REPORTS @SCM

AN ELECTRONIC JOURNAL OF THE SOCIETAT CATALANA DE MATEMÀTIQUES

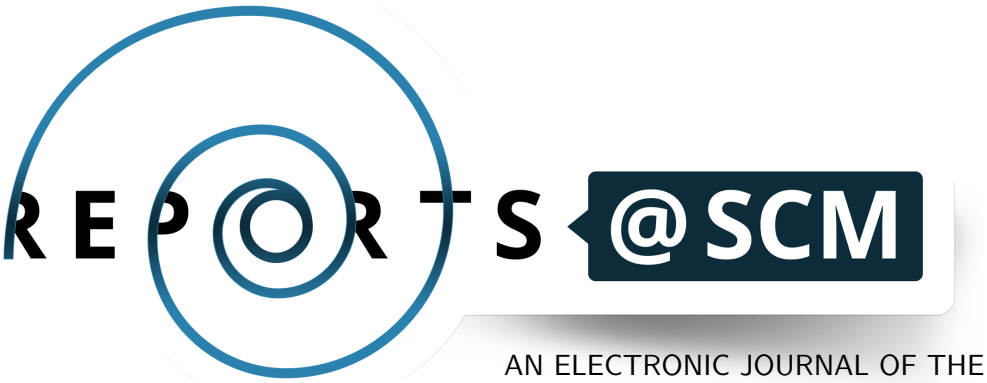
Volume 2, num. 1, 2016

ISSN (electronic edition): 2385 - 4227

http://reportsascm.iec.cat

Institut d'Estudis Catalans

Image: Arnaud Chéritat



AN ELECTRONIC JOURNAL OF THE
SOCIETAT CATALANA DE MATEMÀTIQUES

Volume 2, number 1

April 2016

http://reportsascm.iec.cat
ISSN electronic edition: 2385 - 4227

Societat
Catalana de
Matemàtiques

Institut
d'Estudis
Catalans

Focus and Scope

Reports@SCM is a non-profit electronic research journal on Mathematics published by the Societat Catalana de Matemàtiques (SCM) which originated from the desire of helping students and young researchers in their first steps into the world of research publication.

Reports@SCM publishes short papers (maximum 10 pages) in all areas of pure mathematics, applied mathematics, and mathematical statistics, including also mathematical physics, theoretical computer science, and any application to science or technology where mathematics plays a central role. To be considered for publication in Reports@SCM an article must be written in English (with an abstract in Catalan), be mathematically correct, and contain some original interesting contribution. All submissions will follow a peer review process before being accepted for publication.

Research announcements containing preliminary results of a larger project are also welcome. In this case, authors are free to publish in the future any extended version of the paper elsewhere, with the only condition of making an appropriate citation to Reports@SCM.

We especially welcome contributions from researchers at the initial period of their academic careers, such as Master or PhD students. We wish to give special attention to the authors during the whole editorial process. We shall take special care of maintaining a reasonably short average time between the reception of a paper and its acceptance, and between its acceptance and its publication.

All manuscripts submitted will be **peer reviewed** by at least one reviewer. Final decisions on the acceptance of manuscripts are taken by the editorial board, based on the reviewer's opinion.

Edited by Societat Catalana de Matemàtiques, Institut d'Estudis Catalans (IEC)
Carrer del Carme 47, 08001 Barcelona.
http://scm.iec.cat Telèfon: (+34) 93 324 85 83
scm@iec.cat Fax: (+34) 93 270 11 80

Style revision by Enric Ventura.

Institut d'Estudis Catalans
http://www.iec.cat
informacio@iec.cat

http://reportsascm.iec.cat
ISSN electronic edition: 2385-4227

Reports@SCM
Volume 2, number 1, 2016

Table of Contents

REPORTS@SCM

AN ELECTRONIC JOURNAL OF THE
SOCIETAT CATALANA DE MATEMÀTIQUES

Maximal values for the simultaneous number of null components of a vector and its Fourier transform

*Alberto Debernardi Pinos

Centre de Recerca Matemàtica (CRM), Bellaterra (Barcelona)
adebernardi@crm.cat

∗Corresponding author

Resum *(CAT)*

Motivat pel principi d'incertesa, el propòsit d'aquest treball és el de trobar el valor dels nombres $L(N) = \max_{x\in\mathbb{C}^N\setminus\{0\}} \min\{Z(x), Z(\hat{x})\}$, on $\hat{x}$ i $Z(x)$ denoten la transformada de Fourier discreta i el nombre de components nul·les de x, respectivament. Dit d'una altra manera, ja que el principi d'incertesa ens assegura que $Z(x)$ és inversament proporcional a $Z(\hat{x})$, estudiem el millor balanç que hi pot haver entre aquests dos nombres.

Abstract *(ENG)*

Motivated by the uncertainty principle, the purpose of this work is to find the value of the numbers $L(N) = \max_{x\in\mathbb{C}^N\setminus\{0\}} \min\{Z(x), Z(\hat{x})\}$, where $\hat{x}$ and $Z(x)$ denote the discrete Fourier transform and the number of null components of x, respectively. In other words, since the uncertainty principle ensures that $Z(x)$ is inversely proportional to $Z(\hat{x})$, we study the best possible balance between these two numbers.

Keywords: *Uncertainty principle, Fourier matrix, Fourier submatrix, DFT.*
MSC (2010): *42A99, 15A03.*
Received: *January 30th, 2015.*
Accepted: *June 12th, 2015.*

Acknowledgement

The author acknowledges the support of prof. J. Soria de Diego (Universitat de Barcelona), who advised the Master thesis in which this research was carried out. The research was partially supported by an AGAUR master's grant (course 2013–14) and the grant MTM2014–59174–P.

Societat Catalana de Matemàtiques

Institut d'Estudis Catalans

1. Introduction

In quantum mechanics, the uncertainty principle (due to Heisenberg, 1927) is a very basic result, asserting that we cannot determine the position and the momentum of a particle at the same time; in particular, the more precisely the position of a particle is determined, the less accurate its momentum can be known (and vice versa).

In mathematics there are many versions of this result, but the most remarkable one is the following: suppose that we have a function $f \in L^2(\mathbb{R})$. Then, we cannot arbitrarily concentrate both f and its Fourier transform, namely $\hat{f}$. Concretely, one of its many generalizations states that if f is practically zero outside a measurable set T, and $\hat{f}$ is practically zero outside a measurable set S, then $|T| \cdot |S| \geq 1 - \delta$, where δ is a small number which depends on the meaning of the phrase "practically zero" (for an accurate statement, see [4, Thm. 2]).

The version of this principle that we are going to deal with is the discrete one, i.e., for finite-dimensional vectors $x \in \mathbb{C}^N$. The discrete Fourier transform (DFT) of $x = (x_0, x_1, \dots, x_{N-1})$ is defined term-wise as

$$\hat{x}_j = \sum_{k=0}^{N-1} x_k e^{-2\pi ijk/N}, \quad j = 0, 1, \dots, N-1,$$

or it can also be defined as the linear map

$$\hat{x} = \Omega_N x, \tag{1}$$

where Ω_N is the so-called *N-dimensional Fourier matrix*, defined as $\Omega_N = (\omega_{j,k})$, $\omega_{j,k} = e^{-2\pi ijk/N}$, for $0 \leq j, k \leq N-1$. If we set $H(x) := \big|\{0 \leq n \leq N-1 : x_n \neq 0\}\big|$, then the discrete uncertainty principle states the following:

Theorem 1.1 (Donoho–Stark, [4]). *$H(x) \cdot H(\hat{x}) \geq N$.*

Once we know that we cannot concentrate arbitrarily the nonzero elements of a vector and its discrete Fourier transform (DFT) on very few components, we may be interested on the greatest number of null components that we can find on x and $\hat{x}$.

The goal of this paper is to determine the value of

$$L(N) := \max_{x \in \mathbb{C}^N \setminus \{0\}} \min\big\{Z(x), Z(\hat{x})\big\},$$

where $Z(x)$ is the number of null components of x or, equivalently, $Z(x) = N - H(x)$. The numbers $L(N)$ obviously depend on N but, furthermore, we will see that they strongly depend on the decomposition of N as a product of two numbers (see Theorem 2.4 below). For certain values of N, such as $N = n^2$ or $N = 2^n$, we will be able to give a formula for $L(N)$. However, finding a closed expression for all N is still an open problem. Despite of this fact, we are going to find an algorithm that will allow us to determine $L(N)$ for every N.

2. First approach: bounds for L(N)

We are going to determine upper and lower bounds for $L(N)$; in some very special cases those will coincide, yielding an equality. We start with a trivial upper bound being a direct consequence of Theorem 1.1.

Proposition 2.1. *For all N, we have $L(N) \leq N - \sqrt{N}$.*

Proof. Suppose that there exists $x \in \mathbb{C}^N$ such that $\min\{Z(x), Z(\hat{x})\} > N - \sqrt{N}$. Then,

$$H(x) = N - Z(x) < \sqrt{N}, \qquad H(\hat{x}) = N - Z(\hat{x}) < \sqrt{N},$$

and we conclude that $H(x)H(\hat{x}) < N$, which contradicts Theorem 1.1. □

Our next task is to start finding lower bounds for $L(N)$. The next result will provide a lower bound that will be sharp in general. However, there are special cases in which it will coincide with the bound given in Proposition 2.1.

Theorem 2.2. *(i) Let $N > 3$ be non-prime and suppose that $N = m \cdot n$, with $m \geq n$. Then,*

$$L(N) \geq \max\left\{K \,:\, m|K,\ K \leq N - \sqrt{N}\right\} = m(n-1). \tag{2}$$

(ii) Among all the possible decompositions $N = m \cdot n$, with $m \geq n$, the greatest lower bound of $L(N)$ that can be obtained from equation (2) is given when $m - n$ is minimized.

Proof. Consider the sequence

$$x_j = x_j^n = \begin{cases} 1, & \text{if } j = k \cdot n, \text{ for } k = 0, \dots, m-1, \\ 0, & \text{otherwise.} \end{cases}$$

We observe that $Z(x) = N - m = m(n-1)$. Now let us compute $\hat{x}$. Suppose that $k \in \{0, \dots, N-1\}$ is such that $k = m \cdot l$, i.e., m divides k. Then,

$$\hat{x}_k = \sum_{j=0}^{N-1} x_j e^{-2\pi ijk/N} = \sum_{s=0}^{m-1} x_{s\cdot n} e^{-2\pi isnk/(mn)} = \sum_{s=0}^{m-1} e^{-2\pi isml/m} = \sum_{s=0}^{m-1} 1 = m.$$

On the other hand, if m does not divide k, then $k = m \cdot q + d$, with $d \neq 0$, and

$$\begin{aligned} \hat{x}_k &= \sum_{j=0}^{N-1} x_j e^{-2\pi ijk/N} = \sum_{s=0}^{m-1} x_{s\cdot n} e^{-2\pi isn(mq+d)/(nm)} = \sum_{s=0}^{m-1} e^{-2\pi isnmq/(nm)} e^{-2\pi isnd/(nm)} \\ &= \sum_{s=0}^{m-1} e^{-2\pi isd/m} = \frac{1 - e^{-2\pi idm/m}}{1 - e^{-2\pi id/m}} = \frac{1-1}{1 - e^{-2\pi id/m}} = 0. \end{aligned}$$

In the last expression the denominator can never vanish, since $0 < d \leq m - 1$. We also note that $\hat{x} = \widehat{x^n} = m \cdot x^m$. Therefore, $Z(\hat{x}) = N - n = n(m-1)$. Since we are assuming $m \geq n$, it follows that $m(n-1) \leq n(m-1)$, so we have found a vector $x = x^n \in \mathbb{C}^N \backslash \{0\}$ such that $\min\{Z(x), Z(\hat{x})\} = m(n-1)$. This implies that $L(N) \geq m(n-1)$, proving (i).

To see (ii), suppose that $N = m \cdot n = m_0 \cdot n_0$, with $m \geq n$ and $m_0 \geq n_0$. Also, assume that $m - n \leq m_0 - n_0$. Under these conditions we have that $m_0 \geq m \geq n \geq n_0$. We want to prove that

$$m(n-1) \geq m_0(n_0 - 1).$$

But this happens if and only if $N - m \geq N - m_0$, which is obviously true, since $m_0 \geq m$. □

As an example to illustrate this result, we consider $N = 30 = 6 \cdot 5 = 2 \cdot 15$. Then:

$$Z(x^5) = 30 - 6 = 24, \quad Z(\widehat{x^5}) = Z(x^6) = 30 - 5 = 25,$$
$$Z(x^2) = 30 - 15 = 15, \quad Z(\widehat{x^2}) = Z(x^{15}) = 30 - 2 = 28.$$

We can observe that $\min\{Z(x^5), Z(x^6)\} = 24$ and $\min\{Z(x^2), Z(x^{15})\} = 15$. Hence, $L(30) \geq 24$. Moreover, by Proposition 2.1, it holds that $L(30) \leq \lfloor 30 - \sqrt{30} \rfloor = 24$, where $\lfloor \cdot \rfloor$ denotes the floor function. Therefore, the conclusion is that $L(30) = 24$. As we are going to see, there are certain N for which we can determine $L(N)$ explicitly.

Corollary 2.3. *(i) If $N = n^2$ for some n, then $L(N) = N - \sqrt{N} = n^2 - n = n(n-1)$;*

(ii) if $N = n(n-1)$ for some n, then $L(N) = \lfloor N - \sqrt{N} \rfloor = n(n-2)$.

Proof. By Theorem 2.2, we have that $L(N) \geq N - \sqrt{N} = n(n-1)$. On the other hand, Proposition 2.1 tells us that $L(N) \leq N - \sqrt{N}$. This proves (i).

To see (ii), again by Theorem 2.2, $L(N) \geq n(n-2)$. Moreover, we have that

$$\begin{aligned} \lfloor n(n-1) - \sqrt{n(n-1)} \rfloor = n(n-2) & \\ \Updownarrow & \\ n(n-2) \leq n(n-1) - \sqrt{n(n-1)} < n(n-2) + 1 & \\ \Updownarrow & \\ -n \leq -\sqrt{n(n-1)} < -n + 1 & \\ \Updownarrow & \\ n - 1 < \sqrt{n(n-1)} \leq n, & \end{aligned}$$

and the last expression is trivially true. Hence, using Proposition 2.1, $L(N) = \lfloor N - \sqrt{N} \rfloor = n(n-2)$. □

Before proceeding, we are going to improve the lower bound found in Theorem 2.2:

Theorem 2.4. *Let $N > 3$ be non-prime, and assume $N = m \cdot n$, with $m \geq n$. Define the set*

$$A_n = \{m_0 \cdot n \,|\, 1 \leq m_0 < m,\ m_0 < N - m_0 n \leq m\}.$$

Then, $L(N) \geq \max(A_n \cup \{m(n-1)\})$.

Proof. We already know from Theorem 2.2 that $L(N) \geq m(n-1)$. Thus, it only remains to prove that $L(N) \geq \max A_n$ whenever such set is not empty (otherwise we are done). If A_n is not empty, then fix $m_0 < m$ satisfying $m_0 < N - nm_0 \leq m$. We are going to obtain the stated lower bound by finding $x \in \mathbb{C}^N$ with $Z(x) \geq nm_0$ and such that $Z(\hat{x}) \geq nm_0$. To this end, we choose x to be of the following form

$$x_q = \begin{cases} a_j \in \mathbb{C}, & \text{if } q = j \cdot n, \text{for } j \in \{0, 1, \ldots, N - nm_0 - 1\}, \\ 0, & \text{otherwise.} \end{cases}$$

That is, the nonzero components of x can only be indexed by multiples of n. The first thing to note is that the condition $N - nm_0 \leq m$ is imposed in order to ensure that we do not exceed the number of multiples

of n that are strictly less than N (there are exactly m, since $N = m \cdot n$). Second, we also notice that, by construction, $Z(x) \geq N - (N - nm_0) = nm_0$. Now let $\omega = e^{2\pi i/N}$ and, for simplicity, let us denote $k := N - nm_0$ and $c := m_0 - 1$. Then we can build the following system of equations corresponding to certain components of $\hat{x}$, according to (1):

$$\begin{pmatrix} \hat{x}_0 \\ \hat{x}_1 \\ \hat{x}_2 \\ \vdots \\ \hat{x}_c \\ \hat{x}_m \\ \hat{x}_{m+1} \\ \vdots \\ \hat{x}_{m+c} \\ \vdots \\ \vdots \\ \hat{x}_{(n-1)m+c} \end{pmatrix} = \begin{pmatrix} 1 & 1 & \cdots & 1 \\ 1 & \omega^n & \cdots & \omega^{n(k-1)} \\ 1 & \omega^{2n} & \cdots & \omega^{2n(k-1)} \\ \vdots & \vdots & \ddots & \vdots \\ 1 & \omega^{nc} & \cdots & \omega^{cn(k-1)} \\ 1 & \omega^{nm} & \cdots & \omega^{nm(k-1)} \\ 1 & \omega^{n(m+1)} & \cdots & \omega^{n(m+1)(k-1)} \\ \vdots & \vdots & \ddots & \vdots \\ 1 & \omega^{n(m+c)} & \cdots & \omega^{n(m+c)(k-1)} \\ \vdots & \vdots & \ddots & \vdots \\ \vdots & \vdots & \ddots & \vdots \\ 1 & \omega^{n((n-1)m+c)} & \cdots & \omega^{n((n-1)m+c)(k-1)} \end{pmatrix} \begin{pmatrix} x_0 \\ x_n \\ \vdots \\ x_{n(k-1)} \end{pmatrix}. \tag{3}$$

Now we are going to prove that the homogeneous system associated to (3) has solutions besides the trivial one, or in other words, that there exist nontrivial choices of x such that $Z(\hat{x}) \geq nm_0$, so that $L(N) \geq nm_0$. This will happen if the matrix of the system (3), say Ω, has rank strictly less than the number of variables $N - nm_0$. We observe that Ω is formed by n identical blocks B_q of size $(c+1) \times k$ (or equivalently, $m_0 \times (N - nm_0)$), each one of them consisting on the rows indexed by the values $qm+b$, where $q \in \{0, 1, \ldots n-1\}$ is fixed, and $b \in \{0, 1, \ldots, c\}$ (see (4) below). Then, the rank of Ω is less than or equal to the rank of one B_q, reaching equality if the block has maximum rank, so that rank $\Omega \leq c + 1 = m_0$. Actually, this rank is maximum, since each one of the blocks consists on the rows of a Vandermonde matrix, which is known to have maximum rank (cf. [5, Prop. 3.19]); indeed, if we denote $\beta = e^{2\pi i/m} = \omega^n$, we have, by the exponential periodicity,

$$B_q = \begin{pmatrix} 1 & \omega^{nqm} & \cdots & \omega^{nqm(k-1)} \\ 1 & \omega^{n(qm+1)} & \cdots & \omega^{n(qm+1)(k-1)} \\ 1 & \omega^{n(qm+2)} & \cdots & \omega^{n(qm+2)(k-1)} \\ \vdots & \vdots & \ddots & \vdots \\ 1 & \omega^{n(qm+c)} & \cdots & \omega^{n(qm+c)(k-1)} \end{pmatrix} = \begin{pmatrix} 1 & 1 & \cdots & 1 \\ 1 & \beta & \cdots & \beta^{(k-1)} \\ 1 & \beta^2 & \cdots & \beta^{2(k-1)} \\ \vdots & \vdots & \ddots & \vdots \\ 1 & \beta^c & \cdots & \beta^{c(k-1)} \end{pmatrix}. \tag{4}$$

So, we conclude that rank $\Omega = c + 1 = m_0$. Thus, the system (3) is compatible and indeterminate whenever the number of columns of Ω (or the amount of indeterminates, which is the same number) is strictly greater than its rank, which we are actually assuming with the condition $m_0 < N - nm_0$. Since this procedure works for all $1 \leq m_0 < m$ satisfying $m_0 < N - nm_0 \leq m$, we conclude that $L(N) \geq \max A_n$. □

Example 2.5. In order to illustrate how Theorem 2.4 improves the result from Theorem 2.2, let us compute a lower bound for $L(39)$ using both results. Since the only nontrivial decomposition of 39 is $13 \cdot 3$, Theorem 2.2 tells us that $L(39) \geq 13 \cdot 2 = 26$. On the other hand, the corresponding set to A_n in Theorem 2.4 in this particular case is $A_3 = \{3m_0 : 1 \leq m_0 < m, m_0 < 39 - 3m_0 \leq 13\}$. It is easy to verify

that the only m_0 satisfying $1 \leq m_0 < m$ and $m_0 < 39 - 3m_0 \leq 13$ is $m_0 = 9$. Therefore, we conclude that $L(39) \geq 3 \cdot 9 = 27$, which improves the lower bound obtained previously.

Notice that in the theorems we have presented so far, we have excluded the case of N being a prime number. Our next result deals with the missing case; first of all we will need the following auxiliary theorem.

Theorem 2.6 (Chebotarev)**.** *Let $\Omega_N = \left(\omega_{j,k}\right)$, as defined in (1). If N is prime, then any minor of Ω_N is nonzero.*

There are several proofs for this theorem. We can find one similar to the original made by Chebotarev in [6], and Dieudonné also gave an independent proof for this theorem, cf. [3]. So, if N is prime, whatever submatrix we select from the N-dimensional Fourier matrix will have maximum rank. Using this fact we can compute the exact value of $L(N)$.

Proposition 2.7. *Let $N \geq 3$ be a prime number. Then, $L(N) = \lfloor N/2 \rfloor$.*

Proof. First, we prove that $L(N) \geq \lfloor N/2 \rfloor$. After that, using Chebotarev's theorem, it is easy to see that $L(N) < \lfloor N/2 \rfloor + 1$. Let us define $K = \lfloor N/2 \rfloor$, and let $x = (x_0, x_1, \dots, x_K, 0, \dots, 0) \in \mathbb{C}^N$, with $x_j \neq 0$ for $j = 0, \dots, K$. It is clear that $Z(x) = N - (K+1) = K$, since N is odd. Now let us consider the following homogeneous system of equations:

$$\begin{pmatrix} 0 \\ 0 \\ \vdots \\ 0 \end{pmatrix} = \begin{pmatrix} 1 & 1 & \cdots & 1 \\ 1 & e^{-2\pi i/N} & \cdots & e^{-2\pi iK/N} \\ \vdots & \vdots & \vdots & \vdots \\ 1 & e^{2\pi i(K-1)/N} & \dots & e^{2\pi iK(K-1)/N} \end{pmatrix} \begin{pmatrix} x_0 \\ x_1 \\ \vdots \\ x_K \end{pmatrix}.$$

It is clearly compatible and indeterminate, since the matrix of the system is a Vandermonde matrix (we note that it has rank K, while there are $K+1$ unknowns). Then, it has an infinite number of solutions, and moreover, we note that each row j of the latter system corresponds by definition to the j-th component of the vector $\hat{x}$. Therefore, there exist vectors $x \in \mathbb{C}^N$ with $Z(x) = N - K = K + 1$ and $Z(\hat{x}) = K$, and hence, $L(N) \geq K$.

Now suppose that there exists a vector $x \in \mathbb{C}^N$ with $Z(x) \geq K+1$, $Z(\hat{x}) \geq K+1$. Then, there exists a homogeneous system of equations which is again compatible and indeterminate,

$$\begin{pmatrix} 0 \\ 0 \\ \vdots \\ 0 \end{pmatrix} = \begin{pmatrix} e^{-2\pi i s_1 r_1/N} & e^{-2\pi i s_1 r_2/N} & \dots & e^{-2\pi i s_1 r_K/N} \\ e^{-2\pi i s_2 r_1/N} & e^{-2\pi i s_2 r_2/N} & \dots & e^{-2\pi i s_2 r_K/N} \\ \vdots & \vdots & \vdots & \vdots \\ e^{-2\pi i s_{K+1} r_1/N} & e^{-2\pi i s_{K+1} r_2/N} & \dots & e^{-2\pi i s_{K+1} r_K/N} \end{pmatrix} \begin{pmatrix} x_{r_1} \\ x_{r_2} \\ \vdots \\ x_{r_K} \end{pmatrix},$$

but this happens if and only if the matrix of the system has rank strictly less than K. By Theorem 2.6, we have that this rank is exactly K, which contradicts the existence of such x. This proves $L(N) < K + 1$, and therefore, we conclude that $L(N) = K$. □

3. The algorithm for finding L(N)

In this section we will carry the arguments used before one step further: we have been using homogeneous systems of equations with submatrices of a Fourier matrix in order to place zeros arbitrarily in a vector $\hat{x}$,

with the only restriction that the previous submatrix should have rank less than the number of unknowns, i.e., the components of x that are different from zero (see, for example, the system (3)). Roughly speaking, those matrices have many rows (one for each zero of $\hat{x}$), and only a few columns. Therefore, the matrices we are going to treat from now on will be of size $M \times N$, with $M > N$.

