REPORTS@SCM

AN ELECTRONIC JOURNAL OF THE SOCIETAT CATALANA DE MATEMÀTIQUES

Volume 7, num. 1, 2022

ISSN (electronic edition): 2385 - 4227

http://reportsascm.iec.cat

Institut d'Estudis Catalans

REPORTS@SCM

AN ELECTRONIC JOURNAL OF THE
SOCIETAT CATALANA DE MATEMÀTIQUES

Volume 7, number 1

December 2022

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

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

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

Reports@SCM
Volume 7, number 1, 2022

Table of Contents

REPORTS@SCM

AN ELECTRONIC JOURNAL OF THE
SOCIETAT CATALANA DE MATEMÀTIQUES

On the support of Zygmund measures

*Laia Weisz Font
Universitat Autònoma de Barcelona (UAB)
laia.weisz@gmail.com

*Corresponding author

Resum *(CAT)*
Hem estudiat els compactes que suporten mesures de Zygmund, dels quals no se'n coneix cap caracterització. Hem introduït un concepte anomenat log-porositat, que proporciona una condició suficient per tal que un compacte no pugui ser el suport d'una mesura de Zygmund. Hem vist que aquest resultat no deriva dels treballs de Makarov ni Kaufman. Hem introduït el concepte de capacitat Zygmund d'un compacte i hem proposat una caracterització dels suports de les mesures de Zygmund en termes d'aquesta capacitat. Hem demostrat que els compactes pels quals el límit d'aquesta capacitat és zero, no poden suportar mesures de Zygmund.

Abstract *(ENG)*
We analyse the compact sets that are the support of Zygmund measures, of which no characterisation is known. We introduce the concept of log-porosity which provides a sufficient condition that guarantees that a compact cannot be the support of a Zygmund measure. This result does not derive from the results of Makarov and Kaufman. We introduce the concept of Zygmund capacity of a compact and propose a characterisation of the supports of Zygmund measures in terms of this capacity. We prove that the compact sets for which the limit of this capacity is zero cannot be the support of Zygmund measures.

Societat Catalana de Matemàtiques

Institut d'Estudis Catalans

Keywords: *Zygmund measures, Kaufman's Theorem, log-porosity.*
MSC (2010): *28A12, 28A05, 03E15.*
Received: *March 27, 2022.*
Accepted: *July 27, 2022.*

Acknowledgement
I would like to thank Juan J. Donaire for his invaluable guidance and support.

1. Introduction

A positive finite measure μ in $\mathbb{R}$ is called a Zygmund measure if there exists a constant C such that $|\mu(I) - \mu(I')| \leq C\,|I|$ for all pairs of adjacent intervals I, I' of the same length $|I|$. The infimum of the values of C for which the inequality holds is called the Zygmund norm of μ and is denoted by $\|\mu\|_*$.

It can be seen that a positive measure is a Zygmund measure if and only if its distribution function belongs to the Zygmund space Λ_*, defined as the set of real-valued and bounded functions such that

$$\|f\|_* = \sup_{x,h\in\mathbb{R}^n} \frac{|f(x+h)+f(x-h)-2f(x)|}{\|h\|} < +\infty. \tag{1}$$

Let us recall the definition of Hausdorff measure. A measure function is an increasing continuous function $\varphi\colon [0,\delta) \to \mathbb{R}^+$ such that $\varphi(0) = 0$. Let $E \subset \mathbb{R}$ be a bounded set, we define the Hausdorff measure of E with respect to φ as

$$H_\varphi(E) = \lim_{\varepsilon\to 0} \inf \left\{ \sum_j \varphi(|I_j|) : E \subset \bigcup_j I_j \quad \text{and} \quad |I_j| \leq \varepsilon \right\}.$$

If $\varphi(t) = t^\alpha$ for some $\alpha > 0$, then we will write $H_\alpha(E)$.

Frostman's Lemma ([7, p. 112]) states that a compact set K is the support of a Lip_α measure if and only if $H_\alpha(K) > 0$. As a consequence, compact sets that are the support of Lip_α measures are completely determined. Although Zygmund measures can be considered as the limit of Lip_α measures as $\alpha \to 1$ (see [8]), a characterisation of the compact sets that are the support of Zygmund measures is not known.

2. Preliminary results

Clearly, the Lebesgue measure restricted to a positive measure set is a Zygmund measure. The following theorem proves the existence of a Zygmund measure whose support has zero Lebesgue measure.

Theorem 2.1 (Kahane)**.** *There exists a positive singular Zygmund measure.*

Sketch of proof. In [3], Kahane proved his theorem by geometrically building a Zygmund measure whose support has zero Lebesgue measure.

Let Q_n denote the tetradic intervals of length 4^{-n} and let us define the following succession of simple functions: $s_0 \equiv 1$ and $s_n(x) = s_{n-1}(x) + \varepsilon_n(x)$, where

$$\varepsilon_n(x) = \begin{cases} -1 & \text{if } x \in I_1 \cup I_4, \\ 1 & \text{if } x \in I_2 \cup I_3, \end{cases}$$

and $I_1, I_2, I_3, I_4 \in Q_n$ such that[1] $I_1 \cup I_2 \cup I_3 \cup I_4 = I_j \in Q_{n-1}$ and $x \in I_j$.

[1] We consider I_k to the left of I_{k+1}, for $k = 1, 2, 3$.

Let us consider the stopping time $\tau(x) = \inf\{n : s_n(x) = 0\}$. Cleary, $\tau(x) < \infty$ for almost all x. Let us build the succession of measures defined by $\mathrm{d}\mu_n = s_{n\wedge\tau}\,\mathrm{d}x$, where $n \wedge \tau = \min\{n, \tau\}$ and $\mathrm{d}x$ is the Lebesgue measure. It can be seen that μ_n is a positive probability measure. In addition, if we denote by μ the limit of μ_n as $n \to \infty$, it can be seen that μ is a Zygmund measure and its support has zero Lebesgue measure. □

Theorem 2.2 (Makarov). *If μ is a positive Zygmund measure, then μ is absolutely continuous with respect to H_Φ, where*

$$\Phi(t) = t\sqrt{\log\left(\frac{1}{t}\right)\cdot\log\log\log\left(\frac{1}{t}\right)}. \tag{2}$$

In addition, there exists a Zygmund measure ν which satisfies $0 < H_\Phi(\mathrm{supp}\,(\nu)) < \infty$.

Makarov's Theorem provides the optimal measure function ϕ such that if a compact set K has $H_\phi(K) = 0$, then it cannot be the support of a Zygmund measure. The proof of the first and second statements of Makarov's Theorem can be found in [6] and [5] respectively.

Theorem 2.3 (Kaufman). *For each measure function h such that $\lim_{t\to0}\frac{t}{h(t)} = 0$, there exists a compact set K with $H_h(K) > 0$ which is not the support of a Zygmund measure.*

Kaufman's Theorem implies that it is impossible to characterise the supports of Zygmund measures in terms of Hausdorff measures. In [4], Kaufman proved his theorem by introducing a special class of sets: A compact set E is called *porous* (with a parameter $a > 0$) if, for each $\delta > 0$, there exists a covering of E by disjoint open intervals $I_{\varepsilon_t}(x_t) = (x_t - \varepsilon_t, x_t + \varepsilon_t)$ such that $\varepsilon_t < \delta$ and each interval $I_{\varepsilon_t}(x_t)$ contains an interval $I_{a\varepsilon_t}(x'_t) = (x'_t - a\varepsilon_t, x'_t + a\varepsilon_t) \subset I_{\varepsilon_t}(x_t)$ disjoint from E.

Kaufman proved that porous sets cannot be the support of Zygmund measures and that for each function h that satisfies the hypothesis of Theorem 2.3, there exists a porous set K with $H_h(K) > 0$.

Proof. Let $S = \{n_1 < n_2 < \cdots < n_k < n_{k+1} < \cdots\}$ be a sequence of positive integers whose complement is infinite and let us define

$$E = \left\{\sum_{k=1}^{\infty}\varepsilon_k\,2^{-n_k}\;\middle|\;\varepsilon_k \in \{0,1\}\quad\forall k\right\}.$$

Let m be an integer not in S and let $x \in E$. Then $a \le 2^{m-1}x \le a + \frac{1}{2}$ for some $a \in \mathbb{N}_0$ and consequently, each element of E has distance $\le d = 2^{-m-1}$ from one of the centres $\{(q + \frac{1}{4})2^{1-m} \mid q \in \mathbb{N}_0\}$. The distance between two consecutive centres is $4d$, hence E is porous.

We need to define S so that $H_h(E) > 0$. Let ψ be a positive function such that $\lim_{t\to0^+} t\cdot\psi(t) = 0$ and $\lim_{t\to0^+} h(t)\cdot\psi(t) = \infty$. We build S as $\{\lfloor -\log_2(\psi^{-1}(2^k))\rfloor : k \in \mathbb{N}\}$. It can be seen that, with this definition, $|S^c \cap \mathbb{N}| = \infty$ and $\lim_{k\to\infty} 2^k h(2^{-n_k}) = \infty$. Let ν be a probability measure with support in E such that each interval $I \in \mathfrak{D}_{n_k}$ has measure $\mathcal{O}(2^{-k})$. Let I be an interval of length r small and let k be an integer such that $2^{-n_k} \ge r > 2^{-n_{k+1}}$. We define J as the interval of length 2^{-n_k} and the same centre as I. Therefore,

$$\nu(I) \le \nu(J) \le \mathcal{O}(1)\cdot\nu(I_j^{n_{k+1}}) = \mathcal{O}(1)\cdot 2^{-k-1} < \mathcal{O}(1)\,h(2^{-n_{k+1}}) \le \mathcal{O}(1)\,h(r) = \mathcal{O}(1)\,h(|I|),$$

as $\lim_{k\to\infty} 2^k h(2^{-n_k}) = \infty$. Consequently, $H_h(E) = \lim_{\varepsilon\to0}\inf\{\sum_j h(|I_j|) : E \subset \bigcup I_j,\ |I_j| \le \varepsilon\} > 0$.

We will introduce a generalisation of porosity in order to prove that porous sets cannot be the support of a nontrivial Zygmund measure. □

Given a compact set $K \subset \mathbb{R}$ and a closed interval I, let us denote by I^* the biggest open interval in I, disjoint from K and satisfying $2|I^*| \leq |I|$.

Definition 2.4. Let $K \subset \mathbb{R}$ be a compact set of zero Lebesgue measure. We say K is *log-porous* if

$$\liminf_{\varepsilon\to 0} \left\{ \sum_{j\geq 1} |I_j| \log\left(\frac{|I_j|}{|I_j^*|}\right) : K \subset \bigcup_{j\geq 1} I_j, \text{ with pairwise disjoint interiors and } |I_j| < \varepsilon \right\} = 0.$$

Theorem 2.5. *A log-porous set cannot be the support of a Zygmund measure.*

Theorem 2.6. *There exists a log-porous compact set K such that it is nonporous and $H_\Phi(K) > 0$, where Φ is the function defined in (2).*

It can be easily seen that porous sets are log-porous. Theorem 2.6 implies that Theorem 2.5 is not a consequence of Makarov's Theorem or Kaufman's Theorem. In order to prove Theorem 2.5 we will need the following result.

Proposition 2.7. *Let $f \in \Lambda_*$. Then for any $t \in (0,1)$ and $a, b \in \mathbb{R}$,*

$$|(1-t)f(a) + tf(b) - f((1-t)a + tb)| \leq C\|f\|_* \varphi(t)|b-a|,$$

where C is an absolute constant and $\varphi(t) = t\log\frac{1}{t}$ if $t \leq 1/2$ and $\varphi(t) = \varphi(1-t)$ if $t \geq 1/2$.

The proof of Proposition 2.7 can be found in [1].

Proof of Theorem 2.5. Let $K \subset (0,1)$ be a log-porous compact set. Let μ be a positive Zygmund measure with support in K and let f be its distribution function. Given $\eta > 0$ small, there exists $\varepsilon > 0$ and a covering by closed intervals $\{I_j\}$ given by the definition of log-porosity, such that $|I_j| < \varepsilon$ $\forall j \geq 1$ and $\sum_{j\geq 1} |I_j| \log(\frac{|I_j|}{|I_j^*|}) < \eta$.

Let us denote $I_j = (a_j, b_j)$ and $I_j^* = (c_j, d_j) \subset I_j$. We define $\rho_j = \frac{|I_j^*|}{|I_j|}$ and $x = \frac{c_j - \rho_j a_j}{1-\rho_j} = \frac{d_j - \rho_j b_j}{1-\rho_j}$. Note that $f(d_j) = f(c_j)$ since $K \cap (c_j, d_j) = \emptyset$. By Proposition 2.7, we have

$$\begin{aligned}\rho_j|f(b_j) - f(a_j)| &= |\rho_j f(b_j) - \rho_j f(a_j) + f(c_j) - f(d_j) + (1-\rho_j)f(x) - (1-\rho_j)f(x)| \\ &\leq |\rho_j f(b_j) + (1-\rho_j)f(x) - f(d_j)| + |\rho_j f(a_j) + (1-\rho_j)f(x) - f(c_j)| \\ &\leq C\varphi(\rho_j)(b_j - x) + C\varphi(\rho_j)(a_j - x) = C\varphi(\rho_j)(b_j - a_j),\end{aligned}$$

and, as a consequence, $|f(b_j) - f(a_j)| \leq C|b_j - a_j|\frac{1}{\rho_j}\varphi(\rho_j) = C|b_j - a_j|\log(\frac{1}{\rho_j}) = C|I_j|\log(\frac{|I_j|}{|I_j^*|})$, since $\rho \leq \frac{1}{2}$. Therefore $\mu(K) = f(1) - f(0) \leq C\sum_{j\geq 1}|I_j|\log(\frac{|I_j|}{|I_j^*|}) < C\cdot\eta$.

As η is arbitrarily small, we conclude that $\mu(K) = 0$ and μ must be the trivial measure. □

Proof of Theorem 2.6. Let us consider, in $[0,1]$, the function defined by

$$\varphi(t)=\begin{cases}0 & \text{if } t=0,\\ \dfrac{t}{\sqrt[4]{\log_2(2/t)}} & \text{if } t\in(0,1].\end{cases}$$

Note that φ is increasing and convex, therefore $2\varphi(2^{-n-1})<\varphi(2^{-n})$.

The compact will be constructed inductively in a similar way as the Cantor set. Let $E_0=I_1^0=[0,1]$ and let $E_n=\bigcup_{j=1}^{2^n} I_j^n$. For each closed interval I_j^n, we consider I_{2j-1}^{n+1} and I_{2j}^{n+1} the two closed corner intervals of I_j^n of length $\varphi(2^{-n-1})$. Then we construct E_{n+1} as $E_{n+1}=\bigcup_{j=1}^{2^{n+1}} I_j^{n+1}$. Finally, we set $K=\bigcap_{n\geq 1} E_n$.

Firstly, we will see that K is log-porous. For $n\geq 0$, let us consider the covering of K given by $\bigcup_{j=1}^{2^n} I_j^n$. By construction, for each I in the covering, the length of I^* is $\varphi(2^{-n})-2\varphi(2^{-n-1})$. As a result,

$$\sum_{j=1}^{2^n}|I_j^n|\log\left(\frac{|I_j^n|}{|I_j^{n*}|}\right)=2^n\varphi(2^{-n})\log\left(\frac{\varphi(2^{-n})}{\varphi(2^{-n})-2\varphi(2^{-n-1})}\right)=\frac{1}{\sqrt[4]{n+1}}\log\left(\frac{\sqrt[4]{n+2}}{\sqrt[4]{n+2}-\sqrt[4]{n+1}}\right)$$

which goes to 0 as n tends to infinity. Now we will see that $H_\Phi(K)>0$. Let us define $\lambda_0=1$ and $\lambda_n=\frac{1}{2}\lambda_{n-1}(x)$ if $x\in I_j^n$ for some j and $\lambda_n=0$ otherwise. We consider the succession of measures $d\nu_n=\varphi(2^{-n})\lambda_n\,dx$, and we denote by ν the limit of ν_n. Note that ν is a positive probability measure with support in K and such that $\nu(I_j^n)=2^{-n}$ for all j.

By a similar argument as the one used in the proof of Theorem 2.3, we conclude that $H_{\varphi^{-1}}(K)>0$. Since $\varphi^{-1}(t)=o(\Phi(t))$ as $t\to 0^+$, by the comparison lemma between Hausdorff measures (see [2, p. 60]), we conclude that $H_\Phi(K)=\infty$.

In order to prove that K is nonporous, we assign, to each $x\in K$, a succession $(\delta_n(x))$ of $0's$ and $1's$ in the following way: if $x\in I_{2j-1}^n$ for some j, we set $\delta_n(x)=0$; if $x\in I_{2j}^n$, we set $\delta_n(x)=1$. Let E be the set of $x\in K$ such that there exists $n_0=n_0(x)$ such that, for $n\geq n_0$, $\delta_n(x)\neq\delta_{n+1}(x)$.

We assume that K is porous with parameter $0<\rho<1$ and consider the associated covering of K. If $x\in E$, let I be the interval of the aforementioned covering containing x. We choose $n\in\mathbb{N}$ such that $\varphi(2^{-n-1})<|I|\leq\varphi(2^{-n})$. By election of x, the length of the biggest interval contained in I and disjoint of K is less than $\varphi(2^{-n+2})-2\varphi(2^{-n+1})$. Therefore, $\rho\varphi(2^{-n-1})\leq\rho\cdot|I|\leq\varphi(2^{-n+2})-2\varphi(2^{-n+1})$ and, as a result

$$\frac{\varphi(2^{-n+2})-2\varphi(2^{-n+1})}{\varphi(2^{-n-1})}\geq\rho.$$

We have reached a contradiction, since the left expression goes to 0 as n tends to infinity. □

3. An approach to the characterisation of the supports of Zygmund measures

Our aim is to find a characterisation of the compact sets that support Zygmund measures, *i.e.*, the compact sets K for which $\sup\{\mu(K):\operatorname{supp}(\mu)\subset K,\ \mu\geq 0,\ \|\mu\|_*\leq 1\}>0$, where μ is a Zygmund measure.

To that end, we will approximate K as the union of closed dyadic intervals and we will determine the maximum mass a fixed-norm Zygmund measure can have. We will use the following notation to denote the dyadic intervals of length 2^{-n}:

$$\mathfrak{D}_n = \left\{ \left[\frac{k}{2^n}, \frac{k+1}{2^n}\right) \;\middle|\; k \in \{0, 1, \dots, 2^n - 1\} \right\}.$$

Lemma 3.1. *Let $K \subset (0,1)$ be the union of 2^{-n}-length closed dyadic intervals. Let μ be a positive measure with $\operatorname{supp}\mu \subseteq K$ and constant density over each interval in $\mathfrak{D}_n$. If the following condition holds, μ is a Zygmund measure with support in K and $\|\mu\|_* \sim C$.*

$$|\mu(I) - \mu(I')| \leq C|I|, \text{ where } I,\, I' \in \mathfrak{D}_k \text{ are adjacent and } k \leq n.$$

This lemma can be easily proven using a variation of the proof of Kahane's Theorem and we will use it to attempt to determine a geometrical characterisation of the compact sets that support a Zygmund measure. With that goal, we shall introduce the concept of $\mathcal{Z}$-2^k sequences.

Definition 3.2. Let $n \in \mathbb{N}$ and $0 \leq k \leq n$. A number sequence $x_1\, x_2\, \dots\, x_{2^{n-k}}$ is said to be $\mathcal{Z}$-2^k if the following conditions hold.

$$\begin{cases} x_j \geq 0 & \forall j, \\ |x_j - x_{j-1}| \leq 2^k & j = 2, \dots, 2^{n-k}, \\ x_j \leq 2^k & j = 1,\, 2^{n-k}. \end{cases}$$

Let $K \subseteq [0,1]$ be a compact set and let $n \in \mathbb{N}$. Firstly, we will associate a density $\mathcal{D}_n^{(n)}$ to each interval $I_j^n \in \mathfrak{D}_n$ in order for it to be a $\mathcal{Z}$-1 sequence. Specifically, $\mathcal{D}_n^{(n)}$ will be constructed as the maximal $\mathcal{Z}$-1 sequence such that if $I_j^n \cap K = \varnothing$, then $\mathcal{D}_n^{(n)}(I_j^n) = 0$. Secondly, we will associate, to each interval $I_j^{n-1} \in \mathfrak{D}_{n-1}$, the maximal density $\mathcal{D}_{n-1}^{(n)}$ such that the resulting sequence is $\mathcal{Z}$-2 and that $\mathcal{D}_{n-1}^{(n)}(I_j^{n-1}) \leq \mathcal{D}_n^{(n)}(I_{2j}^n) + \mathcal{D}_n^{(n)}(I_{2j+1}^n)$. Iterating this process we will obtain a density $\mathcal{D}_0^{(n)}([0,1])$. The limit of $2^{-n}\mathcal{D}_0^{(n)}([0,1])$ as $n \to \infty$ bounds the maximum mass a Zygmund measure defined on the intervals of $\mathfrak{D}_n$ that intersect K and with controlled $\|\mu\|_*$ can have.

Let us formally define the densities $\mathcal{D}_k^{(n)}$ associated to each dyadic interval in $\mathfrak{D}_k$ for $k \leq n$. We start by defining $\mathcal{D}_n^{(n)}$. Given an interval $I_j^n \in \mathfrak{D}_n$ for $j = 0, 1, \dots, 2^n - 1$, we define

$$\mathcal{D}_n^-(I_j^n) = \begin{cases} 0 & \text{if } I_j^n \cap K = \emptyset, \\ \mathcal{D}_n^-(I_{j-1}^n) + 1 & \text{if } I_j^n \cap K \neq \emptyset, \end{cases}$$

and analogously,

$$\mathcal{D}_n^+(I_j^n) = \begin{cases} 0 & \text{if } I_j^n \cap K = \emptyset, \\ \mathcal{D}_n^+(I_{j+1}^n) + 1 & \text{if } I_j^n \cap K \neq \emptyset, \end{cases}$$

with the convention $I_{-1}^n = \left[\frac{-1}{2^n}, 0\right)$, $I_{2^n}^n = \left[1, 1 + \frac{1}{2^n}\right)$ and $\mathcal{D}_n^-(I_{-1}^n) = \mathcal{D}_n^+(I_{2^n}^n) = 0$. Finally, we denote

$$\mathcal{D}_n^{(n)}(I_j^n) = \min\{\mathcal{D}_n^-(I_j^n), \mathcal{D}_n^+(I_j^n)\}.$$

Now we define $\mathcal{D}_{n-k}(I_j^{n-k})$ for each dyadic interval $I_j^{n-k} \in \mathfrak{D}_{n-k}$ for $1 \leq k \leq n$. To do so, let us consider the two intervals $J, J' \in \mathfrak{D}_{n-k+1}$ such that $J, J' \subseteq I_j^{n-k}$. We denote

$$S_{n-k}(I_j^{n-k}) = \mathcal{D}_{n-k+1}^{(n)}(J) + \mathcal{D}_{n-k+1}^{(n)}(J'),$$

and we proceed as before, setting

$$\mathcal{D}_{n-k}^{-}(I_j^{n-k}) = \min\{\mathcal{D}_{n-k}^{-}(I_{j-1}^{n-k}) + 2^k,\ S_{n-k}(I_j^{n-k})\},$$
$$\mathcal{D}_{n-k}^{+}(I_j^{n-k}) = \min\{\mathcal{D}_{n-k}^{+}(I_{j+1}^{n-k}) + 2^k,\ S_{n-k}(I_j^{n-k})\},$$

with the convention $\mathcal{D}_{n-k}^{-}(I_{-1}^{n-k}) = \mathcal{D}_{n-k}^{+}(I_{2^{n-k}}^{n-k}) = 0$. Finally, we define

$$\mathcal{D}_{n-k}^{(n)}(I_j^{n-k}) = \min\{\mathcal{D}_{n-k}^{-}(I_j^{n-k}),\ \mathcal{D}_{n-k}^{+}(I_j^{n-k})\}.$$

Hence, we built the densities $\mathcal{D}_n^{(n)}, \mathcal{D}_{n-1}^{(n)}, \dots, \mathcal{D}_0^{(n)}$. Finally, we define the *Zygmund Capacity* as

$$C_n(K) = 2^{-n}\,\mathcal{D}_0^{(n)}([0,1]).$$

Let K' be the compact set formed by the union of the dyadic intervals in $\mathfrak{D}_n$ which intersect K. By construction, $C_n(K)$ is an upper bound to the mass of any Zygmund measure with support in K'.

