition 4.5. *A Higgs bundle $\mathfrak{E}$ of rank one is said to be Higgs-numerically effective (H-nef for short) if it is numerically effective in the usual sense. If* $\mathrm{rk}\,\mathfrak{E} \geq 2$, *we inductively define H-nefness by requiring that*

1. *all Higgs bundles $\mathfrak{Q}_{s,\mathfrak{E}}$ are Higgs-nef, and*

2. the determinant line bundle $\det(E)$ *is nef.*

If both $\mathfrak{E}$ *and* $\mathfrak{E}^*$ *are Higgs-numerically effective,* $\mathfrak{E}$ *is said to be Higgs-numerically flat (H-nflat).*

Definition 4.5 immediately implies that the first Chern class of an H-numerically flat Higgs bundle is numerically equivalent to zero.

It was proved in [4] that numerically flat Higgs bundles make up a neutral Tannakian category. Therefore, after fixing a point $x \in X$, we can define the *Higgs fundamental group* $\pi_1^H(X, x)$ as the group which is Tannaka dual to that category. A numerically flat vector bundle, equipped with the zero Higgs field, is a numerically flat Higgs vector bundle, so that there is a morphism

$$\pi_1^H(X, x) \twoheadrightarrow \pi_1^S(X, x)$$

which is again faithfully flat.

The nature of this fundamental group is related to the validity of Theorem 4.3 for Higgs bundles. The following theorem was proved in [7].

THEOREM 4.6. *If* $\mathfrak{E} = (E, \phi)$ *is semistable, and* $\Delta(E) = 0$*, then* $\mathfrak{E}$ *is curve semistable.*

The question whether the opposite result holds true is an open problem.

CONJECTURE 4.7. *If the Higgs bundle* $\mathfrak{E}$ *is curve semistable, then* $\Delta(E) = 0$.

Conjecture 4.7 is known to hold for certain classes of varieties (varieties whose tangent bundle is numerically effective [11] and K3 surfaces [10], and varieties obtained from these two classes by some simple geometric constructions [11]).

The category of semistable Higgs bundles on X having vanishing Chern classes (semi-harmonic Higgs bundles) is Tannakian (the definition of this category does not require the specification of a polarization since such bundles are semistable with respect to all polarizations). Its Tannaka dual is isomorphic to the proalgebraic completion of the topological fundamental group $\pi_1^{\mathrm{alg}}(X, x)$ [24]. Since such semi-harmonic Higgs bundles are Higgs numerically effective, there is a morphism (again, a faithfully flat morphism)

$$\pi_1^H(X, x) \to \pi_1^{\mathrm{alg}}(X, x). \tag{4}$$

THEOREM 4.8. *The morphism* (4) *is an isomorphism if and only the Conjecture 4.7 holds.*

Proof. If the morphism (4) is an isomorphism, the categories of numerically flat Higgs bundles and semi-harmonic bundles are equivalent. Then a numerically flat Higgs bundle has vanishing Chern classes, which implies the conjecture.

Vice versa, if the conjecture holds, and $\mathfrak{E} = (E, \phi)$ is a numerically flat Higgs bundle, then $\mathfrak{E}$ is curve semistable, and since the Conjecture is assumed to hold, $\Delta(E) = 0$; moreover, $\mathfrak{E}$ is semistable and $c_1(E) = 0$ [7], so that by Theorem 2 in [24], all Chern classes of E vanish, and $\mathfrak{E}$ is semi-harmonic. Thus the two above mentioned categories are isomorphic, and (4) is an isomorphism. $\square$

Acknowledgements

The original results presented in this paper were obtained in collaboration with Indranil Biswas, Beatriz Graña Otero, Sudarshan Gurjar, Daniel Hernández Ruipérez, Valeriano Lanza and Alessio Lo Giudice. I thank them all for their collaboration, in particular B.G.O. with whom I wrote the related paper [8]. I also thank V. Lanza and T. Pantev for very useful discussions on Tannaka duality.

The text of this paper reflects the talk I gave on January 29th, 2018, at the "INdAM day" held at the University of Trieste to celebrate the 50th anniversary of the "Rendiconti dell'Istituto di Matematica dell'Università di Trieste" journal. I am grateful to the organizers for their invitation.

This research was partly supported by GNSAGA–INdAM and PRIN "Geometria delle varietà algebriche."

References

[1] S. Amrutiya, *A note on certain Tannakian group schemes*, `arXiv 1810.10027`.

[2] S. Amrutiya and I. Biswas, *On the F-fundamental group scheme*, Bull. Sci. Math. **134** (2010), 461–474.

[3] H. Bass, A. Lubotzky, A. Magid, and S. Mozes, *The proalgebraic completion of rigid groups*, Geom. Dedicata **95** (2002), 19–58.

[4] I. Biswas, U. Bruzzo, and S. Gurjar, *Higgs bundles and fundamental group schemes*. `arXiv:1607.07207`, to appear in Adv. Geom.

[5] I. Biswas, A. J. Parameswaran, and S. Subramanian, *Monodromy group for a strongly semistable principal bundle over a curve*, Duke Math. J. **132** (2006), 1–48.

[6] U. Bruzzo and B. Graña Otero, *Numerically flat Higgs vector bundles*, Commun. Contemp. Math. **9** (2007), 437–446.

[7] U. Bruzzo and B. Graña Otero, *Semistable and numerically effective principal (Higgs) bundles*, Adv. Math. **226** (2011), 3655–3676.

[8] U. Bruzzo and B. Graña Otero, *Higgs varieties and fundamental groups*, J. Geom. Phys. **128** (2018), 12–19.

[9] U. Bruzzo and D. Hernández Ruipérez, *Semistability vs. nefness for (Higgs) vector bundles*, Differential Geom. Appl. **24** (2006), 403–416.