Definition 3.1. For a matrix $A \in \mathbb{C}^{M\times N}$, we say that A is *rank-deficient* if $\operatorname{rank} A < N$, or equivalently, if there exists $x \in \mathbb{C}\backslash\{0\}$ such that $Ax = 0$.

Let us fix $N \in \mathbb{N}$, and let $x \in \mathbb{C}^N$. Defining I to be the set of indexes where $\hat{x}$ is nonzero, J the set of indexes where x is nonzero, and $\overline{N} := \{0, 1, \dots, N-1\}$, we would like to find submatrices of the Fourier matrix Ω_N such that $\Omega_N\big(\overline{N}\backslash I, J\big)x_{|J} = 0$ and $\Omega_N\big(\overline{N}\backslash I, J\big)$ is rank-deficient, where $x_{|J}$ is the vector x restricted to the set of indexes J and $\Omega_N\big(\overline{N}\backslash I, J\big)$ is the restriction of Ω_N to the rows and columns indexed by $\overline{N}\backslash I$ and J, respectively. If we find one of those submatrices, then automatically $L(N) \geq \min\{N - |I|, N - |J|\}$. However, computing the ranks of every $m \times n$ submatrix of Ω_N is not an option, since the number of ranks to compute in this case has order $m!$. To solve this issue we introduce the following definition, that will lead to an easier reformulation of the problem.

Definition 3.2. For a matrix $A \in \mathbb{C}^{N\times N}$ and an integer $d \in \{1, \dots, N\}$, we define the Hamming number $H_A(d)$ as the minimal cardinality of all index sets I for which $A\big(\overline{N}\backslash I, J\big)$ is rank-deficient, for a suitable J with $|J| \leq d$.

In other words, $H_A(d) = k$ means that we can find $x \in \mathbb{C}^N\backslash\{0\}$ such that $A\big(\overline{N}\backslash I, J\big)x_{|J} = 0$ and $x_{|\overline{N}\backslash J} = 0$, where $|J| \leq d$ and $|I| = k$. Therefore, in terms of Fourier matrices, this would mean that we can find $x \in \mathbb{C}^N\backslash\{0\}$ such that $Z(x) \geq \big|\overline{N}\backslash J\big| \geq N - d$ and $Z(\hat{x}) \geq Z\big(\Omega_N\big(\overline{N}\backslash I, J\big)x_{|J}\big) = \big|\overline{N}\backslash I\big| = N - k$, which implies that

$$L(N) \geq \min\big\{N - d, N - k\big\} = \min\big\{N - d, N - H_{\Omega_N}(d)\big\} = N - \max\big\{d, H_{\Omega_N}(d)\big\}.$$

Remark 3.3. We have defined the Hamming numbers to depend on the complement of the set I instead of the set itself. Doing so, we stay close to the formulation of the uncertainty principle (cf. [1, p. 351]). Indeed, note that we can rewrite it as $d \cdot H_{\Omega_N}(d) \geq N$, for all $1 \leq d \leq N$.

In papers [1, 2], the numbers $H_{\Omega_N}(d)$ are investigated, concluding with an equality for any N and d; see Theorem 3.7 below. Those equalities will become crucial for us to compute the numbers $L(N)$.

Theorem 3.4. *Let $N \in \mathbb{N}$ and $1 \leq k < N$. Then, $L(N) = k$ if and only if*

$$N - H_{\Omega_N}(N - k) \geq k, \tag{5}$$

and

$$N - H_{\Omega_N}(N - (k+1)) \leq k. \tag{6}$$

We would like to make some comments about the meaning of equations (5) and (6) before proving the theorem. The first one means that we can find a vector $z \in \mathbb{C}^N$ with at least k zero components such that $\hat{x}$ also has more than k zero components. On the other hand, the second inequality tells us that we can find no vector $z \in \mathbb{C}^N$ such that both z and $\hat{z}$ have more than k zero components.

Proof of Theorem 3.4. First of all, we observe that, by definition,

$$N - H_{\Omega_N}(d) = \max\left\{Z(\hat{x}) : x \in \mathbb{C}^N, H(x) \leq d\right\} = \max\left\{Z(\hat{x}) : x \in \mathbb{C}^N, Z(x) \geq N - d\right\}. \tag{7}$$

($\Rightarrow$) Suppose that $L(N) = k$. Then, by (7), $N - H_{\Omega_N}(N-k) = \max\left\{Z(\hat{x}) : x \in \mathbb{C}^N, Z(x) \geq k\right\} \geq k$, where the last inequality is our assumption. This proves (5). In order to prove (6), if $x \in \mathbb{C}^N$ is such that $Z(x) \geq k+1$ then, necessarily, $Z(\hat{x}) \leq k$ (otherwise, $L(N) = k$ would be false). Joining this fact along with relation (7), we get $N - H_{\Omega_N}(N-(k+1)) = \max\left\{Z(\hat{x}) : x \in \mathbb{C}^N, Z(x) \geq k+1\right\} \leq k$, which proves the result.

($\Leftarrow$) We prove this implication by contradiction. Suppose that $L(N) \neq k$. Then, either (i) $L(N) < k$, or (ii) $L(N) > k$. In the case of (i), for any vector $x \in \mathbb{C}^N$ such that $Z(x) \geq k$, necessarily $Z(\hat{x}) < k$ (otherwise, $L(N) < k$ would be false). By (7), we deduce that

$$N - H_{\Omega_N}(N-k) = \max\left\{Z(\hat{x}) : x \in \mathbb{C}^N, Z(x) \geq k\right\} < k,$$

i.e., inequality (5) is false. Finally, if (ii) holds true, then there exists $N > \tilde{k} > k$ such that $L(N) = \tilde{k}$. Since $H_{\Omega_N}(d)$ is decreasing on the variable d, it follows that the expression $N - H_{\Omega_N}(N-M)$ is decreasing on the variable M. Now, since $k+1 \leq \tilde{k}$,

$$N - H_{\Omega_N}(N-(k+1)) \geq N - H_{\Omega_N}\left(N-\tilde{k}\right) = \max\left\{Z(\hat{x}) : x \in \mathbb{C}^N, Z(x) \geq \tilde{k}\right\} \geq \tilde{k} > k,$$

so that inequality (6) is false. □

The following results we state are due to S. Delvaux and M. Van Barel; the proofs can be found in the corresponding citations.

Theorem 3.5. *[2, Thm. 9] Let p^m be a power of a prime number p. Let $d \in \{1, 2, \ldots, p^m\}$ be such that $cp^t \leq d < (c+1)p^t$ for certain $c = 1, \ldots, p-1$ and $t = 0, \ldots, m-1$. Then, $H_{\Omega_{p^m}}(d) = (p-c+1)p^{m-t-1}$.*

Theorem 3.6. *[1, Cor. 23] For each divisor d of N, we have that $H_{\Omega_N}(d) = N/d$, i.e., equality in the uncertainty principle is reached.*

Theorem 3.7. *[1, Eq. (4)] Let $1 \leq t < N$. Then,*

$$H_{\Omega_N}(t) = \min\left\{(p-c+1)\frac{N}{pd} : pd \text{ divides } n,\ p \text{ prime},\ c \in \{1, \ldots, p\},\ cd \leq t\right\}.$$

In fact, if we assume $t < N$, then the numbers p, c, d can be chosen to be such that $c < p$, $cd \leq t < (c+1)d$, and p is the smallest prime divisor of n/d.

Finally, the algorithm to find $L(N)$ is done in the following steps (always for N non-prime).

Algorithm 3.8. Let k be the lower bound of $L(N)$ obtained through Theorem 2.4.

1. If k equals the higher bound $\left\lfloor N - \sqrt{N} \right\rfloor$ (given by Proposition 2.1), then we trivially have $L(N) = k$, so we do not need extra computations (as it occurs in Corollary 2.3).
2. If $k < \left\lfloor N - \sqrt{N} \right\rfloor$, then we check the veracity of (5) and (6), where the candidate to be $L(N)$ is k. If both inequalities are true, then by Theorem 3.4, we have $L(N) = k$. In order to compute the hamming numbers appearing in these inequalities, we make use of Theorems 3.5, 3.6 and 3.7.

3. If either (5) or (6) does not hold for k, then $L(N) > k$, so we proceed to check the veracity of (5) and (6) with $k+1$ in place of k.

4. We repeat the previous step until we find $\tilde{k}$ for which (5) and (6) hold. Then, $L(N) = \tilde{k}$.

Example 3.9. Consider $N = 2^{2n+1}$, with $n \in \mathbb{N}$. It can be checked that $2^{2n+1} - 2^{n+1} < \left\lfloor 2^{2n+1} - \sqrt{2^{2n+1}} \right\rfloor$ for all n, so we will have to use the algorithm to compute $L(N)$. So, let $k = 2^{n+1}(2^n - 1) = 2^{2n+1} - 2^{n+1}$, which is the lower bound of $L(N)$ found in Theorem 2.4 (which, in this case, is the same given by Theorem 2.2). Note that we are only considering odd powers of 2, since any even power is covered by the case $N = m^2$. Now we check that inequality (5) holds:

$$N - H_{\Omega_N}\left(N - \left(2^{2n+1} - 2^{n+1}\right)\right) = 2^{2n+1} - H_{\Omega_N}\left(2^{n+1}\right).$$

By Theorem 3.6, we have that $H_{\Omega_N}\left(2^{n+1}\right) = 2^n$. Therefore, $2^{2n+1} - 2^n \geq 2^{2n+1} - 2^{n+1} = k$. Now we prove inequality (6). Note that $N - \left(2^{2n+1} - 2^{n+1} + 1\right) = 2^{n+1} - 1$. Using Theorem 3.5 to compute $H_{F_N}\left(2^{n+1} - 1\right)$, we observe that $t = n$ and $c = 1$. Hence,

$$H_{\Omega_N}\left(2^{n+1} - 1\right) = (2 - 1 + 1) \cdot 2^{2n+1-n-1} = 2 \cdot 2^n = 2^{n+1}.$$

Finally, since $N - H_{\Omega_N}\left(2^{n+1} - 1\right) = N - 2^{n+1} = 2^{2n+1} - 2^{n+1} = k$, we obtain that $L\left(2^{2n+1}\right) = 2^{2n+1} - 2^{n+1}$.

In this example we did not need to go further than step 1 of the algorithm, since the lower bound we started with was the precise number we were looking for. Now, we have the following example that will force us to carry the algorithm one step further, and will as well illustrate how Theorem 2.4 improves Theorem 2.2.

Example 3.10. Let $N = 39 = 13 \cdot 3$. We have already seen in Example 2.5 that $L(39) \geq 9 \cdot 3 = 27$. Since $\left\lfloor 39 - \sqrt{39} \right\rfloor = 32 > 27$, we have to use the algorithm in order to find $L(39)$, i.e., we check whether the inequalities (5) and (6) hold with $k = 27$. As we are going to see, (6), which in this case reads as

$$39 - H_{\Omega_{39}}(11) \geq 27$$

does not hold. In order to prove it, we use Theorem 3.7. In this case, we note that the choice of p, c, and d must be the following:

$$p = 13, \qquad c = 3, \qquad d = 3. \tag{8}$$

Then, $H_{\Omega_{39}}(11) = (13 - 3 + 1) = 11$, so that $39 - H_{\Omega_{39}}(11) = 28 \not\leq 27$. Now we have to apply the second step of the algorithm: let $k = 28$. We can use the previous computations to see that

$$39 - H_{\Omega_{39}}(39 - 28) = 39 - H_{\Omega_{39}}(11) = 28,$$

so inequality (5) holds. Further, we have to compute $H_{\Omega_{39}}(39 - 29) = H_{\Omega_{39}}(10)$. Again, we apply Theorem 3.7 with the choice of p, c, and d as in (8), which is suitable, since $3 \cdot 3 \leq 10 < 3 \cdot 4$. Since the parameters involved in Theorem 3.7 did not change, we have that $H_{\Omega_{39}}(10) = H_{\Omega_{39}}(11) = 11$, and

$$39 - H_{\Omega_{39}}(10) = 39 - 11 = 28,$$

so that inequality (6) holds. Therefore, $L(39) = 28$.

	0	1	2	3	4	5	6	7	8	9
0		0	0	1	2	2	3	3	4	6
10	6	5	8	6	8	10	12	8	12	9
20	15	15	14	11	18	20	16	21	21	14
30	24	15	24	24	22	28	30	18	24	28
40	32	20	35	21	34	36	30	23	40	42
50	40	37	40	26	45	45	48	42	38	29
60	50	30	40	54	56	53	55	33	53	51
70	60	35	63	36	48	65	60	66	66	39
80	70	72	54	41	72	70	56	64	77	44
90	80	78	72	69	62	78	84	48	84	88

Table 1: Values of $L(N)$, where at each cell, N is given by the sum of the top value of its column and the leftmost value of its row. This table can be easily obtained by coding the algorithm 3.8 in any numerical programming language.

To conclude, we present the table 1 with the values of $L(N)$ for $N = 1, \dots, 99$. Observe that we trivially have $L(1) = L(2) = 0$. We observe that for the first 10 natural numbers, $L(N)$ seems to be monotone, but $L(11) = 5 < L(10)$. This is because, as we have already mentioned, $L(N)$ depends strongly on its possible decompositions as a product of two numbers. Since 11 is prime, it cannot have such a decomposition besides the trivial one, while $10 = 5 \cdot 2$ does and, therefore, it has a "better" behavior in terms of getting $L(N)$ as close as possible to $N - \sqrt{N}$. Finally, we remark that the lower bound for the numbers $L(N)$ given in Theorem 2.4 is often optimal (when N is not prime) for the first 99 natural numbers: indeed, this lower bound fails to be equal to $L(N)$ only for the following values of N:

$$27, 39, 44, 51, 65, 68, 75, 87, 95.$$

References

[1] S. Delvaux and M. Van Barel, "Rank-deficient submatrices of Kronecker products of Fourier matrices", *Linear Algebra Appl.* **426** (2007), 349–367.

[2] S. Delvaux and M. Van Barel, "Rank-deficient submatrices of Fourier matrices", *Linear Algebra Appl.* **429** (2008), 1587–1605.

[3] J. Dieudonné, "Une propriété des racines de l'unité", *Rev. Un. Mat. Argentina* **25** (1970/71), 1–3.

[4] D.L. Donoho and P.B. Stark, "Uncertainty principles and signal recovery", *SIAM J. Appl. Math.* **49** (1989), 906–931.

[5] P.A. Fuhrmann, "A polynomial approach to linear algebra", Universitext, Springer, New York, 2012.

[6] P. Stevenhagen and H.W. Lenstra Jr., "Chebotarëv and his density theorem", *Math. Intelligencer* **18** (1996), 26–37.

REPORTS@SCM

AN ELECTRONIC JOURNAL OF THE
SOCIETAT CATALANA DE MATEMÀTIQUES

Some improvements to the Erik+2 method for unbalanced partitions

Óscar Rivero Salgado
Universitat Politècnica de Catalunya
rversal@hotmail.com

*__Pol Torrent i Soler__
Universitat Politècnica de Catalunya
ptorrent@me.com

∗Corresponding author

Resum *(CAT)*

En aquest article proposem millores al mètode Erik+2, que s'empra per obtenir la distribució de les espècies a les fulles d'un arbre filogenètic, suggerint solucions als problemes provocats per la falta de dades experimentals quan es tracta amb un nombre elevat d'espècies. Presentem una nova tècnica per calcular les puntuacions assignades a cada distribució que es basa en aplicar successivament el mètode Erik+2 tenint en compte les files i columnes més plenes de la matriu de dades observades i compensant les puntuacions obtingudes per files i per columnes segons les dimensions de la matriu. Segons aquestes dimensions també proposem normalitzacions de les puntuacions obtingudes.

Abstract *(ENG)*

We aim to improve the Erik+2 method for obtaining the right distributions at the leaves of a phylogenetic tree, by addressing the problems that are due to the lack of enough experimental data when dealing with a high number of species. We introduce a new procedure based on successive applications of the Erik+2 method to take into account the most filled rows and columns of the observed data matrix and on balancing the scores obtained from both rows and columns. We also propose normalizations to compare the scores based on the dimensions of the data matrix.

Keywords: *Phylogenetics, Flattening matrix, Erik+2 Method.*
MSC (2010): *92D15, 60J20.*
Received: *September 17th, 2015.*
Accepted: *October 29th, 2015.*

Acknowledgement

The authors would like to thank Marta Casanellas for her vital guide and support while working on this project.

1. Introduction

Phylogenetics is a classical branch of science whose main aim is to determine evolutionary relationships between species. We typically have DNA sequences from genes of the different species we are studying and the classical approach would be to perform some kind of statistical analysis to determine the tree that fits the best to our data. However, in recent years, the use of tools from algebraic geometry have let to obtain a great progress in this field: we could talk about a new branch, phylogenetic algebraic geometry, that would study algebraic varieties representing statistical models of evolution, mixing that way mathematics, statistics, biology and computation. We take as our starting point the approach of Nicholas Eriksson and others, that uses the computation of the singular value decomposition of a matrix to study the distance to a particular algebraic variety. In recent years, Marta Casanellas and Jesús Fernández-Sánchez developed an improved version, Erik+2, that leaded to better results in the case of four species. Now, we try to extend their idea to the case of more species (here we work with the case of 12), having the necessity of doing some ponderations during the process concerning the size of the submatrices to obtain a result that, and even though our result is not optimal, provides some good approaches.

2. Background

The evolution of species is usually modeled in a phylogenetic tree $\mathcal{T}$. The leaves of the tree represent current species and the root the common ancestor. The aim of phylogenetics is to determine the phylogenetic tree of a set of species from the DNA sequences of current species. Due to its structure, we can deal with DNA sequences as if they were a sequence of nucleotides (A, C, G, T). For this reason, we need a statistical model for the substitutions of nucleotides to face our problem. We will work under the following assumptions:

(i) the trees are binary (which means that two branches come out of the root, if it exists, and that they are divided into another two branches in each node);

(ii) the processes in each branch do only depend on the common father node;

(iii) mutations of the DNA chain occur randomly;

(iv) each position of the DNA sequence evolves independently and under the same mutation probabilities; this means it is enough to model one position of the chain.

Following these assumptions we can think the nucleotide mutation process as a Markov process by assigning to each edge e a transition matrix

$$S_e = \begin{pmatrix} \mathrm{P}(\mathtt{A}\,|\,\mathtt{A},e) & \mathrm{P}(\mathtt{C}\,|\,\mathtt{A},e) & \mathrm{P}(\mathtt{G}\,|\,\mathtt{A},e) & \mathrm{P}(\mathtt{T}\,|\,\mathtt{A},e) \\ \mathrm{P}(\mathtt{A}\,|\,\mathtt{C},e) & \mathrm{P}(\mathtt{C}\,|\,\mathtt{C},e) & \mathrm{P}(\mathtt{G}\,|\,\mathtt{C},e) & \mathrm{P}(\mathtt{T}\,|\,\mathtt{C},e) \\ \mathrm{P}(\mathtt{A}\,|\,\mathtt{G},e) & \mathrm{P}(\mathtt{C}\,|\,\mathtt{G},e) & \mathrm{P}(\mathtt{G}\,|\,\mathtt{G},e) & \mathrm{P}(\mathtt{T}\,|\,\mathtt{G},e) \\ \mathrm{P}(\mathtt{A}\,|\,\mathtt{T},e) & \mathrm{P}(\mathtt{C}\,|\,\mathtt{T},e) & \mathrm{P}(\mathtt{G}\,|\,\mathtt{T},e) & \mathrm{P}(\mathtt{T}\,|\,\mathtt{T},e) \end{pmatrix},$$

where $\mathrm{P}(\mathtt{I}\,|\,\mathtt{J},e)$ is the probability of the nucleotide in the father node J becoming I after the edge e. These entries are unknown and along with the distribution in the root $\pi = (\pi_{\mathtt{A}}, \pi_{\mathtt{C}}, \pi_{\mathtt{G}}, \pi_{\mathtt{T}})$ are the parameters of

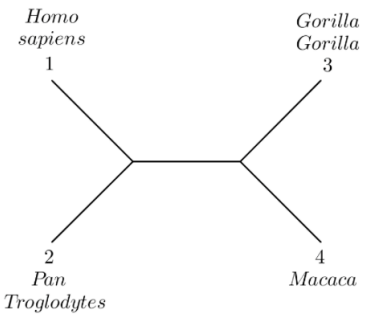


Figure 1: A example of an unrooted 4-leaf phylogenetic tree.

our model. By imposing conditions on the matrix S_e, one obtains different models. We deal with the most general Markov model; see [4].

We define now the random variables X_i as the state of the leaf i for $i \in \{1, \dots, n\}$ so that X_i takes values in $\{\mathtt{A}, \mathtt{C}, \mathtt{G}, \mathtt{T}\} = \mathcal{K}$, where n is the number of leaves of the tree. Now let $p_{x_1 x_2 \cdots x_n} = \mathrm{P}(X_1 = x_1, \dots, X_n = x_n)$ be the joint distribution at the leaves of the tree. Those probabilities can be calculated using only the entries of the transition matrices.

We are now ready to state the main definition and the main theorem we will need to understand Erik+2 method.

Definition 2.1. Let $A|B$ be a partition of the leaves (that is, if $L(\mathcal{T})$ is the set of leaves of the rooted tree $\mathcal{T}$ then $L(\mathcal{T}) = A \cup B$ and $A \cap B = \varnothing$), where we also assume that A and B are ordered sets. Then we define the *flattening matrix* $\mathrm{flat}_{A|B}$ of a joint distribution vector p associated to the partition $A|B$ as the $4^{|A|} \times 4^{|B|}$ matrix

$$\mathrm{flat}_{A|B}(p) = \begin{pmatrix} p_{\mathtt{AA}\cdots\mathtt{AA}} & p_{\mathtt{AA}\cdots\mathtt{AC}} & p_{\mathtt{AA}\cdots\mathtt{AG}} & \cdots & p_{\mathtt{AA}\cdots\mathtt{TT}} \\ p_{\mathtt{AC}\cdots\mathtt{AA}} & p_{\mathtt{AC}\cdots\mathtt{AC}} & p_{\mathtt{AC}\cdots\mathtt{AG}} & \cdots & p_{\mathtt{AC}\cdots\mathtt{TT}} \\ p_{\mathtt{AG}\cdots\mathtt{AA}} & p_{\mathtt{AG}\cdots\mathtt{AC}} & p_{\mathtt{AG}\cdots\mathtt{AG}} & \cdots & p_{\mathtt{AG}\cdots\mathtt{TT}} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ p_{\mathtt{TT}\cdots\mathtt{AA}} & p_{\mathtt{TT}\cdots\mathtt{AC}} & p_{\mathtt{TT}\cdots\mathtt{AG}} & \cdots & p_{\mathtt{TT}\cdots\mathtt{TT}} \end{pmatrix}.$$

That is, each column of the flattening matrix corresponds to a state of the leaves in B and each row to a state of the leaves in A. We will call such a partition an *edge split* if we can remove an edge such that all the leaves in A are in the same connected component and all the leaves in B are in the other one, and we will refer as the *size of the partition* to the pair $(|A|, |B|)$ (though we will usually write it as $|A| \times |B|$).

For instance, in the previous example 12|34 is an edge split partition, while 13|24 is not. Now we are ready to state the following result.

Theorem 2.2 ([1, 2])**.** *Let $A|B$ be a partition of the set of leaves of the tree $\mathcal{T}$ and let p be the joint distribution at the leaves of $\mathcal{T}$ for certain parameters. If that partition is an edge split, then* $\mathrm{rank}\ \mathrm{flat}_{A|B}(p) \leq 4$, *whereas if it is not an edge split partition and the parameters are general enough and $|A|, |B| > 1$, then* $\mathrm{rank}\ \mathrm{flat}_{A|B}(p) > 4$. □

For the case with $n = 4$ species at the leaves if the parameters are "general enough", one can show that the rank of the flattening matrix for partitions which are not an edge split is maximum (i.e., 16) but, since we will be dealing with cases with $n = 12$, we cannot assume this as true (cf., [1]).