Proposition 3.3. *For each compact set $K \subset [0,1]$, $\exists \lim_{n\to\infty} C_n(K)$.*

Proof. By construction, $C_n(K) \geq 0\ \forall n$. In order to prove that the limit exists, it suffices to see that the succession $(C_n(K))_n$ is decreasing. Let $I_j^n \in \mathfrak{D}_n$ be an interval such that $I_j^n \cap K = \emptyset$ and let I_{2j}^{n+1}, I_{2j+1}^{n+1} be the two intervals of $\mathfrak{D}_{n+1}$ contained in I_j^n. Clearly, I_{2j}^{n+1} and I_{2j+1}^{n+1} are disjoint from K. Hence,

$$\mathcal{D}_n^{(n)}(I_j^n) = 0 \Longrightarrow \mathcal{D}_n^{(n+1)}(I_j^n) = 0.$$

Alternatively, if $I_j^n \cap K \neq \emptyset$, then $\mathcal{D}_n^{(n)}(I_j^n) = a > 0$. Therefore, we conclude that $\mathcal{D}_n^{(n+1)}(I_{j-a}^n) = 0$ or that $\mathcal{D}_n^{(n+1)}(I_{j+a}^n) = 0$. Consequently, $\mathcal{D}_n^{(n+1)}(I_j^n) \leq 2a$. This implies that $\mathcal{D}_n^{(n+1)}(I_j^n) \leq 2\mathcal{D}_n^{(n)}(I_j^n)$ for all j, so $\mathcal{D}_0^{(n+1)}([0,1]) \leq 2\mathcal{D}_0^{(n)}([0,1])$. As a consequence, $C_{n+1}(K) \leq C_n(K)$. □

Theorem 3.4. *Let $K \subseteq [0,1]$ a compact set. If $\lim_{n\to\infty} C_n(K) = 0$, K cannot be the support of a nontrivial Zygmund measure.*

Proof. Let μ be a Zygmund measure with support in K. We will prove that μ must be the trivial measure. Let us assume, without loss of generality, that $\|\mu\|_* \leq 1$. Given $n \in \mathbb{N}$, let $0 \leq k \leq n$ be an integer. We will prove by induction on k that $\mu(I_j^{n-k}) \leq 2^{-n}\,\mathcal{D}_{n-k}^{(n)}(I_j^{n-k})$ for all $I_j^{n-k} \in \mathfrak{D}_{n-k}$.

It is clear that if $I_j^{n-k} \cap K = \emptyset$, the inequality holds, so let us assume that $I_j^{n-k} \cap K \neq \emptyset$.

Firstly, we will show that the inequality holds for $k = 0$. Let $I_j^n \in \mathfrak{D}_n$ be a dyadic interval and let $I_{j-\ell}^n \in \mathfrak{D}_n$ be the closest interval to I_j^n such that $I_{j-\ell}^n \cap K = \emptyset$. Note that $-2^n + j \leq \ell \leq j+1$ and we can assume, without loss of generality that $\ell > 0$. Therefore,

$$\frac{\mu(I_j^n)}{|I_j^n|} \leq 1 + \frac{\mu(I_{j-1}^n)}{|I_j^n|} \leq 1 + 1 + \frac{\mu(I_{j-2}^n)}{|I_j^n|} \leq \cdots \leq \ell + \frac{\mu(I_{j-\ell}^n)}{|I_j^n|} = \ell = \mathcal{D}_n^{(n)}(I_j^n)$$

and $\mu(I_j^n) \leq 2^{-n}\mathcal{D}_n^{(n)}(I_j^n)$. Let us assume the inequality holds for $n-k+1$ where $1 \leq k \leq n$ is fixed, and we will prove it holds for $n-k$. On one hand we have

$$\mu(I_j^{n-k}) = \mu(I_{2j}^{n-k+1}) + \mu(I_{2j+1}^{n-k+1}) \leq 2^{-n}(\mathcal{D}_{n-k+1}^{(n)}(I_{2j}^{n-k+1}) + \mathcal{D}_{n-k+1}^{(n)}(I_{2j+1}^{n-k+1})),$$

since μ is a measure and $I_j^{n-k} = I_{2j}^{n-k+1} \cup I_{2j+1}^{n-k+1}$. On the other hand,

$$\frac{\mu(I_j^{n-k})}{|I_j^{n-k}|} \leq 1 + \frac{\mu(I_{j-1}^{n-k})}{|I_j^{n-k}|} \Longrightarrow \mu(I_j^{n-k}) \leq 2^{-n+k} + \mu(I_{j-1}^{n-k}).$$

Analogously, $\mu(I_j^{n-k}) \leq 2^{-n+k} + \mu(I_{j+1}^{n-k})$. As both of these inequalities hold for all j, clearly

$$\mu(I_j^{n-k}) \leq 2^{-n}\min\{\mathcal{D}_{n-k+1}^{(n)}(I_{2j}^{n-k+1}) + \mathcal{D}_{n-k+1}^{(n)}(I_{2j+1}^{n-k+1}), \mathcal{D}_{n-k}^{-}(I_{j-1}^{n-k}) + 2^k, \mathcal{D}_{n-k}^{+}(I_{j+1}^{n-k}) + 2^k\}.$$

Therefore, $\mu(I_j^{n-k}) \leq 2^{-n}\mathcal{D}_{n-k}^{(n)}(I_j^{n-k})$ which implies that $\mu([0,1]) \leq 2^{-n}\mathcal{D}_0^{(n)}([0,1]) = C_n(K) \xrightarrow{n\to\infty} 0$. Consequently, since $K \subset [0,1]$ we have $\mu(K) = 0$ and, as a result, $\mu \equiv 0$. □

Conjecture 3.5. *A compact set K is the support of a nontrivial Zygmund measure if and only if*

$$\lim_{n\to\infty} C_n(K) > 0.$$

By Theorem 3.4, it suffices to prove that if the limit of $C_n(K)$ as n tends to infinity is positive, then there exists a nontrivial Zygmund measure with support in K. In order to do so, we will define a succession of Zygmund measures μ_n with mass equal to $C_n(K)$. As $\lim_{n\to\infty} \|\mu_n\| = \lim_{n\to\infty} C_n(K) > 0$, the limit $\mu = \lim_{n\to\infty} \mu_n$ will be well-defined. Choosing appropriately the support of μ_n, μ will have support in K. Furthermore, if μ_n are uniformly bounded, then μ will be a Zygmund measure.

Given an integer n we denote $\mathcal{D}_0^{(n)} := \mathcal{D}_0^{(n)}([0,1])$. By construction, $\mathcal{D}_0^{(n)} \leq \sum_{j=0}^{2^n-1} \mathcal{D}_n^{(n)}(I_j^n)$ where $I_j^n \in \mathfrak{D}_n$. If both expressions are equal, then considering

$$\mathrm{d}\mu_n = \sum_{j=0}^{2^n-1} \mathcal{D}_n^{(n)}(I_j^n)\chi_{I_j^n}\,\mathrm{d}x,$$

we have that μ_n is a Zygmund measure. Let us assume, that $\mathcal{D}_0^{(n)} < \sum_{j=0}^{2^n-1} \mathcal{D}_n^{(n)}(I_j^n)$. We want to find a Zygmund measure μ_n with mass $2^{-n}\mathcal{D}_0^{(n)}$, constant over dyadic intervals I_j^n and such that $\mu_n(I_j^n) = 0$ if I_j^n is disjoint from K.

Therefore we aim to determine some numbers $d_0, \dots, d_{2^n-1} \geq 0$ such that $\sum_{j=0}^{2^n-1} d_j = \mathcal{D}_0^{(n)}$, that

$$\mathrm{d}\mu_n = \sum_{j=0}^{2^n-1} d_j \chi_{I_j^n}\,\mathrm{d}x \tag{3}$$

is a Zygmund measure and that $d_j = 0$ if $I_j^n \cap K = \emptyset$. Finding a technique to determine this numbers would end the proof, since the limit of μ_n would be a positive Zygmund measure with support in K.

Therefore, we will try to *distribute* $\mathcal{D}_0^{(n)}$ between $d_0, \dots, d_{2^n-1} \geq 0$ in such a way that the aforementioned conditions are met. Clearly, in order for μ_n as defined in (3) to be a Zygmund measure with $\|\mu\|_* \leq 1$, the sequence $d_0 d_1 \dots d_{2^n-1}$ must be $\mathcal{Z}$-1. By construction, the sequence $\mathcal{D}_n^{(n)}(I_0^n) \dots \mathcal{D}_n^{(n)}(I_{2^n-1}^n)$ is the maximal $\mathcal{Z}$-1 sequence such that the j-th number is zero if $I_j^n \cap K = \emptyset$. As a result, we have $d_j \leq \mathcal{D}_n^{(n)}(I_j^n)$ for all j. Let us denote $d_j^1 = d_{2j} + d_{2j+1}$. Then, in order for (3) to be a Zygmund measure, the sequence $d_0^1 \dots d_{2^{n-1}-1}^1$ must be $\mathcal{Z}$-2 and less than $\mathcal{D}_{n-1}^{(n)}(I_0^{n-1}) \dots \mathcal{D}_{n-1}^{(n)}(I_{2^{n-1}-1}^{n-1})$. Denoting, inductively, $d_j^k = d_{2j}^{k-1} + d_{2j+1}^{k-1}$ for $1 \leq k \leq n$ we conclude that the sequences $d_0^k \dots d_{2^{n-k}-1}^k$ must be $\mathcal{Z}$-2^k and less than $\mathcal{D}_{n-k}^{(n)}(I_0^{n-k}) \dots \mathcal{D}_{n-k}^{(n)}(I_{2^{n-k}-1}^{n-k})$. Note that $d_0^n = \mathcal{D}_0^{(n)}$.

We shall say that $d_0, \dots, d_{2^n-1}$ constitute a *distribution* of $\mathcal{D}_0^{(n)}$ if they meet the following conditions:

- $d_j \geq 0$ for all j,
- $\sum_{j=0}^{2^n-1} d_j = \mathcal{D}_0^{(n)}$,
- $d_0^k \dots d_{2^{n-k}-1}^k$ must be a $\mathcal{Z}$-2^k sequence for all $k = 0, \dots, n$,
- $d_j^k \leq \mathcal{D}_{n-k}^{(n)}(I_j^{n-k})$ for all $k = 0, \dots, n$ and for all $j = 0, \dots, 2^{n-k} - 1$,

where $d_j^0 = d_j$ for all j.

If $d_0, \dots, d_{2^n-1}$ is a *distribution* of $\mathcal{D}_0^{(n)}([0,1])$, then the measure defined in (3) meets the hypothesis of Lemma 3.1. This implies that μ_n is a Zygmund measure with mass $\|\mu_n\| = 2^{-n}\mathcal{D}_0^{(n)}([0,1]) = C_n(K)$.

First of all, we are going to prove that given d_j^k for $j = 0, \dots, 2^{n-k} - 1$ that meet certain conditions, we can determine d_j^{k-1} for $j = 0, \dots, 2^{n-k+1} - 1$. After that, we will state a *distribution* method that appears to guarantee that the conditions are met, although we have not been able to prove this result. To that end, let us introduce the concept of sequence of corrected sums.

Definition 3.6. Let $n \in \mathbb{N}$ and $1 \leq k \leq n$. Given an integer $\mathcal{Z}$-2^{k-1} sequence $x_1\, x_2\, \dots\, x_{2^{n-k+1}}$, we determine its *sequence of corrected sums* $t_1\, t_2\, \dots\, t_{2^{n-k}}$ as:

- Firstly, let us consider the integer sequence $s_1\, s_2\, \dots\, s_{2^{n-k}}$ defined as $s_j = x_{2j-1} + x_{2j}$ $\forall j$.
- Secondly, let $e_1\, e_2\, \dots\, e_{2^{n-k}}$ and $d_1\, d_2\, \dots\, d_{2^{n-k}}$ be two integer sequences constructed as follows:

$$\begin{cases} e_1 = \min\{s_1,\, 2^k\}, \\ d_{2^{n-k}} = \min\{s_{2^{n-k}},\, 2^k\}, \\ e_j = \min\{s_j,\, e_{j-1} + 2^k\} & j = 2, \dots, 2^{n-k}, \\ d_j = \min\{s_j,\, d_{j+1} + 2^k\} & j = 1, \dots, 2^{n-k} - 1. \end{cases}$$

- Finally, we define the integer sequence $t_1\, t_2\, \dots\, t_{2^{n-k}}$ as $t_j = \min\{e_j, d_j\}$ $\forall j$.

Note that, given a sequence $\mathcal{Z}$-2^{k-1}, its sequence of corrected sums will be $\mathcal{Z}$-2^k.

Lemma 3.7. *Given $n \in \mathbb{N}$ and $1 \leq k \leq n$, let $x_1 x_2 \dots x_{2^{n-k+1}}$ be an integer $\mathcal{Z}$-2^{k-1} sequence and let us denote by $t_1 t_2 \dots t_{2^{n-k}}$ its sequence of corrected sums. Let $r_1 r_2 \dots r_{2^{n-k}}$ be an integer $\mathcal{Z}$-2^k sequence with $r_j \leq t_j$ $\forall j$. Then there exists $y_1 y_2 \dots y_{2^{n-k+1}}$ an integer $\mathcal{Z}$-2^{k-1} sequence satisfying $y_j \leq x_j$ for $j = 1, \dots, 2^{n-k+1}$ and $y_{2j-1} + y_{2j} = r_j$ for $j = 1, \dots, 2^{n-k}$, if the following conditions are met*

$$\begin{cases} -x_{2j-3} - 2^{k-1} \leq r_j - r_{j-1} \leq x_{2j} + 2^{k-1}, \\ r_j - r_{j-1} \leq 2^k + x_{2j} - y_{2j-4} \end{cases}$$

for all $j = 2, \dots, 2^{n-k}$.

Notation. Let us introduce the following notation: $a_j = x_{2j} - x_{2j-1}$, $\delta_j = x_{2j+1} - x_{2j}$, $b_j = s_j - r_j$ and $\tilde{b}_j = x_j - y_j$.

Proof. We will construct the sequence $y_1 y_2 \dots y_{2^{n-k+1}}$ as $y_j = x_j - \tilde{b}_j$ where $\tilde{b}_{2j-1} + \tilde{b}_{2j} = b_j$. We will take $\tilde{b}_j \in \mathbb{N}$, as a result, $y_j \leq x_j$, $y_j \in \mathbb{N}$ and $y_{2j-1} + y_{2j} = s_j - b_j = r_j$. Consequently, we only need to prove that there exists $\tilde{b}_j$ with the previous definition such that $y_1 y_2 \dots y_{2^{n-k+1}}$ satisfies $y_j \geq 0$ $\forall j$ and $|y_j - y_{j-1}| \leq 2^{k-1}$ for $j = 2, \dots, 2^{n-k+1}$.

Note that $\tilde{b}_{2j-1}$ needs to satisfy the following conditions:

(i) $\tilde{b}_{2j-1} \in I_0 = [b_j - x_{2j},\ x_{2j-1}]$ since $y_j \geq 0$ implies $\tilde{b}_{2j-1} \leq x_{2j-1}$ and $\tilde{b}_{2j} = b_j - \tilde{b}_{2j-1} \leq x_{2j}$.

(ii) $\tilde{b}_{2j-1} \in I_1 = [0, b_j]$ since $y_j \leq x_j$ implies $\tilde{b}_{2j-1} \geq 0$ and $\tilde{b}_{2j-1} = b_j - \tilde{b}_{2j}$.

(iii) $\tilde{b}_{2j-1} \in I_2 = \left[\left\lceil \frac{b_j - a_j}{2} \right\rceil - 2^{k-2}, \left\lfloor \frac{b_j - a_j}{2} \right\rfloor + 2^{k-2}\right]$, where $\lfloor \ell \rfloor$ and $\lceil \ell \rceil$ denote the floor and ceiling of ℓ respectively. This condition derives from the inequality $|y_{2j} - y_{2j-1}| \leq 2^{k-1}$.

(iv) $\tilde{b}_{2j-1} \in I_3 = [\delta_{j-1} + \tilde{b}_{2j-2} - 2^{k-1},\ \delta_{j-1} + \tilde{b}_{2j-2} + 2^{k-1}]$. This condition originates derives from the fact that

$$|y_{2j-1} - y_{2j-2}| = |x_{2j-1} - \tilde{b}_{2j-1} - x_{2j-2} + \tilde{b}_{2j-2}| = |\delta_{j-1} - \tilde{b}_{2j-1} + \tilde{b}_{2j-2}| \leq 2^{k-1}.$$

The remainder of the proof consists on checking that $I_0 \cap I_1 \cap I_2 \cap I_3 \neq \varnothing$ when the conditions of the lemma are met. □

Notation. Let us denote by I_j the intersection of the aforementioned four intervals of the j-th step. By construction, the endpoints of I_j are integer, and we will denote them by $I_j = [\tilde{b}_{2j-1,m}, \tilde{b}_{2j-1,M}]$, with

$$\tilde{b}_{2j-1,m} = \max\left\{b_j - x_{2j}, 0, \left\lceil \frac{b_j - a_j}{2} \right\rceil - 2^{k-2}, \delta_{j-1} + \tilde{b}_{2j-2} - 2^{k-1}\right\},$$

$$\tilde{b}_{2j-1,M} = \min\left\{x_{2j-1}, b_j, \left\lfloor \frac{b_j - a_j}{2} \right\rfloor + 2^{k-2}, \delta_{j-1} + \tilde{b}_{2j-2} + 2^{k-1}\right\}.$$

Note that Lemma 3.7 implies that $I_j \neq \emptyset$. As a result, $\tilde{b}_{2j-1,m} \leq \tilde{b}_{2j-1,M}$ if the conditions are met. Let $\tilde{b}_{2j-1} \in I_j$ be an integer which minimises the distance[2]

$$d(\tilde{b}_{2j-1}) = \max\{|b_j - a_j - 2\tilde{b}_{2j-1}|, |y_{2j-2} - x_{2j-1} + \tilde{b}_{2j-1}|\}$$

subject to the restriction

$$\delta_j + b_j - \tilde{b}_{2j-1} + 2^{k-1} \geq \min\left\{x_{2j+1}, b_{j+1}, \left\lfloor \frac{b_{j+1} - a_{j+1}}{2} \right\rfloor + 2^{k-2}, \delta_j + b_j - \tilde{b}_{2j-1,w} + 2^{k-1}\right\} = \tilde{b}_{2j+1,M}$$

if $r_{j+1} \leq r_{j+2}$; or subject to the restriction

$$\delta_j + b_j - \tilde{b}_{2j-1} - 2^{k-1} \leq \max\left\{b_{j+1} - x_{2j+2}, 0, \left\lceil \frac{b_{j+1} - a_{j+1}}{2} \right\rceil - 2^{k-2}, \delta_j + b_j - \tilde{b}_{2j-1,w} - 2^{k-1}\right\} = \tilde{b}_{2j+1,m}$$

if $r_{j+1} > r_{j+2}$, where

$$\tilde{b}_{2j-1,w} = \begin{cases} \tilde{b}_{2j-1,M} & \text{if } r_j \leq r_{j+1}, \\ \tilde{b}_{2j-1,m} & \text{if } r_j > r_{j+1}. \end{cases}$$

If $j = 2^{n-k}$, the restriction to be satisfied is $x_{2j} - b_j + \tilde{b}_{2j-1} \leq 2^{k-1}$.

It can be seen easily that the aforementioned restrictions and Lemma 3.7 guarantee that the intersection of the intervals will not be empty.

Given two sequences r_j and x_j that satisfy the hypothesis of Lemma 3.7, we apply the aforemention method to obtain an integer sequence, denoted by $\tilde{b}_j^{(e)}$. Then we build a sequence symbolised by $y^{(e)}$, as $y_j^{(e)} = x_j - \tilde{b}_j^{(e)}$.

Let $r'_j = r_{2^{n-k}} \dots r_2 r_1$ and $x'_j = x_{2^{n-k+1}} \dots x_2 x_1$, be the inverted sequences r_j and x_j respectively. Clearly both sequences meet the conditions of Lemma 3.7. Applying the previously mentioned method, we obtain another integer sequence, $\tilde{b}_j^{(d)}$. Then, we build the sequence $y^{(d)}$, as $y_j^{(d)} = x_j - \tilde{b}^{(d)}_{2^{n-k+1}-j}$.

Note that $y^{(e)}$ and $y^{(d)}$ might not be $\mathcal{Z}$-2^{k-1} sequences, however, the following linear combination will be. We define y_j as

$$\begin{cases} \begin{cases} y_{2j-1} = \dfrac{y_{2j-1}^{(e)} + y_{2j-1}^{(d)}}{2} \\ y_{2j} = \dfrac{y_{2j}^{(e)} + y_{2j}^{(d)}}{2} \end{cases} & \text{if } (y_{2j-1}^{(e)} + y_{2j-1}^{(d)}) \equiv 0 \pmod 2, \\ \begin{cases} y_{2j-1} = \dfrac{y_{2j-1}^{(e)} + y_{2j-1}^{(d)} + 1}{2} \\ y_{2j} = \dfrac{y_{2j}^{(e)} + y_{2j}^{(d)} - 1}{2} \end{cases} & \text{if } (y_{2j-1}^{(e)} + y_{2j-1}^{(d)}) \equiv 1 \pmod 2 \quad \text{and } \Delta_1 < \Delta_2, \\ \begin{cases} y_{2j-1} = \dfrac{y_{2j-1}^{(e)} + y_{2j-1}^{(d)} - 1}{2} \\ y_{2j} = \dfrac{y_{2j}^{(e)} + y_{2j}^{(d)} + 1}{2} \end{cases} & \text{if } (y_{2j-1}^{(e)} + y_{2j-1}^{(d)}) \equiv 1 \pmod 2 \quad \text{and } \Delta_1 \geq \Delta_2, \end{cases}$$

[2] If $j = 1$ we take $y_{2j-2} = 0$.

where

$$\Delta_1 = \max\left\{\left|\frac{y^{(e)}_{2j-1}+y^{(d)}_{2j-1}+1}{2}-y_{2j-2}\right|, \left|\frac{y^{(e)}_{2j}+y^{(d)}_{2j}-1}{2}-\frac{y^{(e)}_{2j-1}+y^{(d)}_{2j-1}+1}{2}\right|, \left|\frac{y^{(e)}_{2j+1}+y^{(d)}_{2j+1}}{2}-\frac{y^{(e)}_{2j}+y^{(d)}_{2j}-1}{2}\right|\right\},$$

$$\Delta_2 = \max\left\{\left|\frac{y^{(e)}_{2j-1}+y^{(d)}_{2j-1}-1}{2}-y_{2j-2}\right|, \left|\frac{y^{(e)}_{2j}+y^{(d)}_{2j}+1}{2}-\frac{y^{(e)}_{2j-1}+y^{(d)}_{2j-1}-1}{2}\right|, \left|\frac{y^{(e)}_{2j+1}+y^{(d)}_{2j+1}}{2}-\frac{y^{(e)}_{2j}+y^{(d)}_{2j}+1}{2}\right|\right\}.$$

Although this method yielded promising numerical results, we have not been able to prove that this method of *distribution* guarantees that the conditions of Lemma 3.7 are always met.