[10] U. Bruzzo, V. Lanza, and A. Lo Giudice, *Semistable Higgs bundles on Calabi-Yau manifolds*, `arXiv:1710.03671`.

[11] U. Bruzzo and A. Lo Giudice, *Restricting Higgs bundles to curves*, Asian J. Math. **20** (2016), 399–408.

[12] P. Deligne and J. S. Milne, *Tannakian categories*, in Hodge cycles, motives, and Shimura varieties, P. Deligne, J. S. Milne, A. Ogus, and K.-Y. Shih, eds., vol. 900 of Lecture Notes in Mathematics, Springer, Berlin-New York, 1982, pp. ii+414.

[13] W. Fulton, *Intersection theory*, vol. 2 of Ergebnisse der Mathematik und ihrer Grenzgebiete. 3. Folge, Springer, Berlin, second ed., 1998.

[14] A. Grothendieck and M. Raynaud, *Revêtements étales et groupe fondamental*, Séminaire de Géométrie Algébrique du Bois Marie – 1960-61 (SGA 1) (Documents Mathématiques 3), Société Mathématique de France, Paris, 1971–2003, pp. xviii+327. Exp. V, IX, X.

[15] A. Hogadi and V. Mehta, *Birational invariance of the S-fundamental group scheme*, Pure Appl. Math. Q. **7** (2011), 1361–1369.

[16] A. Langer, *On the S-fundamental group scheme*, Ann. Inst. Fourier (Grenoble) **61** (2011), 2077–2119.

[17] A. Langer, *On the S-fundamental group scheme. II*, J. Inst. Math. Jussieu **11** (2012), 835–854.

[18] V. Lanza and A. Lo Giudice, *Bruzzo's conjecture*, J. Geom. Physics **118** (2017), 181–191.

[19] H. Lenstra, *Profinite groups*, available at `http://websites.math.leidenuniv.nl/algebra/Lenstra-Profinite.pdf`.

[20] N. Nakayama, *Normalized tautological divisors of semi-stable vector bundles*, Sūrikaisekikenkyūsho Kōkyūroku, (1999), pp. 167–173. Kyoto University, Research Institute for Mathematical Sciences.

[21] M. V. Nori, *On the representations of the fundamental group*, Compositio Math. **33** (1976), 29–41.

[22] M. V. Nori, *The fundamental group-scheme*, Proc. Indian Acad. Sci. Math. Sci. **91** (1982), 73–122.

[23] S. Otabe, *An extension of Nori fundamental group*, Comm. Algebra **45** (2017), 3422–3448.

[24] C. T. Simpson, *Higgs bundles and local systems*, Inst. Hautes Études Sci. Publ. Math. **75** (1992), 5–95.

[25] C. T. Simpson, *Local systems on proper algebraic V-manifolds*, Pure Appl. Math. Q. **7** (2011), 1675–1759.

Author's address:

Ugo Bruzzo
SISSA (Scuola Internazionale Superiore di Studi Avanzati)
Via Bonomea 265
34136 Trieste, Italia;
Istituto Nazionale di Fisica Nucleare, Sezione di Trieste;
IGAP (Institute for Geometry and Physics), Trieste;
Arnold-Regge Centre for Algebra, Geometry and Theoretical Physics, Torino
E-mail: `bruzzo@sissa.it`

Received October 28, 2018
Accepted November 7, 2018

Editorial Note

The Managing Editors would like to remind the authors of the ethical rules set out in the document of IMU "Best current practices for Journals" (see www.mathunion.org). In particular, we stress the following:

"Authors must abide by high standards of research integrity and good scholarship. It is the responsibility of authors to submit a well written, mathematically correct article, if necessary seeking advice if it is not written in their native language, to clearly describe any novel and non-trivial content, and to suitably acknowledge the contributions of others, including referees. Submission of a paper to a journal implies that it is not currently under consideration by any other journal, and that any substantial overlap with other published or submitted papers is duly acknowledged. In addition authors should be responsive to correspondence with the journal. Multiple authors should communicate fully, speak with one voice, and accept mutual responsibility in their communications with the journal. All authors are expected to have materially contributed to the paper, and to be familiar with its content."

Alessandro Fonda
Emilia Mezzetti
Pierpaolo Omari
Maura Ughi

Contents

Editorial Policy
The journal *Rendiconti dell'Istituto di Matematica dell'Università di Trieste* publishes original articles in all areas of mathematics, with emphasis to the area of expertise of the Editorial Board. 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 special issues related to some events, like conferences, workshops, and advanced schools.

The members of the Editorial Board may propose to publish articles under their own responsibility, in which case the work will be published with the wording "Communicated by ...".

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.

Open Access
Rendiconti dell'Istituto di Matematica dell'Università di Trieste - An International Journal of Mathematics - is a no-fee open access journal: all published articles are freely available on-line and no fee is charged to authors. Authors also retain the copyright of their work.

Instructions for Authors
Authors are invited to submit their papers by e-mail directly to one of the Managing Editors in PDF format.

The authors may indicate the name of one or more members of the Editorial Board whose scientific interests are closer to the topic treated in their article.

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 specific templates which can be downloaded from the web page.

Any figure should be recorded in a single PDF file.

EDITORIAL OFFICE ADDRESS

Rendiconti dell'Istituto di Matematica dell'Università di Trieste
Dipartimento di Matematica e Geoscienze, Via Valerio 12/1, 34127 Trieste, Italia

Web page http://rendiconti.dmi.units.it

Direttore Responsabile: ALESSANDRO FONDA

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, via Edoardo Weiss 21, 34128 Trieste, Italia - mail: eut@units.it

ISSN 0049-4704 — eISSN 2464-8728

Contents

ISSN 0049-4704
eISSN 2464-8728