2.1 The Erik+2 method

We start off with a set of ordered nucleotide sequences (one for each leaf in our tree, as they are the observed DNA chains of current species) which we will assume that have no gaps and have all the same length. We think of this set of nucleotide sequences as an *alignment*, that is, nucleotides at the same position of the different sequences are supposed to have evolved from the same nucleotide of the common ancestor.

From this experimental data we can calculate the relative frequencies $\tilde{p}_{x_1x_2\cdots x_n}$, which we will use as estimators for the true probabilities $p_{x_1x_2\cdots x_n}$ (in fact it can be shown that those are the maximum likelihood estimators for the true probabilities, see [3]). Given a partition of the leaves $A|B$, we can build the estimated flattening matrix $\widetilde{\text{flat}}_{A|B}$ just like we did above, but this time using the relative frequencies instead of the true probabilities. We aim to determine the right topology of the tree (i.e., to determine which species is at each leaf) by studying which partitions of the leaves are an edge split according to the experimental data and which are not.

By the theorem we stated above, if that matrix was exactly the flattening matrix we should be able to distinguish between edge splits and the other ones because edge splits would have exactly rank 4 or less and the other ones would not. This could be done easily by checking whether all 5×5 minors vanish or not, but since we only have the estimated matrices we have to develop a method to decide which one is "closer" to rank 4 matrices and to do so we will take the distance induced by the Frobenius norm.

Lemma 2.3. *If M is an $m \times n$ matrix and $\{\sigma_i\}$ are its singular values (ordered from big to small), the Frobenius distance of M to $\mathcal{V}$ (the set of rank 4 or lower matrices) in the Frobenius norm is*

$$d(M, \mathcal{V}) = \sum_{i=5}^{\min\{m,n\}} \sigma_i^2.$$

□

The ErikSVD method (see [1]) uses this fact to give a score to each flattening matrix. Indeed, it works as follows: given an alignment and a partition $A|B$, it computes the estimated flattening matrix and then it obtains the singular value decomposition of the matrix and computes the distance $d(\widetilde{\text{flat}}_{A|B}, \mathcal{V})$ which is the score assigned to the partition. Hence the partition which is estimated to be an edge split is the one having the lowest score.

The Erik+2 method (see [3]) slightly modifies the previous procedure by taking into account that the rank of the flattening matrix could be affected by the presence of long-branch attraction situations. The solution given by the Erik+2 method is to normalize first rows and then columns so each one sums up to 1. Scores obtained after normalizing by both rows and columns are taken into account to compute the final score.

One has to take into account that if we are dealing with a case with $n = 4$ then the flattening matrices for 2×2 partitions will have dimension 16×16. But in our case we used the algorithm to treat cases with 12 species, which leads to flattening matrices with dimensions $4^2 \times 4^{10}$ for 2×10 (actually the dimensions of the matrix we were dealing with computationally were about 16×60000 since we were only taking into account nonempty rows and columns) and $4^5 \times 4^7$ for 5×7 partitions. This explains why alignments with size 100000 work fine with 4 species but often are not enough to fill bigger flattening matrices so as to give a closer approach to the theoretical situation.

Since the number of singular values depends on the dimensions of the matrix and these dimensions depend on the cardinal of the subsets that form the partition, another interesting point is to ensure that

we can compare scores obtained from partitions whose subsets have different cardinals (and hence their flattening matrices have different dimensions).

3. Our proposed modifications

In this section we describe some of the most successful modifications out of the ones we tried. We start off with the observation that for the 2×10 sized partitions the flattening matrices have lots of columns which contain a single element due to the lack of data and that this fact can easily alter the rank of the matrix. Since the theoretical model stated that we should be dealing with matrices of rank approximately 4, we conjectured that there should be an important amount of data in a few rows and columns.

First of all we looked at how data should be distributed if the alignment was completely random (in this paper we will always assume that a random alignment is an alignment such that the distribution of its columns is uniform) to compare it to the actual flattening matrices. The following lemmas will allow us to make those estimations.

Lemma 3.1. *In a randomly generated alignment of length n, the expected number of nonempty columns of a flattening matrix of c columns is*

$$a_n = -c\left(\frac{c-1}{c}\right)^n + c.$$

Proof. We can easily build a recurrence by noticing that, when we have an alignment of length i, then a_{i+1} is simply the probability of the new datum being on an already occupied column times the current number of occupied columns plus the probability of it being on an empty column times the current number of occupied columns plus one. Noticing that the number of currently occupied columns is a_i (so a_{i+1} can only take the values a_i and $a_i + 1$ each one with its probability) we can write

$$a_{i+1} = \mathrm{P}(\text{new datum is in occupied column}) \cdot a_i + \mathrm{P}(\text{new datum not in occupied column}) \cdot (a_i + 1) =$$

$$= \frac{c - a_i}{c}(a_i + 1) + \frac{a_i}{c}a_i.$$

Simplifying, one obtains $ca_{i+1} - (c-1)a_i = c$. Then, we just need to resolve the recurrence. Putting it in an homogeneous form we obtain $ca_{i+2} - (2c-1)a_{i+1} + (c-1)a_i = 0$ so, the characteristic polynomial has roots 1 and $(c-1)/c$ and we get a solution of the form

$$a_n = \alpha\left(\frac{c-1}{c}\right)^n + \beta.$$

By setting initial conditions one obtains the result. □

Lemma 3.2. *In a randomly generated alignment of length n, the expected number of columns with a single matrix of a flattening matrix of c columns is*

$$b_n = a_n\left(\frac{a_n - 1}{a_n}\right)^{n-a_n},$$

where a_n is defined as in the previous lemma.

Proof. We have that a_n columns are occupied so we can focus in the case where each one of them has a single entry and that we have $n - a_n$ data left to distribute. Since having a single entry is now equivalent to not getting any of those remaining data, we can apply the previous lemma with $n = n - a_n$ and $c = a_n$, so the number of occupied columns is now

$$X = -a_n \left(\frac{a_n - 1}{a_n} \right)^{n-a_n} + a_n,$$

and the number of not occupied (and hence with a single entry) columns $a_n - X$ is the one stated above. □

Assuming alignments of size 10^5 as the ones we had, we obtained, for instance, that for the 2×10 partition there would be on average 95380 nonempty columns, where 90869 of them have only one entry. The actual matrices have about 60000 nonempty columns, 40000 of them having a single entry, hence dispersion is lower than in the random model but not much lower. For 5×7 partitions, we observed that random matrices have entries in almost all columns (we computed an average of 16347 nonempty columns out of $4^7 = 16384$ possible, and we expected that just 98 columns had one entry). In this case we observed that, on average, we had 9000 nonempty columns so dispersion was also lower than in the random case. This data is obtained from the following lemmas and completed in table 1.

Partition	Number of columns	Expected nonempty	Expected single entry
2 vs 10	$4^{10} = 1048576$	95380	90869
3 vs 9	$4^9 = 262144$	83137	67874
4 vs 8	$4^8 = 65536$	51287	19837
5 vs 7	$4^7 = 16384$	16347	98

Table 1: Expected number of nonempty columns and columns with a single entry assuming alignments of length $n = 10^5$.

We can also use recurrences to estimate the number of entries in the most populated rows. To normalize, we will need to look at the number of entries at the most populated half according to our proposed method that will be explained below (since the most populated half has a greater weight in the final score). Taking into account that half, we will look at how many entries we have in the most populated sub-half, and so on (this works since the number of rows is always a power of 2). We will treat the problem of determining the number of entries in the most populated half as the problem of looking for the expected cardinality of the most populated half (tails or heads) when we toss n times a coin (this is equivalent to our problem since the data distribution is uniform). Let c_n be that number. Clearly $c_1 = 1$ and $c_n = c_{n-1} + 1/2$ if n is even (since the new coin will result in the result which is currently most frequent with probability $1/2$), and if n is odd we see that c_n is $1/2$ plus the previous number of coins in the most populated half, as before, but we have to take into account the existence of draws by adding an extra term that takes care of this probability, resulting in

$$c_n = c_{n-1} + \frac{1}{2} + \frac{1}{2} \frac{\binom{n-1}{(n-1)/2}}{2^{n-1}},$$

for odd n. By Stirling's approximation we get

$$c_n \approx c_{n-1} + \frac{1}{2} + \frac{1}{\sqrt{2\pi(n-1)}}$$

so, for n big enough, by adding up both results we get

$$c_n \approx \frac{n+1}{2} + \frac{1}{2\sqrt{\pi}} \left(\sum_{i=1}^{(n-1)/2} \frac{1}{\sqrt{i}} \right).$$

We also looked with detail to some cases and found out the following patterns for flattening matrices coming from an edge split. They usually (respect to flattening matrices not coming from a partition which is an edge split) have a lower amount of nonempty rows and columns, have less rows and columns with only 1 entry, and have more entries in the most populated rows.

This led us to think that it would be convenient to reorder rows and columns according to their number of entries, in order to have the most populated (and hence most significant) rows and columns in the first place. Then we consider the sub-matrices obtained by taking the m rows and the first k columns, where m is the number of rows of the matrix and k is a parameter of the method (we used $k = 1000$). We apply the Erik+2 method to those sub-matrices and then we extend the sub-matrix with k more columns, compute the score again and so on, and finally we add up all the scores. In order to compare the scores between partitions of different size, it is convenient to divide the score by the number of total SVDs done. However, when dealing with partitions of the same size, this does not help since it would decrease the score for wrong matrices which usually have more nonempty columns.

We also considered to do an analogous procedure for both rows and columns, i.e., considering sub-matrices of size $k_1 \times k_2$, and then increase both k_1 and k_2 but, since we are usually dealing with matrices which have $m \ll n$, we did not see a significant improvement of the results. Due to this fact we also need to multiply by m the score obtained by normalizing the columns, and by n the score obtained by normalizing the rows, in order to have the same order of magnitude.

Since we are adding up scores of matrices with different dimensions, the next step is to give estimates for the value of those scores so we can normalize. If we have an $m \times n$ matrix and we normalize the rows so as the elements of each row sum up to 1, we get

$$\sqrt{\frac{\sum\sum a_{ij}^2}{mn}} \geq \frac{\sum\sum a_{ij}}{mn} = \frac{m}{mn} = \frac{1}{n}$$

since each row adds up to one (we assume that in each row there is at least one entry since the method does only take into account nonempty rows). We obtain

$$\sqrt{\sum\sum a_{ij}^2} \geq \sqrt{\frac{m}{n}} \Longrightarrow n\sqrt{\sum\sum a_{ij}^2} \geq \sqrt{mn}$$

and, by symmetry, we obtain the same result when we normalize columns and multiply by m. To get an upper bound notice that, since $(\sum b_i)^2 = 1$ (where the b_i are elements of a row or a column which has been normalized), we obtain $\sum b_i^2 \leq 1$. By proceeding this way, we get $\sqrt{\sum\sum a_{ij}^2} \leq \sqrt{m}$ and, multiplying by n and arguing analogously for rows and columns, we finally get that

$$n \cdot \text{rownorm} + m \cdot \text{colnorm} \in [2\sqrt{mn}, (\sqrt{m} + \sqrt{n})\sqrt{mn}].$$

The experimental results tell us that neither of those bounds is sharp enough.

We try another approximation: assuming $\hat{e}_i = e/m$ where e is the total number of entries of the matrix and $\hat{e}_i$ is an estimator for the number of entries in a row then, for the ML estimator properties (we can view that estimator as an estimator for a binomial distribution), $\hat{1/e_i} = m/e$. This way we see that there is approximately one datum for each one of the e_i entries and we have

$$\sqrt{\sum\sum a_{ij}^2} = \sqrt{\sum \frac{1}{e_i^2} e_i} = \sqrt{\sum \frac{1}{e_i}} \sim \sqrt{\sum \frac{m}{e}} = \sqrt{\frac{m^2}{e}} = \frac{m}{\sqrt{e}}.$$

Hence, after multiplying by n, we get a value of $mn/\sqrt{e}$ (the value obtained for the other normalization is the same, by symmetry). These are in good agreement with this value so it results a nice normalization. This is the expected value for a random matrix which has only ones at $e \ll mn$, but not a bound (as the matrix gets further away from the random model, the value gets also more different from this one).

This normalization is interesting because it makes the sub-scores of the sub-matrices have similar values as we increase the number of rows of the sub-matrices instead of having an increasing sequence as we prove in the following lemma.

Lemma 3.3. *Consider the sequence of values (x_n) corresponding to the Frobenius norm of the matrix obtained by taking into account the first n columns, and then normalizing by rows and columns. For n sufficiently big, (x_n) becomes increasing.*

Proof. If we normalize by columns, the result follows trivially since the other columns remain unchanged and we add a new positive term to the computation of the norm. If we normalize by rows, it suffices to show that, if the matrix has s data in the row and the new column adds d data, then

$$\sqrt{\frac{\sum a_i^2}{s^2}} \leq \sqrt{\frac{\sum a_i^2 + d}{(s+d)^2}}$$

which is equivalent to $\sum a_i^2 \leq ds^2/(2sd + d^2)$. By using the inequality between the arithmetic mean and the quadratic mean, we get

$$\sum a_i^2 \leq \frac{\sum a_i}{n} = \frac{s}{n}$$

so we need

$$\frac{s}{n} \leq \frac{ds^2}{2ds + s^2} \iff s \geq \frac{d}{n-2},$$

which is true for n big enough (and in general for our matrices n will almost always be big enough). □

After this discussion, since dispersion is high, we assume that our data will be closer to the random model and hence the score we assign to a $m \times n$ sub-matrix is the following (the overall score is obtained after adding up all the scores given to sub-matrices):

$$\text{score}(M) = \frac{n \cdot \text{rowscore} + m \cdot \text{colscore}}{mn/\sqrt{e}}. \tag{1}$$

After computing the overall score, we can divide by either the number of SVDs done (so as to compare our score to scores coming from partitions with different size) or by the expected number of SVDs for that size of the partition, in order to keep a penalty to flattening matrices which require a higher number of SVDs, because they have a higher number of columns.

4. Performance tests for several methods

To test the performance of our method and to compare it to the original Erik+2 method, we considered a set of 100 data files corresponding to trees with 12 leaves with the same topology but with random branch lengths. For every data set, we obtained the scores for 9 partitions, 3 of size 2×10, 3 of size 3×9 and 3 of size 5×7, where one partition of each size was an edge split and the rest were not.

The following tables 2 and 3 contain the information of the performance (success[1] and scores assigned to both edge splits and other partitions) for the methods corresponding to the following scores: sc_1 is the number of nonempty rows of the flattening matrix, sc_2 is the number of nonempty columns of the flattening matrices, sc_3 is the original Erik+2 score, sc_4 the Erik+2 score using the $mn/\sqrt{e}$ normalization, sc_5 the score given by the variant of our method[2] without dividing by the number of SVDs computed, sc_6 the score given by our method taking the arithmetic mean of the scores obtained for each sub-matrix, and sc_7 the score of our method taking a pondered mean of the sub-scores.

Partition	sc_1	sc_2	sc_3	sc_4	sc_5	sc_6	sc_7
2 vs 10	100	64	33	40	70	65	67
3 vs 9	100	50	39	31	42	35	36
5 vs 7	97	76	58	21	47	24	26

Table 2: Percentage of success of the different methods (where each method is represented by its score).

Partition		sc_1	sc_2	sc_3	sc_4	sc_5	sc_6	sc_7
2 vs 10	(ES)	16	57418	3209	181	9972	176	177
2 vs 10	(NES)	16	59347	3206	181	10502	179	183
3 vs 9	(ES)	62	38453	13926	296	10929	291	291
3 vs 9	(NES)	64	39422	14396	293	11134	288	288
5 vs 7	(ES)	890	8745	75489	589	4530	620	677
5 vs 7	(NES)	954	9816	87601	560	4814	584	638

Table 3: Average of the score given to edge splits (rows labelled with (ES)) and to partitions which are not edge splits (labelled with (NES)) by each method.

5. Conclusions

We can see that our method (score 5) works significantly better than the original Erik+2 method for 2 vs 10 partitions, since it recognizes the edge split of the three partitions 70 out of 100 times, and the original method worked fine only 33% of the time. This could be explained by the fact that the Erik+2 method

[1] We consider a test successful if the score the method assigned to the edge split is lower or equal than the score it assigned to two partitions which were not an edge split of that size. Notice that with our data set we could make 100 test for each size.

[2] We will refer as "our method" to the method that implements the modifications proposed above: reorder rows and columns, consider sub-matrices formed by the first ik columns in the i-th iteration, compute the score for each sub-matrix using (1) and add up all the scores, resulting in score 5. Scores 6 and 7 slightly modify this method by taking the mean of the sub-scores.

computes a single SVD where only 16 singular values are obtained (notice that the Erik+2 method is more accurate as the partition is more balanced), and that the dispersion present in flattening matrices coming from unbalanced partitions fits nicely with the assumptions we made to obtain the $mn/\sqrt{e}$ normalization.

For the 3 vs 9 case, our method turns out to be slightly better but not significantly; neither the original method nor ours provided a satisfactory result, so we think new ideas should be introduced to deal with this problem. In the 5 vs 7 case the most effective score turns out to be the original Erik+2 method, but we should notice that the percentage of success in taking the score as simply the number of columns is really high and the averaged difference of columns between edge splits and the other partitions is percentage-wise the most significant. The score sc_1 is not reliable for unbalanced partitions as the number of rows of the flattening matrices of unbalanced partitions is small and almost never there is an empty row (we can see in table 3 that, for 2 vs 10 and 3 vs 9 partitions, that averages for both edge splits and the rest of partitions are really close to 16 and 64, the number of rows of the flattening matrices). Nevertheless, when we consider balanced partitions (e.g., 5 vs 7) and hence the number of rows of the flattening matrices is higher, we can take it into account since for those cases the difference of scores between edge splits and the other partitions is noticeable and it has a huge percentage of success.

We can also see that, while we have reduced the relative difference between scores when averaging (although by doing this we are decreasing the percentage of success), those scores are not yet comparable. A noticeable fact is that for the method that works better (without averaging) scores obtained for the first two sizes are really close, but for the 5 vs 7 it reduces to less than one half (this is due to the fact that we make much less SVDs, as one can see looking at the averaged score), while for the original Erik+2 the score shows a steady increasing trend when the partition gets more balanced. We should also note that we worked estimations for the norm of the matrix and not for the distance to rank 4 flattening itself (they differ in the square of the first 4 singular values) and this could affect the success of our method in some cases.

References

[1] N. Eriksson, "Tree construction using singular value decomposition", in *Algebraic Statistics for computational biology*, 347–358, Cambridge University Press, New York, 2005.

[2] M. Casanellas and J. Fernández-Sánchez, "Relevant phylogenetic invariants of equivariant models", *J. de mathématiques Pures et Appliquées* **96** (2010), 207–229.

[3] J. Fernández-Sánchez, M. Casanellas, "Invariant versus quartet inference when evolution is heterogeneous across sites and lineages", *Systematic Biology* (2015), to appear.

[4] E. Allman, J.A. Rhodes, "Phyogenetic ideals and varieties for the general Markov model", *Adv. in Appl. Math.* **40** (2008), 127–148.

REPORTS@SCM

AN ELECTRONIC JOURNAL OF THE
SOCIETAT CATALANA DE MATEMÀTIQUES

A geometric application of Runge's theorem

*Ildefonso Castro-Infantes
Universidad de Granada.
icastroinfantes@ugr.es

∗Corresponding author

Resum *(CAT)*

En aquest article donem una demostració simple de l'existència d'aplicacions harmòniques de qualsevol superfície de Riemann cap al pla complex $\mathbb{C} \equiv \mathbb{R}^2$. La nostra eina principal és la teoria d'aproximació per funcions holomorfes en superfícies de Riemann.

Abstract *(ENG)*

In this article we give a simple proof of the existence of proper harmonic maps from any open Riemann surface into the complex plane $\mathbb{C} \equiv \mathbb{R}^2$. Our main tool will be the Approximation Theory by holomorphic functions on Riemann surfaces.

Societat Catalana de Matemàtiques

Institut d'Estudis Catalans

Keywords: *Harmonic map, proper map, Riemann surface, Runge theorem.*
MSC (2010): *Primary 30F15.*
Received: *March 3th, 2015.*
Accepted: *December 28th, 2015.*

Acknowledgement

The author is partially supported by a "Beca Iniciación a la Investigación de la Universidad de Granada".

1. Introduction

The Runge and Mergelyan Theorems are the main results of Approximation Theory in one complex variable. The former, proved in 1885, asserts that every holomorphic function defined on an open neighbourhood of a compact set K of $\mathbb{C}$ can be uniformly approximated on K by entire functions, provided that the complement of K in $\mathbb{C}$ has no relatively compact connected components, see [11]. In the same line, Mergelyan [10] proved in 1951 that a continuous function $K \to \mathbb{C}$, which is holomorphic on K°, can be uniformly approximated on K by holomorphic functions on an open neighbourhood of K. Later, Bishop [5] extended these results to the context of open Riemann surfaces.

Theorem 1.1 (Runge–Mergelyan Theorem)**.** *Let $\mathcal{R}$ be an open Riemann surface and let $K \subset \mathcal{R}$ be a compact subset such that $\mathcal{R} \setminus K$ has no relatively compact connected components in $\mathcal{R}$. For any continuous function $f\colon K \to \mathbb{C}$ which is holomorphic on K° and any $\epsilon > 0$, there exists a holomorphic function $F\colon \mathcal{R} \to \mathbb{C}$ such that $\|F(p) - f(p)\| < \epsilon$ for all $p \in K$.*

The Runge and Mergelyan Theorems are useful in many different areas, e.g., complex analysis or surface theory. In particular, these tools have been exploited in the construction of minimal surfaces in the three-dimensional euclidean space $\mathbb{R}^3$. Recall that this class of surfaces is closely related to complex analysis through the Enneper–Weierstrass representation.

A fundamental problem in minimal surface theory is to understand how the conformal type (i.e., the type of the underlying Riemann surface) influences the global geometry of minimal surfaces. From an analytical point of view, an open Riemann surface is hyperbolic if and only if it admits negative non-constant subharmonic functions, and it is parabolic otherwise. This classification can also be explained in terms of Brownian motion of a particle over the surface; parabolicity is equivalent to the property that the Brownian motion visits any open set at arbitrarily large moments of time with probability 1. See the book of Grigor'yan [8] for more details.

Up to biholomorphisms, the only simply connected open Riemann surfaces are the unit disk $\mathbb{D}$ (of hyperbolic type) and the complex plane $\mathbb{C}$ (of parabolic type). Heinz [9] proved in 1952 that there do not exist harmonic diffeomorphisms between $\mathbb{D}$ and $\mathbb{C}$ with the euclidean metrics, extending the classical theorems by Riemann and Liouville. As a generalization of this result, Schoen–Yau [12, p. 18] conjectured in 1985 the nonexistence of proper harmonic maps $\mathbb{D} \to \mathbb{R}^2$. Schoen and Yau related this conjecture with the problem of existence of minimal surfaces in $\mathbb{R}^3$ having hyperbolic conformal type and proper projection into $\mathbb{R}^2$; recall that the coordinate functions of a conformal minimal immersion from a Riemann surface into $\mathbb{R}^3$ are harmonic. In 2001, Forstnerič–Globevnik [7, Theorem 1.4] disproved Schoen–Yau's conjecture. In 2011, Alarcón–Gálvez [1] extended this result to surfaces with finite topology. Although the Schoen–Yau conjecture was solved, its version for minimal surfaces was still open. This problem was settled in the most general and optimum form by Alarcón–López [2, 3], who proved the following result.

Theorem 1.2 (Alarcón–López [2, 3])**.** *Every open Riemann surface $\mathcal{R}$ admits a conformal minimal immersion $X = (X_1, X_2, X_3)\colon \mathcal{R} \to \mathbb{R}^3$, such that $(X_1, X_2)\colon \mathcal{R} \to \mathbb{R}^2$ is a proper map.*

The proof of Alarcón–López is based on a Runge–Mergelyan type theorem for minimal surfaces [2], a powerful tool in the construction of minimal surfaces which has found many applications. Since the coordinate functions of a conformal minimal immersion are harmonic, the full answer to the Schoen–Yau conjecture is immediately derived from Theorem 1.2:

Theorem 1.3. *Every open Riemann surface $\mathcal{R}$ admits a proper harmonic map $\mathcal{R} \to \mathbb{R}^2$.*

An alternative proof of this result was given later by Andrist–Wold [4].