References

[1] J.J. Carmona, J.J. Donaire, "On removable singularities for the analytic Zygmund class", *Michigan Math. J.* **43**(1) (1996), 51–65.

[2] J. Garnett, *Analytic Capacity and Measure*, Lecture Notes in Mathematics **297**, Springer-Verlag, Berlin-New York, 1972.

[3] J.-P. Kahane, "Trois notes sur les ensembles parfaits linéaires", *Enseign. Math. (2)* **15** (1969), 185–192.

[4] R.P. Kaufman, "Smooth functions and porous sets", *Proc. Roy. Irish Acad. Sect. A* **93**(2) (1993), 189–191.

[5] N.G. Makarov, "Probability methods in the theory of conformal mappings" (Russian), *Algebra i Analiz* **1**(1) (1989), 3–59; translation in: *Leningrad Math. J.* **1**(1) (1990), 1–56.

[6] N.G. Makarov, "A class of exceptional sets in the theory of conformal mappings" (Russian), *Mat. Sb.* **180**(9) (1989), 1171–1182, 1296; translation in: *Math. USSR-Sb.* **68**(1) (1991), 19–30.

[7] P. Mattila, *Geometry of Sets and Measures in Euclidean Spaces*, Fractals and rectifiability, Cambridge Studies in Advanced Mathematics **44**, Cambridge University Press, Cambridge, 1995.

[8] A. Zygmund, "Smooth functions", *Duke Math. J.* **12** (1945), 47–76.

REPORTS@SCM
AN ELECTRONIC JOURNAL OF THE
SOCIETAT CATALANA DE MATEMÀTIQUES

Bijective enumeration of constellations in higher genus

***Jordi Castellví**
Universitat Politècnica de Catalunya (UPC)
jordi.castellvi@upc.edu

∗Corresponding author

Resum *(CAT)*
Bousquet-Mélou i Schaeffer donaren el 2000 una enumeració bijectiva de certs mapes anomenats constel·lacions. El 2019, Lepoutre va descriure una bijecció entre mapes biacoloribles de gènere arbitrari i certs mapes unicel·lulars del mateix gènere. Presentem una bijecció entre constel·lacions de gènere superior i certs mapes unicel·lulars que generalitza les dues bijeccions existents alhora.
Fent servir aquesta bijecció, enumerem una subclasse de constel·lacions sobre el tor, demostrant que llur funció generadora és una funció racional de la funció generadora de certs arbres.

Abstract *(ENG)*
Bousquet-Mélou and Schaeffer gave in 2000 a bijective enumeration of some planar maps called constellations. In 2019, Lepoutre described a bijection between bicolorable maps of arbitrary genus and some unicellular maps of the same genus. We present a bijection between constellations of higher genus and some unicellular maps that generalizes both existing bijections at the same time.
Using this bijection, we manage to enumerate a subclass of constellations on the torus, proving that its generating function is a rational function of the generating function of some trees.

Keywords: *combinatorics, maps, enumeration, bijection, blossoming map, constellation, rationality.*
MSC (2010): *05C10, 05C30.*
Received: *July 2, 2022.*
Accepted: *July 30, 2022.*

Acknowledgement
The author thanks Marie Albenque and Éric Fusy for their guidance during the elaboration of his bachelor's thesis, on which this paper is based. He also wishes to thank the anonymous referee for her/his useful comments.

Societat Catalana de Matemàtiques
Institut d'Estudis Catalans

1. Introduction

A *map* M of genus g is a proper embedding of a graph in $\mathcal{S}_g$, the torus with g holes, such that the maximal connected components of $\mathcal{S}_g \setminus M$ are contractible. These components are called *faces*. Multiple edges and loops are allowed. Maps are considered up to orientation preserving homeomorphisms. A *unicellular map* is a map with a single face.

Maps of genus 0 are called *planar maps*. They receive this name because embedding graphs in the sphere or in the plane is essentially the same. The stereographical projection, for instance, can produce a plane embedding from a sphere embedding. All the faces of a planar map embedded in the plane are contractible except for one, the *exterior face*, which is homeomorphic to the complement of a disk.

A *corner* of a map is a couple of consecutive edges around a vertex. Equivalently, a corner can be seen as an incidence between a face and a vertex. The *degree* of a vertex or face is its number of corners.

A *rooted map* is a map with a marked corner, which is called the *root corner* (or, simply, *root*). This root corner naturally defines a *root vertex* and a *root face*. The maps we consider here will always be rooted.

If a map is rooted, we have a notion of clockwise and counterclockwise when following contractible cycles. Precisely, we say that a tour around a contractible cycle is clockwise (resp. counterclockwise) if the root face lies on the left (resp. right) side of it.

In maps, edges join two (possibly equal) vertices and separate two (possibly equal) faces. Thus, given a map M, we can define its *dual map* M^* in the following way. The faces (resp. vertices) of M become the vertices (resp. faces) of M^* and the dual of an edge e joining vertices v_1 and v_2 and separating faces f_1 and f_2 is an edge e^* joining vertices f_1^* and f_2^* and separating faces v_1^* and v_2^*. Note that the dual of a corner is "itself" (i.e., the same vertex-face incidence) and that dualization is involutive: $(M^*)^* = M$.

Figure 1 contains an example of a planar map and its dual. Figure 2 contains an example of a map on the torus with some additional structure that is presented later. Maps on the torus are drawn on a square the parallel sides of which have to be identified.

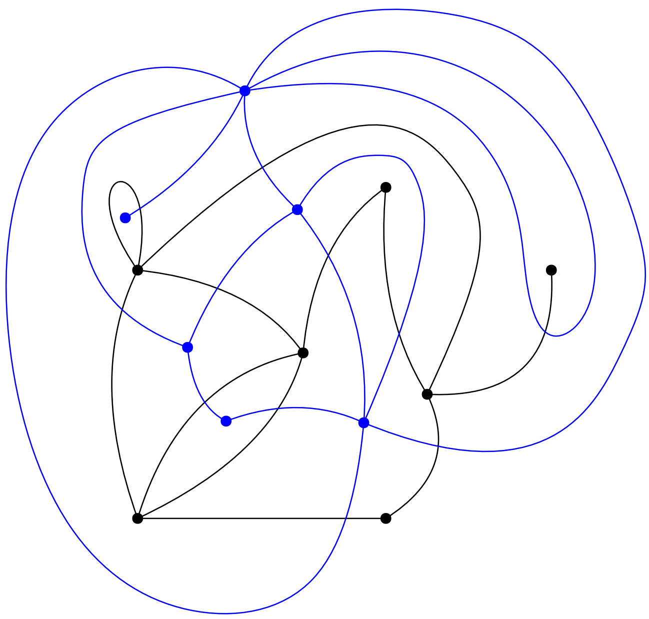


Figure 1: A planar map and its dual.

Maps are fundamental combinatorial objects that appear in many other fields of mathematics such as algebra and mathematical physics. The enumeration of planar maps began with the work of Tutte in the sixties [10]. In his work, Tutte enumerated a variety of families of maps, obtaining remarkably simple formulas. For example, he showed that the number of rooted planar maps with n edges is

$$\frac{2(2n)!3^n}{n!(n+2)!}.$$

His methods are based on the recursive combinatorial decomposition of maps and the equations obtained usually require the introduction of additional parameters called catalytic variables. In the late eighties, these techniques were extended to maps on surfaces of higher genus by Bender and Canfield [1, 2].

The simplicity of the formulas obtained by Tutte called for bijective demonstrations. Cori and Vauquelin gave the first bijective proof of the enumeration of planar maps in 1981 [7]. After them, many others continued this work, starting with Schaeffer, who gave numerous bijective constructions in the late nineties. In 1997, he introduced blossoming trees to formulate a new bijection for planar maps [9]. In 2000, Bousquet-Mélou and Schaeffer gave a bijection between planar constellations and some blossoming trees, which allowed them to prove enumerative formulas for constellations [3]. It should be mentioned that there is a second trend of bijections of maps based on trees decorated with some integers that encode metric properties of the maps. These bijections were applied to planar constellations in [4] and were later extended to higher genus in [5].

In positive genus, the natural equivalent of trees are unicellular maps. Chapuy, Marcus and Schaeffer introduced in [6] some techniques to analyse these unicellular maps by decomposing them into schemes with branches. In 2019, Lepoutre gave a blossoming bijection for bicolorable maps of any genus, which are a particular case of constellations [8].

Inspired by the work of Lepoutre, we reformulate the planar blossoming bijection of [3] in a way that naturally extends to higher genus. Thus, we obtain a blossoming bijection between constellations and some blossoming unicellular maps that also extends the bijection of [8]. Using this bijection, we are able to enumerate a particular case of constellations on the torus.

2. Constellations and m-bipartite unicellular maps

2.1 Constellations

Definition 2.1. Let $m \geq 2$. We say that a map whose faces are bicolored (black and white) is an *m-constellation* (Figure 2a) if

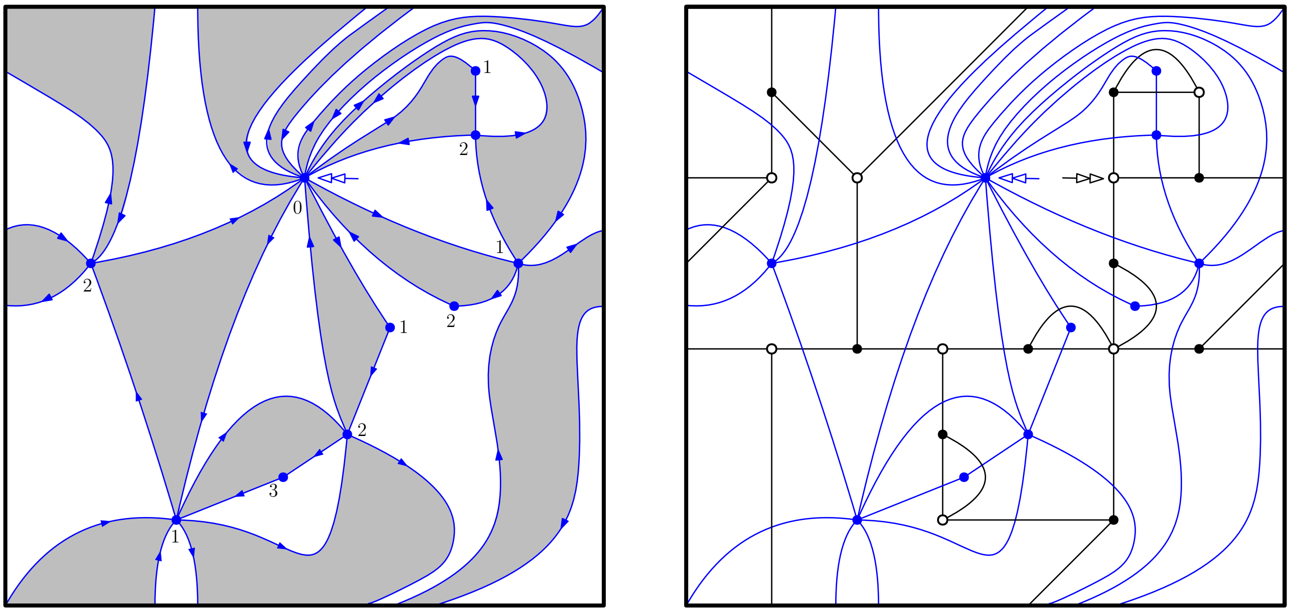
(i) adjacent faces have different colors,

(ii) black faces have degree m and white faces have degree mi for some integer $i \geq 1$ (which can be different among white faces),

(iii) vertices can be labeled with integers in $\{1, 2, \dots, m\}$ in such a way that turning clockwise around any black face the labels read $1, 2, \dots, m$.

A *rooted constellation* is a constellation that is rooted on a white corner. The first edge found when turning counterclockwise around the root vertex starting from the root corner is called the *root edge*. Note that the root corner can be recovered from the root edge, so it is equivalent to root a constellation on a white corner or on an edge.

The dual of an m-constellation is called an *m-Eulerian map* (Figure 2b). The dual of a rooted m-constellation (resp. rooted m-Eulerian map) is a rooted m-Eulerian map (resp. rooted m-constellation) with the "same" root. In other words, the root vertex, the root face and the root edge become, respectively, the root face, the root vertex and the root edge through dualization.



(a) A rooted 3-constellation of genus 1 endowed with its canonical orientation and labelling. The root is pointed by the double arrow.

(b) A rooted 3-constellation of genus 1 (blue) with its dual rooted 3-Eulerian map (black). Their roots are pointed by the double arrows.

Figure 2: A constellation and its dual map.

Consider a rooted m-constellation. The *canonical orientation* of its edges is the orientation for which its edges turn clockwise around black faces. When endowed with this orientation, the *canonical labelling* (Figure 2a) of its vertices is obtained by labelling every vertex with the length of the shortest oriented path to it from the root vertex.

This orientation and labelling was introduced by Bouttier, Di Francesco and Guitter in [4] for planar constellations to define what is now known as the BGD bijection.

2.2 Blossoming unicellular maps

Blossoming bijections were introduced by Schaeffer in [9] to put some classes of planar maps in bijection with decorated trees. These bijections consist in selecting a canonical spanning tree (or, more generally, a canonical spanning submap) and cut into two half-edges the edges not belonging to it. The resulting map is said to be a blossoming map, which can be *closed* back into the original map.

A *blossoming map* is a map with *stems* (that can be viewed as half-edges) attached to its vertices. There are two types of stems: *outstems*, which are outgoing stems, and *instems*, which are ingoing stems. Stems separate corners as if they were edges, which means that they count towards the (total) degree of their vertex. We will use the term *inner degree* when we want to ignore stems, i.e., when we only count the number of incident edges to a vertex (loops are counted twice).

A *rooted blossoming map* is a blossoming map with a marked instem, which is called the *root (instem)*. The vertex to which the root is attached is called the *root vertex* and the face incident to the root is called the *root face*. The corner on the right side of the root is called the *root corner*.

From now on we only consider blossoming maps that are unicellular.

The *good orientation* of a rooted unicellular blossoming map is the orientation for which every edge is, first, followed backwards and, then, forwards in tour around the unique face starting at the root corner. Note that it does not matter whether the face is followed clockwise or counterclockwise.

Given a rooted unicellular blossoming map which has m more instems than outstems, we can label its corners in the following way (Figure 3a). We make a counterclockwise tour around its unique face starting at the first corner after the root corner. Along this tour, we will visit every corner once and we will label it with the value of a counter that starts at $m-1$, increases by 1 every time we encounter an outstem and decreases by 1 every time we encounter an instem. The result of this procedure is called the *good labelling* of the unicellular blossoming map.

We say that an edge or stem *increases* (resp. *decreases*) by d if the value of its left label(s) minus the value of its right label(s) is d (resp. $-d$). Observe that, since there are m more instems than outstems, the last corner to label, which is the root corner, has good label 0.

2.3 m-bipartite unicellular maps

In [3], Bousquet-Mélou and Schaeffer define some objects called m-Eulerian trees and they construct a bijection between them and planar constellations. Here, we give a generalization of these objects to higher genus (m-bipartite unicellular maps) that we will show to be in bijection with constellations of higher genus.

Definition 2.2. Let $m \geq 2$. We say that a rooted unicellular blossoming map with m more instems than outstems and whose vertices are bicolored is an *m-bipartite unicellular map* (Figure 3a) if

(i) neighbouring vertices have different colors, instems are attached to white vertices and outstems are attached to black vertices,

(ii) black vertices have degree m,

(iii) white vertices have degree mi for some integer $i \geq 1$ (which can be different among white vertices),

and, when endowed with its good labelling,

(iv) the edges whose origin is a black vertex either decrease by 1 or increase by $m-1$,

(v) the edges whose origin is a white vertex decrease by $m-1$.

Given an m-bipartite unicellular blossoming map, consider the cyclic word formed by its stems in the order they appear in a counterclockwise tour around the face. Outstems are represented by the letter o and instems are represented by the letter i. Now we match letters o and i as if they were opening and closing parentheses, respectively. First, every letter o immediately followed by a letter i is *matched* with it. Then, all matched letters are removed and this procedure is repeated until no more matchings are possible (Figure 3b). Since there are exactly m more instems than outstems, m instems remain unmatched. We call these instems *single*. Note that the matching described is the only possible one, since, in a correct parenthesis word, an opening parentheses next to a closing one always have to be matched and can be ignored from that point on.

An m-bipartite unicellular map is *well-rooted* if its root instem is single. Well-rootedness can be characterized in the following way.

Proposition 2.3. *An m-bipartite unicellular map U is well-rooted if and only if its good labels are non negative.*

3. The bijection between m-constellations and m-bipartite unicellular maps

In this section we present our main result:

Theorem 3.1. *Rooted m-constellations of genus g with d_i white faces of degree mi are in bijection with well-rooted m-bipartite unicellular maps of genus g with d_i white vertices of degree mi.*

3.1 The closure Φ

We first describe how a well-rooted m-bipartite unicellular map can be closed to obtain an m-Eulerian map.

Definition 3.2. Let U be a well-rooted m-bipartite unicellular map. Let r be its root vertex. We define the *closure* $\Phi(U)$ of U in the following way (Figure 3).

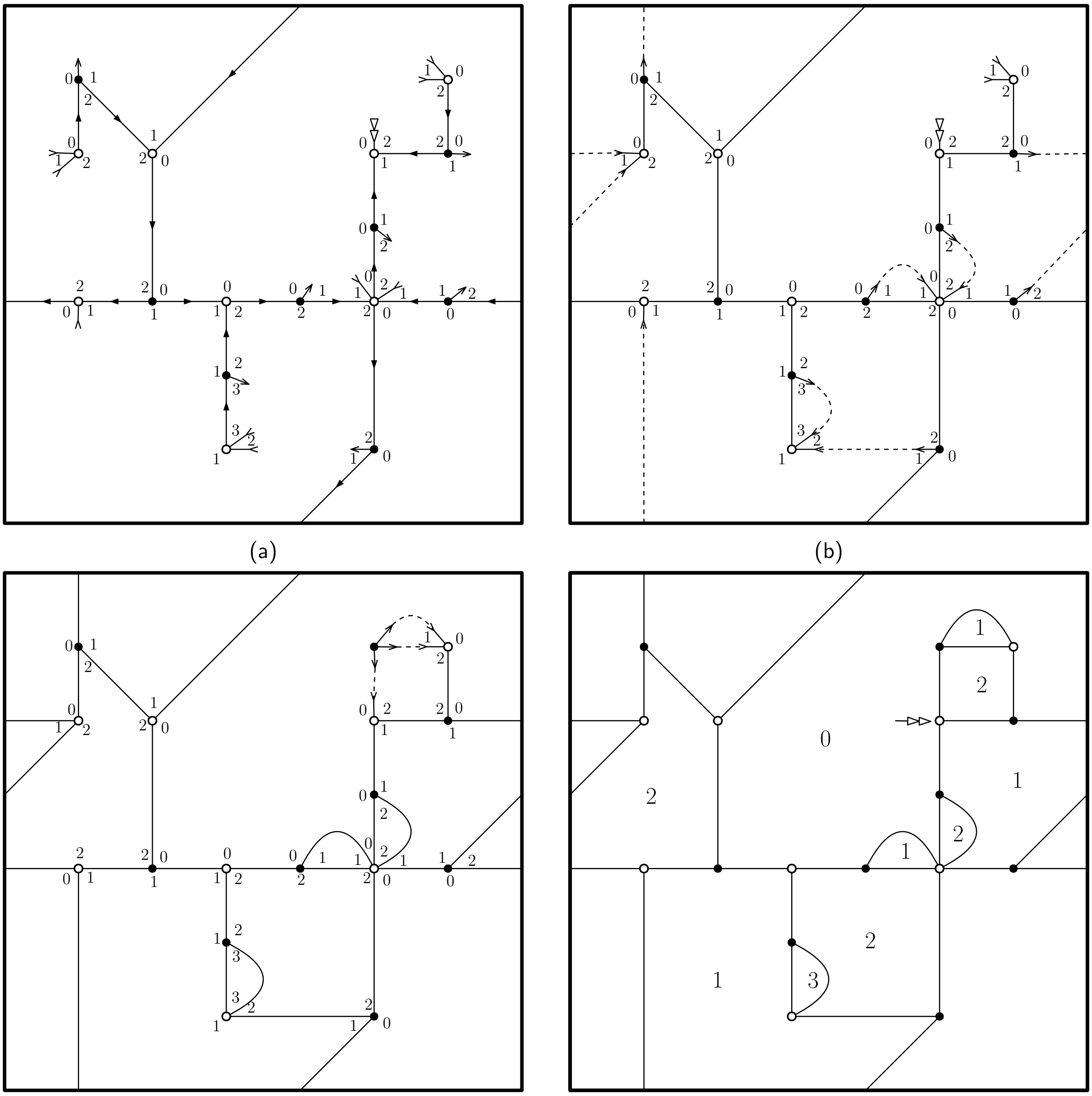
First, every pair of matched stems b, l is connected to form a complete edge. The fact that the matched stems of U form a valid parentheses word ensures that these new edges can be drawn without intersections.

After this, there are m unmatched instems, including the root. Place a black vertex s with m outstems attached to it in the unique face and connect each of the outstems to a different unmatched instems. It is clear that this can also be done without intersections.

The final result, $\Phi(U)$, is a map. We choose to root it on the same corner as U or, equivalently, on the edge joining r and s.