The goal of this article is to give a simple proof of Theorem 1.3. Our proof combines the ideas of Alarcón–López with the classical Runge–Mergleyan Theorem 1.1. Roughly speaking, given an open Riemann surface $\mathcal{R}$, we will construct an expansive sequence of compact sets $\{M_n\}_{n\in\mathbb{N}}$ on $\mathcal{R}$ and harmonic maps $\{h_n\colon M_n \to \mathbb{R}^2\}_{n\in\mathbb{N}}$ satisfying $h_{n+1} \approx h_n$ on M_n, $n \geq 1$, and $\{h_n(\partial M_n)\}_{n\in\mathbb{N}} \to \infty$. We will ensure that the limit map $h := \lim_{n\to\infty} h_n$ exists and is proper and harmonic.

2. Background

We denote by $\|\cdot\|$ the euclidean norm in $\mathbb{R}^n$. Given a compact topological space K and a continuous function $f\colon K \to \mathbb{R}^n$, we denote by $\|f\|_K = \max_{a\in K} \|f(a)\|$ the maximum norm of f on K. Given $\zeta \in \mathbb{C}$ we denote by $\Re(\zeta)$ and $\Im(\zeta)$ its real and imaginary parts, respectively.

Let S be a topological surface. We denote by ∂S the topological boundary of S; recall that ∂S is a 1-dimensional topological manifold. Hence, we say that the surface S is *open* if it is not compact and $\partial S = \emptyset$. Given a subset $A \subset S$, we denote by $\overline{A}$ and A° the closure and the interior of A in S, respectively. Given subsets $A, B \subset S$, we write $A \Subset B$ when $\overline{A} \subset B^\circ$. A subset $A \subset S \setminus \partial S$ is called a *bordered region* in S if A is a compact topological surface with the topology induced by S and $\partial A \neq \emptyset$; in particular, ∂A consists of a finite family of pairwise disjoint Jordan curves. If S is a differentiable surface, a bordered region A on S is called *differentiable* if ∂A is differentiable.

Let X and Y be two topological spaces. A map $f\colon X \to Y$ is called *proper* if $f^{-1}(C)$ is a compact subset of X for any compact subset $C \subset Y$. If f is continuous and Y is Hausdorff, then f is proper if and only if for any divergent sequence $\{x_n\}_{n\in\mathbb{N}}$ in X (i.e., leaving any compact set), the sequence $\{f(x_n)\}_{n\in\mathbb{N}}$ is divergent in Y.

Recall that a *Riemann surface* (without boundary) $\mathcal{R}$ is a 1-dimensional complex manifold and every open set of a Riemann surface is canonically a Riemann surface by restriction of charts.

From now on, $\mathcal{R}$ will denote an open Riemann surface.

A function $\phi\colon \mathcal{R} \to \mathbb{C}$ is called *holomorphic* if the composition with any chart of $\mathcal{R}$ is a holomorphic function; equivalently, if for any point $p \in \mathcal{R}$ there exists a chart around $p \in \mathcal{R}$ such that the composition with ϕ is again a holomorphic function.

Definition 2.1. Let $\mathcal{R}$ be an open Riemann surface. A function $h\colon \mathcal{R} \to \mathbb{R}$ is called *harmonic* if its composition with any chart is harmonic; equivalently, if for any point $p \in \mathcal{R}$ there exists a chart around $p \in \mathcal{R}$ such that the composition is harmonic. A map $(h_1, \dots, h_n)\colon \mathcal{R} \to \mathbb{R}^n$, $n \in \mathbb{N}$, is called *harmonic* if $h_j\colon \mathcal{R} \to \mathbb{R}$ is harmonic for all $j = 1, \dots, n$.

Recall that, since the changes of charts in a Riemann surface are biholomorphisms and the composition of a harmonic function with a biholomorphism is again harmonic, the notion of harmonicity is well-defined on a Riemann surface. Furthermore, a function $h\colon \mathcal{R} \to \mathbb{R}$ is harmonic if and only if for any simply connected open set $D \subset \mathcal{R}$ there exists a holomorphic function $\phi\colon D \to \mathbb{C}$ such that $h|_D = \Re(\phi)$.

A compact subset $K \subset \mathcal{R}$ is called *Runge* if $\mathcal{R} \setminus K$ has no relatively compact connected components in $\mathcal{R}$.

Theorem 2.2 (Runge–Mergelyan). *Let $\mathcal{R}$ be an open Riemann surface and let $K \subset \mathcal{R}$ be a compact Runge subset. Given a continuous function $f\colon K \to \mathbb{C}$ which is holomorphic on K° and given $\epsilon > 0$, there exists a holomorphic function $F\colon \mathcal{R} \to \mathbb{C}$ such that $\|F - f\|_K < \epsilon$.*

3. Proof of Theorem 1.3

Theorem 1.3 is a consequence of the following more general result, concerning the existence of holomorphic functions into $\mathbb{C}^2$.

Theorem 3.1. *Let $\mathcal{R}$ be an open Riemann surface. Then there exists a holomorphic function $H = (H_1, H_2)\colon \mathcal{R} \to \mathbb{C}^2$ such that $\Re(H) = (\Re(H_1), \Re(H_2))\colon \mathcal{R} \to \mathbb{R}^2$ is proper.*

This is the main result of the paper; since $\Re(H)$ is harmonic, Theorem 1.3 follows directly. Before going into the proof of Theorem 3.1 we need some preparations.

Lemma 3.2. *For any open Riemann surface $\mathcal{R}$ there exists a sequence of bordered regions $\{M_n\}_{n\in\mathbb{N}}$ in $\mathcal{R}$ such that*

(i) M_n is a differentiable bordered region and it is Runge and connected for all $n \in \mathbb{N}$;

(ii) $\{M_n\}_{n\in\mathbb{N}}$ is an exhaustive sequence, that is, $M_n \Subset M_{n+1}$ $\forall n \in \mathbb{N}$ and $\bigcup_{n\in\mathbb{N}} M_n = \mathcal{R}$;

(iii) $\chi(M_{n+1} \setminus M_n^\circ) \in \{-1, 0\}$ $\forall n \in \mathbb{N}$, where $\chi(M)$ denotes the Euler characteristic of the region M.

Proof. Let $\{U_n\}_{n\in\mathbb{N}}$ be a exhaustive sequence of $\mathcal{R}$ by connected (differentiable) bordered regions; such sequences trivially exists. Let us first show that we can find a new exhaustion $\{V_n\}_{n\in\mathbb{N}}$ of $\mathcal{R}$ by connected Runge regions. Indeed, if U_1 is Runge we define $V_1 = U_1$; otherwise, we define V_1 as the union of U_1 with all the bounded connected components of $\mathcal{R} \setminus U_1$. Therefore V_1 is Runge and connected. Inductively, for any $n \geq 2$ let V_n be the union of V_{n-1}, U_n and all the bounded connected components of $\mathcal{R} \setminus U_n$. This implies that V_n is Runge and connected. As $U_n \subset V_n$ and $V_n \Subset V_{n+1}$ $\forall n \in \mathbb{N}$, the sequence $\{V_n\}_{n\in\mathbb{N}}$ is an exhaustion of $\mathcal{R}$ by Runge connected bordered regions.

The properties (i) and (ii) of the lemma are formally satisfied by $\{V_n\}_{n\in\mathbb{N}}$. The second step of the proof consists of adding convenient terms to the exhaustion $\{V_n\}_{n\in\mathbb{N}}$ in order to guarantee property (iii).

We consider now two consecutive regions V_m and V_{m+1}, $m \in \mathbb{N}$. Set $A := V_m$, $B := V_{m+1}$ and recall that $A \Subset B$. Let $n := -\chi(B \setminus A^\circ)$.

Claim 3.3. There exist compact sets $N_1, \dots, N_{n-1}$ in $\mathcal{R}$ such that

- $A \Subset N_1 \Subset N_2 \Subset \cdots \Subset N_{n-1} \Subset B$;
- $\chi(N_1 \setminus A^\circ)$, $\chi(N_i \setminus N_{i-1}^\circ)$ and $\chi(B \setminus N_{n-1}^\circ)$ take values in $\{-1, 0\}$, for $i = 2, \dots, n-1$.

Proof. We proceed by induction on n. If $n \in \{-1, 0\}$ there is nothing to prove. Suppose the claim is true when $-\chi(B \setminus A^\circ) \leq n$, $n \in \mathbb{N}$, and let us prove it in the case $-\chi(B \setminus A^\circ) = n + 1$. Recall that A and B are connected Runge regions and $A \Subset B$. Hence, (I) A and $B \setminus A^\circ$ have at least one common boundary component γ_1, thus satisfying $\gamma_1 \subseteq \partial A \cap \partial(B \setminus A^\circ)$; and (II) $B \setminus A^\circ$ has at least one boundary component γ_2 which does not intersect A (in particular, $\gamma_1 \neq \gamma_2$ and so, $\partial(B \setminus A^\circ)$ is not connected).

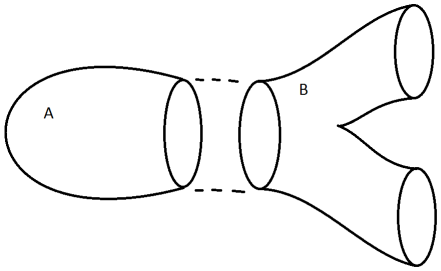


Figure 1: Possibility 1: adding a boundary component.

Let us call g the genus of $B \setminus A^\circ$, and $k \geq 2$ the number of connected components of $\partial(B \setminus A^\circ)$. It follows that $\chi(B \setminus A^\circ) = 2 - 2g - k$. Since $-\chi(B \setminus A^\circ) = n + 1 > 1$ (that is, $2g + k \geq 4$), properties (I) and (II) ensure the existence of a compact region W in $\mathcal{R}$ such that:

(i) W has genus 0 and three boundary components;

(ii) $W \subset B$, $\gamma_2 \subseteq \partial W$ and $W \cap A = \emptyset$;

(iii) if $\gamma \subset \partial W$ is a boundary component of W, then either $\gamma \subset \partial B$ or $\gamma \subset B^\circ$.

Property (iii) is equivalent to the fact that $\partial W \cap \partial B$ has either one or two connected components.

Finally, we define $B_* := \overline{B \setminus W}$. Now we observe that B_* is Runge and connected and also, $A \Subset B_* \Subset B$, $\chi(B_* \setminus A^\circ) = -n$ and $\chi(B \setminus B_*^\circ) = -1$. By the induction hypothesis applied to the pair $A \Subset B_*$, there exist connected Runge compact sets $N_1, \ldots, N_{n-2}$ in $\mathcal{R}$ such that $A \Subset N_1 \Subset \cdots \Subset N_{n-2} \Subset B_*$, $\chi(N_1^\circ \setminus A) \in \{-1, 0\}$, $\chi(B_*^\circ \setminus N_{n-2}) \in \{-1, 0\}$, and $\chi(N_i^\circ \setminus N_{i-1}) \in \{-1, 0\}$ for $i = 2, \ldots, n-2$. Setting $N_{n-1} := B_*$, the sequence of connected Runge compact sets $N_1, \ldots, N_{n-1}$ proves the inductive step and concludes the proof of the claim. □

The sequence $\{M_n\}_{n\in\mathbb{N}}$ that satisfies the statement of the lemma is generated by the process described in the Claim 3.3 applied to each pair $V_m \Subset V_{m+1}$ with $-\chi(V_{m+1} \setminus V_m^\circ) > 1$. We only have to add the new necessary terms and re-enumerate the arising sequence accordingly. □

Remark 3.4. It is interesting to think on the topological operations used in Lemma 3.2. The way we change to the next term is with an Euler characteristic change of value -1 or 0. We can study both in detail:

- Case $\chi(B \setminus A^\circ) = 0$. The compact set B has the same genus and the same number of boundary components than A, hence $B^\circ \setminus A$ is a finite union of pairwise disjoint annuli.
- Case $\chi(B \setminus A^\circ) = -1$. We put $W := B \setminus A^\circ$ and recall that W must have at least two boundary components ($k \geq 2$), one of them contained in ∂A and the other ones disjoint to A and contained in ∂B. Since $\chi(W) = 2 - 2g - k = -1$, where g is the genus of W, we have that $g = 0$ and $k = 3$. This case is possible only in two different topological situations, illustrated in Figures 1 and 2.

We continue with the following lemma whose proof is based on the Runge–Mergelyan Theorem.

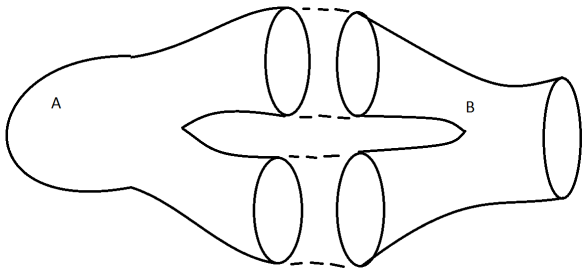


Figure 2: Possibility 2: adding a handle and removing a boundary component.

Lemma 3.5. *Let $\mathcal{R}$ be an open Riemann surface. Let A and B be bordered regions of $\mathcal{R}$ such that $A \Subset B$ and $\chi(B \setminus A^\circ) \in \{-1, 0\}$. Let $\tau > 0$ be a positive number and let $f = (f^1, f^2) \colon A \to \mathbb{C}^2$ be a continuous function, which is holomorphic on A°, and such that $\max\{\Re(f^1), \Re(f^2)\} > \tau$ on ∂A. Then, for any $\delta > 0$, there exists a continuous function $F \colon B \to \mathbb{C}^2$, which is holomorphic on B°, and satisfying the following properties:*

(a) $\|F - f\|_A < \delta$;

(b) $\max\{\Re(F^1), \Re(F^2)\} > \tau$ *on* $B \setminus A$;

(c) $\max\{\Re(F^1), \Re(F^2)\} > \tau + 1$ *on* ∂B.

Proof. We distinguish cases depending on the value of the Euler characteristic of $B \setminus A^\circ$.

Case 1: $\chi(B \setminus A^\circ) = 0$.

In this case $B \setminus A^\circ = A_1 \cup \cdots \cup A_m$, where each A_i is an annulus for all $i \in \{1, \dots, m\}$. In order to simplify notation we will suppose that $m = 1$. The general case is almost identical and consists of applying the same argument to each annulus A_i. Therefore, $B \setminus A^\circ$ is an annulus.

So, $\partial(B \setminus A^\circ)$ consists of two disjoint connected components, that is, $\partial(B \setminus A^\circ) = c \cup d$ where $c = \partial A$ and $d = \partial B$. Since $\max\{\Re(f^1), \Re(f^2)\} > \tau$ on ∂A, we can find an open cover $\widetilde{\Sigma} = \{O_\lambda : \lambda \in \Lambda\}$ of c such that

$$\Re(f^1) > \tau \quad \text{or} \quad \Re(f^2) > \tau, \text{ on each } O_\lambda, \ \lambda \in \Lambda. \tag{1}$$

Since ∂A is compact, there exists a finite subcover $\Sigma = \{O_1, \dots, O_k\}$ of ∂A contained in $\widetilde{\Sigma}$. Take arcs $\alpha_1, \dots, \alpha_n$ in c such that the following properties are satisfied:

- $\alpha_j \subset O_{h(j)}$ for some $h(j) \in \{1, \dots, k\}$, $\forall j = 1 \dots, n$;
- $\cup_{j=1}^n \alpha_j = c$;
- $\alpha_{j_1}^\circ \cap \alpha_{j_2}^\circ = \emptyset, \ \forall j_1, j_2 \in \{1, \dots, n\}, \ j_1 \neq j_2$.

We denote by $p_j \in c$ the initial point of the curve α_j, $\forall j = 1, \dots, n$. We relabel the arcs α_j, $j = 1, \dots, n$, in order to ensure that the final point of α_{j-1} is the initial point p_j of α_j, for any $j > 1$. We adopt the cyclic notation, $p_1 = p_{n+1}$, to identify the initial point of α_1 and the final point of α_n.

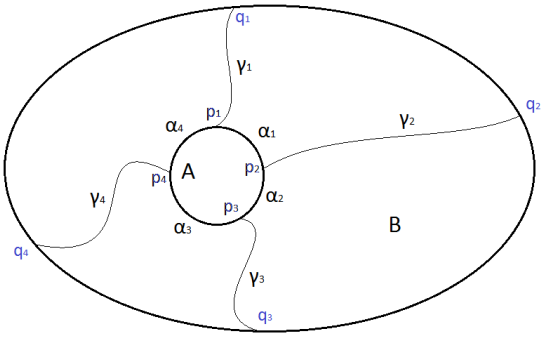


Figure 3: $B \setminus A^{\circ}$.

Let (I_1, I_2) be subsets of $\{1, \dots, n\}$ satisfying: (1) $I_1 \cup I_2 = \{1, \dots, n\}$ and $I_1 \cap I_2 = \emptyset$; and (2) if $j \in I_\mu$ then $\Re(f^\mu) > \tau$ on α_j, $\mu \in \{1, 2\}$. We consider now a family of non-intersecting simple curves from c to d that we denote $\{\gamma_j\}_{j=1,\dots,n}$. We suppose that the initial point of γ_j is $p_j \in c$ and we call its final point $q_j \in d$. It follows that $\gamma_j \cap (c \cup d) = \{p_j, q_j\}$ for any $j = 1, \dots, n$; see Figure 3.

We define now an auxiliary continuous function $g = (g^1, g^2)\colon A \cup \gamma_1 \cup \cdots \cup \gamma_n \to \mathbb{C}^2$, which is holomorphic on A and satisfies the following properties:

(i) $g|_A = f$;

(ii) if $j - 1 \in I_\mu$ then $\Re(g^\mu) > \tau$ on γ_j and $\Re(g^\mu(q_j)) > \tau + 1$ $\forall j = 1, \dots, n$ (here, we call $\alpha_0 = \alpha_n$ and $q_0 = q_n$);

(iii) if $j \in I_\mu$ then $\Re(g^\mu) > \tau$ on γ_j and $\Re(g^\mu(q_j)) > \tau + 1$ $\forall j = 1, \dots, n$;

where $\mu \in \{1, 2\}$. Such a function g exists due to properties (1) and (2) above.

Since A is Runge, the set $M = A \cup \gamma_1 \cup \cdots \cup \gamma_n$ is also Runge and the Runge–Melgerlyan Theorem gives a continuous function $G\colon B \to \mathbb{C}^2$, which is holomorphic on B° and satisfies the following properties:

(iv) $\|G - g\|_A < \delta/2$ on A; here, δ is the positive number given in the statement of the lemma;

(v) $\max\{\Re(G^1), \Re(G^2)\} > \tau$ on $\partial A = c$;

(vi) if $j - 1 \in I_\mu$ for $\mu \in \{1, 2\}$, then $\Re(G^\mu) > \tau$ on γ_j and $\Re(G^\mu(q_j)) > \tau + 1$, $\forall j = 1, \dots, n$;

(vii) if $j \in I_\mu$ for $\mu \in \{1, 2\}$, then $\Re(G^\mu) > \tau$ on γ_j and $\Re(G^\mu(q_j)) > \tau + 1$, $\forall j = 1, \dots, n$.

Summarizing, the function G formally satisfies properties (a), (b), and (c) on the set M and, by continuity, in a neighbourhood of M, but not necessary in the whole B.

Given $j \in \{1, \dots, n\}$, there is an open neighbourhood Γ_j on B of $\gamma_j \cup \alpha_j \cup \gamma_{j+1}$ such that G still satisfies (a), (b), and (c) in any set Γ_j, $j \in \{1, \dots, n\}$. More concretely, if $j \in I_\mu$ for $\mu \in \{1, 2\}$ then

$$\Re(G^\mu) > \tau \text{ on } \Gamma_j \qquad \text{and} \qquad \Re(G^\mu) > \tau + 1 \text{ on } \Gamma_j \cap \partial B, \tag{2}$$

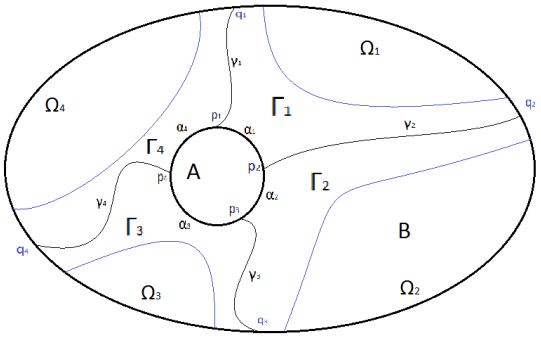


Figure 4: $B \setminus A^\circ$.

$\forall j = 1, \dots, n$. We introduce some more notation. For any fixed $j \in \{1, \dots, n\}$, we consider the topological closed disk in $B \setminus A^\circ$ whose boundary contains the set $\gamma_j \cup \alpha_j \cup \gamma_{j+1}$ and is disjoint from α_l, $l \neq j$. We call Ω_j the complement of Γ_j in this disk; see Figure 4.

Set:

(viii) $\Omega^1 = \bigcup_{j \in I_1} \Omega_j$;

(ix) $\Omega^2 = \bigcup_{j \in I_2} \Omega_j$;

(x) $\Gamma^1 = \bigcup_{j \in I_1} \Gamma_j$;

(xi) $\Gamma^2 = \bigcup_{j \in I_2} \Gamma_j$.

It follows that

$$B = A \cup \Gamma^1 \cup \Gamma^2 \cup \Omega^1 \cup \Omega^2. \tag{3}$$

Consider the functions

$$\widetilde{G}^1 \colon A \cup \Gamma^1 \cup \Omega^2 \to \mathbb{C}, \quad \widetilde{G}^1 = \begin{cases} G^1 & \text{in } A \cup \Gamma^1, \\ \tau + 2 & \text{in } \Omega^2, \end{cases} \tag{4}$$

$$\widetilde{G}^2 \colon A \cup \Gamma^2 \cup \Omega^1 \to \mathbb{C}, \quad \widetilde{G}^2 = \begin{cases} G^2 & \text{in } A \cup \Gamma^2, \\ \tau + 2 & \text{in } \Omega^1, \end{cases} \tag{5}$$

and recall that $(A \cup \Gamma^1) \cap \Omega^2 = \emptyset$ and $(A \cup \Gamma^2) \cap \Omega^1 = \emptyset$.

The sets $A \cup \Gamma^1 \cup \Omega^2$ and $A \cup \Gamma^2 \cup \Omega^1$ are Runge in B, whereas the functions $\widetilde{G}^1$ and $\widetilde{G}^2$ are continuous, and holomorphic on the interior. Hence, by the Runge Theorem, there exist two holomorphic functions on B, which we call F^1 and F^2, that approach $\widetilde{G}^1$ and $\widetilde{G}^2$, respectively. Then, if the approximation is close enough, $F = (F^1, F^2) \colon B \to \mathbb{C}^2$ is holomorphic and satisfies:

(xii) $\|F - \widetilde{G}\|_A < \delta/2$;

(xiii) F^1 approaches $\widetilde{G^1}$ in $A \cup \Gamma^1 \cup \Omega^2$;

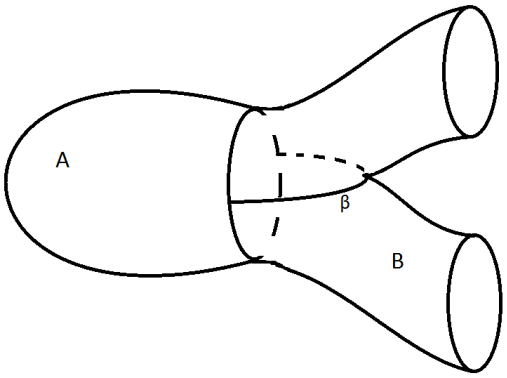


Figure 5: The arc β.

(xiv) F^2 approaches $\widetilde{G^2}$ in $A \cup \Gamma^2 \cup \Omega^1$.

Let us check that F solves the Lemma. Indeed:

- By properties (iv) and (v), $\tilde{G} = G$ on A, and so (xii) gives $\|F - G\|_A < \delta/2$. Thus, taking into account (i) and (iv), we get $\|F - f\|_A < \|F - G\|_A + \|G - f\|_A < \delta$.
- In Γ^1, we have $\Re(G^1) > \tau$ (by equation (2) and property (x)) and so, $\Re(F^1) > \tau$ on Γ^1 provided the approximation in (xiii) is close enough; take into account equation (4). Hence, $\max\{\Re(F^1), \Re(F^2)\} > \tau$ on Γ^1.
- In Γ^2, we have $\Re(G^2) > \tau$ (by equation (2) and property (xi)) and so, $\Re(F^2) > \tau$ on Γ^2 provided the approximation in (xiv) is close enough; use equation (5). Therefore, $\max\{\Re(F^1), \Re(F^2)\} > \tau$ on Γ_2.
- In Ω^1, we have $\Re(\widetilde{G^2}) > \tau$ (by equation (2) and property (viii)) and so, $\Re(F^2) > \tau$ on Ω^1 provided the approximation in (xiv) is close enough. Thus $\max\{\Re(F^1), \Re(F^2)\} > \tau$ on Ω^1.
- In Ω^2, we have $\Re(\widetilde{G^1}) > \tau$ (by equation (2) and property (ix)) and so, $\Re(F^1) > \tau$ on Ω^2 provided the approximation in (xiii) is close enough. Hence $\max\{\Re(F^1), \Re(F^2)\} > \tau$ on Ω^2.

We finish the discussion with the set ∂B. On the one hand, we have $\Re(\widetilde{G^2}) > \tau + 1$ on Ω^1 (by equations (2) and (5)). On the other hand, $\Re(\widetilde{G^1}) > \tau + 1$ on Ω^2 (by equations (2) and (4)). Finally, $\max\{\Re(G^1), \Re(G^2)\} > \tau + 1$ on $\partial B \setminus (\Omega^1 \cup \Omega^2) = (\Gamma^1 \cup \Gamma^2) \cap \partial B$. Indeed, $G^1 > \tau + 1$ on $\partial B \cap \Gamma^1$ and $G^2 > \tau + 1$ on $\partial B \cap \Gamma^2$; see equation (2). Thus, $\max\{\Re(F^1), \Re(F^2)\} > \tau + 1$ in ∂B.

This concludes the proof in case 1.

Case 2: $\chi(B \setminus A^\circ) = -1$.

By Remark 3.4, B can be described as a neighbourhood of the set that we obtain when we add an arc in $B \setminus A$ to A with initial point and final point in ∂A. We call this arc β and we observe that $A \cup \beta$ is a deformation retract of B; see Figures 5 and 6.

Consider a continuous function $g\colon A \cup \beta \to \mathbb{C}^2$, which is holomorphic on A° and satisfies $g = f$ on A and $\max\{\Re(g^1), \Re(g^2)\} > \tau$ on β. By Runge's Theorem we may approximate g on $A \cup \beta$ by holomorphic

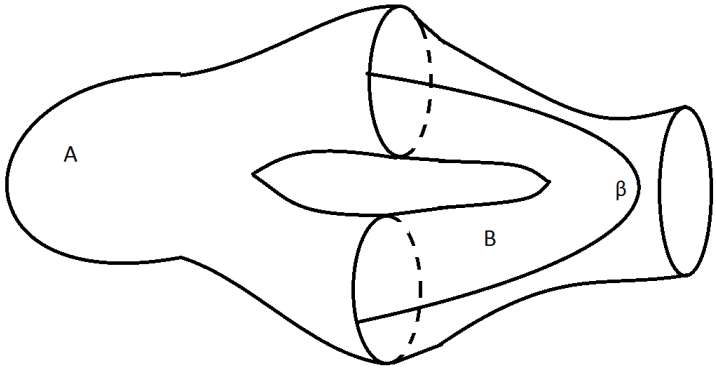


Figure 6: The arc β.

functions $\hat{f}$ on B. If we take a closed neighbourhood $\hat{A}$ of $A \cup \beta$ on B° with $\chi(B \setminus \hat{A}^\circ) = 0$ (recall that $A \cup \beta$ is a deformation retract of B), and the approximation is close enough, the function $\hat{f}|_{\hat{A}}$ formally satisfies the hypothesis (a), (b), and (c) of the lemma. This reduces the proof to Case 1. □

Proof of Theorem 3.1. Let $\{M_n\}_{n\in\mathbb{N}}$ be an exhaustive sequence of Runge and connected bordered regions in $\mathcal{R}$ such that $\chi(M_{n+1} \setminus M_n) \in \{-1, 0\}$ for all $n \in \mathbb{N}$. Such sequences exists by Lemma 3.2. Given a sequence of real numbers $\{\epsilon_n\}_{n\in\mathbb{N}}$, $\epsilon_n > 0$, a recursive use of Lemma 3.5 supplies a sequence of continuous functions $f_n = (f_n^1, f_n^2)\colon M_n \to \mathbb{C}^2$, $n \in \mathbb{N}$, satisfying:

(a) f_n is holomorphic on M_n°, $\forall n \in \mathbb{N}$;

(b) $\|f_{n+1} - f_n\|_{M_n} < \epsilon_n$, $\forall n \in \mathbb{N}$;

(c) $\max\{\Re(f_n^1), \Re(f_n^2)\} > n$ on ∂M_n, $\forall n \in \mathbb{N}$;

(d) $\max\{\Re(f_{n+1}^1), \Re(f_{n+1}^2)\} > n$ on $M_{n+1} \setminus M_n^\circ$, $\forall n \in \mathbb{N}$.

Indeed, for the basis of the induction, choose any continuous function $f_1\colon M_1 \to \mathbb{C}^2$, which is holomorphic on M_1°, and satisfies $\max\{\Re(f_1^1), \Re(f_1^2)\} > 1$ on ∂M_1. For instance, we may take f_1 to be a suitable constant in $\mathbb{C}^2$. For the inductive step, let $n \in \mathbb{N}$ and suppose that we have functions $f_1\colon M_1 \to \mathbb{C}^2, \ldots, f_n\colon M_n \to \mathbb{C}^2$ satisfying formally the above properties. Since $M_n \Subset M_{n+1}$ and $\chi(M_{n+1}^\circ \setminus M_n) \in \{-1, 0\}$, Lemma 3.5 applied to $\tau = n$ and $\delta = \epsilon_n$ gives a continuous function $F = (F^1, F^2)\colon M_{n+1} \to \mathbb{C}^2$, which is holomorphic in M_{n+1}°, and satisfies $\|F - f_n\| < \epsilon_n$. In addition, $\max\{\Re(F^1), \Re(F^2)\} > n$ on $M_{n+1} \setminus M_n$ and $\max\{\Re(F^1), \Re(F^2)\} > n+1$ on ∂M_{n+1}. Obviously, we finish the induction setting $f_{n+1} = F$.

Let $\{f_n\colon M_n \to \mathbb{C}^2\}_{n\in\mathbb{N}}$ be the sequence we have already found satisfying (a)–(d). Let us see first that, up to a suitable choice of the numbers ϵ_n, the sequence f_n converges uniformly on compact sets of $\mathcal{R}$. It is enough to prove that (for a good choice of the ϵ_n) given $\epsilon > 0$ and a compact set $K \subset \mathcal{R}$, there exists $n_0 \in \mathbb{N}$ such that if $p, q \geq n_0$ then $\|f_p - f_q\|_K \leq \epsilon$. It is required that n_0 is large enough to satisfy that $K \subset M_{n_0}$, so that f_p and f_q are well defined on K. Indeed, if we take the sequence $\{\epsilon_n\}_{n\in\mathbb{N}}$ such that

$\sum_{n=1}^{\infty} \epsilon_n < +\infty$, we can consider n_0 such that $\sum_{n=n_0+1}^{\infty} \epsilon_n < \epsilon$ and $K \subset M_{n_0}$. Then, given $p, q \geq n_0,\ p > q$,

$$\|f_p - f_q\|_K = \|\sum_{k=1}^{p-q} (f_{q+k} - f_{q+k-1})\|_K \leq \sum_{k=1}^{p-q} \|f_{q+k} - f_{q+k-1}\|_{M_{q+k-1}} < \sum_{k=1}^{p-q} \epsilon_{q+k-1} = \sum_{n=q}^{p} \epsilon_n < \epsilon.$$

Therefore, $\{f_n\}_{n\in\mathbb{N}}$ is a Cauchy sequence with the maximum norm and, consequently, it converges uniformly on compact sets to a function $f\colon \mathcal{R} \to \mathbb{C}^2$. Furthermore, the convergence Harnack theorem asserts that f is holomorphic and so, $\Re(f)\colon \mathcal{R} \to \mathbb{R}^2$ is harmonic. On the other hand, if we fix ϵ and n_0 as above and we take limits in the previous estimation, we obtain that

$$\|f - f_n\|_{M_n} \leq \epsilon, \quad \forall n \geq n_0. \tag{6}$$

To finish the proof, let us check that $\Re(f)\colon \mathcal{R} \to \mathbb{R}^2$ is proper. Let $\{x_n\}_{n\in\mathbb{N}} \subset \mathcal{R}$ be a divergent sequence. Then for each $n \in \mathbb{N}$ there exists $m_n \in \mathbb{N}$ such that $x_n \in M_{m_n} \setminus M_{m_n-1}$ and, by (d), $\|\Re(f_n(x_n))\| > m_n$. Hence, using (6), we deduce that $\|\Re(f(x_n))\| > m_n - \epsilon$. But now, as $\{x_n\}_{n\in\mathbb{N}}$ is divergent on $\mathcal{R}$ and $\{M_n\}_{n\in\mathbb{N}}$ is increasing, we have that m_n depends on n in such a way that if $n \to +\infty$, then $m_n \to +\infty$. Therefore, it is clear that $\|\Re(f(x_n))\|_{M_n} \to +\infty$ as $n \to +\infty$, and $\{\Re(f(x_n))\}$ is a divergent sequence. Thus, $\varphi = \Re(f)\colon \mathcal{R} \to \mathbb{R}^2$ is proper, which concludes the proof. □

References

[1] A. Alarcón and J.A. Gálvez, "Proper harmonic maps from hyperbolic Riemann surfaces into the euclidean plane", *Results Math.* **60**(1–4) (2011), 487–505.

[2] A. Alarcón and F.J. López, "Minimal surfaces in $\mathbb{R}^3$ properly projecting into $\mathbb{R}^2$", *J. Differential Geom.* **90**(3) (2012), 351–381.

[3] A. Alarcón and F.J. López, "Properness of associated minimal surfaces", *Trans. Amer. Math. Soc.* **366**(10) (2014), 5139–5154.

[4] R.B. Andrist and E.F. Wold, "Riemann surfaces in stein manifolds with density property", preprint. arXiv:1106.4416v1.

[5] E. Bishop, "Subalgebras of functions on a Riemann surface", *Pacific J. Math.* **8** (1958), 29–50.

[6] H.M. Farkas and I. Kra, "Riemann surfaces" (second edition), *Graduate Texts in Mathematics* **71**. Springer-Verlag, New York, 1992.

[7] F. Forstnerič and J. Globevnik, "Proper holomorphic disks in $\mathbb{C}^2$", *Math. Res. Lett.* **8**(3) (2001), 257–274.

[8] A. Grigor'yan, "Analytic and geometric background of recurrence and non-explosion of the Brownian motion on Riemannian manifolds", *Bull. Amer. Math. Soc.* **36**(2) (1999), 135–249.

[9] E. Heinz, "Uber die Lösungen der Minimalflächengleichung" (german), *Nachr. Akad. Wiss. Göttingen. Math.-Phys. Kl. Math.-Phys.-Chem. Abt.* **1952** (1952), 51–56.

[10] S.N. Mergelyan, "On the representation of functions by series of polynomials on closed sets", *Doklady Akad. Nauk SSSR* **78** (1951), 405–408.

[11] C. Runge, "Zur Theorie der Analytischen Functionen" (german), *Acta Math.* **6** (1885), 245–248.

[12] R. Schoen and S.T. Yau, "Lectures on harmonic maps", *Conference Proceedings and Lecture Notes in Geometry and Topology, II*. International Press, Cambridge, MA, 1997.

REPORTS @SCM

AN ELECTRONIC JOURNAL OF THE
SOCIETAT CATALANA DE MATEMÀTIQUES

On the concept of fractality for groups of automorphisms of a regular rooted tree

*Jone Uria-Albizuri

University of the Basque Country, UPV/EHU, Department of Mathematics.
jone.uria@ehu.eus

∗Corresponding author

Resum *(CAT)*

L'objectiu d'aquest article és discutir i aclarir la noció de fractalitat per a subgrups del grup d'automorfismes d'un arbre arrelat i regular. Per això, definim tres tipus de fractalitat i demostrem, donant contraexemples, que no són equivalents. També presentem alguns resultats que ajuden a determinar el tipus de fractalitat d'un grup donat.

Abstract *(ENG)*

The aim of this article is to discuss and clarify the notion of fractality for subgroups of the group of automorphisms of a regular rooted tree. For this purpose, we define three types of fractality. We show that they are not equivalent, by giving explicit counter-examples. Furthermore, we present some tools that are helpful in order to determine the fractality of a given group.

Societat Catalana de Matemàtiques

Institut d'Estudis Catalans

Keywords: *Rooted tree, automorphism, fractal.*
MSC (2010): *Primary 20E08.*
Received: *September 15th, 2015.*
Accepted: *February 8th, 2016.*

Acknowledgement

The author is supported by the Basque Government research project IT753-13 and by the Basque Goverment predoctoral grant PRE-2014-1-347.

1. Introduction

The subgroups of the group of automorphisms of the d-adic tree T (i.e., a regular rooted tree with d branches going down at every vertex) are an important source of groups with interesting properties. For example, finitely generated torsion infinite groups can be constructed easily, giving a negative answer to the General Burnside Problem. The large amount of articles about this topic in the last years shows their interest.

Given a subgroup G of $\mathrm{Aut}\, T$, the section of an element $g \in G$ at a vertex u is an automorphism which represents how g acts on the subtree of T hanging from the vertex u (the formal definition is given in Section 2). We say that G is *self-similar* if, for each element $g \in G$ and each vertex $u \in T$, the section of g at the vertex u belongs to G again. This is a natural property that a majority of the most interesting subgroups of $\mathrm{Aut}\, T$ possess.

It is usual to work with vertex and level stabilizers of G, i.e., the subgroups of all automorphisms in G that fix a vertex u or a whole level L_n of the tree, denoted by $\mathrm{Stab}_G(u)$ and $\mathrm{Stab}_G(L_n)$, respectively. Then one can consider the homomorphisms ψ_u, which sends each $g \in \mathrm{Stab}_G(u)$ to its section at the vertex u, and ψ_n, which sends each $g \in \mathrm{Stab}_G(L_n)$ to the d^n-tuple of its sections at the n-th level. Observe that in these cases the sections are just the restrictions to the corresponding subtrees.

If $G = \mathrm{Aut}\, T$, then the homomorphisms ψ_u and ψ_n are surjective onto $\mathrm{Aut}\, T$ and $\mathrm{Aut}\, T \times \overset{d^n}{\cdots} \times \mathrm{Aut}\, T$, respectively. On the other hand, if G is self-similar then the images of ψ_u and ψ_n are contained in G and $G \times \overset{d^n}{\cdots} \times G$, and we will consider these sets to be the codomains of those maps. It is natural to ask whether ψ_u and ψ_n are also onto in this case. For many interesting groups, ψ_u is known to be onto, i.e., $\psi_u(\mathrm{Stab}_G(u)) = G$ for each $u \in T$, and the group G is then called *fractal*, *recurrent* or *self-replicating* (see [3, 9]). However, in general, it is too strong to ask ψ_n to be surjective, and we content ourselves with the image of ψ_n being a subdirect product of $G \times \overset{d^n}{\cdots} \times G$, namely that $\psi_u(\mathrm{Stab}_G(L_n)) = G$ for each $u \in L_n$. In some papers, this condition is only required for $n = 1$; however, as we shall see, it is not always inherited by the rest of the levels. Thus it is necessary to make a distinction between these two concepts. Following terminology from previous papers, G is said to be *strongly fractal* or *strongly self-replicating* if $\psi_u(\mathrm{Stab}_G(L_1)) = G$ for all $u \in L_1$. And we say that G is *super strongly fractal* if $\psi_u(\mathrm{Stab}_G(L_n)) = G$ for each $n \in \mathbb{N}$ and $u \in L_n$.

Obviously, every super strongly fractal group is also strongly fractal, and every strongly fractal group is fractal, but there is some confusion in the literature about the converses. In several papers, fractal groups are claimed to be the same as strongly fractal groups, or else fractal groups are simply introduced by using the definition of strong fractality (see [1, 3, 4, 5, 6]). In some other papers, a distinction is made between these two concepts (see [2, 9]), but no examples can be found in the literature where a certain fractal group is shown not to be strongly fractal. On the other hand, strongly fractal and super strongly fractal groups have not been clearly distinguished either. Since a self-similar group that acts transitively on each level can be checked to be fractal by looking only at the vertices on the first level, one may think that the same holds for the property of being strongly fractal, see for example the paragraph after [9, Def. 3.6]. This would mean that being strongly fractal and super strongly fractal are equivalent. However, as we shall see, this is not the case.

Our aim in this article is to fill this gap. On the one hand, for every $d \geq 3$, we give explicit examples of groups that are fractal but not strongly fractal. More specifically, we show that a certain subgroup of the

Hanoi Towers group is of this type. We remark that the restriction to $d \geq 3$ is necessary for these examples to exist, since one can easily show that for $d = 2$ a fractal group is always strongly fractal. In proving that those groups are not strongly fractal, we have obtained a couple of results that allow us to estimate the image of a level stabilizer under ψ_u, which may have some interest of their own. On the other hand, we also give examples of groups which are strongly fractal but not super strongly fractal, and examples of super strongly fractal groups.

These examples belong to the class of the so-called Grigorchuk–Gupta–Sidki groups (GGS-groups, for short), which are a natural generalisation of the Grigorchuk group [8], and the Gupta–Sidki examples from [11].

2. Preliminaries

Let us consider a set X with d elements. The d-adic tree T is a tree whose set of vertices is the free monoid X^*, where a word u is a descendant of v if $u = vx$ for some $x \in X$. The only word of length zero is the empty word $\emptyset$, which is the root of the tree T. If we consider the words of length at most n we have a finite subtree T_n, and the words whose length is exactly n form the n-th level of the tree, L_n.

An automorphism of the d-adic tree is a map preserving incidence between vertices. All automorphisms of T form a group $\mathrm{Aut}\, T$ under composition, where we write fg for $g \circ f$. Thus $(fg)(u) = g(f(u))$ for every vertex u of T.

Let us consider the natural projection $\pi_n \colon \mathrm{Aut}\, T \longrightarrow \mathrm{Aut}\, T_n$, which sends every automorphism to its restriction to T_n. Observe that the stabilizer $\mathrm{Stab}(L_n)$ of the n-th level is the kernel of π_n, so it is a normal subgroup in $\mathrm{Aut}\, T$, and we have $\mathrm{Aut}\, T_n \cong \mathrm{Aut}\, T / \mathrm{Stab}(L_n)$.

An important observation is that every automorphism $g \in \mathrm{Aut}\, T$ can be fully described by saying for each vertex $u \in T$ how g permutes the d vertices hanging from u. So, there is a permutation α of X (which clearly depends on u) such that $g(ux) = g(u)\alpha(x)$. We say that α is the *label* of g at the vertex u, and we denote it by $g_{(u)}$.

Since $T \cong T_u$, where T_u denotes the subtree hanging from a vertex u, we have $\mathrm{Aut}\, T \cong \mathrm{Aut}\, T_u$. We speak about the *section* of g at the vertex u and we denote it by g_u, to refer to the automorphism defined by $g(uv) = g(u)g_u(v)$ for each vertex v. Then we have the following formulas:

$$(f^{-1})_u = (f_{f^{-1}(u)})^{-1}, \qquad (fg)_u = f_u g_{f(u)}, \qquad f_{uv} = (f_u)_v, \tag{1}$$

and

$$(f^g)_u = (g_{g^{-1}(u)})^{-1} f_{g^{-1}(u)} g_{g^{-1}f(u)}. \tag{2}$$

Also, we define the homomorphism $\psi_n \colon \mathrm{Stab}(L_n) \longrightarrow \mathrm{Aut}\, T \times \overset{d^n}{\cdots} \times \mathrm{Aut}\, T$ which sends $g \in \mathrm{Stab}(L_n)$ to the d^n-tuple of its sections $(g_{u_1}, \ldots, g_{u_{d^n}})$, with $u_i \in L_n$. In the same way, for the stabilizer $\mathrm{Stab}(u)$ of the vertex u, we have a homomorphism denoted by ψ_u which sends $g \in \mathrm{Stab}(u)$ to g_u.

Sometimes it is useful to think of $\mathrm{Aut}\, T$ as a semidirect product.

Proposition 2.1. *Let T be the d-adic tree and let us consider the following subgroup for each $n \in \mathbb{N}$:*

$$H_n = \{h \in \mathrm{Aut}\, T \mid h_u = 1 \ \forall u \in L_n\}.$$

Then, we have $\mathrm{Aut}\, T = H_n \ltimes \mathrm{Stab}(L_n)$.

Observe that, for $f \in \mathrm{Stab}(L_n)$ and $g = hg' \in \mathrm{Aut}\, T$, with $h \in H_n$ and $g' \in \mathrm{Stab}(L_n)$, we have

$$(f^g)_u = (f_{h^{-1}(u)})^{g_u} = (f_{h^{-1}(u)})^{g'_u} \text{ for all } u \in L_n. \tag{3}$$

Let now $G \leq \mathrm{Aut}\, T$. Then we can consider the stabilizers in G of each vertex, $\mathrm{Stab}_G(u) = \mathrm{Stab}(u) \cap G$, and the level stabilizers $\mathrm{Stab}_G(L_n) = \cap_{u \in L_n} \mathrm{Stab}_G(u) = \mathrm{Stab}(L_n) \cap G$. So we have the restrictions of ψ_n and ψ_u to $\mathrm{Stab}_G(L_n)$ and $\mathrm{Stab}_G(u)$, respectively. Since we are interested in those groups for which the images under ψ_u and ψ_n are in G and $G \times \overset{d^n}{\cdots} \times G$, we give the following definition.

Definition 2.2. We say that a group $G \leq \mathrm{Aut}\, T$ is *self-similar* if for each element of G its sections are also elements of G; in other words, if

$$\{g_u \mid g \in G, \ \ u \in T\} \subseteq G. \tag{4}$$

It is easy to prove by induction on the length of a vertex, and using the first two formulae in (1), that if (4) is satisfied by the vertices of the first level then the group is self-similar (see [10, Prop. 3.1]).

Lemma 2.3. *A group $G = \langle S \rangle \leq \mathrm{Aut}\, T$ is self-similar if and only if $s_x \in G$ for each $s \in S$ and $x \in X$.*

Observe that, even if in the case of the whole group of automorphisms $\mathrm{Aut}\, T$ the homomorphisms ψ_n and ψ_u are surjective, this might not be true in general. According to this, we have the following definitions.

Definition 2.4. Let $G \leq \mathrm{Aut}\, T$ be a self-similar group. Then,

(i) we say that G is *fractal* if $\psi_u(\mathrm{Stab}_G(u)) = G$ for each vertex $u \in T$;

(ii) we say that G is *strongly fractal* if $\psi_x(\mathrm{Stab}_G(L_1)) = G$ for each $x \in X$;

(iii) we say that G is *super strongly fractal* if $\psi_u(\mathrm{Stab}_G(L_n)) = G$ for each $u \in L_n$ and each $n \in \mathbb{N}$.

Notice that the definition of being super strongly fractal does not imply that ψ_n is surjective from G to $G \times \overset{d^n}{\cdots} \times G$, but only that $\psi_n(\mathrm{Stab}_G(L_n))$ is a subdirect product in $G \times \overset{d^n}{\cdots} \times G$. The same remark applies to strongly fractal groups with $n = 1$.

There is a special case in which the first two definitions are equivalent.

Lemma 2.5. *Let $G \leq \mathrm{Aut}\, T$ and consider a d-cycle $\sigma \in S_X$. If for each $g \in G$ we have $g_{(\emptyset)} = \sigma^k$ for some $k \in \mathbb{N}$ and G is fractal, then G is strongly fractal.*

Proof. Let $g \in \mathrm{Stab}_G(x)$ for $x \in X$. Then $\sigma^k(x) = x$ which only happens if $k \equiv 0 \pmod d$. This implies that $g \in \mathrm{Stab}_G(L_1)$ so, $\mathrm{Stab}_G(x) = \mathrm{Stab}_G(L_1)$ and we are done. □

Observe that for $d = 2$ the label at the root must be 1 or (1 2) so, according to the previous lemma, in this case being fractal and being strongly fractal are equivalent.

This can be generalised, to obtain another important corollary that follows from the previous lemma in the case $d = p$, where p is a prime. If we consider T to be the p-adic tree, $\mathrm{Aut}\, T$ is a profinite group which has a standard Sylow pro-p subgroup consisting of automorphisms which have powers of a fixed p-cycle as a label in every vertex. Then, the previous lemma shows that, for every subgroup of the Sylow pro-p

subgroup, being fractal and strongly fractal are equivalent. For example, this happens for the GGS-groups (for the definition see Section 4).