Lemma 3.3. *The closure $\Phi(U)$ of a well-rooted m-bipartite unicellular map U of genus g with d_i white vertices of degree mi is a rooted m-Eulerian map of genus g with d_i white vertices of degree mi. Moreover, the good labelling of the corners of U corresponds to the canonical labelling of the faces of $\Phi(U)$.*

Proof sketch. The rules of good labels around stems ensure that the closure is a rooted m-Eulerian map. Moreover, the canonical labels are at least as large as the good ones because when turning clockwise around black vertices the good labels either increase by one or decrease by $m-1$, and the equality holds because there is a path from the root face to any other face that crosses only edges created by joining stems. □



(a) (b) (c) (d)

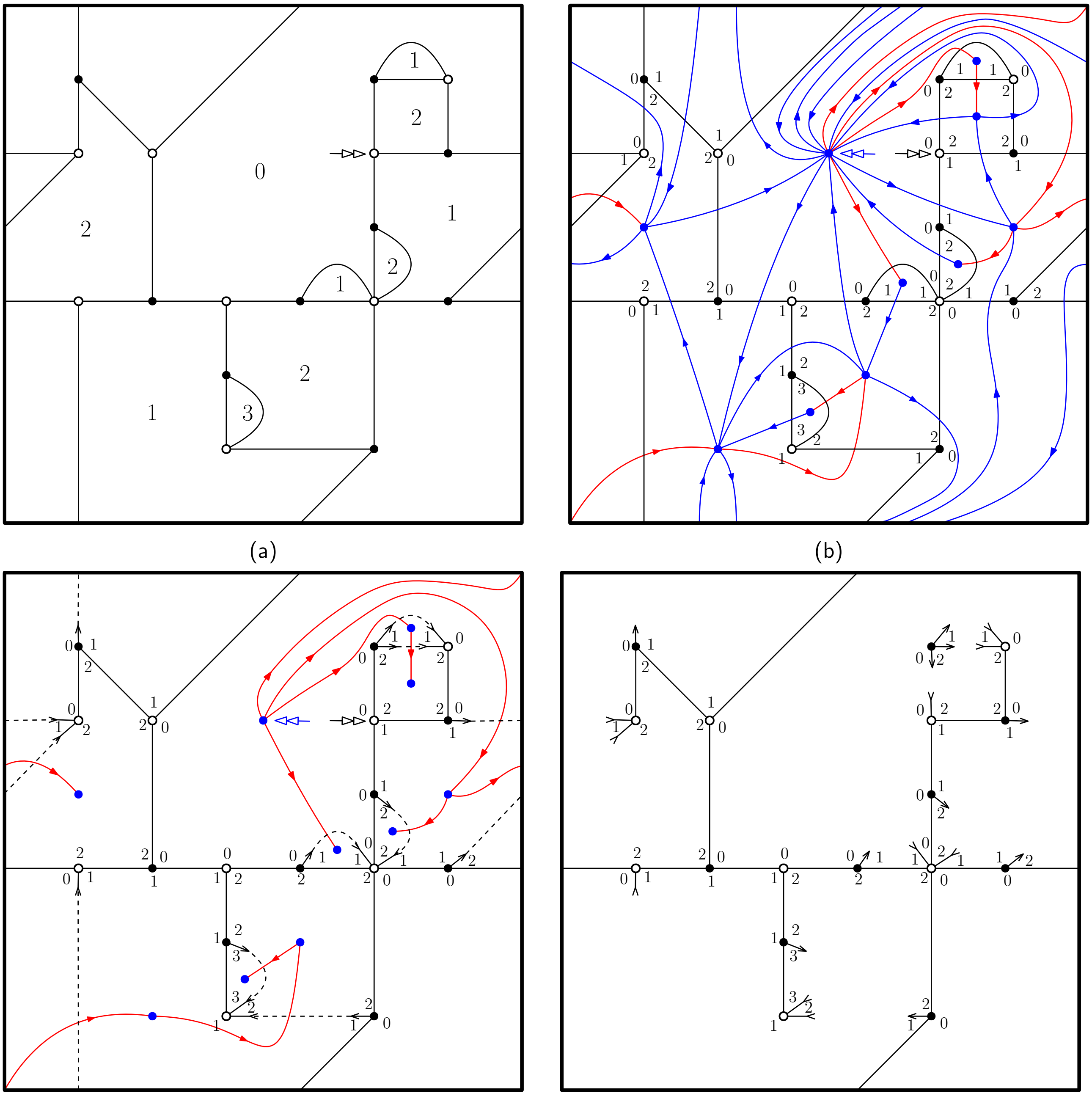
Figure 3: The closure of a well-rooted 3-bipartite unicellular map.

3.2 The opening Ψ

Here we do the inverse transformation, that is, starting from a rooted *m*-Eulerian map, cut some of its edges into stems so that the result is a well-rooted *m*-bipartite unicellular map.

Definition 3.4. Let M be a rooted m-Eulerian map. We define its *opening* $\Psi(M)$ in the following way (Figure 4). First, consider the dual map C of M, which is a rooted m-constellation. Endow C with its canonical orientation and labelling and take its leftmost Breadth-First Search (BFS) exploration tree T. For every edge of M whose dual belongs to T, cut it into two stems: an instem attached to the white vertex and an outstem attached to the black one. Finally, cut the root edge and remove s.

We root the result of this, $\Psi(M)$, at the instem created when cutting the root edge.



(a) (b) (c) (d)

Figure 4: The opening of a rooted 3-Eulerian map.

Lemma 3.5. *The opening $\Psi(M)$ of a rooted m-Eulerian map M of genus g with d_i white vertices of degree mi is a well-rooted m-bipartite unicellular map of genus g with d_i white vertices of degree mi. Moreover, the canonical labelling of M corresponds to the good labelling of $\Psi(M)$ and*

$$\Phi(\Psi(M)) = M.$$

Proof sketch. The only difficulty here is to show that, after the opening, there can be no edge oriented from white to black that increases by 1. If there was one such edge, the leftmost BFS tree would have seen, first, its left side and, then, its right side, which would be a contradiction. □

So far we have shown that the closure of a well-rooted *m*-bipartite unicellular map is a rooted *m*-Eulerian map and that the opening of a rooted *m*-Eulerian map is a well-rooted *m*-bipartite unicellular map whose closure is the original map. To show that Φ and Ψ are inverse operations and, thus, to prove Theorem 3.1, we just need the following lemma.

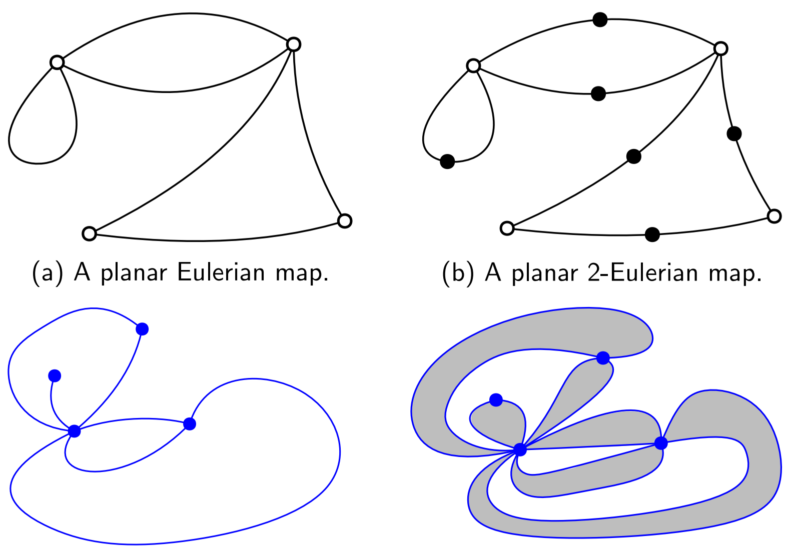
Lemma 3.6. *Let U be a well-rooted m-bipartite unicellular map. Then,*

$$\Psi(\Phi(U)) = U.$$

Proof sketch. Similarly, here one needs to prove that the duals of the edges that are created during the closure by joining stems form a leftmost BFS tree. □

Remark 3.7. The *m*-Eulerian trees described in [3] by Bousquet-Mélou and Schaeffer are the planar instances of the *m*-bipartite unicellular trees we have introduced here. We use the same closing operation as they do, but flipping the orientation of the surface, which amounts to swapping the notions of left-right and clockwise-counterclockwise. Thus, when we restrict our bijection to the sphere, we recover their bijection.

Remark 3.8. In [8], Lepoutre gives a bijection between bicolorable maps of arbitrary genus and an adequate family of blossoming unicellular maps. It is easy to convince oneself that bicolorable maps are, in fact, 2-Eulerian maps whose black vertices have been replaced by a single edge connecting their two white neighbours (Figure 5).



(a) A planar Eulerian map. (b) A planar 2-Eulerian map.

(c) The dual map of 5a. (d) The dual map of 5b: a planar 2-constellation.

Figure 5: The relation between bicolorable maps and 2-Eulerian maps.

In his bijection, Lepoutre opens bicolorable maps in a way that he shows to be equivalent to the following one. First, take the dual of the bicolorable map, which is a bipartite map, and endow it with its geodesic orientation. Then, consider the leftmost BFS exploration tree of this oriented bipartite map and, finally, cut all the edges of the bicolorable map whose dual does not belong to the tree. This is essentially the same way in which we open 2-Eulerian maps: the difference between our canonical orientation of 2-constellations (we orient vertices clockwise around black faces) and Lepoutre's geodesic orientation of the bipartite map is explained by the fact that he has collapsed the black faces of the 2-constellation into edges to obtain the bipartite map. Therefore, we can say that our bijection also generalizes the one given by Lepoutre in [8].

4. Rerooting an m-bipartite unicellular map

After Theorem 3.1 has been established, one can try to enumerate rooted constellations by enumerating well-rooted m-bipartite unicellular maps.

The first problem we run into when trying to count well-rooted m-bipartite unicellular maps is precisely the fact that they are well-rooted. As Lepoutre explains in [8], well-rootedness is a global notion, since it requires the positivity of all the good labels of a map. This complicates the task of counting these objects and, thus, we would like to get rid of it. In order to do so, we use the technique of rerooting first introduced in [9] and which was successfully used in [3] and [8]. Specifically, we provide an algorithm to reroot a well-rooted map on any instem, which will later yield an enumerative relation between m-bipartite unicellular maps and well-rooted m-bipartite unicellular maps.

Definition 4.1. Let U be a rooted m-bipartite unicellular map, let r be its root and let t be a distinguished instem of U. We endow U with its good orientation and its good labelling.

The *rerooting* algorithm is defined as follows. If $t = r$, we do nothing. Otherwise, we first join r and t to create an edge. This divides the single face of U into faces f_L and f_R, where f_L is the one containing the root corner of U. We then add m to all labels of f_L and we reverse the orientation of all the edges that separate f_L and f_R. Finally, we cut the edge joining r and t back into two instems and we swap the roles of r and t: t becomes the root and r becomes the distinguished instem.

The rerooting procedure always produces a valid m-bipartite unicellular map. This is why we say that these maps are stable under rerooting. Furthermore, it allows us to prove the following:

Proposition 4.2. *m-bipartite unicellular maps with a distinguished single instem are in bijection with well-rooted m-bipartite unicellular maps with a distinguished instem.*

5. Enumeration of bipartite 3-face-colorable cubic maps on the torus

In this section, we prove our second theorem:

Theorem 5.1. *Bipartite 3-face-colorable cubic maps of genus 1 are enumerated by*

$$C(z) = \frac{T(z)^3}{(1 - T(z))(1 - 4T(z))^2},$$

where z marks the number of white vertices and $T(z)$ is the unique generating function satisfying $T(z) = z + 2T(z)^2$. In particular, $C(z)$ is a rational function of $T(z)$.

Bipartite 3-face-colorable cubic maps of genus 1 are 3-Eulerian maps of genus 1 whose white vertices all have degree 3. This is a very particular case compared to the general m-constellations of arbitrary genus for which we have built a bijection, but it allows for relatively simple calculations that can be done by hand.

Generating functions are formal power series whose n-th coefficient equals the number of objects of size n in some combinatorial class. For example, if $\mathcal{C}$ is the class of rooted bipartite 3-face-colorable cubic maps of genus 1, counted by their number of white vertices, then $C(z) = \sum_{n\geq 0} c_n z^n$ is their generating function in the sense that there are exactly c_n such maps with n white vertices. We use the *Symbolic Method* to translate the relations between the combinatorial classes (classes of graphs in our setting) into equations involving their generating functions.

Let $\mathcal{O}$ be the class of well-rooted 3-bipartite unicellular maps of genus 1 whose white vertices have degree 3, counted by their number of instems. Since the number of instems of a map $o \in \mathcal{O}$ is equal to the number of white vertices of its closure $c \in \mathcal{C}$, $C(z) = O(z)$.

Let $\mathcal{U}$ be the class of 3-bipartite unicellular maps of genus 1 whose white vertices have degree 3 counted by their number of instems different from the root. By Proposition 4.2, we have the following.

Lemma 5.2. *The generating functions of $\mathcal{O}$ and $\mathcal{U}$ satisfy the relation*

$$O(t) = 3\int_0^t U(z)\, dz.$$

5.1 The pruned maps and their enumeration

We follow the framework introduced by Chapuy, Marcus and Schaeffer in [6] to study unicellular maps.

The *extended scheme* of a map $u \in \mathcal{U}$ is the map obtained by, first, removing all its stems and, then, iteratively removing all its vertices of degree 1. This procedure only removes stems and treelike parts from the map, so an extended map is also a unicellular map. In fact, any map $u \in \mathcal{U}$ can be decomposed into an extended scheme and some attached stems and treelike parts.

An extended scheme can only have vertices of degree 2, which we call *branch vertices*, and vertices of degree 3, which we call *scheme vertices*. The treelike parts can only be attached to white branch vertices. In our setting, there are always exactly two scheme vertices and they are black.

Let $\mathcal{T}$ be the class of these attachable treelike parts, counted by their number of instems. For the sake of simplicity, we will consider that a single instem is a treelike part and belongs to $\mathcal{T}$. It is easy to see that the generating function of $\mathcal{T}$ satisfies the following recursive relation:

$$T(z) = z + 2T(z)^2.$$

Let $u \in \mathcal{U}$ be a 3-bipartite unicellular map whose white vertices have degree 3. Its *pruned map* p is obtained by replacing all its treelike parts by instems. The treelike part containing the root is replaced by a root instem. Let $\mathcal{P}$ be the class obtained by pruning every map in $\mathcal{U}$. The pruned maps of $\mathcal{P}$ are counted by their number of instems different from the root. Observe that, if we keep the good labels on the pruned map, the rules of the labelling still apply. In other words, $\mathcal{P} \subset \mathcal{U}$.

Lemma 5.3. *The generating functions of $\mathcal{U}$, $\mathcal{P}$ and $\mathcal{T}$ satisfy the relation*

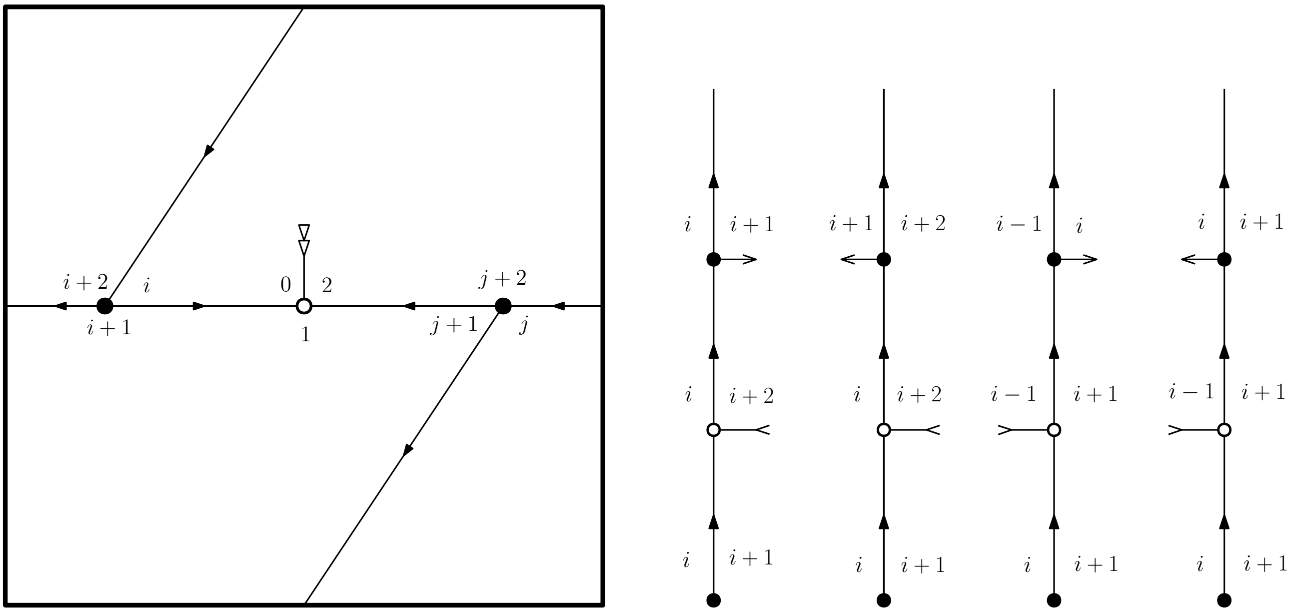
$$U(z) = \frac{\partial T}{\partial z} P(T(z)).$$

Proof sketch. Each stem in the pruned map is replaced by a tree and the root stem is replaced by a tree with a marked leaf that becomes the new root. □

We would now like to enumerate $\mathcal{P}$.

The *labelled scheme* of a pruned map $p \in \mathcal{P}$ is obtained by removing all its branch vertices except for the one where the treelike part containing the root was attached. The good labels of the remaining corners are kept.

It is clear that labelled schemes are uniquely determined by the lowest label on each of its scheme vertices (Figure 6a). There is, thus, a correspondance between labelled schemes and pairs $(i,j) \in \mathbb{Z}^2$.



(a) A generic labelled scheme.

(b) The first step of a branch.

Figure 6: Counting the pruned maps.

The labelled scheme associated to the pair (i,j) will be denoted $l_{i,j}$, and the subclass of pruned maps that have $l_{i,j}$ as labelled scheme will be denoted $\mathcal{P}_{i,j}$. Given $(i,j) \in \mathbb{Z}^2$, we want to compute $P_{i,j}(z)$. To do so, we replace every edge of $l_{i,j}$ by a valid branch whose labels agree with the labels of $l_{i,j}$. A branch starts at a black vertex. There are four ways to place the stems of the first two vertices (Figure 6b).

The generating functions of branches are obtained by using *weighted Motzkin paths*. Multiplying the four branches in a given $l_{i,j}$ and summing over all pairs (i,j) gives the following:

$$P = \sum_{i,j \in \mathbb{Z}} P_{i,j} = \cdots = \frac{z^2(2z-1)}{(z-1)^2(4z-1)^3}.$$

We can finally conclude the proof of Theorem 5.1:

$$\begin{aligned} C(z) = O(t) &= 3\int_0^t U(z)\,dz = 3\int_0^t \frac{\partial T}{\partial z} P(T(z))\,dz \\ &= 3\left(\int_0^z P(\eta)\,d\eta\right)_{|z=T(t)} = \frac{T(z)^3}{(1-T(z))(1-4T(z))^2}. \end{aligned}$$

In view of Theorem 5.1, we can formulate the following conjecture.

Conjecture 5.4. *Bipartite* 3*-face-colorable cubic maps of arbitrary genus are enumerated by a generating function which is a rational function of* $T(z)$.

Since blossoming bijections in [8] produce enumerative results in which there is scheme by scheme rationality, we hope that will also be the case for these maps.

References

[1] E.A. Bender, E.R. Canfield, "The asymptotic number of rooted maps on a surface", *J. Combin. Theory Ser. A* **43**(2) (1986), 244–257.

[2] E.A. Bender, E.R. Canfield, "The number of rooted maps on an orientable surface", *J. Combin. Theory Ser. B* **53**(2) (1991), 293–299.

[3] M. Bousquet-Mélou, G. Schaeffer, "Enumeration of planar constellations", *Adv. in Appl. Math.* **24**(4) (2000), 337–368.

[4] J. Bouttier, P. Di Francesco, E. Guitter, "Planar maps as labeled mobiles", *Electron. J. Combin.* **11**(1) (2004), Research Paper 69, 27 pp.

[5] G. Chapuy, "Asymptotic enumeration of constellations and related families of maps on orientable surfaces", *Combin. Probab. Comput.* **18**(4) (2009), 477–516.

[6] G. Chapuy, M. Marcus, G. Schaeffer, "A bijection for rooted maps on orientable surfaces", *SIAM J. Discrete Math.* **23**(3) (2009), 1587–1611.

[7] R. Cori, B. Vauquelin, "Planar maps are well labeled trees", *Canadian J. Math.* **33**(5) (1981), 1023–1042.

[8] M. Lepoutre, "Blossoming bijection for higher-genus maps", *J. Combin. Theory Ser. A* **165** (2019), 187–224.

[9] G. Schaeffer, "Bijective census and random generation of Eulerian planar maps with prescribed vertex degrees", *Electron. J. Combin.* **4**(1) (1997), Research Paper 20, 14 pp.

[10] W.T. Tutte, "A census of planar maps", *Canadian J. Math.* **15** (1963), 249–271.

REPORTS@SCM
AN ELECTRONIC JOURNAL OF THE
SOCIETAT CATALANA DE MATEMÀTIQUES

Bhargava cubes and elliptic curves

*Martí Oller Riera
University of Cambridge
mo512@cam.ac.uk

∗Corresponding author

Resum *(CAT)*
En les seves cèlebres Disquisitiones Arithmeticae, Gauss va descobrir una llei de composició que confereix una estructura de grup al conjunt de classes de formes quadràtiques binàries amb discriminant fixat. Dos segles més tard, Bhargava va donar una reinterpretació d'aquesta llei a través de cubs $2 \times 2 \times 2$ d'enters, ara coneguts com a cubs de Bhargava. El plantejament d'aquest article rau en utilitzar la mateixa idea dels cubs de Bhargava però en cubs $3 \times 3 \times 3$, que donen lloc a corbes planes projectives de grau 3. L'objectiu és determinar lleis de composició anàlogues que involucrin aquestes corbes. A tal fi, es desenvoluparan els coneixements matemàtics pertinents, incloent cohomologia de Galois i geometria algebraica, fent èmfasi en corbes ellptiques i, més en general, en les propietats de corbes de gènere 1.

Abstract *(ENG)*
In his celebrated Disquisitiones Arithmeticae, Gauss discovered a composition law that gives a group structure to the set of classes of binary quadratic forms of a given discriminant. Two centuries later, Bhargava gave a reinterpretation of this law through $2 \times 2 \times 2$ cubes of integers, now known as Bhargava cubes. In this article, we aim to use the same idea of Bhargava cubes but in $3 \times 3 \times 3$ cubes, that yield projective plane curves of degree 3. Our aim is to determine analogous composition laws involving these curves. To this end, we will review the needed mathematical knowledge, including Galois cohomology and algebraic geometry, with an emphasis on elliptic curves and, more generally, in the properties of genus one curves.

Keywords: *Bhargava cubes, genus one curves, elliptic curves, Galois cohomology, arithmetic geometry, number theory.*
MSC (2010): *Primary 11D09, 11G05. Secondary 11R34.*
Received: *July 21, 2022.*
Accepted: *September 28, 2022.*

Acknowledgement
The author wants to thank Jordi Guàrdia and Santi Molina for their guidance during the development of this project, and the anonymous referee for their helpful comments.

Societat Catalana de Matemàtiques
Institut d'Estudis Catalans

1. Introduction

In 1801, Gauss published his Disquisitiones Arithmeticae [10], which among many other topics study the composition of binary quadratic forms. More specifically, he found a group law between the classes of binary quadratic forms of a given discriminant. 200 years later, in his PhD thesis, Bhargava studied whether there were higher analogues of this law that could help interpret other number rings and their class groups. He did that by considering different-sized cubes of integers and the forms arising from them. Most notable is his approach in [2] using $2 \times 2 \times 2$ cubes of integers, which yield an elegant reinterpretation of Gauss composition and allows to obtain higher composition laws. His work led to a bigger understanding of parametrizations of quartic and quintic rings and the density of their discriminants.

The next obvious step would be to consider $3 \times 3 \times 3$ cubes. In [3], it is explained that $3 \times 3 \times 3$ cubes give rise to a composition law on general ternary cubic forms, but this composition doesn't directly give information on the corresponding cubic rings. In fact, cubic rings are most naturally related to binary cubic forms, obtained by $2 \times 3 \times 3$ cubes. This is the explanation given by Bhargava to focus on $2 \times 3 \times 3$ cubes rather than on the $3 \times 3 \times 3$ case.

The aim of this article is to explore the behaviour of $3 \times 3 \times 3$ cubes in a more geometrical setting. We will consider cubes with entries in some field K, which will give rise to genus one curves in the projective plane, and we will see how there is an analogous group law satisfied by these curves.

This article will begin with a brief exposition Gauss' composition and Bhargava's work in $2 \times 2 \times 2$ cubes. We will later introduce concepts in arithmetic geometry that will be necessary for us later. This includes a brief introduction to elliptic curves and more generally to genus one curves, and also Galois cohomology and its relation to elliptic curves. We will conclude by explaining results in the aforementioned $3 \times 3 \times 3$ cubes, in parallel with the results in [4].

2. Gauss' composition law and Bhargava cubes

Definition 2.1. A binary quadratic form is a polynomial of the form $f(x, y) = ax^2 + bxy + cy^2$, with $a, b, c \in \mathbb{Z}$. We say that f is primitive if $\gcd(a, b, c) = 1$. The discriminant of a binary quadratic form is defined to be $D := b^2 - 4ac$.