One of our goals is to give examples of subgroups of Aut T for $d \geq 3$ which are fractal but are not strongly fractal. We next give the definition of being level transitive, because the examples that we present are of this type and also because in this case it is easier to check if a group is fractal or not.

Definition 2.6. Let $G \leq \mathrm{Aut}\, T$. We say that G is *level transitive* or that acts spherically transitively on T, if it is transitive on each level.

In a similar way to Lemma 2.3, in some cases, to check whether a group is fractal it is enough to look at the vertices on the first level (for a reference, see [9, Sect. 3]).

Lemma 2.7. *If $G \leq \mathrm{Aut}\, T$ is transitive on the first level and $\psi_x(\mathrm{Stab}_G(x)) = G$ for some $x \in X$, then G is fractal and level transitive.*

Since we will want to prove that a group is not strongly fractal, we are interested in identifying which is the first level stabilizer. We present a tool that we have developed in order to do this in the following lemma. Let us denote by ρ the homomorphism from G to S_d sending each $g \in G$ to the label of g at the root, $g_{(\emptyset)}$. We use the notation $\langle S \rangle^G$ for the normal closure in G of the subgroup generated by the set S.

Lemma 2.8. *Let $G \leq \mathrm{Aut}\, T$ and put $J = \rho(G)$. Suppose that we have a presentation $J = \langle Y \mid R \rangle$ and let $\theta \colon F \longrightarrow J$ be the epimorphism corresponding to this presentation, where F is the free group generated by Y. If there exists a surjective homomorphism $\phi \colon F \longrightarrow G$ making the following diagram commutative*

$$\begin{array}{ccc} F & \xrightarrow{\phi} & G \\ & \searrow^{\theta} & \downarrow{\scriptstyle \rho} \\ & & J, \end{array}$$

then $\mathrm{Stab}_G(L_1) = \langle \phi(R) \rangle^G$.

Proof. We know that $\ker \theta = \langle R \rangle^F$. On the other hand, since ϕ is surjective, every $g \in G$ can be written as $g = \phi(x)$ for some $x \in F$, and then $g \in \ker \rho$ if and only if $x \in \ker(\rho \circ \phi)$. Consequently,

$$\mathrm{Stab}_G(L_1) = \ker \rho = \phi(\ker(\rho \circ \phi)) = \phi(\ker \theta) = \phi(\langle R \rangle^F) = \langle \phi(R) \rangle^G,$$

which completes the proof. □

Notice that the actual condition we are asking about ϕ is to be surjective since, by the universal property of free groups, we are always able to construct some ϕ making the diagram commutative. In other words, the point is whether for each $y \in Y$ we can choose an element $g_y \in \rho^{-1}(\theta(y))$, in such a way that $\{g_y \mid y \in Y\}$ generates the whole group G or not.

Now, in the following lemma we present another new result, which will help us proving that the image of a level stabilizer under ψ_u is strictly contained in G.

Lemma 2.9. *Let $G \leq \mathrm{Aut}\, T$ be a self-similar group. If $K = \langle S \rangle^G \subseteq \mathrm{Stab}_G(L_n)$ for some $n \in \mathbb{N}$ and $\psi_u(S) \subseteq N$ for each $u \in L_n$, where $N \trianglelefteq G$, then $\psi_u(K) \subseteq N$ for each $u \in L_n$.*

Proof. Consider $k \in K$ and let us write $k = (s_1^{\epsilon_1})^{g_1} \cdots (s_r^{\epsilon_r})^{g_r}$, where $\epsilon_i \in \{-1, 1\}$, $s_i \in S$ and $g_i \in G$ for each $i = 1, \dots, r$. Let $u \in L_n$. Since $K \leq \mathrm{Stab}_G(L_n)$ we know that $k \in \mathrm{Stab}_G(u)$ and we have

$$\psi_u(k) = \psi_u(s_1^{g_1})^{\epsilon_1} \cdots \psi_u(s_r^{g_r})^{\epsilon_r}.$$

Thus, it is enough to see that $\psi_u(s^g) \in N$ for each $s \in S$, $g \in G$. Since $G \leq \mathrm{Aut}\, T$ and $\mathrm{Aut}\, T = H_n \ltimes \mathrm{Stab}(L_n)$, we write each $g = ht$ where $h \in H_n$ and $t \in \mathrm{Stab}(L_n)$. Now by (3) we have $\psi_u(s^g) = (s_{h^{-1}(u)})^{t_u}$ for each $u \in L_n$, and since $\psi_{h^{-1}(u)}(S) \subseteq N$ and N is normal in G, it is enough to check that t_u belongs to G. We know that G is self-similar so, $g_v \in G$ for each $v \in T$ and, in particular, for $v = h^{-1}(u)$. But $g_{h^{-1}(u)} = h_{h^{-1}(u)} t_u = t_u$ because $h \in H_n$, so we are done. □

Now, let us introduce a stronger version of the previous lemma that will help us checking whether a strongly fractal group is super strongly fractal or not.

Lemma 2.10. *Let G be level transitive and super strongly fractal. If $K = \langle S \rangle^G \subseteq \mathrm{Stab}_G(L_n)$ for some $n \in \mathbb{N}$, then $\psi_u(K) = \langle \psi_v(S) \mid v \in L_n \rangle^G$ for any $u \in L_n$.*

Proof. Let us denote $N = \langle \psi_v(S) \mid v \in L_n \rangle^G$. Since $\psi_u(S) \subseteq N$ for every $u \in L_n$, which is a normal subgroup, the inclusion $\psi_u(K) \subseteq N$ follows from the previous lemma.

Now, let $g = (\psi_{u_1}(s_1)^{\epsilon_1})^{g_1} \cdots (\psi_{u_r}(s_r)^{\epsilon_r})^{g_r} \in N$. Since G is level transitive for every $u_i \in \{u_1, \dots, u_r\}$, there is some $f_i \in G$ such that $f_i(u_i) = u$. Then, by (3),

$$\psi_u(s_i^{f_i})^{((f_i)_{u_i})^{-1}} = \psi_{u_i}(s_i).$$

Then we can write $g = (\psi_u(s_1^{f_1})^{\epsilon_1})^{g_1'} \cdots (\psi_u(s_r^{f_r})^{\epsilon_r})^{g_r'}$, where $g_i' = ((f_i)_{u_i})^{-1} g_i \in G$. From the fact that G is super strongly fractal, we know that there are some $h_i \in \mathrm{Stab}_G(L_n)$ such that $\psi_u(h_i) = g_i'$ for $i = 1, \dots, r$. Since

$$\begin{aligned} g &= (\psi_u(s_1^{f_1})^{\epsilon_1})^{g_1'} \cdots (\psi_u(s_r^{f_r})^{\epsilon_r})^{g_r'} = (\psi_u(s_1^{f_1})^{\epsilon_1})^{\psi_u(h_1)} \cdots (\psi_u(s_r^{f_r})^{\epsilon_r})^{\psi_u(h_r)} \\ &= \psi_u((s_1^{\epsilon_1})^{f_1 h_1} \cdots (s_r^{\epsilon_r})^{f_r h_r}) \in \psi_u(K), \end{aligned}$$

we are done. □

In particular, we have the following result when the group is strongly fractal.

Corollary 2.11. *Let G be a strongly fractal group acting transitively on the first level. If $K = \langle S \rangle^G$ and $K \subseteq \mathrm{Stab}_G(L_1)$ then $\psi_x(K) = \langle \psi_y(S) \mid y \in X \rangle^G$ for any $x \in X$.*

Finally, let us introduce another lemma that will help us proving that a group is super strongly fractal. This lemma tells us that in some cases, it suffices to check whether in each level stabilizer there are elements whose sections at vertices on this level generate the whole group.

Lemma 2.12. *Let $G \leq \mathrm{Aut}\, T$ be a self-similar group such that there is a rooted automorphism $a \in G$, with $a_{(\emptyset)}$ being a d-cycle. If for each $n \in \mathbb{N}$ we have $\langle \psi_{u_n}(\mathrm{Stab}_G(L_n)) \mid u_n \in L_n \rangle = G$, then G is super strongly fractal.*

Proof. The proof works by induction on the length of the vertices. Let $x \in X$ and $g \in G$. We know that there are some $y_1, \dots, y_r \in X$ such that $g = \psi_{y_1}(g_1)^{\epsilon_1} \cdots \psi_{y_r}(g_r)^{\epsilon_r}$, where $g_i \in \mathrm{Stab}_G(L_1)$ and $\epsilon_i \in \{1, -1\}$. Then for each $i = 1, \dots, r$ we have $a^{j_i}(y_i) = x$ for some $j_i \in \{0, \dots, d-1\}$. Then, considering $g_i^{a^{j_i}}$, we get an element on the first level stabilizer such that $(g_i^{a^{j_i}})_x = (g_i)_{y_i}$. Then the element $h = (g_1^{a^{j_1}})^{\epsilon_1} \cdots (g_r^{a^{j_r}})^{\epsilon_r} \in \mathrm{Stab}_G(L_1)$ satisfies $h_x = g$ so, $\psi_x(\mathrm{Stab}_G(L_1)) = G$.

Now let us suppose that we know the result for length $n-1$ and let us see it for n. Let $v = x_1 \cdots x_n$ and $g \in G$. By assumption, we know that $g = \psi_{w_1}(g_1)^{\epsilon_1} \cdots \psi_{w_r}(g_r)^{\epsilon_r}$ where $w_i \in L_n$, $g_i \in \mathrm{Stab}_G(L_n)$ and $\epsilon_i \in \{1, -1\}$ for each $i = 1, \dots, r$. It suffices to show that for $i = 1, \dots, r$ there is some $h_i \in \mathrm{Stab}_G(L_n)$ such that $(h_i)_v = (g_i)_{w_i}$, because then $h = h_1^{\epsilon_1} \cdots h_r^{\epsilon_r} \in \mathrm{Stab}_G(L_n)$ and $h_v = g$, as desired.

Let w be an arbitrary vertex in L_n. Then $w = y_1 \cdots y_n$ with $y_i \in X$. For each $k = 1, \dots, n$ there is some $j_k = 0, \dots, d-1$ such that $a^{j_k}(y_k) = x_k$. By the inductive assumption, $a \in \psi_u(\mathrm{Stab}_G(L_k))$ for every $u \in L_k$, with $k = 1, \dots, n-1$. Thus, for each $k = 1, \dots, n-1$ there is some $f_k \in \mathrm{Stab}_G(L_k)$ such that $(f_k)_{y_1 \dots y_k} = a^{j_{k+1}}$. Then, if we consider the element $f = a^{j_1} f_1 \cdots f_{n-1}$, which belongs to H_n, we obtain that $f(w) = v$. Thus, in particular for each $i = 1, \dots, r$, there is some $t_i \in H_n$ such that $t_i(w_i) = v$. Then $h_i = g_i^{t_i} \in \mathrm{Stab}_G(L_n)$ and, by (3), $(h_i)_v = (g_i)_{t_i^{-1}(v)} = (g_i)_{w_i}$. □

Remark 2.13. In particular, in the conditions of the previous lemma, it is enough for a group G to be super strongly fractal having one vertex $u_n \in L_n$ such that $\psi_{u_n}(\mathrm{Stab}_G(L_n)) = G$ for each $n \in \mathbb{N}$.

3. Fractal groups which are not strongly fractal

In this section we present an example for each $d \geq 3$ which is fractal but not strongly fractal. Even more, that example is a group acting spherically transitively on T. We denote by $x_1, \dots, x_d$ the elements of X, namely, the vertices of the first level.

The example that we consider is a subgroup of the Hanoi Towers group, which is defined as follows for each $d \geq 3$. For $1 \leq i < j \leq d$, we define the element a_{ij} which has the permutation $(x_i\, x_j)$ at the root and, for each vertex on the first level,

$$(a_{ij})_{x_k} = \begin{cases} 1 & \text{if } k = i, j, \\ a_{ij} & \text{otherwise.} \end{cases}$$

The Hanoi Towers group is $H = \langle a_{ij} \mid 1 \leq i < j \leq d \rangle$. Although H is strongly fractal (see [10, pag. 13]), we are going to show that it has a subgroup which is fractal but not strongly fractal.

Consider the subgroup $G = \langle a_{i,i+1} \mid i = 1, \dots, d-1 \rangle \leq H$. To simplify the notation, we write $b_i = a_{i,i+1}$. As a consequence of Lemma 2.3, it is clear that G is self-similar, since $(b_j)_{x_i} \in G$ for each $j = 1, \dots, d-1$ and $i = 1, \dots, d$. Let us see that G is fractal. First observe that, since the element $b_{d-1} b_{d-2} \cdots b_1$ has label $(x_1 x_2 \cdots x_d)$ at the root, G is transitive on the first level so, by Lemma 2.7, it is enough to show that $\psi_{x_1}(\mathrm{Stab}_G(x_1)) = G$.

It then suffices to check that each $b_i \in \psi_{x_1}(\mathrm{Stab}_G(x_1))$. Since $b_i \in \mathrm{Stab}_G(x_1)$ for $i \neq 1$ and in this case $\psi_{x_1}(b_i) = b_i$, it only remains to check that $b_1 \in \psi_{x_1}(\mathrm{Stab}_G(x_1))$. To show this, consider the element $b_1^{b_2 b_1}$. First of all observe that $(b_1^{b_2 b_1})_{(\emptyset)} = (x_1 x_2)^{(x_1 x_2 x_3)} = (x_2 x_3)$ so, $b_1^{b_2 b_1}$ belongs to $\mathrm{Stab}_G(x_1)$. On the

other hand, using (2) we have

$$\begin{aligned}(b_1^{b_2b_1})_{x_1} &= ((b_2b_1)_{(b_2b_1)^{-1}(x_1)})^{-1}(b_1)_{(b_2b_1)^{-1}(x_1)}(b_2b_1)_{(b_2b_1)^{-1}b_1(x_1)} = ((b_2b_1)_{x_3})^{-1}(b_1)_{x_3}(b_2b_1)_{x_3}\\ &= ((b_2)_{x_3}(b_1)_{x_2})^{-1}b_1(b_2)_{x_3}(b_1)_{x_2} = b_1.\end{aligned}$$

We obtain that $\psi_{x_1}(b_1^{b_2b_1}) = b_1$. Thus, we conclude that $\psi_{x_1}(\mathrm{Stab}_G(x_1)) = G$ as desired.

Let us now calculate $\mathrm{Stab}_G(L_1)$. We have $\rho(G) = \langle \rho(b_i) \mid i = 1, \dots, d-1\rangle = S_d$. We know that a presentation of the group S_d can be obtained by considering $\{\tau_i = (i\, i+1)\}_{i=1,\dots,d-1}$ as generators, and taking the relations:

$$\begin{aligned}\tau_i^2 &= 1, && i = 1, \dots, d-1,\\ \tau_i\tau_j &= \tau_j\tau_i, && |i-j| > 1,\\ (\tau_i\tau_{i+1})^3 &= 1, && i = 1, \dots, d-2.\end{aligned}$$

In order to apply Lemma 2.8, let F be the free group generated by $\{\tau_1, \dots, \tau_{d-1}\}$ and $\theta\colon F \longrightarrow S_d$ the epimorphism corresponding to the presentation above. Thus $\ker\theta = \langle \tau_i^2, [\tau_i, \tau_j], (\tau_i\tau_{i+1})^3 \mid i, j = 1, \dots, d-1, |i-j| > 1\rangle^F$. For each $i = 1, \dots, d-1$ we have $b_i \in \rho^{-1}(\theta(\tau_i))$ and the b_i generate the whole group G. We can define $\phi\colon F \longrightarrow G$ by sending τ_i to b_i for each $i = 1, \dots, d-1$. Then ϕ is a surjective homomorphism that makes the diagram commutative. Now, applying the lemma, if

$$S = \{\{b_i^2\}_{i=1,\dots,d-1}, \{(b_ib_{i+1})^3\}_{i=1,\dots,d-2}, \{[b_i, b_j]\}_{|i-j|>1}\},$$

then we obtain that $\mathrm{Stab}_G(L_1) = \langle S\rangle^G$.

To conclude, let us see that $\psi_{x_k}(\mathrm{Stab}_G(L_1)) \neq G$ for some $k = 1, \dots, d$. In fact we will see that this happens for any $k \in \{1, \dots, d\}$. One can check that

$$(b_i^2)_{x_k} = \begin{cases} b_i^2 & \text{if } k \neq i, i+1,\\ 1 & \text{if } k = i, i+1,\end{cases}$$

$$((b_ib_{i+1})^3)_{x_k} = \begin{cases} (b_ib_{i+1})^3 & \text{if } k \neq i, i+1, i+2,\\ b_ib_{i+1} & \text{if } k = i,\\ b_{i+1}b_i & \text{if } k = i+1,\\ b_ib_{i+1} & \text{if } k = i+2,\end{cases}$$

and, for $|i-j| > 1$,

$$([b_i, b_j])_{x_k} = \begin{cases} [b_i, b_j] & \text{if } k \neq i, i+1, j, j+1,\\ 1 & \text{otherwise.}\end{cases}$$

To see the importance of the condition $|i-j| > 1$ in the last case, let us calculate, for example, $[b_i, b_j]_{x_i}$:

$$\begin{aligned}[b_i, b_j]_{x_i} &= (b_i^{-1}b_i^{b_j})_{x_i} = (b_i^{-1})_{x_i}(b_i^{b_j})_{x_{i+1}} = ((b_j)_{b_j^{-1}(x_{i+1})})^{-1}(b_i)_{b_j^{-1}(x_{i+1})}(b_j)_{b_j^{-1}b_i(x_{i+1})}\\ &= ((b_j)_{x_{i+1}})^{-1}(b_i)_{x_{i+1}}(b_j)_{x_i} = b_j^{-1}b_j = 1.\end{aligned}$$

Here, it is important that b_j does not move x_i and x_{i+1}, which happens since $|i-j| > 1$. On the other hand, observe that b_i^2 and $[b_i, b_j]$ when $|i-j| > 1$ are the identity automorphism, because they belong to the first level stabilizer and the sections at the first level are just themselves or the identity.

If $\sigma\colon S_d \longrightarrow \{1,-1\}$ is the homomorphism sending each permutation to its signature, observe that for any $s \in S$ and $k = 1,\dots,d$ we have $\sigma(\psi_{x_k}(s)_{(\emptyset)}) = 1$ because $\psi_{x_k}(s)$ is always a product of an even number of b_i. Then, if we consider $N = \langle \psi_{x_k}(S) \mid k = 1,\dots,d\rangle^G$ we still have that $\sigma(n_{(\emptyset)}) = 1$ for any $n \in N$.

Now, we have $\mathrm{Stab}_G(L_1) = \langle S\rangle^G$ and $\psi_{x_k}(S) \subseteq N$, where N is normal in G. So, by Lemma 2.9, we conclude that $\psi_{x_k}(\mathrm{Stab}_G(L_1)) \subseteq N$. But N cannot be the whole group G because each $n \in N$ has an even permutation at the root and, consequently, $b_i \notin N$ for each $i = 1,\dots,d-1$. In other words, $\rho(N) \subseteq A_d$ while $\rho(G) = S_d$, so $N \neq G$.

4. Strongly fractal groups which are not super strongly fractal

In order to see an example of a group which is strongly fractal but not super strongly fractal, we have to introduce the GGS-groups. These groups are subgroups of Aut T where T is the d-adic tree for $d \geq 2$.

Definition 4.1. Let us consider the rooted automorphism corresponding to $(1\dots d)$ and denote it by a. Given a non-zero vector $e = (e_1,\dots,e_{d-1}) \in (\mathbb{Z}/d\mathbb{Z})^{d-1}$, we define an automorphism $b \in \mathrm{Stab}(L_1)$ by means of $\psi(b) = (a^{e_1},\dots,a^{e_{d-1}},b)$. Then, a *GGS-group* is the group G generated by these two automorphisms a and b.

From now on we consider $d = p$ where p is a prime. First of all, let us see that every GGS-group is strongly fractal. For these groups $\mathrm{Stab}_G(L_1) = \langle b\rangle^G = \langle b, b^a,\dots,b^{a^{p-1}}\rangle$. To simplify notation, we write $b_i = b^{a^i}$.

Lemma 4.2. *Let G be a GGS-group. Then G is strongly fractal.*

Proof. Let us see that G is fractal. Since G is in the Sylow pro-p subgroup of Aut T corresponding to the cycle $(1\dots p)$, this is enough to show that G is strongly fractal because of the discussion after Lemma 2.5. Since $\langle a\rangle$ acts transitively on the first level, according to Lemma 2.7 it suffices to show that $\psi_x(\mathrm{Stab}_G(x)) = G$ for some x in the first level. Observe that conjugating b by powers of a permutes the sections of b at the first level. In other words,

$$\psi(b_i) = (a^{e_{p-i+1}},\dots,a^{e_{p-1}},b,a^{e_1},\dots,a^{e_{p-i}}).$$

Then, since e is non-zero, there is some $e_{p-i+1} \neq 0$ and since $b_1, b_i \in \mathrm{Stab}_G(x_1)$ we obtain that $\psi_{x_1}(\mathrm{Stab}_G(x_1)) \geq \langle b, a^{e_{p-i+1}}\rangle = G$. We conclude that G is strongly fractal. □

Let us consider a GGS-group with constant defining vector. By replacing b with a suitable power of b, we may assume that $e = (1,\dots,1)$.

Proposition 4.3. *Let G be a GGS-group with constant defining vector. Then G is strongly fractal but not super strongly fractal.*

Proof. By the previous lemma it is enough to show that G is not super strongly fractal. In [7, Thm. 2.4] it is shown that $|G : \mathrm{Stab}_G(L_2)| = p^{t+1}$, where t is the rank of the circulant matrix whose first row is $(1, \dots, 1, 0)$. In this case the rank is p. It is also proved in [7, Thm. 2.14] that $|G : \mathrm{Stab}_G(L_1)'| = p^{p+1}$. The mentioned paper is written for p an odd prime, but these two results are also true for $p = 2$. Since $\mathrm{Stab}_G(L_1)/\mathrm{Stab}_G(L_2)$ is abelian we know that $\mathrm{Stab}_G(L_1)' \subseteq \mathrm{Stab}_G(L_2)$, so we conclude that $\mathrm{Stab}_G(L_2) = \mathrm{Stab}_G(L_1)'$. Now $\mathrm{Stab}_G(L_1)' = \langle [b_i, b_j] \mid i, j = 1, \dots, p \rangle^G$. Observe that $\psi([b_i, b_j]) = (1, \dots, 1, \underset{j}{[a, b]}, 1, \dots 1, \underset{i}{[b, a]}, 1 \dots, 1)$. By Corollary 2.11, we conclude that

$$\psi_{x_1}(\mathrm{Stab}_G(L_2)) = \psi_{x_1}(\mathrm{Stab}_G(L_1)') = \langle [a, b], [b, a] \rangle^G = G'.$$

Now again, $\psi([a, b]) = \psi(b_1^{-1}b) = (b^{-1}a, 1, \dots, 1, a^{-1}b)$. By the same argument as before, we have

$$\psi_{x_1}(G') = \psi_{x_1}(\langle [a, b] \rangle^G) = \langle b^{-1}a \rangle^G.$$

But then, for the vertex $u = x_1x_1 \in L_2$, we have that $\psi_u(\mathrm{Stab}_G(L_2)) = \langle b^{-1}a \rangle^G$. It is not hard to see that $G/G' \cong C_p \times C_p$ (see [7, Thm. 2.1]). Since the image of $\langle ba^{-1} \rangle^G$ in G/G' is cyclic, we have $\langle ba^{-1} \rangle^G \neq G$, and G is not super strongly fractal. □

5. Groups which are super strongly fractal

In the same family of GGS-groups, we have examples of groups which are super strongly fractal. More specifically, the GGS-groups which are periodic (or, equivalently, those having defining vector e such that $e_1 + \cdots + e_{p-1} \equiv 0 \pmod{p}$, see [12, Thm. 1]) are examples of super strongly fractal groups.

Proposition 5.1. *Let G be a GGS-group with defining vector $e = (e_1, \dots, e_{p-1})$ such that $e_1 + \cdots + e_{p-1} \equiv 0 \pmod{p}$. Then G is super strongly fractal.*

Proof. By [7, Lem. 3.3] we know that, for $n \geq 3$, $\psi(\mathrm{Stab}_G(L_n)) = \mathrm{Stab}_G(L_{n-1}) \times \overset{p}{\cdots} \times \mathrm{Stab}_G(L_{n-1})$. Since we also know that $\psi_x(\mathrm{Stab}_G(L_1)) = G$ for every $x \in X$, it only remains to show that $\psi_x(\mathrm{Stab}_G(L_2)) = \mathrm{Stab}_G(L_1)$ for each $x \in X$. Since G contains the rooted automorphism a, by Remark 2.13, it is enough to check the condition in one vertex.

Let us consider the element $g = b_1 b_2 \cdots b_{p-1} b$. We have

$$\begin{aligned} \psi(g) &= (ba^{e_1+\cdots+e_{p-1}}, a^{e_1+\cdots+e_{p-1}} b_{e_2+\cdots+e_{p-1}}, a^{e_1+\cdots+e_{p-1}} b_{e_3+\cdots+e_{p-1}}, \dots, a^{e_1+\cdots+e_{p-1}} b) \\ &= (b, b_{e_2+\cdots+e_{p-1}}, b_{e_3+\cdots+e_{p-1}}, \dots, b) \end{aligned}$$

so, we conclude that $g \in \mathrm{Stab}_G(L_2)$. On the other hand, since G is strongly fractal by Corollary 2.11, we have $\psi_{x_1}(\mathrm{Stab}_G(L_2)) \geq \psi_{x_1}(\langle g \rangle^G) = \langle b, b_{e_2+\cdots+e_{p-1}}, b_{e_3+\cdots+e_{p-1}}, \dots, b \rangle^G = \mathrm{Stab}_G(L_1)$, and we have finished. □

In [9, pag. 85], it is said that being strongly fractal implies being super strongly fractal, and also that the first Grigorchuk group is an example of this. It is true that the first Grigorchuk group is super strongly fractal, but it is not a direct consequence of being strongly fractal. The proof of this is similar to the previous example.

Definition 5.2. Let T be the 2-adic tree. The first Grigorchuk group, denoted by $\mathcal{G}$, is the group generated by the automorphisms a, b, c, d defined as:

$$\begin{aligned} a_{(\emptyset)} = (12), \quad \psi(a) = (1,1), \\ b, c, d \in \mathrm{Stab}(L_1), \quad \psi(b) = (a,c), \quad \psi(c) = (a,d), \quad \psi(d) = (1,b). \end{aligned}$$

Proposition 5.3. *The group $\mathcal{G}$ is super strongly fractal.*

Proof. In [1, Thm. 4.3] it is shown that $\psi(\mathrm{Stab}_{\mathcal{G}}(L_n)) = \mathrm{Stab}_{\mathcal{G}}(L_{n-1}) \times \mathrm{Stab}_{\mathcal{G}}(L_{n-1})$, for $n \geq 4$. Since $a \in G$, by Lemma 2.12 it suffices to show that $\langle \psi_{u_n}(\mathrm{Stab}_{\mathcal{G}}(L_n)) \mid u_n \in L_n \rangle = \mathcal{G}$ when $n = 1, 2, 3$.

For $n = 1$ this follows directly from the definition of the elements b, c, d. To see the cases $n = 2$ and $n = 3$, it is easy to calculate and check that $d, (ab)^4, (ac)^4 \in \mathrm{Stab}_{\mathcal{G}}(L_2)$ and that

$$\begin{aligned} \psi_{x_2x_1}(d) &= a, \\ \psi_{x_2x_2}(d) &= c, \\ \psi_{x_2x_2}((ab)^4) &= ad, \\ \psi_{x_2x_2}((ac)^4) &= b. \end{aligned}$$

To conclude, the element $g = (ab)^4(adabac)^2$ belongs to $\mathrm{Stab}_{\mathcal{G}}(L_3)$ and

$$\begin{aligned} \psi_{x_1x_2x_1}(g) &= d, \\ \psi_{x_1x_2x_2}(g) &= ba, \\ \psi_{x_2x_2x_1}(g) &= a, \\ \psi_{x_2x_2x_2}(g) &= c. \end{aligned}$$

This proves that $\mathcal{G}$ is super strongly fractal. □

References

[1] L. Bartholdi and R.I. Grigorchuk, "On parabolic subgroups and Hecke algebras of some fractal groups", *Serdica Math. J.* **28**(1) (2002), 47–90.

[2] L. Bartholdi, R.I. Grigorchuk, and Z. Šuniḱ, "Branch groups", in *Handbook of algebra* **3**, 989–1112, North-Holland, Amsterdam, 2003.

[3] A.M. Brunner and S.N. Sidki, "Abelian state-closed subgroups of automorphisms of m-ary trees", *Groups Geom. Dyn.* **4**(3) (2010), 455–472.

[4] F. Dahmani, "An example of non-contracting weakly branch automaton group", in *Geometric methods in group theory*; *Contemp. Math.* **372**, 219–224. Amer. Math. Soc., Providence, RI, 2005.

[5] D. D'Angeli and A. Donno, "Self-similar groups and finite Gelfand pairs", *Algebra Discrete Math.* **2** (2007), 54–69.

[6] A. Donno, "Gelfand Pairs: from self-similar groups to Markov chains", *PhD thesis, Università degli studi di Roma, La Sapienza*, 2008.

[7] G.A. Fernández-Alcober and A. Zugadi-Reizabal, "GGS-groups: order of congruence quotients and Hausdorff dimension", *Trans. Amer. Math. Soc.* **366**(4) (2014), 1993–2017.

[8] R.I. Grigorchuk, "On Burnside's problem on periodic groups", *Funktsional. Anal. i Prilozhen.* **14**(1) (1980), 53–54.

[9] R.I. Grigorchuk, "Some topics in the dynamics of group actions on rooted trees", *Proceedings of the Steklov Institute of Mathematics* **273**(1) (2011), 64–175.

[10] R.I. Grigorchuk and Z. Šunić, "Self-similarity and branching in group theory", in *Groups St. Andrews 2005*; *London Math. Soc. Lecture Note Ser.* **339**, 36–95, Cambridge Univ. Press, Cambridge, 2007.

[11] N. Gupta and S.N. Sidki, "On the Burnside problem for periodic groups", *Math. Z.* **182**(3) (1983), 385–388.

[12] T. Vovkivsky, "Infinite torsion groups arising as generalizations of the second Grigorchuk group", in *Algebra (Moscow, 1998)*, 357–377, de Gruyter, Berlin, 2000.

REPORTS@SCM

AN ELECTRONIC JOURNAL OF THE
SOCIETAT CATALANA DE MATEMÀTIQUES

Stochasticity conditions for the general Markov model

*Marina Garrote López
Universitat Politècnica de Catalunya
marinagarrotelopez@gmail.com

*Corresponding author

Resum *(CAT)*

En filogenètica sovint es modelitza l'evolució mol·lecular mitjanant models estadístics paramètrics. Usant aquests models es poden deduir relacions polinomials (*invariants filogenètics*) entre les probabilitats teòriques de caràcters observats a les fulles de l'arbre. En aquest article estudiem i corregim alguns resultats teòrics que ens proporcionaran condicions en l'estocasticitat dels paràmetres d'aquests models i que utilitzem per tal de trobar nous invariants filogenètics.

Abstract *(ENG)*

In phylogenetics it is useful to model evolution adopting parametric statistic models. Using these models one is able to deduce polynomial relationships between the theoretical probabilities of characters at the leaves of a phylogenetic tree, known as *phylogenetic invariants*. We revisit and correct some results on stochasticity conditions of the parameters of these models and we find new phylogenetic invariants.

Keywords: *Phylogenetic tree, phylogenetic invariants, topology invariants, general Markov model, joint distribution, tensor.*
MSC (2010): *92D15, 92D20, 14P10, 60J20, 62P10.*
Received: *February 3th, 2016.*
Accepted: *March 9th, 2016.*

Acknowledgement

I would like to thank my advisors Marta Casanellas and Jesús Fernández-Sánchez for having invested a lot of their time in this research. I am very grateful to them for sharing with me their scientific knowledge and for their unconditional support.

Societat Catalana de Matemàtiques

Institut d'Estudis Catalans

1. Introduction

Strong evidences suggest that all living organisms share a common ancestor and therefore, are related by evolutionary relationships. These relationships are usually expressed in the form of a phylogenetic tree.

Nowadays there are more and more mathematicians and statisticians who collaborate with biologists in order to solve the major problems of phylogenetics. Many different areas of mathematics, like statistics, probability, algebra, combinatorics and numerical methods are involved in phylogenetic studies. Even more, recently developed techniques from algebraic geometry have already been used in the study of phylogenetics.

The main goal of phylogenetic reconstruction is recovering the ancestral relationships among a group of current species. In order to reconstruct phylogenetic trees it is necessary to model evolution adopting a parametric statistic model. Using these models one is able to deduce polynomial relationships between the parameters of the model, known as *phylogenetic invariants*. Mathematicians have recently begun to be interested in the study of these polynomials and the geometry of the algebraic varieties that arise in this setting. Furthermore, they have started to use some phylogenetic invariants called *topology invariants* to reconstruct phylogenetic trees; see [4, 8].

The aim of this paper is to understand the relationship between phylogenetics and these algebraic techniques to recover phylogenetic trees from real data. Our main goal is to study and to analyze the characterizations of stochasticity of the points in the algebraic varieties mentioned above, and provided in [5].

The paper is divided into two parts. In the first one, we explain basic concepts on phylogenetics that are already known. We explain what *phylogenetic trees* are from the mathematical standpoint, we describe the general Markov model, and we explain then what *phylogenetic invariants* and *topology invariants* are. Moreover, we define *joint distributions* of a tree and its representation as a tensor. We will define some operations among tensors that will be useful, and their meaning in terms of phylogenetic trees. This part will be developed in Section 2. After that, in Section 3, we will revisit results related to the stochasticity of the parameters of the general Markov model on a tree. One of these results, [5, Theorem 3.2.4], has been restated and the proof rewritten since the statement of the original theorem is not completely correct. We also provide a counterexample to show this; see Counterexample 3.8. Finally, in Theorem 3.11 we present new topology invariants that can be used to design original methods for phylogenetic reconstruction; see [10] for further details.

2. Preliminaries

2.1 Biological preliminaries

Phylogenetics is the study of relationships between different species or biological entities. It studies how species evolve and where contemporary species come from. According to the theory of the biological evolution developed by Darwin (s.XIX), all species of organisms evolve through the natural selection of small variations that increase the individual's ability to compete, survive, and reproduce. We can model these specialization processes with phylogenetic trees. The nodes of these trees represent different species and every branch is an evolutionary process between two species. The leaves of the tree are contemporary species and the root of the tree is the common ancestor of all the species represented on the tree.

Genetic information of each individual is encoded in the DNA of the nucleus of its cells, which is composed of four different simpler units named *nucleotides*. According to the bases forming the nucleotides, they are called adenine (A), cytosine (C), guanine (G) and thymine (T).

Heredity information in a genome is thought to be contained in genes. But DNA sequences of a same gene may look quite different for different species. They contain similar parts but they can also contain some other parts that can not be compared. For that reason the first problem is identifying which part of DNA sequences of different species can be compared. This information is collected in an *alignment*. A sequence alignment is a way of arranging the sequences of DNA to identify regions of similarity that may be a consequence of functional, structural, or evolutionary relationships between the species. We can represent an alignment with a table whose rows are DNA sequences of the species and whose columns correspond to nucleotides evolved from the same nucleotide at the common ancestor of all the species in the table. Alignments are used in many contexts, phylogenetics among them, to see relationships between some species and to reconstruct the phylogenetic tree relating them.

One of the basic objects in a phylogenetic model is a tree T encoding the evolutionary relationships among a given set of species. In this section we introduce some concepts that allow us to deal with these phylogenetic trees following the approach in [2, 3, 6].

Definition 2.1. A *tree* T is a connected graph with no cycles. The *degree* of a vertex is the number of edges incident to it. The vertices of degree 1 are called *leaves* and the set of leaves of T is denoted by $L(T)$. All the other vertices, which have degree at least 2, are *interior nodes* and are designated by the set $Int(T)$. $E(T)$ is the set of the edges of the tree. If all nodes in $Int(T)$ have degree 3, then T is called a *trivalent tree*. A tree is called a *rooted tree* if one vertex has been labelled as "root", and the edges are oriented away from it. A *phylogenetic tree* is a pair (T, ϕ), where T is a tree and $\phi\colon X \to L(T)$ is a one-to-one correspondence between the set of leaves and a finite set of labels denoted be X. The *tree topology* of a phylogenetic tree is the topology of the tree as a labelled graph.

In a phylogenetic tree, the set X represents a set of living species and the tree T shows the ancestral relationships among them. Every edge represents an evolutionary process between two species and if it is rooted, then the root represents the common ancestor to the set of species X. For our purposes, usually X will be taken as the set $\{1, 2, \dots, n\}$. Moreover, two phylogenetic trees T_1 and T_2, with the same set of labels X at the leaves, have the same topology if there is a one-to-one correspondence φ between their vertices respecting adjacency and leaf labelling. If r_1, r_2 are the roots of T_1 and T_2, respectively, then we need to impose $\varphi(r_1) = r_2$.

Remark 2.2. For the rest of the paper, we denote by T_n the set of all possible tree topologies for n-leaf unrooted trivalent trees. Note that n has to be greater or equal than 3 ($|T_3| = 1$). We will denote the three possible topologies of T_4 by $T_{12|34}$, $T_{13|24}$, and $T_{14|23}$; see left hand side of Figure 1.

2.2 Evolutionary models

Evolution is usually modeled adopting a parametric statistical model. That is, evolution is assumed to be a stochastic process, in which nucleotides mutate randomly over time according to certain probabilities. Moreover we assume that DNA substitutions occur randomly and the nucleotides observed in the DNA sequences are independent and identically distributed.

We associate a discrete random variable X_i to each node i of T such that X_i can take κ different states. We denote by $\mathcal{K}$ this set of states. Usually $\mathcal{K}$ is the set of the four nucleotides in DNA, which are

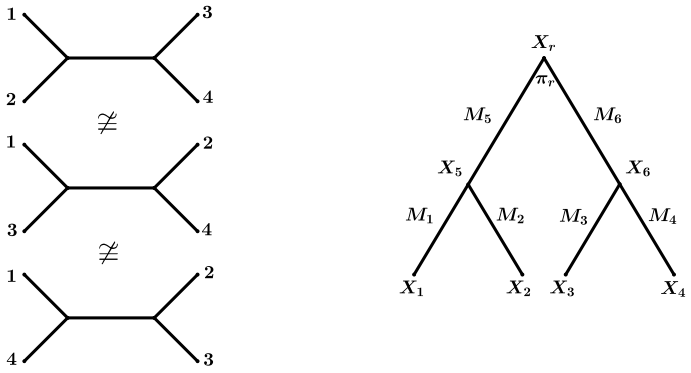


Figure 1: *Left:* the three topologies of T_4, say $T_{12|34}$, $T_{13|24}$, and $T_{14|23}$. *Right:* a Markov process on a rooted 4-leaf tree given by a distribution vector π and transition matrices $\{M_1, \dots, M_6\}$.

denoted by their first letter, so $\mathcal{K} = \{\mathtt{A}, \mathtt{C}, \mathtt{G}, \mathtt{T}\}$ and $\kappa = 4$. Since DNA sequences of the contemporary species are known, we say that random variables at the leaves are observed. However, we do not have any information about the ancestral species, that is why random variables at the interior nodes are hidden. For a tree T with leaves $1, 2, \dots, n$, $X = (X_1, X_2, \dots, X_n)$ represents the joint distribution vector of the leaves. Each column of an alignment is an observation of this vector of random variables.

Hereafter we introduce a Markov process in a rooted tree T. First, we define a vector $\pi = (\pi_1, \dots, \pi_\kappa)$, the distribution of X_r which is the random variable associated to the root r and satisfying that all entries are nonnegative and $\sum_i \pi_i = 1$. If $\mathcal{K} = \{\mathtt{A}, \mathtt{C}, \mathtt{G}, \mathtt{T}\}$, we interpret these entries as the probabilities that an arbitrary site in the DNA sequence at the root is occupied by the corresponding base. A second set of parameters is associated to the evolutionary process that occurs in every edge. For each edge e we associate a $\kappa \times \kappa$ matrix M_e, called *substitution* or *transition* matrix.

Definition 2.3. A *transition matrix* is a $\kappa \times \kappa$ matrix M_e associated to each edge of a phylogenetic tree. Every entry is the conditional probability $P(x|y, e)$ that a state y at the parent node of e had been substituted by a state x at its child, during the evolutionary process along the edge e. Since each row contains the probabilities of the κ possible changes that can occur in an evolutionary process, rows of M_e sum up to 1. These matrices M_e are also called *Markov matrices* or *row stochastic matrices*.

We consider a Markov process on T given by π and the matrices $\{M_E\}_{e \in E(T)}$. In particular, the substitutions on two adjacent branches at a node v are independent given the state at v.

The substitution probabilities on a given edge depend only on the state at the parent node. Besides, we only have observations of the random variables at the leaves so, ours is a *hidden* Markov process. According to the shape of the transition matrices different *models* are defined, but in this paper we focus on the general Markov model, that is, transition matrices do not satisfy any other restriction.

Example 2.4. On the right hand side of Figure 1, a Markov process on a phylogenetic tree is represented. The $X_i's$ are random variables associated to the leaves, the $M_i's$ are the transition matrices, and π_r is the root distribution. Under the general Markov model, we have 3×4 free parameters for each transition matrix and 3 free parameters for the vector π_r. Therefore, this model has $3 \cdot 4 \cdot 6 + 3 = 75$ free parameters.

In what follows we describe how to compute the joint probability of observing states $x_1, x_2, \dots, x_n$ at the leaves according to the Markov process we have described.

We denote by $p_{x_1,\dots,x_n}$ the joint distribution at the leaves of a rooted phylogenetic tree T, $p_{x_1,\dots,x_n} = \text{Prob}(X_1 = x_1, X_2 = x_2, \dots, X_n = x_n)$. We define P as a κ^n-dimensional vector whose components are the joint probabilities $p_{x_1,\dots,x_n}$, $P = (p_{x_1,\dots,x_n})_{x_1,\dots,x_n \in \mathcal{K}}$.

Since the evolutionary processes follow a Markov process, they are independent and we can express $p_{x_1,\dots,x_n}$ in terms of the transition matrices,

$$p_{x_1,\dots,x_n} = \sum_{x_r,(x_v)_{v\in \text{Int}(T)}} \prod_{e\in E(T)} M_e(x_{a(e)}, x_{d(e)}), \tag{1}$$

where $x_r \in \mathcal{K}$ is a state of the root, $x_{a(e)} \in \mathcal{K}$ is a state of the parent node of the edge e, and $x_{d(e)} \in \mathcal{K}$ is the state of the descendant node of the edge e. If e is a terminal edge ending at the leaf i then $x_{d(e)} = x_i$. Every entry of P can be seen as a polynomial with the parameters of the model $\mathcal{M}$ as variables.

Example 2.5. We compute now the joint distribution p_{x_1,x_2,x_3,x_4} of the tree presented on the right hand side of Figure 1. Using equation (1) we get

$$p_{x_1,x_2,x_3,x_4} = \sum_{x_r\in\mathcal{K}} \sum_{x_5\in\mathcal{K}} \sum_{x_6\in\mathcal{K}} \pi_{x_r} \cdot M_5(x_r, x_5) \cdot M_1(x_5, x_1) \cdot M_2(x_5, x_2) \cdot M_6(x_r, x_6) \cdot M_3(x_6, x_3) \cdot M_4(x_6, x_4).$$

2.3 Phylogenetic invariants and flattening

It is known that there exist many algebraic relations among the components of the joint distribution P; see [4, 6, 7, 9].

Since components of P are polynomials in the model parameters, we can associate to the tree a polynomial map $\varphi_T \colon \mathbb{C}^d \to \mathbb{C}^{\kappa^n}$ mapping any d-tuple of parameters to a distribution vector of the κ^n possible observations at the leaves of T. More precisely, we define the map

$$\begin{array}{rcl} \varphi_T \colon \mathbb{C}^d & \longrightarrow & \mathbb{C}^{\kappa^n} \\ (\pi, \{M_e\}_{e\in E(T)}) & \mapsto & P = (p_{x_1,x_1,\dots,x_1}, p_{x_1,x_1,\dots,x_2}, p_{x_1,x_1,\dots,x_3}, \dots, p_{x_\kappa,x_\kappa,\dots,x_\kappa}), \end{array} \tag{2}$$

where d is the number of free parameters of the model and each component $p_{x_1,\dots,x_n}$ is expressed in terms of the root distribution π and the transition matrices M_e according to the expression (1).

Remark 2.6. Notice that, to read the parameters as probabilities, we should restrict to nonnegative real numbers. Analogously, the points in the image of φ_T represent a joint distribution only if they lie in the standard $(\kappa^n - 1)$-simplex. However, in order to use techniques from algebraic geometry, we abandon temporally these restrictions and work over the complex field. We will consider *complex parameters* and complex parametrization map in general, but we will refer to *stochastic parameters* to the ones coming from the original probabilistic model (that is, all the components of π and the entries of the transition matrices M_i are nonnegative).

We introduce now an algebraic variety in $\mathbb{C}^{\kappa^n}$ which contains the set of image points of φ_T.

Definition 2.7. The *phylogenetic variety* associated to a tree T, denoted by $\mathcal{V}(T)$, is the smallest algebraic variety containing the image $\text{Im}\,\varphi_T$.

Remark 2.8. The image set $\text{Im}\,\varphi_T$ is not, in general, an algebraic variety, but it defines a dense open subset in $\mathcal{V}(T)$ under Zariski topology. The ideal $I(\text{Im}\,\varphi_T)$ of all polynomial relations in $\mathbb{C}[P_{x_1,\dots,x_n}]$ of the points in $\text{Im}(\varphi_T)$ coincides with the ideal of the variety $\mathcal{V}(T)$. We will denote it by $I(T)$. It can be proved that $\mathcal{V}(T)$ is independent from the node chosen as root in T; see [1] for a complete proof.

Definition 2.9. The polynomials in $I(T)$ are called *phylogenetic invariants* of T. If f is a polynomial in $I_{\mathcal{M}}(T)$ that does not belong to $I(T')$ for some other tree topology T' on n leaves, then f is called a *topology invariant* of T.

Definition 2.10. Let $A|B$ be a partition of the leaves of a tree T, that is $A, B \subseteq L(T)$, with $|A|, |B| \geq 2$ such that $L(T) = A \cup B$ and $A \cap B = \emptyset$. Let $\tilde{X}_A = (x_i)_{i\in A}$ and $\tilde{X}_B = (x_j)_{j\in B}$ be the random variables associated to A and B. Then $\tilde{X}_A$ and $\tilde{X}_B$ can take $a := \kappa^{|A|}$ and $b := \kappa^{|B|}$ states, respectively. Given a vector $P \in \mathbb{C}^{\kappa^n}$ we define the *flattening* $Flatt_{A|B}(P)$ as the $a \times b$ matrix whose entries are the joint distributions of all possible observations of $\tilde{X}_A$ and $\tilde{X}_B$:

$$Flatt_{A|B}(P) = \begin{array}{c} \\ \text{States of } \tilde{X}_A \end{array} \begin{array}{c} \text{States of } \tilde{X}_B \\ \begin{pmatrix} p_{u_1v_1} & p_{u_1v_2} & \cdots & p_{u_1v_b} \\ p_{u_2v_1} & p_{u_2v_2} & \cdots & p_{u_2v_b} \\ \vdots & \vdots & \ddots & \vdots \\ p_{u_av_1} & p_{u_av_2} & \cdots & p_{u_av_b} \end{pmatrix} \end{array}.$$

This matrix allows us to state the following result, which gives us some topology invariants associated to a 4-leaf tree.

Theorem 2.11 (Casanellas–Fernández-Sánchez, [8])**.** *Let T be a tree, $A|B$ a bipartition of $L(T)$ and $P = \varphi_T(\pi, \{M_e\}_{e\in E(T)})$. Then the $(\kappa+1)\times(\kappa+1)$ minors of $Flatt_{A|B}(P)$ vanish if $A|B$ is induced by removing an edge of T. Otherwise, $Flatt_{A|B}(P)$ has rank $\geq \kappa^2$ for general P. Therefore, the $(\kappa+1)\times(\kappa+1)$ minors of $Flatt_{A|B}(P)$ are topology invariants for the tree T.*

There is a more algebraic way of viewing the joint distribution at the leaves of a phylogenetic tree, which will be really useful in this article.