Definition 2.2. We say that two binary quadratic forms f, g are equivalent if there exists a matrix $S = \left(\begin{smallmatrix} r & s \\ t & u \end{smallmatrix}\right) \in \mathrm{SL}_2(\mathbb{Z})$ such that $g(x, y) = f(rx + sy, tx + uy)$. We will denote this as $f \sim g$.

It is not difficult to see that equivalence of binary quadratic forms is an equivalence relation, and that any two equivalent binary quadratic forms have the same discriminant.

Definition 2.3. Let f, g be two primitive binary quadratic forms with the same discriminant. A binary quadratic form h is a composition of f and g if the following conditions hold:

$$\begin{cases} f(x, y) \cdot g(z, w) = h(B_1(x, y, z, w), B_2(x, y, z, w)); \\ p_1 q_2 - p_2 q_1 = f(1, 0); \\ p_1 r_2 - p_2 r_1 = g(1, 0); \end{cases}$$

where $B_i(x, y, z, w) = p_i xz + q_i xw + r_i yz + s_i yw$ $(i = 1, 2)$ are two bilinear forms with integer coefficients.

Gauss famously proved that, in fact, composition gives a group law to the set of equivalence classes primitive binary quadratic forms of fixed discriminant D. More precisely:

Theorem 2.4 (Gauss). (i) *Given two primitive binary quadratic forms f, g of given discriminant D, there always exists a composition h of f and g. Moreover, this composition is unique and well-defined up to equivalence, meaning:*

(1) *If h_1, h_2 are two compositions of f and g, then $h_1 \sim h_2$.*
(2) *If h_i is the composition of f_i and g_i for $i = 1, 2$, satisfying $f_1 \sim f_2$ and $g_1 \sim g_2$, then $h_1 \sim h_2$.*

(ii) *The equivalence classes of primitive binary quadratic forms of fixed discriminant D constitute an abelian group under composition.*

(iii) *The identity is given by*

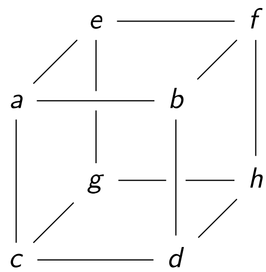
$$Q_{\text{id},D}(x,y) = \begin{cases} \left[x^2 - \dfrac{D}{4}\right], & \text{if } D \equiv 0 \pmod 4, \\ \left[x^2 + xy - \dfrac{D-1}{4}\right], & \text{if } D \equiv 1 \pmod 4, \end{cases}$$

where $[f]$ denotes the equivalence class of f.

There is a reinterpretation of Gauss composition due to Dirichlet, which relates the composition of binary quadratic forms with the multiplication of fractional ideals in orders of number fields; see [6] for more details.

We now present Bhargava's reinterpretation of the Gauss composition law through $2 \times 2 \times 2$ cubes.

Definition 2.5. A Bhargava cube is an element $A \in \mathbb{Z}^2 \otimes \mathbb{Z}^2 \otimes \mathbb{Z}^2$. If A is represented by (a, b, c, d, e, f, g, h) under a basis of $\mathbb{Z}^2 \otimes \mathbb{Z}^2 \otimes \mathbb{Z}^2$, then it can visualized as:



This cube can be partitioned into two 2×2 matrices in three different ways, according to the three orientations of the cube. Namely, the corresponding matrices are:

$$M_1 = \begin{bmatrix} a & b \\ c & d \end{bmatrix}, \quad N_1 = \begin{bmatrix} e & f \\ g & h \end{bmatrix};$$

$$M_2 = \begin{bmatrix} a & c \\ e & g \end{bmatrix}, \quad N_2 = \begin{bmatrix} b & d \\ f & h \end{bmatrix};$$

$$M_3 = \begin{bmatrix} a & e \\ b & f \end{bmatrix}, \quad N_3 = \begin{bmatrix} c & g \\ d & h \end{bmatrix}.$$

Given any such partition, we may obtain a binary quadratic form through:

$$Q_i^A(x,y) = -\det(M_i x - N_i y).$$

Under this setting, a natural question to ask is: how are these three binary quadratic forms related?

Theorem 2.6 (Bhargava)**.** *Let A be a Bhargava cube giving rise to three primitive binary quadratic forms Q_1, Q_2, Q_3. Then,*

(i) *The forms Q_1, Q_2, Q_3 have the same discriminant D.*

(ii) *The three forms satisfy*

$$[Q_1] + [Q_2] + [Q_3] = [Q_{\mathrm{id},D}],$$

where $+$ corresponds to Gauss composition and $Q_{\mathrm{id},D}$ is the identity form defined in Theorem 2.4.

(iii) *Conversely, given any three forms satisfying $[Q_1]+[Q_2]+[Q_3] = [Q_{\mathrm{id},D}]$, there exists a cube A giving rise to $[Q_1]$, $[Q_2]$, $[Q_3]$ (which is unique modulo a suitable action of $\mathrm{SL}_2(\mathbb{Z})$).*

Here, there is an action of $\Gamma = (\mathrm{SL}_2(\mathbb{Z}))^3$ on a cube $A \in \mathbb{Z}^2 \otimes \mathbb{Z}^2 \otimes \mathbb{Z}^2$. In terms of the partition, the action of the i-th matrix $\left(\begin{smallmatrix} r & s \\ t & u \end{smallmatrix}\right)$ replaces (M_i, N_i) for $(rM_i + sN_i, tM_i + uN_i)$.

3. Genus one curves

The main goal of this article will be to find an analogue to Theorem 2.6 but for $3 \times 3 \times 3$ cubes. To do that, we first need to introduce some concepts related to genus one curves.

3.1 Preliminaries in algebraic geometry

We will assume some familiarity with the basics of algebraic geometry. For further context, the reader may wish to consult [9] or the first two chapters of [12].

Fix throughout a perfect field K with algebraic closure $\overline{K}$. Let $C \subseteq \mathbb{P}^2$ be a curve, that is, the vanishing locus of an irreducible homogeneous polynomial $f(x, y, z)$ of degree d. We will denote by $C(K)$ the set of K-points of C, and we will typically denote the $\overline{K}$-points just by C.

Definition 3.1. The divisor group of C is the free abelian group generated by the $\overline{K}$-points of C. In other words, a divisor D of C is a formal sum

$$D = \sum_{P \in C(\overline{K})} n_P P,$$

where $n_P \in \mathbb{Z}$ and $n_P = 0$ for all but finitely many P. The degree of a divisor is defined by

$$\deg D = \sum_{P \in C(\overline{K})} n_P.$$

Finally, a principal divisor is of the form

$$\operatorname{div} f = \sum_{P \in C(\overline{K})} \operatorname{ord}_P(f) P,$$

for some $f \in \overline{K}(C)$.

The principal divisors of C form a subgroup of the divisor class group, since for any $f, g \in \overline{K}(C)$:

$$\operatorname{div}(fg) = \operatorname{div}(f) + \operatorname{div}(g), \quad \operatorname{div}(1/f) = -\operatorname{div}(f).$$

Definition 3.2. The Picard group of C is the quotient of its divisor group by the subgroup of principal divisors.

Another fundamental concept in our study is the genus $g(C)$ of the curve C. In our particular case where $C \subseteq \mathbb{P}^2$ is given by the vanishing locus of a homogeneous polynomial of degree d, the genus of C can be computed to be

$$g(C) = \frac{(d-1)(d-2)}{2},$$

if the curve C is non-singular (see e.g. [9, Chap. 8, Prop. 5]). Note in particular that if $d = 3$, then $g(C) = 1$.

3.2 Elliptic curves

Definition 3.3. An elliptic curve is a genus one curve E/K with a distinguished K-rational point $O_E \in E(K)$.

Proposition 3.4. *Let* char $K \neq 2, 3$. *Then,* E/K *is isomorphic to a projective plane curve of the form*

$$y^2 z = x^3 + axz^2 + bz^3,$$

where the point O_E *corresponds to the point at infinity* $(0 : 1 : 0)$. *The coefficients satisfy* $4a^3 + 27b^2 \neq 0$.

The points of an elliptic curve are known to have a natural group structure. Given two points $P, Q \in E(K)$, we define $P + Q$ with the following procedure, which is represented in Figure 1:

- If $P \neq Q$, the line passing through P and Q intersects E in another point R. Then, the line passing through O_E and R intersects E at a third point, which we define to be $P + Q$.
- If $P = Q$, we choose the first line to be the tangent line of E at P.
- If $R = O_E$, set $P + Q := O_E$.

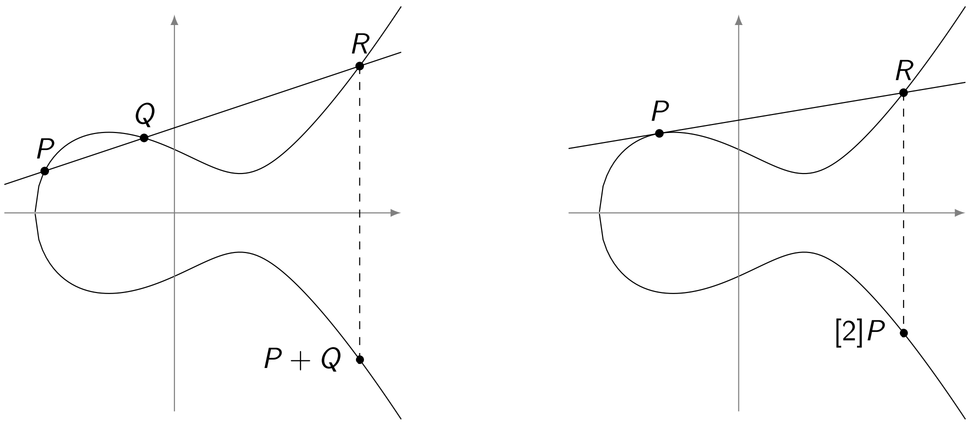


Figure 1: The group law of an elliptic curve. Figure obtained from [8].

Theorem 3.5. *The operation* $+$ *defines an abelian group structure on* E.

We denote by $E[n]$ the group of n-torsion points, that is, the group of points $P \in E$ such that $nP = 0_E$.

3.3 The Jacobian of a genus one curve

Given any genus one curve, there is a natural way to associate to it an elliptic curve, called the Jacobian. This can be made more precise (see [5, Chap. 20, Theor. 1]):

Proposition 3.6. *Let C/K be a genus one curve. Then, there exists an elliptic curve E/K together with an isomorphism $\phi\colon C \to E$ with the property that for every $\sigma \in \mathrm{Gal}(\overline{K}/K)$ the isomorphism $\varphi_\sigma\colon E \to E$ defined by $\varphi_\sigma = ({}^\sigma\phi) \circ \phi^{-1}$ is a translation by a point P_σ, for some $P_\sigma \in E(\overline{K})$. Moreover, E is unique up to K-isomorphism.*

We define the Jacobian of the curve C/K to be the elliptic curve E/K appearing in Proposition 3.6.

Proposition 3.7. *The group structure of the points of the Jacobian E is isomorphic to the degree-0 Picard group of C (which is the group defined in Definition 3.2 restricted to the divisors of degree 0).*

3.4 Models of genus one curves

If a genus one curve has a rational point, then it can be brought to a Weierstrass form, which is the form given by Proposition 3.4. However, if the curve does not have a rational point, we have to seek other models for the curve. We will follow the exposition in [1].

Assume that a genus one curve C/K has a K-rational divisor D, meaning that $\sigma D = D$ for all $\sigma \in \mathrm{Gal}(\overline{K}/K)$. Assume $\deg D = n > 0$, and define

$$\mathcal{L}(D) := \{f \in K(C) \mid \mathrm{div}(f) + D \geq 0\} \cup \{0\}.$$

It is a K-vector space, and the Riemann–Roch theorem in this case tells us that $\dim_K \mathcal{L}(D) = n$ (see [9]).

Let us focus on the case $n = 3$ (the cases $n = 2, 4$ are covered in [1]). Since $\dim_K \mathcal{L}(D) = 3$, we choose a K-basis of $\mathcal{L}(D)$, say $\{x, y, z\}$. Then, the ten elements x^3, x^2y, x^2z, xy^2, xyz, xz^2, y^3, y^2z, yz^2, z^3 all belong to the 9-dimensional space $\mathcal{L}(3D)$, so there exists a linear relation between these elements. In other words, there exists a ternary cubic form U such that

$$U(x, y, z) = 0.$$

In [1], there is an expression for two invariants of U, which we will call c_4 and c_6. We can further define $\Delta = \frac{c_4^3 - c_6^2}{1728}$.

Theorem 3.8. *The equation $U(x, y, z) = 0$ defines a genus one curve if and only if $\Delta \neq 0$. In that case, and if $\mathrm{char}\, K \neq 2, 3$, the Jacobian of the curve is*

$$y^2 = x^3 - 27c_4x - 57c_6.$$

3.5 Galois cohomology and elliptic curves

Let G be a topological group (i.e. G has a topology where the group operation and the inverse are continuous).

Definition 3.9. An abelian group M is a G-module if there is an action $G \times M \to M$ satisfying, for all $g, g' \in G$, $m, m' \in M$:

(i) $g(m + m') = gm + gm'$.

(ii) $(gg')m = g(g'm)$.

(iii) $1m = m$.

(iv) The G-action is continuous with respect to the topology on G and the discrete topology on M.

Definition 3.10. A morphism of G modules is a group morphism $\alpha\colon M \to N$ respecting the G-action on M and N.

Now, let K be a perfect field and set $G_K = \mathrm{Gal}(\overline{K}/K)$. We note that G_K is naturally a topological group under the Krull topology.

Definition 3.11. Let M be a G_K-module. Then, its 0-th cohomology group is

$$H^0(K, M) := M^{G_K} = \{m \in M \mid gm = m \text{ for all } g \in G_K\}.$$

Definition 3.12. Let M be a G_K-module. The group of 1-cocycles is given by

$$Z^1(K, M) = \{\xi\colon G_K \to M \mid \xi(gh) = g(\xi(h)) + \xi(g),\ \xi \text{ continuous}\},$$

Its subgroup of 1-coboundaries $B^1(K, M)$ consists of the cocycles $\xi \in Z^1(K, M)$ such that ξ is of the form $\xi(g) = gm - m$ for some $m \in M$. Then, the 1st cohomology group is

$$H^1(K, M) = \frac{Z^1(K, M)}{B^1(K, M)}.$$

Proposition 3.13. *Consider the exact sequence of G_K-modules given by*

$$0 \longrightarrow P \longrightarrow M \longrightarrow N \longrightarrow 0.$$

Then, there is a long exact sequence

$$\begin{array}{l} 0 \longrightarrow H^0(K, P) \longrightarrow H^0(K, M) \longrightarrow H^0(K, N) \\ \xrightarrow{\ \delta\ } H^1(K, P) \longrightarrow H^1(K, M) \longrightarrow H^1(K, N). \end{array}$$

Remark 3.14. We could define higher cohomology groups $(H^2, H^3 \dots)$ that would continue the long exact sequence in an analogous manner.

Let us return to the setting of elliptic curves. For an elliptic curve E/K, there is a natural Galois action defined component-wise. Let us consider the exact sequence

$$0 \longrightarrow E[n] \longrightarrow E \xrightarrow{\times n} E \longrightarrow 0.$$

Then, Proposition 3.13 gives us a long exact sequence:

$$0 \longrightarrow E(K)[n] \longrightarrow E(K) \xrightarrow{\times n} E(K) \xrightarrow{\delta} H^1(K, E[n]) \longrightarrow H^1(K, E) \xrightarrow{\times n} H^1(K, E).$$

From this long exact sequence we can extract a short exact sequence:

$$0 \longrightarrow \frac{E(K)}{nE(K)} \xrightarrow{\delta} H^1(K, E[n]) \xrightarrow{\iota} H^1(K, E)[n] \longrightarrow 0. \quad (1)$$

This sequence is known as the Kummer exact sequence for E/K.

3.6 The twisting principle

We will conclude this section by interpreting what $H^1(K, E[n])$ and $H^1(K, E)$ are. We will assume that char $K \nmid n$, a harmless assumption given that later we will deal with $n = 3$ and $\mathrm{char}(K) \neq 2, 3$. We will follow the exposition in [7].

We will make use of the twisting principle, which says that if X/K is an object defined over K, then K-isomorphism classes of twists of X (other objects Y/K isomorphic to X over $\overline{K}$) are parametrized by $H^1(K, \mathrm{Aut}(X))$, where $\mathrm{Aut}(X)$ is the automorphism group of X.

Here, the twisting principle is stated rather loosely, but it will be true for all our applications. See [11] and [7] for further details.

In view of the principle, if we are able to find a suitable object such that $\mathrm{Aut}(X)$ is E or $E[n]$, then we will be able to interpret the objects arising in the Kummer sequence.

Definition 3.15. A torsor under E is a pair (C, μ), where C is a smooth projective curve of genus one defined over K, and $\mu\colon E \times C \to C$ is a morphism defined over K that induces a simple transitive action on $\overline{K}$-points.

An isomorphism of torsors $(C_1, \mu_1) \cong (C_2, \mu_2)$ is an isomorphism of the underlying curves that respects the E-action.

Lemma 3.16. *Every torsor under E is a twist of $(E, +)$, where $(E, +)$ is the trivial torsor given by the group law. Moreover,* $\mathrm{Aut}(E, +) = E$.

Hence, by the twisting principle:

Theorem 3.17. *The group $H^1(K, E)$ parametrizes the torsors of E.*

Definition 3.18. A torsor divisor class pair $(C, [D])$ is a pair consisting of a torsor C of E and a K-rational divisor class $[D]$ of degree n. Here, rationality means that $\sigma(D) \sim D$ for all $\sigma \in \mathrm{Gal}(\overline{K}/K)$.

Two such pairs $(C_1, [D_1])$ and $(C_2, [D_2])$ are isomorphic if there is an isomorphism of torsors $\phi\colon C_1 \to C_2$ such that $\phi^* D_2 \sim D_1$.

Lemma 3.19. *Every torsor divisor class pair is a twist of $(E, [nO_E])$, O_E is the point at infinity of E. Moreover,* $\mathrm{Aut}(E, [nO_E]) = E[n]$.

The twisting principle in this case gives

Theorem 3.20. *The group* $H^1(K, E[n])$ *parametrizes the* K*-isomorphism classes of torsor divisor class pairs.*

In particular, it can be shown that in the Kummer sequence

$$0 \longrightarrow \frac{E(K)}{nE(K)} \stackrel{\delta}{\longrightarrow} H^1(K, E[n]) \stackrel{\imath}{\longrightarrow} H^1(K, E)[n] \longrightarrow 0,$$

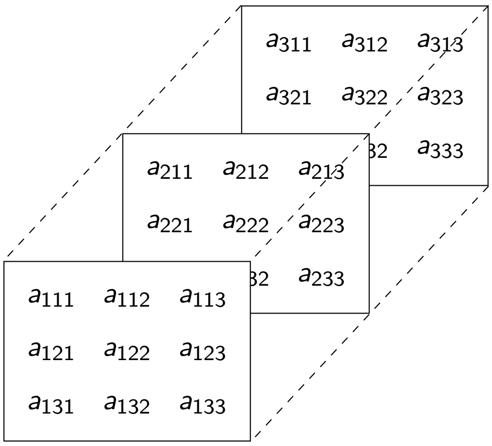
the maps are defined as

$$\delta(P) = (E, [(n-1)O_E + P]) \quad \text{and} \quad \imath(C, [D]) = C.$$

4. $3 \times 3 \times 3$ Bhargava cubes

Most of the results that we present here appear in [4]. However, this article takes a slightly different point of view more focused in the group law of Theorem 4.4.

Assume that K is a perfect field and $\mathrm{char}(K) \neq 2, 3$. Let us consider a $3 \times 3 \times 3$ cube (a_{ijk}) with entries in K. It can be drawn as follows:



Analogously to the $2 \times 2 \times 2$ case, we can partition this cube in three different ways to obtain the "front" section, the "top" section and the "side" section:

$$A_{f,i} = \begin{pmatrix} a_{i11} & a_{i12} & a_{i13} \\ a_{i21} & a_{i22} & a_{i23} \\ a_{i31} & a_{i32} & a_{i33} \end{pmatrix}, \quad A_{t,i} = \begin{pmatrix} a_{1i1} & a_{1i2} & a_{1i3} \\ a_{2i1} & a_{2i2} & a_{2i3} \\ a_{3i1} & a_{3i2} & a_{3i3} \end{pmatrix}, \quad A_{s,i} = \begin{pmatrix} a_{11i} & a_{21i} & a_{31i} \\ a_{12i} & a_{22i} & a_{32i} \\ a_{13i} & a_{23i} & a_{33i} \end{pmatrix},$$

for $i = 1, 2, 3$. These three partitions yield three polynomials using:

$$P_\bullet(X, Y, Z) = \det(A_{\bullet,1}X + A_{\bullet,2}Y + A_{\bullet,3}Z),$$

with $\bullet = f, s, t$.

We note that P_f, P_t, P_s are homogeneous polynomials in X_1, X_2, X_3 of degree 3, and hence define algebraic sets in $\mathbb{P}^2$:

$$C_\bullet = \{(X_1, X_2, X_3) \in \mathbb{P}^2 \mid P_\bullet(X_1, X_2, X_3) = 0\}.$$

By the discussion in Theorem 3.8, any of these algebraic sets define a smooth genus one curve if and only if their discriminant Δ defined in Subsection 3.4 is different from 0.

Lemma 4.1. *All curves C_f, C_s, C_t share the same invariants c_4, c_6, and hence the same discriminant Δ. In particular, if one of the curves is smooth, then they all are smooth.*

The proof can be done with an explicit computation, which we omit. Assume from now on that all three curves are smooth.

Theorem 4.2. *Let C_f, C_s, C_t be the curves arising from a $3 \times 3 \times 3$ cube, and assume they are all smooth. Then, all three curves are isomorphic over K.*

Sketch of proof. Let $(x_1, x_2, x_3) \in C_f$, and consider the matrix $M_f(x_1, x_2, x_3) = A_{f,1}x_1 + A_{f,2}x_2 + A_{f,3}x_3$. Then, the columns $c_{f,i}(x_1, x_2, x_3)$ of this matrix are linearly dependent, say by some coefficients $(X_1, X_2, X_3) \in \mathbb{P}^2$. Then, a quick computation shows that

$$\begin{aligned} 0 &= X_1 c_{f,1}(x_1, x_2, x_3) + X_2 c_{f,2}(x_1, x_2, x_3) + X_3 c_{f,3}(x_1, x_2, x_3) \\ &= x_1 c_{s,1}(X_1, X_2, X_3) + x_2 c_{s,2}(X_1, X_2, X_3) + x_3 c_{s,3}(X_1, X_2, X_3). \end{aligned} \tag{2}$$

The assignment $\varphi_{fs} \colon C_f \to C_s$ given by sending $(x_1, x_2, x_3) \mapsto (X_1, X_2, X_3)$ can be seen to be an isomorphism of algebraic curves. $\square$

We can analogously choose φ_{ft}, φ_{sf}, φ_{st}, φ_{tf} and φ_{ts}. It holds that $\varphi_{ij} = \varphi_{ji}^{-1}$ for any choice of i, j; but in general it is not true that $\varphi_{ki} \circ \varphi_{jk} \circ \varphi_{ij}$ is the identity.

We can interpret Theorem 4.2 as the analogue of Theorem 2.6, item (i). Both results restrict how "different" the arising objects can be: the binary quadratic forms have the same discriminant and the genus one curves are isomorphic.

In particular, given that the three curves are isomorphic, they have the same Jacobian curve, which we will call E.

We still need to find out whether these three curves obey some suitable group law. To this end, consider the divisor at infinity D_f of C_f, given by the intersection of C_f with any hyperplane (if we change the hyperplane, we get a linearly equivalent divisor). Similarly, consider the divisors at infinity D_s and D_t of C_s and C_t, let $\Delta_f = D_f$, and let Δ_s and Δ_t be the pullbacks of D_s, D_t with respect to φ_{fs} and φ_{ft}, respectively. Then, define

$$\alpha_f = (C_f, [\Delta_f]), \quad \alpha_s = (C_s, [\Delta_s]), \quad \alpha_t = (C_t, [\Delta_t]).$$

By Theorem 3.20, the elements α_f, α_s, α_t can be interpreted in $H^1(K, E[3])$. Additionally, it can be shown that there exist points P_f, P_s, P_t in C_f such that $3P_\bullet \sim \Delta_\bullet$ for $\bullet = f, s, t$.

Lemma 4.3. *Assume Δ_f, Δ_s, Δ_t arise from a cube. Then,*

$$2\Delta_f \sim \Delta_s + \Delta_t,$$

and Δ_f is not linearly equivalent to either of Δ_s or Δ_t.

Theorem 4.4. *The three cocycles $\alpha_f, \alpha_s, \alpha_t \in H^1(K, E[3])$ satisfy*

$$\alpha_f + \alpha_s + \alpha_t = 0.$$

Proof. Recall that a degree 0 divisor can be identified as a point in the Jacobian (see Subsection 3.3). We claim that:

$$Q = P_t + P_s - 2P_f \in E[3].$$

Indeed, remembering that $3P_\bullet \sim \Delta_\bullet$, for $\bullet = f, s, t$, and using the previous lemma:

$$3Q = 3P_t + 3P_s - 6P_f \sim \Delta_t + \Delta_s - 2\Delta_f \sim 0.$$

We conclude that the cocycle $\alpha_f + \alpha_s + \alpha_t$ is given by

$$\begin{aligned}\alpha_f(\sigma) + \alpha_s(\sigma) + \alpha_t(\sigma) &= \sigma(P_f + P_s + P_t) - (P_f + P_s + P_t) \\ &= (\sigma Q - Q) + \sigma(3P_f) - 3P_f \\ &\sim (\sigma Q - Q) + \sigma\Delta_f - \Delta_f = \sigma Q - Q,\end{aligned}$$

since Δ_f is a K-rational divisor. Thus, $\alpha_f + \alpha_s + \alpha_t$ is a coboundary and the result follows. □

4.1 Converse results

Now, we are interested in the converse to Theorem 4.4, namely: given any three cocycles $\alpha_1, \alpha_2, \alpha_3 \in H^1(K, E[3])$, does there exist a cube giving rise to them? To start answering the question, we first state the converse result for divisors.

Theorem 4.5. *There is a bijection between:*

(i) $3 \times 3 \times 3$ *cubes (modulo a suitable action of* $\mathrm{GL}_3(K)$*).*

(ii) *Isomorphism classes of* $(C, \Delta_f, \Delta_s, \Delta_t)$*, where C is a genus one curve and Δ_f, Δ_s, Δ_t are K-rational divisors of degree 3 satsifying* $2\Delta_f \sim \Delta_s + \Delta_t$ *and* $\Delta_f \nsim \Delta_s, \Delta_t$.

See [4] for the proof.

Now, we recall again the Kummer exact sequence

$$0 \longrightarrow \frac{E(K)}{nE(K)} \xrightarrow{\ \delta\ } H^1(K, E[n]) \xrightarrow{\ \imath\ } H^1(K, E)[n] \longrightarrow 0.$$

By the definition of the map $\imath$, if we have any three cocycles $\alpha_1, \alpha_2, \alpha_3 \in H^1(K, E[n])$ we need to have $\imath(\alpha_1) = \imath(\alpha_2) = \imath(\alpha_3)$.

However, there is still one more consideration to make, which is that an element of $H^1(K, E[3])$ is not necessarily represented by a projective cubic plane curve. Given a torsor divisor class pair $(C, [D])$, the divisor D does not necessarily satisfy $\sigma D = D$ for every $\sigma \in \mathrm{Gal}(\overline{K}/K)$, but rather that $\sigma D \sim D$. By the discussion in Subsection 3.4, the curve C needs to have a K-rational divisor D in order to be represented by a projective plane cubic curve.

In [7], an obstruction map Ob is defined, so that $\mathrm{Ob}(\alpha) = 0$ for $\alpha \in H^1(K, E[n])$ if and only if the cocycle α can be represented by $(C, [D])$, with $\sigma D = D$ for all σ in the Galois group.

Finally, observe that the action of a translation by a point $P \in E(K)$ does not change the cocycle. In other words, $(C,[D]) = (C,[D'])$, where D' is obtained by the action of $3P$ on D. If $3E(K) = \{O_E\}$, then in the case where $\alpha_1 = \alpha_2 = \alpha_3$ we would not be able to change the divisor corresponding to the cocycles and hence we would not be able to guarantee the conditions in Lemma 4.3.

Theorem 4.6. *Assume that* $\alpha_1, \alpha_2, \alpha_3 \in H^1(K, E[3])$ *satisfy*

(i) $\alpha_1 + \alpha_2 + \alpha_3 = 0$.

(ii) $\imath(\alpha_1) = \imath(\alpha_2) = \imath(\alpha_3)$.

(iii) $Ob(\alpha_1) = Ob(\alpha_2) = Ob(\alpha_3) = 0$.

(iv) $3E(K) \neq \{O_E\}$ *if* $\alpha_1 = \alpha_2 = \alpha_3$.

Then, there exists a cube giving rise to α_1, α_2, α_3.

As a final remark, observe that we are dealing with elements $\alpha_1, \alpha_2, \alpha_3 \in H^1(K, E[3])$ with $\imath(\alpha_1) = \imath(\alpha_2) = \imath(\alpha_3)$. By looking at the Kummer sequence, we see that the group law is actually taking place more naturally in $E(K)/3E(K)$. Therefore, Theorem 4.6 can be restated more naturally:

Corollary 4.7. *Assume we have a genus one curve* C/K *with Jacobian* E/K, *and suppose given three points* $P_1, P_2, P_3 \in E(K)/3E(K)$ *such that* $P_1 + P_2 + P_3 = 0$ *in* $E(K)/3E(K)$. *Then, there exists a cube giving rise to this information as long as we avoid the case where* $P_1 = P_2 = P_3$ *and* $3E(K) = \{O_E\}$.

References

[1] An Sang Yook et al., "Jacobians of genus one curves", *J. Number Theory* **90(2)** (2001), 304–315.

[2] M. Bhargava, "Higher composition laws. I. A new view on Gauss composition, and quadratic generalizations", *Ann. of Math. (2)* **159(1)** (2004), 217–250.

[3] M. Bhargava, "Higher composition laws. II. On cubic analogues of Gauss composition", *Ann. of Math. (2)* **159(2)** (2004), 865–886.

[4] M. Bhargava, W. Ho, "Coregular spaces and genus one curves", *Camb. J. Math.* **4(1)** (2016), 1–119.

[5] J.W.S. Cassels, *Lectures on Elliptic Curves*, London Mathematical Society Student Texts **24**, Cambridge University Press, Cambridge, 1991.

[6] D.A. Cox, *Primes of the Form* $x^2 + ny^2$, Fermat, class field theory and complex multiplication, A Wiley-Interscience Publication, John Wiley & Sons, Inc., New York, 1989.

[7] J.E. Cremona, T.A. Fisher, C. O'Neil, D. Simon, M. Stoll, "Explicit *n*-descent on elliptic curves. I. Algebra", *J. Reine Angew. Math.* **615** (2008), 121–155.

[8] L. De Feo, "Mathematics of isogeny based cryptography", Preprint (2017), `http://arxiv.org/abs/1711.04062`.

[9] W. Fulton, *Algebraic Curves. An Introduction to Algebraic Geometry*, Notes written with the

collaboration of Richard Weiss, Reprint of 1969 original, Advanced Book Classics, Addison-Wesley Publishing Company, Advanced Book Program, Redwood City, CA, 1989.

[10] C.F. Gauss, *Disquisitiones Arithmeticae*, Translated and with a preface by Arthur A. Clarke, Revised by William C. Waterhouse, Cornelius Greither and A. W. Grootendorst and with a preface by Waterhouse, Springer-Verlag, New York, 1986.

[11] J.-P. Serre, *Galois Cohomology*, Springer-Verlag, 1964.

[12] J.H. Silverman, *The Arithmetic of Elliptic Curves*, Second edition, Graduate Texts in Mathematics **106**, Springer, Dordrecht, 2009.

REPORTS@SCM

AN ELECTRONIC JOURNAL OF THE
SOCIETAT CATALANA DE MATEMÀTIQUES

Dynamics of a family of meromorphic functions

*Àlex Rodríguez Reverter
Universitat de Barcelona
alexrodriguezreverter@gmail.com

*Corresponding author

Resum *(CAT)*

En aquest projecte estudiem el comportament dinàmic de la família de funcions transcendents meromorfes $f_\lambda(z) = \lambda(\frac{e^z}{z+1} - 1)$, la qual es pot veure com l'anàleg meromorf de la ben coneguda família de polinomis cúbics de Milnor $P_a(z) = z^2(z-a)$ [12] o la seva versió entera $\lambda z^2 e^z$ [7, 8]. Contràriament a aquests dos casos, les conques d'atracció de f_λ no són simplement connexes. De fet, en aquest document es demostra que sota certes condicions, la conca d'atracció de $z = 0$ és infinitament connexa.

Abstract *(ENG)*

In this paper we analyze the dynamical behavior of the family of transcendental meromorphic maps $f_\lambda(z) = \lambda(\frac{e^z}{z+1} - 1)$. This family is the meromorphic analogue of the well-known Milnor family of cubic polynomials $P_a(z) = z^2(z-a)$ [12] or its entire version $\lambda z^2 e^z$ [7, 8]. Opposed to these two cases, the basins of attraction of f_λ are not simply connected. In fact, we prove that under certain conditions, the basin of attraction of $z = 0$ is infinitely connected.

Societat Catalana de Matemàtiques

Institut d'Estudis Catalans

Keywords: *holomorphic dynamics, complex analysis, dynamical systems.*
MSC (2020): *37F10, 30D45, 30D30.*
Received: *July 11, 2022.*
Accepted: *September 29, 2022.*

Acknowledgement

I would like to express my most sincere gratitude to Núria Fagella, my mentor and advisor.

1. Introduction

In this work we focus on some dynamical aspects of transcendental meromorphic functions, i.e., we study the dynamical systems given by the iterates of meromorphic functions $f\colon \mathbb{C} \to \mathbb{C}_\infty$ with an essential singularity at ∞, where $\mathbb{C}_\infty$ denotes $\mathbb{C} \cup \{\infty\}$ or the Riemann sphere. Here the n-th iterate of a point $z \in \mathbb{C}$ is denoted by $f^n(z) = (f \circ \overset{(n)}{\cdots} \circ f)(z)$, and the sequence of iterates $\{f^n(z)\}_{n\in\mathbb{N}}$ is well-defined for all $z \in \mathbb{C}$ except for the countable set of poles and prepols of f of any order.

The interest for these functions is twofold: The essential singularity, on the one hand, adds a lot of chaos to the dynamical system, mainly because of Picard's Theorem, which states that in each punctured neighborhood of ∞, these functions assume each value of the Riemann sphere $\mathbb{C}_\infty$, with at most two exceptions (such exceptional values are known as omitted values), infinitely often. Hence, given a point $z \in \mathbb{C}$, if its orbit $\mathcal{O}_f^+(z) = \{f^n(z) : n \in \mathbb{N}\}$ is near ∞ at some moment, after one iteration it can land at almost any place of the plane. On the other hand, the presence of poles allows for more generality, when compared to entire functions, since ∞ is not required to be an omitted value.

The phase space (also called dynamical plane) of a meromorphic function f splits into two completely invariant sets: The Fatou set $F(f)$, which is the set of points $z \in \mathbb{C}$ such that the sequence of iterates $\{f^n\}_{n\in\mathbb{N}}$ is defined and normal in some neighborhood of z; and its complement, the Julia set $J(f)$.

It follows trivially from the definition that the Fatou set is open and hence the Julia set is closed. The first consists of components known as Fatou components, each of them might be either simply or multiply connected (including the infinitely connected case as we will see here). Let $U = U_0$ be a Fatou component, then $f^n(U)$ is contained in another component of $F(f)$ that we denote by U_n. We say that U_0 is preperiodic if $U_n = U_m$ for some $n > m \geq 0$ (if $m = 0$, we say that its periodic and if $n = 1$, we say that it is fixed or forward invariant), otherwise we say that U is wandering. Periodic Fatou components are classified according to the following celebrated result of Fatou [2], which for simplicity we state for fixed components.

Theorem 1.1 (Classification Theorem for fixed Fatou components)**.** *Let U be a fixed Fatou component. Then we have one of the following possibilities:*

(i) *U contains an attracting fixed point z_0 and $f^n(z) \xrightarrow[n\to\infty]{} z_0$ for all $z \in U$, which is called the* immediate attractive basin *of z_0.*

(ii) *∂U contains a fixed point z_0 and $f^n(z) \xrightarrow[n\to\infty]{} z_0$ for all $z \in U$. Moreover, $f'(z_0) = 1$ if $z_0 \in \mathbb{C}$ and U is called a* parabolic (or Leau) domain*.*

(iii) *There exists $\phi\colon U \to \mathbb{D}$ conformal such that $\phi(f(\phi^{-1}(z))) = e^{2\pi i\alpha}z$ for some $\alpha \in \mathbb{R} \setminus \mathbb{Q}$. Moreover, U is called a* Siegel disk*.*

(iv) *There exists $\phi\colon U \to A$ conformal where $A = \{z : 1 < |z| < r\}$, $r > 1$, is an annulus such that $\phi(f(\phi^{-1}(z))) = e^{2\pi i\alpha}z$ for some $\alpha \in \mathbb{R} \setminus \mathbb{Q}$. Moreover, U is called a* Herman ring*.*

(v) *There exists $z_0 \in \partial U$ such that $f^n(z) \xrightarrow[n\to\infty]{} z_0$ for all $z \in U$ but $f(z_0)$ is not defined. Moreover, U is called a* Baker domain*.*

In order to study the Fatou components we introduce the notion of the singularities of the inverse, which are the points $a \in \mathbb{C}$ where some branch of f^{-1} is not well-defined (holomorphic and injective) in a neighborhood of $a \in \mathbb{C}$. Two different cases can arise: either there exists $z \in \mathbb{C}$ such that $f(z) = a$ and $f'(z) = 0$ ($z \in \mathbb{C}$ is then said to be a critical point and $a \in \mathbb{C}$ a critical value); or there exists a curve $\gamma\colon [0,\infty) \to \mathbb{C}$ such that $\gamma(t) \xrightarrow[t\to\infty]{} \infty$ and $f(\gamma(t)) \xrightarrow[t\to\infty]{} a$ ($a \in \mathbb{C}$ is then said to be an asymptotic value and the curve γ an asymptotic path).

Singular values (critical or asymptotic) play a fundamental role in the dynamical behavior of holomorphic (or meromorphic) functions. For example, any immediate attractive or parabolic basin of attraction needs to contain a singular value. In the remaining cases they are also relevant (see [1, 2, 4, 5, 11]), but in this paper we will focus in the study of an attractive basin of the family of maps

$$f_\lambda(z) = \lambda\left(\frac{e^z}{z+1} - 1\right),$$

where $\lambda \in \mathbb{C}\setminus\{0\}$ is a complex parameter.

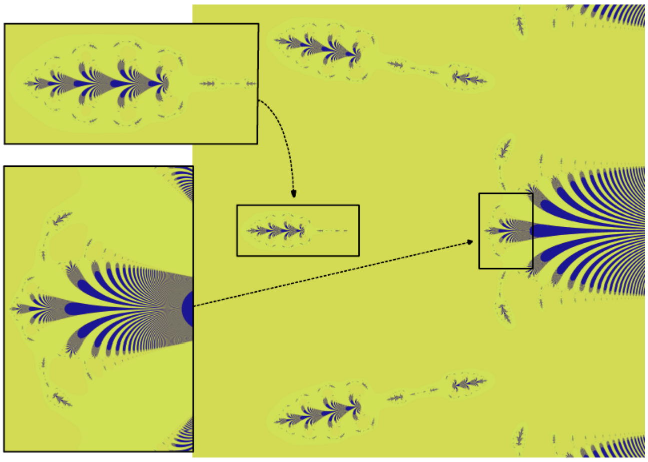


Figure 1: In green, $F(f_{0.89})$. Range $(-5,7)\times(-6,6)$.

Maps in this family are the simplest meromorphic maps with two singular values: $z = 0$ which is a fixed critical value (and fixed point), and $z = -\lambda$, which is an asymptotic value whose orbit depends on λ. It has also one single pole $z = -1$, which is not omitted except for $\lambda = 1$. Since $z = 0$ is a superattracting fixed point (i.e., a critical point which is a fixed point), its basin of attraction $\mathcal{A}_\lambda(0)$ is non-empty for all values of λ.

This family can be viewed as the meromorphic analogue to the well-known Milnor family of cubic polynomials $P_a(z) = z^2(z-a)$ [12] or its entire version $\lambda z^2 e^z$ [7, 8], both having also a superattracting fixed point and a free second singular value, which may or may not be captured by the attracting basin of 0. In both cases all components of the Fatou set are simply connected. In contrast, in this paper we prove that the basin of attraction of $z = 0$ for f_λ is infinitely connected for some parameter values.

Additionally, it is well-known that functions with only finitely many singular values do not have Wandering nor Baker domains [3, 4, 6, 10], hence $F(f_\lambda)$ does not have any of these components. Moreover, since any attractive basin or rotation domain needs a singular value, we can have at most two periodic cycles of Fatou components for every parameter value, one of which is always the basin of $z = 0$.

Hence it is to our interest to study the main capture component $\mathcal{C}_0 = \{\lambda \in \mathbb{C}^* : -\lambda \in \mathcal{A}^*_\lambda(0)\}$, where $\mathcal{A}^*_\lambda(0)$ denotes the immediate basin of attraction of $z = 0$. In this case there is only one Fatou component and we can draw an accurate picture of $F(f)$ by considering the points whose orbit is attracted to $z = 0$.

After addressing the study of the dynamical properties of f_λ for $\lambda \in \mathcal{C}_0$, we prove:

Theorem A. *If $-\lambda \in \mathcal{A}^*_\lambda(0)$, then $\mathcal{A}_\lambda(0)$ is connected and infinitely connected. Moreover, the set $\mathcal{C}_0 := \{\lambda \in \mathbb{C}^* : -\lambda \in \mathcal{A}^*_\lambda(0)\}$ contains a punctured disk of center 0 and radius $1/2$ (see Figure 2).*

The paper is structured as follows. In Section 2 we prove some estimates which are useful in Section 3, where we prove Theorem A.

2. Preliminaries

Our goal in this section is to prove some estimates that will be useful in the later section. We start by estimating the radius of the largest disk contained in $\mathcal{A}_\lambda(0)$, which we denote by

$$r_\lambda = \sup\{r > 0 : D(0, r) \subset \mathcal{A}_\lambda(0)\} < 1,$$

where the last inequality follows from the fact that $z = -1$ is a pole and hence belongs to the Julia set.

Proposition 2.1 (Maximum inner disk)**.** *For every $\lambda \in \mathbb{C}^*$,*

$$r_\lambda \geq \varepsilon(\lambda) := \frac{1}{2}\left(2 + |\lambda| - \sqrt{|\lambda|^2 + 4|\lambda|}\right) \in (0, 1).$$

Consequently $D(0, \varepsilon(\lambda)) \subset \mathcal{A}_\lambda(0)$.

Proof. For $0 < \varepsilon < 1$ and $|z| < \varepsilon$ we have

$$|f_\lambda(z)| = |f_\lambda(z) - f_\lambda(0)| \leq |\lambda| \left(\max_{|z|=\varepsilon} \left|\frac{ze^z}{(z+1)^2}\right|\right) |z|,$$

where we have used the Maximum Modulus Principle for f'_λ. For $z = \varepsilon e^{i\theta}$ we have

$$g_\lambda(\varepsilon, \theta) := |\lambda| \left|\frac{ze^z}{(z+1)^2}\right| = |\lambda| \frac{\varepsilon e^{\varepsilon \cos(\theta)}}{1 + \varepsilon^2 + 2\varepsilon \cos(\theta)}.$$

The goal is to obtain the maximum ε such that f_λ is a strict contraction in $D(0, \varepsilon)$, because then all points in $D(0, \varepsilon)$ converge to $z = 0$ under iteration, i.e., we want to obtain $\sup\{\varepsilon \in (0, 1) : g_\lambda(\varepsilon, \theta) < 1,\ \theta \in [0, 2\pi)\}$ since then it follows that $|f_\lambda(z)| < |z|$. We split it in two cases depending on θ.

- For $\theta \in [-\pi/2, \pi/2)$, we have $g_\lambda(\varepsilon, \theta) \leq |\lambda|\varepsilon e^\varepsilon/(1 + \varepsilon^2) =: g_{\lambda,1}(\varepsilon)$.
- For $\theta \in [\pi/2, 3\pi/2)$, we have $g_\lambda(\varepsilon, \theta) \leq |\lambda|\varepsilon/(1 - \varepsilon)^2 =: g_{\lambda,2}(\varepsilon)$.

Observe now that for $0 < \varepsilon < 1$, we always have $g_{\lambda,1}(\varepsilon) \leq g_{\lambda,2}(\varepsilon)$. Moreover, for $0 < \varepsilon < 1$,

$$|\lambda| \frac{\varepsilon}{(1 - \varepsilon)^2} < 1 \iff \varepsilon^2 - (2 + |\lambda|)\varepsilon + 1 > 0,$$

and this last polynomial has roots

$$\varepsilon(\lambda) = \frac{1}{2}\left(2 + |\lambda| - \sqrt{|\lambda|^2 + 4|\lambda|}\right) \quad \text{and} \quad \frac{2 + |\lambda| + \sqrt{|\lambda|(|\lambda| + 4)}}{2}.$$

Then $\varepsilon(\lambda) \in (0, 1)$ and the result follows. ∎

As a consequence of Proposition 2.1 we see now that for a disk of parameters of definite size, the free asymptotic value $z = -\lambda$ belongs to the immediate basin of $z = 0$. The set of parameters with this property is called the main capture component $\mathcal{C}_0$, which was defined in the introduction.

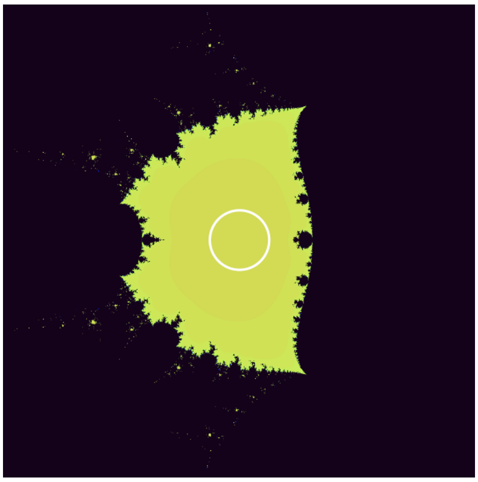


Figure 2: In green, $\mathcal{C}_0$. In white, $\partial D(0, 1/2)$. Range $(-4, 4) \times (-4, 4)$.

Corollary 2.2. $D^*(0, 1/2) = D(0, 1/2) \setminus \{0\} \subset \mathcal{C}_0$.