Let $\mathcal{W} := \mathbb{C}^{\kappa}$ be regarded as a vector space. We identify the canonical basis of $\mathcal{W}$ with the set $\mathcal{K}$. Then, the natural basis of $\mathcal{W} \otimes \overset{n)}{\cdots} \otimes \mathcal{W}$ is $\{x_1 \otimes \cdots \otimes x_n\}_{x_1,\ldots,x_n\in\mathcal{K}}$. For instance, if $\mathcal{K} = \{\mathtt{A}, \mathtt{C}, \mathtt{G}, \mathtt{T}\}$, the natural basis of $\mathcal{W} \otimes \mathcal{W} \otimes \mathcal{W}$ is $\{\mathtt{A}\otimes\mathtt{A}\otimes\mathtt{A},\ \mathtt{A}\otimes\mathtt{A}\otimes\mathtt{C},\ \ldots,\ \mathtt{T}\otimes\mathtt{T}\otimes\mathtt{T}\}$. Back to the description of the joint distribution $P = (p_{x_1,\ldots,x_n})_{x_1,\ldots,x_n\in\mathcal{K}}$ in the phylogenetic framework, we can think of P as a n-tensor in $\mathcal{W} \otimes \overset{n)}{\cdots} \otimes \mathcal{W}$ whose components in the natural basis above are $P = (p_{x_1,\ldots,x_n})_{x_1,\ldots,x_n\in\mathcal{K}}$:

$$P = \sum_{x_1,\ldots,x_n\in\mathcal{K}} p_{x_1,\ldots,x_n} x_1 \otimes \cdots \otimes x_n.$$

Each factor in $\mathcal{W} \otimes \overset{n)}{\cdots} \otimes \mathcal{W}$ corresponds to one specie so, in order to make species apparent in this tensor product, we denote it as $\mathcal{W}_1 \otimes \cdots \otimes \mathcal{W}_n$, where $\mathcal{W}_i = \mathcal{W}$ for every $i = 1, \ldots, n$. If we view the vector of joint distribution P as a tensor in $\mathcal{W}_1 \otimes \cdots \otimes \mathcal{W}_n$ then, keeping the notation of Definition 2.10, the flattening $Flatt_{A|B}(P)$ is the image of P via the isomorphism

$$\begin{array}{ccccc} \mathcal{W}_1 \otimes \cdots \otimes \mathcal{W}_n & \cong & Hom(\bigotimes_{i\in A} \mathcal{W}_i, \bigotimes_{j\in B} \mathcal{W}_j) & \cong & M_{a\times b}(\mathbb{C}), \\ P & & \longmapsto & & Flatt_{A|B}(P) \end{array}$$

where $M_{a\times b}(\mathbb{C})$ is the space of all $a \times b$ matrices with complex entries.

Notation 2.12. For the rest of the paper, given a vector $\mathbf{v} \in \mathbb{C}^\kappa$, $\mathbf{v}(i)$ will be the i-th component of $\mathbf{v}$ relative to the canonical basis $\{\mathbf{e_1}, \dots, \mathbf{e_\kappa}\}$ of $\mathbb{C}^\kappa$, and we will write $\mathbf{1}$ for $(1, \dots, 1)$. Moreover, we will call an *n-tensor* to the tensors $P \in \mathbb{C}^\kappa \otimes \overset{n)}{\cdots} \otimes \mathbb{C}^\kappa$, and it will be convenient to write $P(x_1, \dots, x_n)$ for the component $p_{x_1,\dots,x_n}$.

Definition 2.13. Given an n-tensor P, an integer $i \in \{1, \dots, n\}$ and a vector $\mathbf{v} \in \mathbb{C}^\kappa$, we define $P *_i \mathbf{v}$ the $(n-1)$-tensor given by $(P *_i \mathbf{v})(j_1, \dots, j_{i-1}, j_{i+1}, \dots, j_n) = \sum_{j_i=1}^{\kappa} v(j_i) P(j_1, \dots, j_i, \dots, j_n)$. We also define the *l-th slice of P in the i-th index* by $P_{\dots l \dots} = P *_i \mathbf{e_l}$, The *i-th marginalization of P* is defined as $P_{\dots + \dots} = P *_i \mathbf{1}$. Given a $\kappa \times \kappa$ matrix M, we define the *n-tensor* $P *_i M$ by

$$(P *_i M)(j_1, \dots, j_n) = \sum_{l=1}^{\kappa} P(j_1, \dots, j_{i-1}, l, j_{i+1}, \dots, j_n) M(l, j_i). \tag{3}$$

Remark 2.14. From now on, we consider the 2-tensors as $\kappa \times \kappa$ matrices via the isomorphism

$$P = \sum P(j_1, j_2) \mathbf{e_{j_1}} \otimes \mathbf{e_{j_2}} \leftrightarrow (P(j_1, j_2))_{j_1, j_2},$$

where rows of the matrix are indexed by the first component, and columns by the second.

3. Theoretical results

3.1 Transforming tensors

In this section we state some technical results related to marginalizations and slices of tensors that arise from stochastic parameters of the general Markov model on a tree T. For a complete proof of these results see [10].

Lemma 3.1. *Let P be a 3-tensor in the image of parameters for the general Markov model, $P = \varphi(\pi, \{M_1, M_2, M_3\})$, where T is a trivalent 3-leaf tree. Then, the three possible marginalizations of P are given by*

$$P_{..+} = M_1^t diag(\pi) M_2, \quad P_{.+.} = M_1^t diag(\pi) M_3, \quad P_{+..} = M_2^t diag(\pi) M_3. \tag{4}$$

And the slices of P are

$$P_{..i} = M_1^T diag(M_3 \mathbf{e_i}) diag(\pi) M_2, \quad P_{.i.} = M_1^T diag(M_2 \mathbf{e_i}) diag(\pi) M_3, \quad P_{i..} = M_2^T diag(M_1 \mathbf{e_i}) diag(\pi) M_3. \tag{5}$$

Corollary 3.2. *Let P be a tensor arising from parameters of the general Markov model on T with tree topology $T_{12|34}$, $P = \varphi_{T_{12|34}}(\pi; M_1, M_2, M_3, M_4, M_5)$ (see the left hand side of Figure 2). Then the double marginalizations $P_{+..+}$, $P_{+.+.}$, $P_{.+.+}$ and $P_{.++.}$ can be computed in terms of the transition matrices as follows:*

$$P_{+..+} = M_2^T diag(\pi) M_5 M_3, \quad P_{+.+.} = M_2^T diag(\pi) M_5 M_4,$$
$$P_{.+.+} = M_1^T diag(\pi) M_5 M_3, \quad P_{.++.} = M_1^T diag(\pi) M_5 M_4.$$

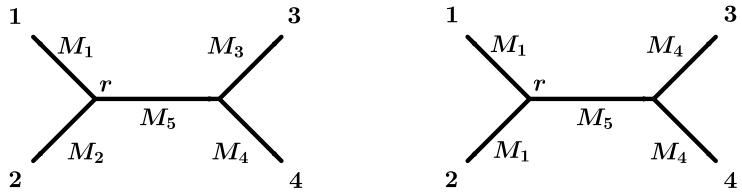


Figure 2: *Left:* Rooted 4-leaf tree $T_{12|34}$ with transition matrices $\{M_1, M_2, M_3, M_4, M_5\}$. *Right:* Rooted 4-leaf tree $T_{12|34}$ with transition matrices $\{M_1, M_1, M_4, M_4, M_5\}$.

The following lemma describes how, given a tensor in the image of φ_T for a 4-leaf tree T, we can produce a new tensor still in $\mathrm{Im}\varphi_T$. This is done by multiplying the original tensor with a matrix (in the sense of (3)), which has the effect of changing the transition matrix of an exterior edge of the tree.

Lemma 3.3. *Let P be a 4-tensor for the general Markov model, $P = \varphi_T(\pi; M_1, \dots, M_5)$. If M_i is non singular for some $i = 1, 2, 3, 4$, then the tensor $\bar{P} = P *_i (M_i^{-1} M)$ is the image of the same parameters as P except for M_i which has been replaced by M.*

3.2 Stochasticity conditions

In this section we will discuss some theoretical results that will allow us to provide some conditions to ensure that a tensor of a joint distribution comes from stochastic parameters.

Definition 3.4. A set $\{\pi, \{M_e\}_{e \in E(T)}\}$ of stochastic parameters for the general Markov model on a tree T with root r is called *nonsingular* if

(i) at every node j of T the distribution of the random variable X_j has no zero entry;

(i) the matrix M_e of every edge e is nonsingular.

Remark 3.5. For stochastic parameters and assuming (ii), condition (i) in the previous definition is equivalent to requiring that the root distribution π_r has no zero entry.

The following result has been proved in [5]. As we do not use it specifically, we do not include the proof here.

Theorem 3.6 (Allman–Rhodes–Taylor, [5])**.** *Let P be a (either real or complex) 3-tensor. Then, P arises from nonsingular parameters for the general Markov model with κ parameters on the 3-leaf tree if and only if the following conditions hold:*

(i) $f_i(P; x) \neq 0$ for an arbitrary vector x and some $i = 1, 2, 3$, where $f_i(P; x) = \det H_x((\det(P *_i x)))$ and H_x denotes the Hessian operator;

(ii) $det(P *_i 1) \neq 0$ for $i = 1, 2, 3$.

We want to find a similar characterization of P for stochastic parameters. That is, we want to find some conditions allowing us to distinguish when a tensor P is the image of positive real parameters.

Theorem 3.7. *Let $P = \varphi_T(\pi, \{M_1, M_2, M_3\})$ be a 3-tensor with π, $\{M_i\}_i$ having real entries. Then,*

(1) P is the image of nonsingular stochastic parameters for the general Markov model on the 3-leaf tree if and only if its components are nonnegative, they sum up to 1, conditions (i) and (ii) from Theorem 3.6 are satisfied, and

(iii) the matrix

$$det(P_{..+})P_{+..}^T adj(P_{..+})P_{.+.} \tag{6}$$

is positive definite, and the following matrices are positive semidefinite for $i = 1, \dots, \kappa$

$$det(P_{..+})P_{i..}^T adj(P_{..+})P_{.+.}, \quad det(P_{..+})P_{+..}^T adj(P_{..+})P_{.i.}, \quad det(P_{+..})P_{.+.} adj(P_{+..})P_{..i}^T. \tag{7}$$

(2) P is the image of nonsingular real positive parameters if and only if its components are positive, they sum up to one, conditions (i) and (ii) are satisfied, and

(iii') all matrices in (6) and (7) are positive definite.

In both cases, the nonsingular parameters are unique up to label swapping.

Proof. The proof of this theorem is essentially the same as in [5], but for real parameters. Let P be an arbitrary nonnegative 3-tensor whose components sum up to 1. Assuming (i) and (ii) and using Theorem 3.6, P is the image of nonsingular parameters. We want to see that condition (iii) is equivalent to these parameters being nonnegative. To this aim, we are going to analyze what is the meaning of expressions (6) and (7).

Let $\bar{P} = P_{+..}P_{..+}^{-1}P_{.+.}$, using expressions proved in Lemma 3.1 we compute

$$\begin{aligned} \bar{P} = P_{+..}^T P_{..+}^{-1} P_{.+.} &= (M_2^T diag(\pi) M_3)^T (M_1^T diag(\pi) M_2)^{-1} (M_1^T diag(\pi) M_3) \\ &= M_3^T diag(\pi) M_3. \end{aligned} \tag{8}$$

This is a well defined symmetric matrix since $P_{..+}$ is nonsingular. Since M_3 is real, $\bar{P}$ is a positive definite matrix if and only if

$$x^T \bar{P} x = x^T M_3^T \mathrm{diag}(\pi) M_3 x = (M_3 x)^T \mathrm{diag}(\pi)(M_3 x) > 0, \quad \forall x \neq 0.$$

Since M_3 is nonsingular, it can be understood as a change of basis and hence $\bar{P}$ is positive semidefinite if and only if the entries of $\mathrm{diag}(\pi)$ are all positive. We clear denominators and obtain an algebraic expression multiplying this matrix by the square of the appropriate nonzero determinant. It follows that (6) is positive definite if and only if π is positive.

Using the expressions in Lemma 3.1, we have

$$\begin{aligned} P_{i..}^T P_{..+}^{-1} P_{.+.} &= (M_2^T diag(M_1 \mathbf{e_i}) M_3)^T (M_1^T diag(\pi) M_2)^{-1} (M_1 diag(\pi) M_3) = \\ &= M_3^T diag(\pi) diag(M_1 \mathbf{e_i}) M_3. \end{aligned}$$

This matrix is also symmetric, and it is positive semidefinite if and only if the entries of $diag(\pi)diag(M_1\mathbf{e_i})$ are nonnegative. Since π is a positive vector, we need the i-th column of M_1 being nonnegative. Using the

matrices $P_{+..}^T P_{..+}^{-1} P_{.i.}$ and $P_{..+}^T P_{+..}^{-1} P_{..i}$ we can also impose the conditions of the i-th column of M_2 and M_3 being nonnegative. This proves (1).

If the matrices of (6) and (7) are positive definite, we can repeat this proof but requiring positiveness of the parameters. This proves (2).

In order to clear denominators and obtain an algebraic expression, we multiply all these matrices by the square of the appropriate nonzero determinant which does not change the sign and gives us expressions (6) and (7). □

Counterexample 3.8. In paper [5], Theorem 3.7 is announced for general tensors P, that is, for $P = \varphi_T(\pi, \{M_1, M_2, M_3\})$ where π, M_1, M_2 and M_3 are complex. But we provide here a counterexample to show that if M_3 is not real, $diag(\pi)$ being positive does not imply $\bar{P} = M^T diag(\pi) M$ being positive definite; see (8). For $\kappa = 2$ let us consider the matrices

$$D = \frac{1}{2}\begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix} \quad \text{and} \quad M = \frac{1}{4}\begin{pmatrix} 2+i & 2-i \\ 2-i & 2+i \end{pmatrix}.$$

However, the matrix $M^T D M = \frac{1}{16}\begin{pmatrix} 3 & 5 \\ 5 & 3 \end{pmatrix}$ is not positive definite.

Moreover, the reverse implication is not true either. For instance, for the positive definite matrix

$$\bar{P} = M^T D M = \begin{pmatrix} 8 & 0 \\ 0 & 8 \end{pmatrix},$$

we have the following decomposition, where D is not positive: $D = \begin{pmatrix} -1 & 0 \\ 0 & 4 \end{pmatrix}$, $M = \begin{pmatrix} 2i & -2i \\ 1 & 1 \end{pmatrix}$.

Due to this counterexample we are forced to restrict the statement of the above theorem to the case of real matrices.

Assuming now that an n-tensor P arises from nonsingular parameters on a tree, we would like to give some semialgebraic conditions that are satisfied if and only if P comes from stochastic parameters. If we consider marginalizations of P to three variables and using Theorem 3.7, we can derive conditions that hold when the root distribution and the product of matrices associated to any path from an interior node to a leaf are stochastic. Nevertheless, we need some extra conditions to guarantee matrices of the interior edges being stochastic.

The following result gives us a condition for all parameters of the 12|34 tree being stochastic.

Theorem 3.9 (Allman–Rhodes–Taylor, [5]). *Let P be a 4-tensor. Suppose P arises from nonsingular real parameters for the general Markov model on $T_{12|34}$. If the marginalizations $P_{+...}$ and $P_{...+}$ arise from stochastic parameters and, moreover, the $\kappa^2 \times \kappa^2$ matrix*

$$det(P_{+..+}) det(P_{.+.+}) Flatt_{13|24}\Big(P *_2 \big(adj(P_{+..+}^T) P_{.+.+}^T\big) *_3 \big(adj(P_{.+.+}) P_{.++.}\big)\Big) \tag{9}$$

is positive semidefinite, then P arises from stochastic parameters.

Proof. The root r is placed at the interior node near leaves 1 and 2, as we can see in the tree presented on the left of Figure 2. Let M_i, $i = 1, 2, 3, 4$, be the complex matrix associated to the edges leading to

leaves, M_5 the matrix on the internal edge, and π the root distribution. The rows of these matrices sum up to 1. We define the four matrices

$$\begin{array}{ll} N_{32} = P^T_{+..+} = M_3^T M_5^T \, diag(\pi) M_2, & N_{31} = P^T_{.+.+} = M_3^T M_5^T \, diag(\pi) M_1, \\ N_{14} = P_{.++.} = M_1^T \, diag(\pi) M_5 M_2, & N_{13} = P_{.+.+} = M_1^T \, diag(\pi) M_5 M_3. \end{array} \tag{10}$$

We define now a tensor $\bar{P}$ arising from the same parameters as P except that M_2 has been replaced by M_1 (see Lemma 3.3) and, similarly, a tensor $\tilde{P}$ arising from the same parameters as $\bar{P}$ but with M_4 instead of M_3:

$$\bar{P} = P *_2 N_{32}^{-1} N_{31} = P *_2 M_2^{-1} M_1, \quad \tilde{P} = \bar{P} *_3 N_{13}^{-1} N_{14} = \bar{P} *_3 M_3^{-1} M_4. \tag{11}$$

We can express

$$Flat_{13|24}(P) = (M_1 \otimes M_3)^T D (M_2 \otimes M_4), \tag{12}$$

where D is the diagonal matrix containing the κ^2 entries of $diag(\pi)M_5$; see [10] for further details. Since $\tilde{P}$ arises from the same parameters that P except that M_2 has been replaced by M_1 and M_3 by M_4, we can write $Flat_{13|24}(\tilde{P}) = (M_1 \otimes M_4)^T D (M_1 \otimes M_4)$.

Since the 3-marginalization arises from stochastic parameters, M_1 and M_4 are nonsingular and the components of π are positive. Thus, $M_1 \otimes M_4$ is also nonsingular. All principal minors of $Flat_{13|24}(\tilde{P})$ are nonnegative if and only if $Flat_{13|24}(\tilde{P})$ is positive semidefinite. Then we have to require the entries of D to be nonnegative and so, since π has positive components, we can ensure that M_5 has nonnegative entries. By multiplying $Flat_{13|24}(\tilde{P})$ by the square of the appropriate nonzero determinant, we clear denominators and obtain the algebraic expressions stated in the theorem. □

Remark 3.10. The theoretical results proved in this section complement the algebraic description of the model (given by topology invariants) with a semialgebraic description of the points with stochastic sense. In other words, as well as finding polynomials vanishing on the image of the parametrization map, we have found polynomial inequalities sufficing to characterize the stochastic image.

The conditions of matrices being positive definite/semidefinite can be expressed as semialgebraic conditions using Sylvester's criterion, which claims that a real symmetric matrix is positive definite (resp., positive semidefinite) if and only its *leading* principal minors are positive (resp., nonnegative).

On the other hand, the replacements of inverses in (11) by adjoint matrices in (9) is not only done in order to have semialgebraic conditions, but also to avoid dealing with the inverse of ill conditioned matrices.

Let P be the tensor used in Theorem 3.9 and $\tilde{P}$ the one constructed in (11). Since $\tilde{P}$ arises from the same parameters that P except that M_2 has been replaced by M_1 and M_3 by M_4, it is the joint distribution of the tree presented on the right hand side of Figure 2. Observing the symmetry of the exterior transition matrices we can state the following result.

Theorem 3.11. *Let P be a 4-tensor whose components sum up to 1. Suppose that*

$$P = \varphi_T(\pi, M_1, M_2, M_3, M_4, M_5),$$

with $T = T_{12|34}$, and let $\tilde{P}$ be constructed as in (11). Then,

$$Flat_{13|24}(\tilde{P}) = Flat_{14|23}(\tilde{P}) \quad \text{and} \quad Flatt_{12|34}(\tilde{P}) \neq Flatt_{13|24}(\tilde{P}). \tag{13}$$

In particular, the equality of matrices

$$\begin{aligned} & det(P_{+..+})det(P_{.+.+})Flatt_{13|24}\Big(P *_2 \left(adj(P_{+..+}^T)P_{.+.+}^T\right) *_3 \left(adj(P_{.+.+})P_{.++.}\right)\Big) = \\ = \; & det(P_{+..+})det(P_{.+.+})Flatt_{14|23}\Big(P *_2 \left(adj(P_{+..+}^T)P_{.+.+}^T\right) *_3 \left(adj(P_{.+.+})P_{.++.}\right)\Big) \end{aligned}$$

gives rise to 256 *topology invariants of degree* 17*.*

Proof. Using (12), and the fact that, in $\tilde{P}$, M_2 has been replaced by M_1, and M_3 by M_4, we have

$$Flat_{13|24}(\tilde{P}) = (M_1 \otimes M_4)^T D(M_1 \otimes M_4) = Flat_{14|23}(\tilde{P}). \tag{14}$$

In contrast, $Flatt_{12|34}(\tilde{P}) = \bar{M}_1^T diag(\pi)\bar{M}_4$, where $\bar{M}_1(x_i,(x_j,x_k)) = M_1(x_i,x_j)M_1(x_i,x_k)$, $\tilde{M}_4(x_i,(x_j,x_k)) = \sum_{l=1}^{\kappa} M_5(x_i,x_l)M_4(x_l,x_j)M_4(x_l,x_k)$, is, in general, not equal to (14).

The expression $Flat_{13|24}(\tilde{P}) = Flat_{14|23}(\tilde{P})$ provides 16×16 equalities between entries. By (9), these entries are algebraic expressions in terms of components of P. Moreover, because of (13), these equalities are not satisfied by any distribution arising from a tree and then they are topology invariants.

Finally, regarding (9), we infer the degree of these expressions in the components of P:

(i) the two determinants have degree 4 each, which makes degree 8;

(ii) the components of the tensors $adj(P_{+..+}^T)P_{.+.+}^T$ and $adj(P_{.+.+})P_{.++.}$ have degree 4.

The $*$ operation adds degrees, so we obtain a tensor of degree $1+4+4=9$ before applying $Flat_{13|24}(\cdot)$. Altogether gives a tensor with components of degree $8+9=17$. □

4. Conclusions

In this paper, we have seen that the conditions of stochasticity on the parameters from Theorem 3.9 are enough to ensure that the 4-tensor arising from real nonsingular parameters under the general Markov model comes from stochastic parameters. From these conditions we have been able to find new topology invariants. So, we can extract the following conclusions:

(i) we have disentangled the theoretical results of stochastic conditions of the parameters and we have provided a counterexample to an error in a proof of [5] as well;

(ii) using the ideas from the proof of Theorem 3.9 we have provided 256 topology invariants of degree 17.

However, there is still further research to do:

(i) check whether the new topology invariants we found are sufficient to describe the phylogenetic algebraic variety;

(ii) check if these conditions can be used with real data, in order to give new information that can be used in some phylogenetic reconstruction method.

References

[1] E.S. Allman and J.A. Rhodes, "Phylogenetic invariants for the general Markov model of sequence mutation", *Math. Biosci.* **186**(2) (2003), 113–144.

[2] E.S. Allman and J.A. Rhodes, "Mathematical models in biology, an introduction", Cambridge University Press (2004). ISBN 0-521-52586-1.

[3] E.S. Allman and J.A. Rhodes, "The mathematics of phylogenetics", University of Alaska Fairbanks (2005).

[4] E.S. Allman and J.A. Rhodes, "Phylogenetic invariants", in *Reconstructing evolution*, Oxford Univ. Press, Oxford (2007), 108–146.

[5] E.S. Allman, J.A. Rhodes, and A. Taylor, "A semialgebraic description of the general Markov model on phylogenetic trees", *Preprint* (2012), `http://adsabs.harvard.edu/abs/2012arXiv1212.1200A`.

[6] M. Casanellas, "Algebraic tools for evolutionary biology", *La Gaceta de la RSME* **15** (2012), 521–536.

[7] M. Casanellas and J. Fernández-Sánchez, "Reconstrucción filogenética usando geometría algebraica", *Arbor. Ciencia, pensamiento, cultura* **96** (2010), 207–229.

[8] M. Casanellas and J. Fernández-Sánchez, "Relevant phylogenetic invariants of evolutionary models", *Journal de Mathématiques Pures et Appliquées* **96** (2011), 207–229.

[9] N. Eriksson, "Tree construction using singular value decomposition", in L. Pachter and B. Sturmfels, editors, "Algebraic Statistics for computational biology" (chapter 19), Cambridge University Press (2005), 347–358.

[10] M. Garrote, "Multilinear algebra for phylogenetic reconstruction", Master's thesis, Universitat Politècnica de Catalunya (2015).

Reports@SCM
Volume 2, number 1, 2016

Table of Contents