Proof. From the lower bound on r_λ given by Proposition 2.1, we obtain that $-\lambda \in D(0, \varepsilon(\lambda))$ if $\varepsilon(\lambda) - \lambda > 0$, or equivalently if

$$2 - |\lambda| > \sqrt{|\lambda|^2 + 4|\lambda|}.$$

It is easy to verify that this inequality holds for $|\lambda| < 1/2$. □

3. Connectivity of the basin of $z = 0$: Proof of Theorem A

We prove Theorem A in two parts. Assume in what follows that the asymptotic value $z = -\lambda$ belongs to the immediate basin of attraction of $z = 0$, or equivalently the parameter $\lambda \in \mathcal{C}_0$. We first show that the basin of attraction of $z = 0$ is connected, that is, $\mathcal{A}_\lambda(0) = \mathcal{A}^*_\lambda(0)$, and hence totally invariant (Theorem 3.2). Then we prove that under the same hypothesis, $\mathcal{A}^*_\lambda(0)$ is infinitely connected (Theorem 3.4).

Both results follow from two technical lemmas.

Lemma 3.1. *Let $\lambda \in \mathcal{C}_0$. Then, all asymptotic paths of $z = -\lambda$ intersect the same Fatou component of $F(f_\lambda)$.*

Proof. Given an asymptotic path Γ of $z = -\lambda$, i.e.,

$$\Gamma(t) \xrightarrow[t\to\infty]{} \infty \quad \text{and} \quad f_\lambda(\Gamma(t)) \xrightarrow[t\to\infty]{} -\lambda$$

then $\mathrm{Re}(\Gamma(t))$ must be bounded from above, i.e., exists $M_\Gamma < \infty$ such that

$$\sup_{t\geq 0} \mathrm{Re}(\Gamma(t)) \leq M_\Gamma.$$

Since $-\lambda \in \mathcal{A}_\lambda^*(0)$, there exists $\varepsilon > 0$ such that $D(-\lambda, \varepsilon) \subset \mathcal{A}_\lambda^*(0)$. Now, the preimages of $D(-\lambda, \varepsilon)$ must belong to $\mathcal{A}_\lambda(0)$. In particular, there exists $\nu < 0$ and a half-plane, $\Pi_\nu = \{z \in \mathbb{C} : \mathrm{Re}(z) < \nu\}$ such that $f_\lambda(\Pi_\nu) \subset D(-\lambda, \varepsilon)$, and hence Π_ν belongs to one component of $\mathcal{A}_\lambda(0)$. But now, all asymptotic paths must have unbounded negative real part and hence they must all intersect Π_ν. □

We now can prove that in this case the basin of $z = 0$ is connected.

Theorem 3.2. *If $\lambda \in \mathcal{C}_0$, then $\mathcal{A}_\lambda(0) = \mathcal{A}_\lambda^*(0)$ is connected. In particular, $\mathcal{A}_\lambda(0)$ is totally invariant and is the whole Fatou set.*

Proof. Suppose that $-\lambda \in \mathcal{A}_\lambda^*(0)$. From Proposition 2.1, we can consider the disk $U_0 = D(0, \varepsilon(\lambda)) \subset \mathcal{A}_\lambda^*(0)$.

Now we pull-back U_0 in order to obtain the whole immediate basin $\mathcal{A}_\lambda^*(0)$:

Consider, for $N > 0$, U_N as the connected component of $f_\lambda^{-1}(U_{N-1})$ that contains U_{N-1}. This recurrence defines a sequence of subsets $\{U_N\}_{N\geq 0}$ such that:

- $U_N \subset \mathcal{A}_\lambda^*(0)$ for all $N \geq 0$.
- $U_N \subset U_{N+1}$ for all $N \geq 0$.
- $\mathcal{A}_\lambda^*(0) = \bigcup_{N\geq 0} U_N$.

Since $-\lambda \in \mathcal{A}_\lambda^*(0)$, there exists $N > 0$ such that $-\lambda \in U_N$ (i.e., $f_\lambda^N(-\lambda) \in U_0$), and we can find a path $\gamma \subset U_N$ that joins $-\lambda$ and 0.

So U_{N+1} is unbounded, because $z = -\lambda$ is an asymptotic value (a Picard Value), hence the preimage of γ must contain a path that joins 0 and ∞ (which is contained in U_{N+1}). Using Lemma 3.1 we obtain that, in fact, when $-\lambda \in \mathcal{A}_\lambda^*(0)$ all asymptotic tracts intersect $\mathcal{A}_\lambda^*(0)$.

Now suppose that $\mathcal{A}_\lambda(0)$ is not connected, then we must have at least two connected components, $\mathcal{A}_\lambda^*(0)$ and U. Furthermore,

$$f_\lambda(U) = \mathcal{A}_\lambda^*(0) \setminus \{-\lambda\}.$$

So U must contain a tail of an asymptotic path, but by Lemma 3.1 and the previous observation, this tail must be contained in $\mathcal{A}_\lambda^*(0)$ and the claim follows. □

Our next goal is to prove that $\mathcal{A}_\lambda^*(0)$ is infinitely connected (Theorem 3.4). To that end we first construct a closed curve in $\mathcal{A}_\lambda^*(0)$ which surrounds the pole $z = -1$. The Böttcher coordinates are the key ingredient. Given a closed curve γ, we denote by $\mathrm{ind}(\gamma, p)$ the winding number of γ with respect to the point $p \in \mathbb{C}$.

Lemma 3.3. *Let $-\lambda \in \mathcal{A}_\lambda^*(0)$. Then there exists a closed, simple curve β, contained in $\mathcal{A}_\lambda(0)$, such that $0 \notin \beta$ and $\mathrm{ind}(f_\lambda(\beta), 0) = -1$.*

Proof. Let U be a neighborhood of $z = 0$ and $\varphi\colon U \to D(0,r)$ be the Böttcher map which locally conjugates f_λ to $Q_0(w) = w^2$.

Consider $\varepsilon < r < 1$ and define, $D = \varphi^{-1}(D(0,\varepsilon))$ and $D' = \varphi^{-1}(D(0,\varepsilon^2))$. The curves, $\tilde{r}_1(t) = i\sqrt{t}$, $\tilde{r}_2(t) = -i\sqrt{t}$, for $t \in [0,\varepsilon)$, are mapped by Q_0 to $\tilde{r}_0(t) = -t = Q_0(\tilde{r}_j(t))$, $j = 1,2$. Now set $r_j(t) = \varphi^{-1}(\tilde{r}_j(t))$, $j = 1,2$.

Since $-\lambda \in \mathcal{A}^*_\lambda(0)$, there exists a disk V centered at $z = -\lambda$ such that $\bar{V} \subset \mathcal{A}^*_\lambda(0)$ and, by Lemma 3.1, $f_\lambda^{-1}(\bar{V})$ contains a half-plane $\{z \in \mathbb{C} : \mathrm{Re}(z) < \nu\}$.

Furthermore, since by Theorem 3.2, $\mathcal{A}^*_\lambda(0) = \mathcal{A}_\lambda(0)$ is connected, we can find a simple curve $\alpha_0 \subset \mathcal{A}^*_\lambda(0)$ such that $\alpha_0(0) = 0$, $\alpha_0(1) = -\lambda$ and $(r_0)_{|[0,\varepsilon]} = (\alpha_0)_{|[0,\varepsilon]}$.

Define $s \in (\varepsilon, 1)$ such that $\alpha_0(s) \in \partial V$. Observe that the preimage of α_0 by f_λ are two simple curves, α_1, α_2 (because the preimage of $\tilde{r}_0$ by Q_0 consists of two disjoint curves), which are asymptotic paths, such that:

- $(r_j)_{|[0,\varepsilon]} = (\alpha_j)_{|[0,\varepsilon]}$ for $j = 1,2$.
- $(\alpha_1)_{|[\varepsilon,s)} \cap (\alpha_2)_{|[\varepsilon,s)} = \emptyset$, that is, because $0 \notin f_\lambda((\alpha_1)_{|[\varepsilon,s)}) = f_\lambda((\alpha_2)_{|[\varepsilon,s)}) = (\alpha_0)_{|[\varepsilon,s)}$ and hence, f_λ is conformal for every $z \in (\alpha_1)_{|[\varepsilon,s)} \cup (\alpha_2)_{|[\varepsilon,s)}$.

Now define.

- $\gamma_j = (\alpha_j)_{|[\varepsilon,s]}$ for $j = 0,1,2$.
- $\gamma_3 \subset f_\lambda^{-1}(\partial V)$ the simple curve that joins $\gamma_1(s)$ and $\gamma_2(s)$.
- $\tilde{\gamma}_4(t) = \varepsilon e^{-2\pi i t}$ and $\gamma_{4,1} = \varphi^{-1}((\tilde{\gamma}_4)_{|[1/4,3/4]})$, $\gamma_{4,2} = \varphi^{-1}((\tilde{\gamma}_4)_{|[-1/4,1/4]})$.

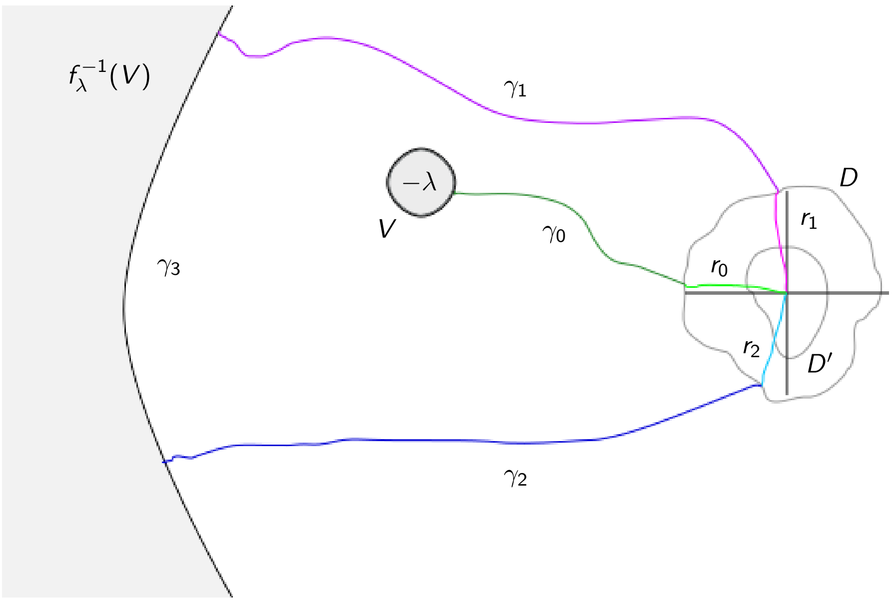


Figure 3: Representation of the curves and domains in the proof of Lemma 3.3.

Then we can define the curves,

$$\beta_1 = \gamma_1 \cup \gamma_2 \cup \gamma_3 \cup \gamma_{4,1} \quad \text{and} \quad \beta_2 = \gamma_1 \cup \gamma_2 \cup \gamma_3 \cup \gamma_{4,2},$$

which by construction are closed, simple curves contained in $\mathcal{A}_\lambda(0)$ that omit $z = 0$. Observe that β_1 (resp. β_2) can be parametrized as a simple closed curve preserving the orientation that $\gamma_{4,1}$ (resp. $\gamma_{4,2}$) inherits from $\tilde{\gamma}_4$.

Finally,

$$f_\lambda(\beta_j) = \partial V \cup \gamma_0 \cup \partial D', \quad j = 1, 2.$$

Hence, since $\partial V \cup \gamma_0$ does not contribute to $\operatorname{ind}(f_\lambda(\beta_j), 0)$, we have $\operatorname{ind}(f_\lambda(\beta_j), 0) = \operatorname{ind}(\partial D', 0)$, thus

$$\operatorname{ind}(f_\lambda(\beta_1), 0) = \operatorname{ind}(\partial D', 0) = -1 \quad \text{or} \quad \operatorname{ind}(f_\lambda(\beta_2), 0) = \operatorname{ind}(\partial D', 0) = -1$$

(we want the curve β_1 or β_2 to be oriented counterclockwise), so we can take $\beta = \beta_1$ or $\beta = \beta_2$ so that $\operatorname{ind}(f_\lambda(\beta), 0) = -1$. □

Finally, we prove the remaining part of the theorem.

Theorem 3.4. *If $\lambda \in \mathcal{C}_0 = \{\lambda \in \mathbb{C}^* : -\lambda \in \mathcal{A}^*_\lambda(0)\}$, then $\mathcal{A}_\lambda(0) = F(f_\lambda)$ is infinitely connected.*

Proof. By Theorem 3.2 we know that $\mathcal{A}_\lambda(0) = \mathcal{A}^*_\lambda(0) = F(f_\lambda)$ is connected. Let $Z(f_\lambda)$ denote the discrete set of zeros of f_λ and $P(f_\lambda)$ the set of poles of f_λ.

Consider the simple closed curve provided by Lemma 3.3. By the Argument Principle (see [9]),

$$\operatorname{ind}(f_\lambda(\beta), 0) = -1 = \sum_{a \in Z(f_\lambda)} m(a) \operatorname{ind}(\beta, a) - \sum_{a \in P(f_\lambda)} m(a) \operatorname{ind}(\beta, a),$$

where $m(a)$ denotes the order of the zero or the pole.

Since β is a simple closed curve oriented counterclockwise, the equation reads

$$-1 = \sum_{a \in Z(f_\lambda)} m(a) \operatorname{ind}(\beta, a) - \operatorname{ind}(\beta, -1),$$

which can only be satisfied if β surrounds no zeros of f_λ and the unique pole $z = -1$ is surrounded by β.

So, $\beta \subset \mathcal{A}_\lambda(0) = \mathcal{A}^*_\lambda(0)$ and $-1 \in \operatorname{int}(\beta)$. Then, the successive preimages of $\overline{\operatorname{int}(\beta)}$ contain points $w \in \mathcal{O}^-_{f_\lambda}(\infty) \subset J(f_\lambda)$ which lie in the interior of a closed curve contained in $\mathcal{A}_\lambda(0)$. Hence, since the backward orbit of ∞ is an infinite set (the points that are eventually mapped to ∞ under iteration by f_λ), $\mathcal{A}_\lambda(0)$ is infinitely connected. □

References

[1] A.F. Beardon, *Iteration of Rational Functions*, Complex analytic dynamical systems, Graduate Texts in Mathematics **132**, Springer-Verlag, New York, 1991.

[2] W. Bergweiler, "Iteration of meromorphic functions", *Bull. Amer. Math. Soc. (N.S.)* **29(2)** (1993), 151–188.

[3] W. Bergweiler, A. Eremenko, "On the singularities of the inverse to a meromorphic function of finite order", *Rev. Mat. Iberoamericana* **11(2)** (1995), 355–373.

[4] B. Branner, N. Fagella, *Quasiconformal Surgery in Holomorphic Dynamics*, With contributions by Xavier Buff, Shaun Bullett, Adam L. Epstein, Peter Haïssinsky, Christian Henriksen, Carsten L. Petersen, Kevin M. Pilgrim, Tan Lei and Michael Yampolsky, Cambridge Studies in Advanced Mathematics **141**, Cambridge University Press, Cambridge, 2014.

[5] L. Carleson, T.W. Gamelin, *Complex Dynamics*, Universitext: Tracts in Mathematics, Springer-Verlag, New York, 1993.

[6] A.È. Erëmenko, M.Yu. Lyubich, "Dynamical properties of some classes of entire functions", *Ann. Inst. Fourier (Grenoble)* **42(4)** (1992), 989–1020.

[7] N. Fagella, A. Garijo, "Capture zones of the family of functions $\lambda z^m \exp(z)$", *Internat. J. Bifur. Chaos Appl. Sci. Engrg.* **13(9)** (2003), 2623–2640.

[8] N. Fagella, A. Garijo, "The parameter planes of $\lambda z^m \exp(z)$ for $m \geq 2$", *Comm. Math. Phys.* **273(3)** (2007), 755–783.

[9] T.W. Gamelin, *Complex Analysis*, Undergraduate Texts in Mathematics, Springer-Verlag, New York, 2001.

[10] L.R. Goldberg, L. Keen, "A finiteness theorem for a dynamical class of entire functions", *Ergodic Theory Dynam. Systems* **6(2)** (1986), 183–192.

[11] J. Milnor, *Dynamics in One Complex Variable*, Third edition, Annals of Mathematics Studies **160**, Princeton University Press, Princeton, NJ, 2006.

[12] J. Milnor, "Cubic Polynomial Maps with Periodic Critical Orbit, Part I", Stony Brook Institute for Mathematical Sciences (2008). `http://www.math.stonybrook.edu/~jack/PREPRINTS/cpm.pdf`.

REPORTS@SCM
AN ELECTRONIC JOURNAL OF THE
SOCIETAT CATALANA DE MATEMÀTIQUES

A Finite Element Method solution to the Phase-Field model of the Stefan problem

*Max Orteu
Freie Universität Berlin
maxorteu@zedat.fu-berlin.de

Esther Sala-Lardies
Universitat Politècnica de Catalunya
esther.sala-lardies@upc.edu

Sonia Fernàndez-Méndez
Universitat Politècnica de Catalunya
sonia.fernandez@upc.edu

∗Corresponding author

Resum *(CAT)*
Aquest treball es centra en el model de camp de fase per al problema d'Stefan, així com en la solució numèrica amb el mètode dels elements finits per a simular exemples en dues dimensions.

Abstract *(ENG)*
This work is concerned with the Phase-Field model of the Stefan problem, as well as the numerical solution with the Finite Element Method to simulate examples in two dimensions.

Societat Catalana de Matemàtiques

Institut d'Estudis Catalans

Keywords: *heat diffusion, numerical methods, PDEs, modeling, FEM, Phase-Field models, Stefan, solidification.*
MSC (2010): *65N30, 80A22, 80M10.*
Received: *July 28, 2022.*
Accepted: *September 29, 2022.*

Acknowledgement
This work was supported by Generalitat de Catalunya (2017-SGR-1278).

1. Introduction

The Stefan problem, first introduced by Josef Stefan in [12] in the context of polar ice caps, is a boundary value problem used to model the state of systems where phase transitions occur. It is a free boundary problem; that is, a partial differential equation (PDE) to be solved for an unknown domain, or in a domain with unknown interface. In this work, biphasic systems modeling a Liquid-Solid coexistence, and featuring a pure substance are considered. A solution to the problem should be able to describe the position of the moving interface $\gamma(t)$ between the two different states of the given material, as well as the temperature at each point in the space-time domain, $u(x, t)$; all this, depending on a given a set of initial and boundary conditions. A classical example could be the melting of an ice cube in a glass of water, where the equations should provide the distribution of ice for any given time t, and the temperature at each point in space and time.

The model we have chosen to approximate solutions to the Stefan problem is the so-called Phase-Field model (PF). Phase-Field models are widely used in transformation problems, since they facilitate the tracking of the interface. The main idea is to consider what is called the Phase-Field function or order parameter (in this work noted as ϕ), a scalar function that equals a fixed constant in each phase and varies rapidly, but smoothly (in our case from -1 to $+1$), in the interface region. The phase transformation occurs inside a finite-width region, whose thickness is a parameter in the model, here noted as ξ, and where in general the transport properties are assumed to vary with u. It is worth mentioning that the solution tends to a Sharp-Interface model as $\xi \to 0$. An important resource for the study of this modeling technique has been [11].

When it comes to the numerical method used to obtain the solutions, the Finite Element Method (FEM) has been chosen. It is a powerful and popular method, used in areas such as solid mechanics [9], electromagnetic potentials [10], or heat transfer – as it is here the case. One of its most salient advantages is that it allows modeling complex and irregular geometries, given that it is capable of working under non-uniform meshes. This is important because it makes it possible to capture local effects by calculating the solution on a partially-refined domain, a process which helps produce accurate yet not so computationally-demanding solutions.

This work serves as a stepping stone to implementing an adaptive solution to the Phase-Field model of the Stefan problem. Such a procedure, inspired by the work of Alba Muixí in the context of crack propagation in [9], would constitute an efficient solution that would drastically decrease the running time involved and capture much better the local phenomena. This would in turn make it possible to calculate approximations with finer meshes (and subsequently smaller time steps), allowing for $\xi \to 0$, conditions under which, as mentioned, the Phase-Field model converges to the Sharp-Interface model.

2. The Phase-Field model for the Stefan problem

The classical Stefan problem arises when we consider two phases of a material undergoing a phase transformation. A heat equation must be solved for each phase, but with the added difficulty of having a boundary on a moving interface, where the temperature is fixed. That is why we say that we have a free boundary problem. A further equation, the Stefan condition (1c), is needed in order to have a mathematically closed system. With this, the general form of the equations of the Stefan problem for a domain Ω are

$$\frac{\partial u}{\partial t} = \nabla \cdot (c_l \nabla u), \quad x \in \Omega_l(t), \tag{1a}$$

$$\frac{\partial u}{\partial t} = \nabla \cdot (c_s \nabla u), \quad x \in \Omega_s(t), \tag{1b}$$

$$lv = c_s \frac{\partial u}{\partial n}\Big|_{x \downarrow \gamma(t)} - c_l \frac{\partial u}{\partial n}\Big|_{x \uparrow \gamma(t)}. \tag{1c}$$

Here u is the temperature and the domain is divided into Ω_l and Ω_s, where the material is in liquid and solid phase, respectively. For each of these subdomains, c_l and c_s are the diffusivity constants of the material. Lastly, l is the latent heat per unit volume, while $\gamma(t)$ is the position of the interface and v its normal velocity.

Equations (1) are referred to here as Sharp-Interface model. The main difficulty of the FEM for solving them comes from the fact that the computational mesh must be fitted to the interface $\gamma(t)$ to represent the weak discontinuity of the temperature. This requires a continuous mesh adaptation that is cumbersome and computationally expensive.

To circumvent this, Phase-Field models consider a smooth variation of the temperature across the interface. To do so, a Phase-Field variable ϕ is introduced, taking the value $+1$ in the liquid domain, -1 in the solid, and varying smoothly between these values in a thin transition region, with the value 0 at the interface.

2.1 Phase-Field models

The derivation of mathematical Phase-Field models from physical principles is built upon Landau–Ginzburg theory of phase transitions, which can be found in [7]. The Phase-Field model presented here is based on the work by Cahn and Hilliard in [3] and features the variables u and ϕ, which account for the temperature and the phase state at each point in the space-time domain. There are several Phase-Field models for the Stefan problem in the literature, most notably the ones based on the Kobayashi and the Caginalp potentials. For the Kobayashi potential, originally introduced in [6], we have followed Fabbri and Voller, but with a change of sign to fix the typo in equation (2.7) in [5]. An important reference for this section has been [8], where a thorough explanation of the Phase-Field model for the Stefan problem can be found.

Given a domain Ω and its boundary $\Gamma = \partial\Omega$, the Phase-Field model for the Stefan problem is a coupled system of two PDEs. If we consider the system in a time interval $[0, T]$, these are:

$$\begin{cases} \dfrac{\partial u}{\partial t} = c\Delta u - \dfrac{l}{2}\dfrac{\partial \phi}{\partial t} & \text{in } \Omega \times [0, T], \quad (2a) \\ \alpha\xi^2 \dfrac{\partial \phi}{\partial t} = \xi^2 \Delta\phi - \dfrac{\partial F}{\partial \phi} & \text{in } \Omega \times [0, T], \quad (2b) \end{cases}$$

with initial conditions $u(x, 0) = u_0(x)$, $\phi(x, 0) = \phi_0(x)$ and boundary conditions.

Equation (2a) is a heat equation, modified by the Phase-Field term. Here, $u(x, t)$ represents $T(x, t) - T_M$, where $T(x, t)$ is the temperature of the material and T_M the temperature at which the change of phase takes place – in Liquid-Solid equilibrium, the fusion/solidifying temperature. In practice, we can think of solidification/fusion in an ice-water system, where $T_M = 0$, and so $T(x,t) = u(x,t)$. An important part of the model is the recovering potential F appearing in (2b), which is a double-well potential that has its minima at the values of ϕ corresponding to the two different phases. The two recovering potentials previously mentioned are

$$\text{Caginalp:}\;\; F(u,\phi) = \frac{1}{8a}(\phi^2-1)^2 - \frac{\Delta s}{2}\phi u, \tag{3}$$

$$\text{Kobayashi:}\;\; F(u,\phi) = \frac{W}{16}\int_0^{\phi}(r^2-1)(r+2b(u))\,dr, \tag{4}$$

where a is a parameter exclusive of the Caginalp potential that should be chosen so that $\frac{\partial F}{\partial \phi}$ has minima near 0, -1 and $+1$, W is a constant with units of energy per unit volume, and Δs is the entropy scale, defined as the difference between the entropies of the liquid and solid phases at the melting temperature.

From a theoretical standpoint, both expressions (3) and (4) result in a double-well potential, with minima at ± 1 when taking sufficiently small values of u, and considering $F(u,\phi)$ as a function of ϕ. However, while for the Kobayashi potential these minima are fixed, in the Caginalp potential they shift from ± 1 as u moves away from 0 (see Figure 1). This effect fades away as $\xi \to 0$ and $a \to 0$, nonetheless, this potential in practice produces values of $|\phi| > 1$, which hinders the interpretation of the results. Consequently, we employ the Kobayashi potential for practical examples.

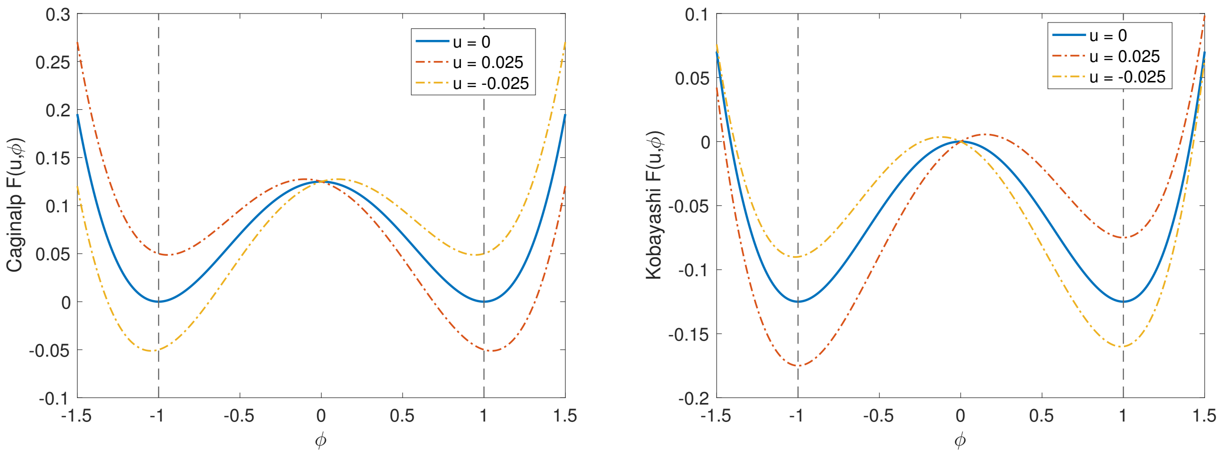


Figure 1: Caginalp (left) and Kobayashi (right) potentials as functions of ϕ, for three different values of u $(-0.025, 0, 0.025)$. In the first case, the minima depart from ± 1 as u drifts away from 0, while in the second they remain in the same place. The parameters of the models are taken as mentioned in Subsection 2.2, $a = 1$, $W = 8$, $\Delta s = 4$.

The reader should note that c has been taken in (2a) as constant and equal in both phases, as will be the case in the whole of this work. Regarding the new parameters in the Phase-Field model, in (2b) α is a timescale parameter and ξ is the Phase-Field energy parameter or gradient-energy coefficient, and it controls the interface width. Sometimes the notation $\tau = \alpha\xi^2$ is used, with τ being the time relaxation parameter. A detailed outline of parameters can be found in the work by Caginalp in [2].

2.2 Interpretation of the model and assumptions

The dynamical interpretation for equations (2a) and (2b) is as follows. Equation (2a) is a Fourier heat equation in which a source term has been added to factor in the latent heat release at the moving interface. On the other hand, the phase equation (2b) is a linear time evolution that features the imbalance between the excess interface free-energy and the restoring potential $F(u, \phi)$.

Throughout the examples, we have considered a constant diffusivity coefficient $c = 1$ and a latent heat $l = 1$. Although these coefficients usually vary between phases and with temperature, respectively, this approximation is valid to study the behavior of the equations. Indeed, this same assumption is found in an important part of the literature (see, for instance, [2] and [8]). Also, in all cases we have considered a pure substance, so that no gradients of concentration appear in the equations and only one set of material parameters is needed.

Regarding the parameters of the Phase-Field model with the Kobayashi potential, we have used $\alpha = 1$, $W = 8$, $\Delta s = 4$, and following Wheeler et al. in [14], $b(u)$ (which should be a monotonic increasing function of$u$ satisfying $|b(u)| < 1/2$) is taken as $b(u) = 6u\Delta s/W$.

Taking into account all of the above, the system of equations (2) with the Kobayashi potential can be re-written as:

$$\begin{cases} \dfrac{\partial u}{\partial t} = c\Delta u - \dfrac{l}{2}\dfrac{\partial \phi}{\partial t} & \text{in } \Omega \times [0, T], \\ \alpha\xi^2 \dfrac{\partial \phi}{\partial t} = \xi^2 \Delta\phi + G(u, \phi) & \text{in } \Omega \times [0, T], \end{cases} \tag{5}$$

with $G(u, \phi) = \frac{1}{16}W\phi + \frac{3}{4}u\Delta s - \frac{1}{16}W\phi^3 - \frac{3}{4}u\phi^2\Delta s$.

One final comment to be made is that, in most cases, we have considered homogeneous Neumann boundary conditions for both u and ϕ, simulating an isolated system. An important caveat is that, when doing so, one has to be careful that the interface (i.e. $\{x \in \Omega : \phi(x, t) = 0\}$) remains far away enough from the boundary with Neumann conditions, since the interface is by definition a region with strong gradients in the Phase-Field variable ϕ, and both conditions cannot coexist.

3. Examples

3.1 2D analog of a known 1D Stefan problem

The goal of this first example is to validate the Phase-Field model and its numerical solution. To this end, we consider the work by Surana et al. in [13], which features a 1D Stefan problem (with sharp interface) with a known analytical solution.

$$u_a(x, t) = \begin{cases} C_1 \dfrac{\operatorname{erf}\left(\frac{\beta}{2}\right) - \operatorname{erf}\left(\frac{x}{2\sqrt{t+t_0}}\right)}{\operatorname{erf}\left(\frac{\beta}{2}\right)}; & x \leq \gamma(0), \\ C_2 \dfrac{\operatorname{erf}\left(\frac{\beta}{2}\right) - \operatorname{erf}\left(\frac{x}{2\sqrt{t+t_0}}\right)}{\operatorname{erfc}\left(\frac{\beta}{2}\right)}; & x > \gamma(0), \end{cases}$$

$$\gamma(t) = \beta\sqrt{t + t_0}.$$

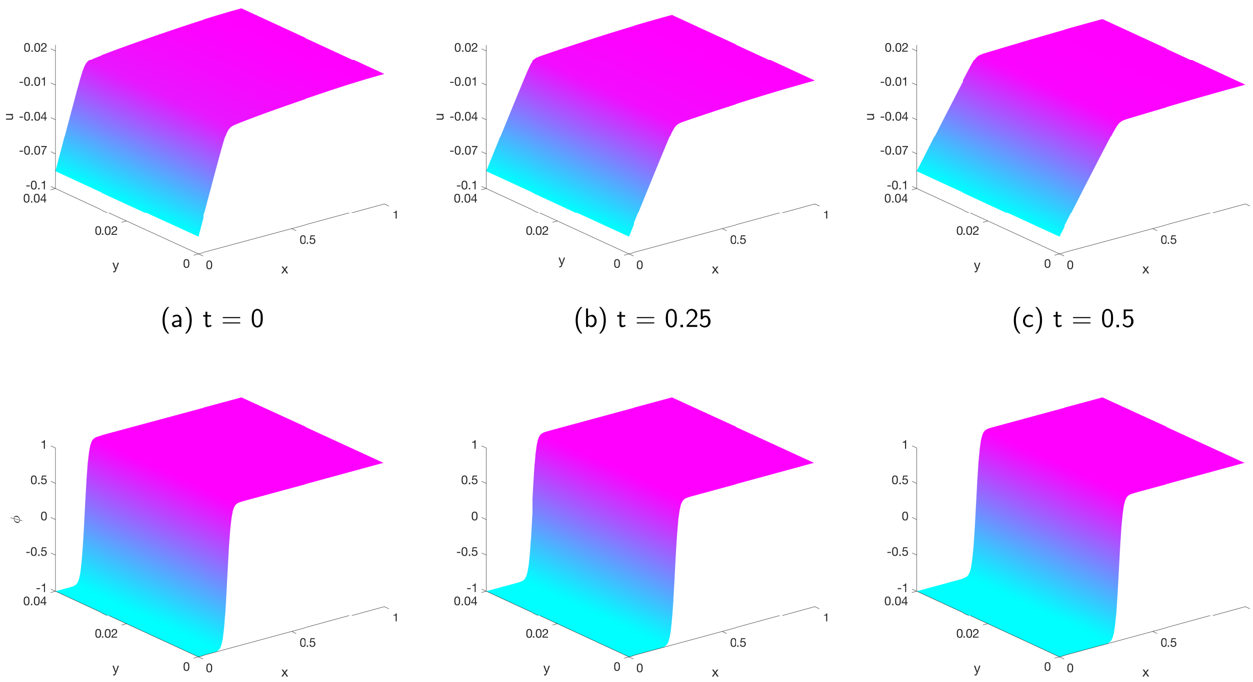
The problem is stated in $\Omega = [0,1]$ with constant Dirichlet boundary conditions at $x = 0$: $u(0,t) = u_a(0,t)$. The initial condition for the PF function $\phi(x,0)$ is calculated here as the stationary solution to equation (5) neglecting the laplacian term, i.e., $G(u(x,0),\phi(x,0)) = 0$. At $x = 1$, homogeneous Neumann boundary conditions are imposed for u.

In our case, we consider a rectangular domain $\Omega = [0,1] \times [0,0.04]$, and use a 2D analog of the boundary and initial conditions employed by Surana et al., treating them as functions of x and independent of the y-variable.

Indeed, for the boundary conditions, we use $u(0,y,t) = u_a(0,t)$, and calculate $\phi(0,y,t)$ such that $G(u,\phi) = 0$. For the rest of the boundary, homogeneous Neumann boundary conditions are employed for both u and ϕ.

As in [13], the parameters are chosen as $C_1 = -0.085$, $C_2 = -0.015$, $t_0 = 0.1246$, $\beta = 0.396618$. This translates into $\gamma(0) = 0.14$ and we can see that at this point the initial condition has a discontinuity in the derivative of the temperature. It can also be observed in Figures 2 and 3 that the 0-level set of ϕ is a good approximation to the position of the interface.

With respect to the numerical parameters of the models, $m = 150$ space intervals have been used, which makes an element size $h = 0.0067$. The limit $\xi \geq 1.25h$ as proposed by Fabbri and Voller in [5] when using fixed-grid methods is employed and, consequently, $\xi = 0.0083$ has been chosen, to have the minimum possible ξ without instability in the solution. In this case, we choose a final time of $T = 0.5$.



(a) t = 0 (b) t = 0.25 (c) t = 0.5

(d) t = 0 (e) t = 0.25 (f) t = 0.5

Figure 2: Temperature $u(x,y,t)$ (above) and PF $\phi(x,y,t)$ (below) obtained with the Kobayashi computational model of the 2D analog of the problem considered in Subsection 3.1.

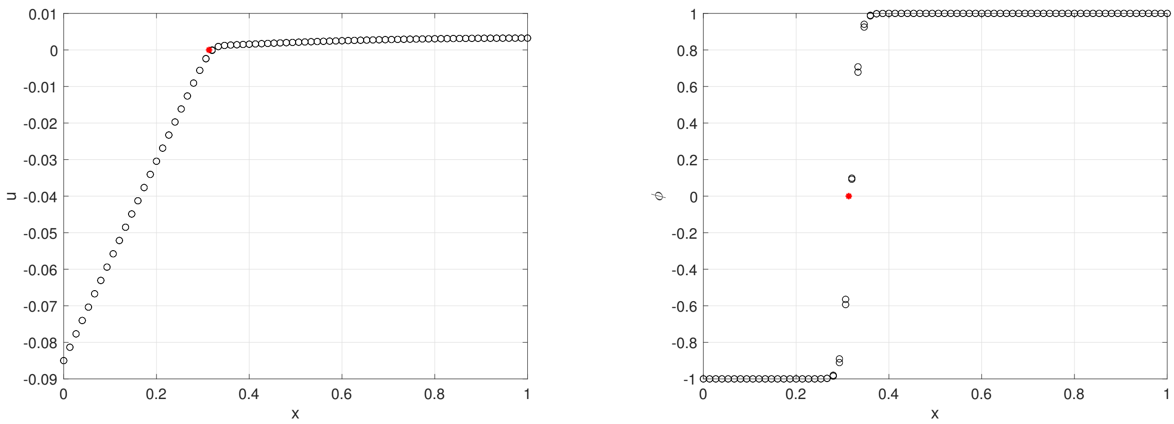


Figure 3: Temperature u (left) and PF ϕ (right) as a function of x for $t = 0.5$. In each case, a red star shows the analytical position of the interface, whose behavior is successfully replicated by the calculated solution.

3.2 Freezing with circular symmetry

In this section, we use the 2D FEM code to calculate a solution to the Stefan problem in a square domain with a circular hole, i.e., $\Omega = (0, 1)^2 \backslash B((0.5, 0.5), 0.05)$. The interest of this example, apart from its physical meaning, is that it motivates the need for an unstructured mesh. In addition to this, for the first time, a partial refinement of the domain is used (see Figure 4). This is, only the elements in which a phase transformation can occur are refined.

We first start with Dirichlet boundary conditions on the boundary of the circle, $u = -0.015$, while we use homogeneous Neumann conditions on the outer rectangle, simulating an isolated system. Regarding the initial condition, we use

$$u(x, y, 0) = \begin{cases} -0.015, & \text{if } (x - 0.5)^2 + (y - 0.5)^2 \leq 0.075^2, \\ 0.015, & \text{otherwise,} \end{cases}$$

and again calculate ϕ from the value of u. Figure 5 shows the results obtained in the simulation using the model parameters $\alpha = 1$ and $\xi = 0.0075$, element size $h = 0.006$, and final time $T = 0.025$.

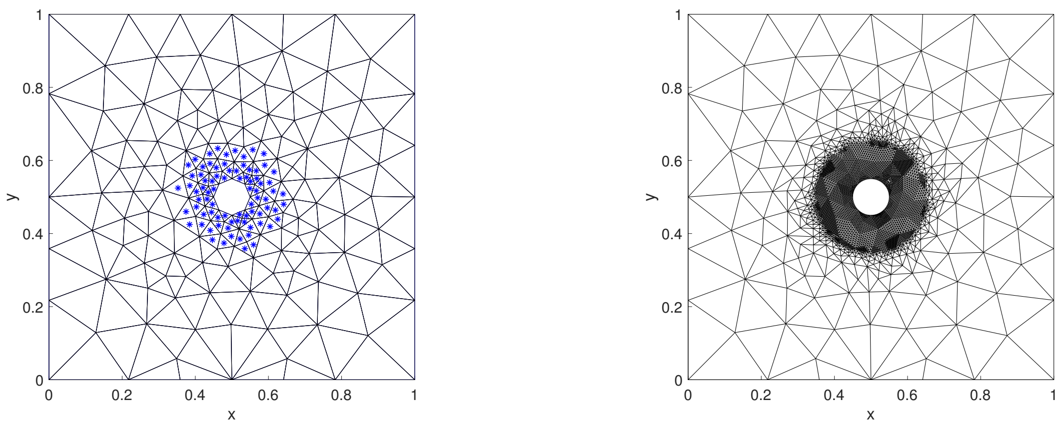
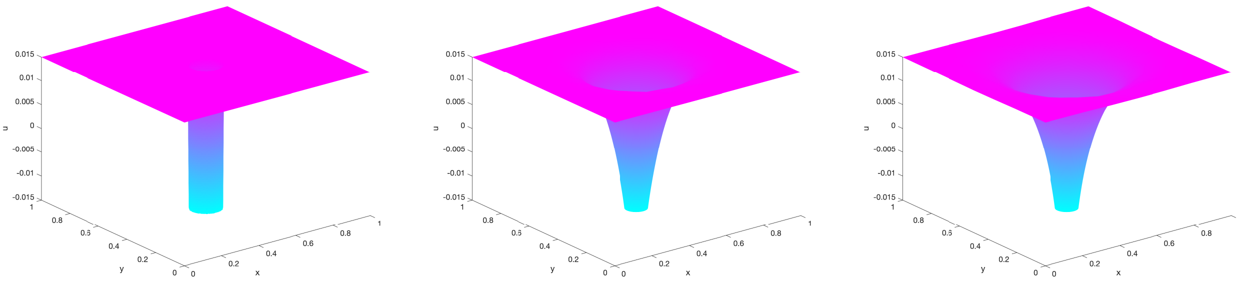
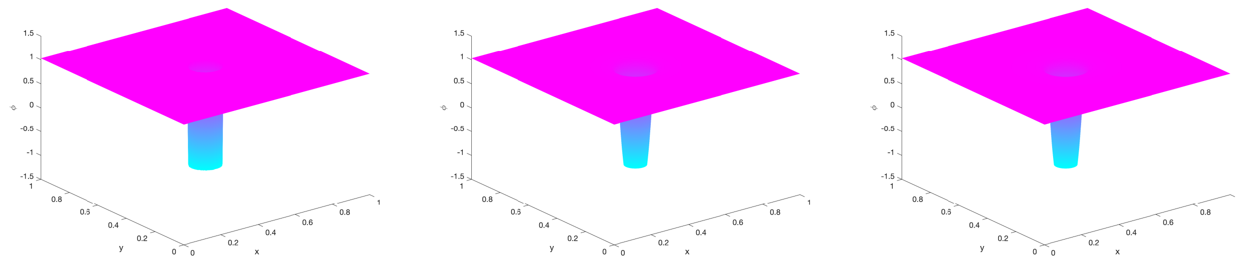


Figure 4: The initial mesh (left) is refined repeatedly until all elements close to the initial interface are small enough to capture a subsequent phase transition. The right image shows the final mesh used in the calculations.

If we take a close look as in the left of Figure 6, where the values of u are plotted as a function of the radius to the center of the domain, we can see that there is now no spike in the temperature. This is thanks to the smoothing effect of the laplacian term. Finally, on the right of Figure 6, there is a contour plot of the points of the domain for which $u = 0$ at four different moments in time. As expected from a physical point of view, this shows a circle that expands with time but maintains its shape, a direct consequence of the circular symmetry of the example. We can conclude that the Phase-Field model excels at tracking the evolution of an existing interface.



(a) t = 0 (b) t = 0.0125 (c) t = 0.025



(d) t = 0 (e) t = 0.0125 (f) t = 0.025

Figure 5: FEM solution for $u(x,t)$ (above) and $\phi(x,t)$ (below) for a domain with one hole and initial interface. The following numerical parameters are used: $h = 0.006$, $\xi = 0.0075$, $T = 0.025$, $\Delta t = 6.21e-08$.

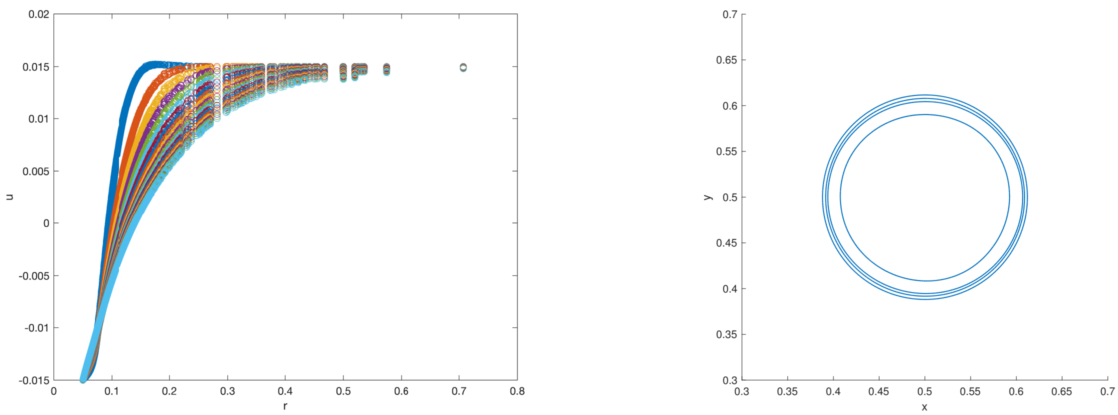


Figure 6: Behavior of the solution when an initial interface is provided. On the left, temperature is represented as a function of the radius, for twenty evenly spaced instants. On the right, the projected view of the level curves for which $u = 0$ is shown at times $t_0 = 0$, $t_1 = 0.008$, $t_2 = 0.017$ and $t_3 = 0.025$.

4. Numerical solution

4.1 Weak form and discretization

We start from the system of equations (2) of the Phase-Field model for the Stefan problem in a general domain Ω, with given Dirichlet boundary conditions $u(\Gamma_u) = u_D$ and $\phi(\Gamma_\phi) = \phi_D$, and homogeneous Neumann boundary conditions $\partial u/\partial n(\partial\Omega \backslash \Gamma_u) = 0$ and $\partial\phi/\partial n(\partial\Omega\backslash\Gamma_\phi) = 0$.

To solve this system of PDEs using the Finite Element Method, we must first write the weak form of the problem. First, we multiply both sides by test functions v, w such that $v(x) = 0$ on Γ_u and $w(x) = 0$ on Γ_ϕ, we integrate by parts the terms that contain a laplacian operator and impose the homogeneous Neumann boundary conditions to obtain:

$$\begin{cases} \displaystyle\int_\Omega v\frac{\partial u}{\partial t} = -c\int_\Omega \nabla v\cdot\nabla u - \frac{l}{2}\int_\Omega v\frac{\partial\phi}{\partial t}, \\ \displaystyle\alpha\xi^2\int_\Omega w\frac{\partial\phi}{\partial t} = -\xi^2\int_\Omega\nabla w\cdot\nabla\phi + \int_\Omega wG(u,\phi). \end{cases} \tag{6}$$

Now, the problem consists on finding $u(x,t), \phi(x,t) \in H^1(\Omega)$ such that $u(x,t) = u_D$ on Γ_u, $\phi(x,t) = \phi_D$ on Γ_ϕ and so that (6) is satisfied for all $v, w \in H^1(\Omega)$ such that $v = 0$ on Γ_u and $w = 0$ on Γ_ϕ.

We discretize (6) with a piece-wise polynomial base and approximate $u(x,t)$ and $\phi(x,t)$ as follows:

$$u(x,t) \simeq u^h(x,t) = \sum_{i=1}^n u_i(t)N_i(x), \quad \phi(x,t)\simeq\phi^h(x,t) = \sum_{i=1}^n \phi_i(t)N_i(x).$$

We denote by $\mathcal{B}_u$, $\mathcal{B}_\phi$ the set of indexes of all nodes on the Dirichlet boundaries Γ_u and Γ_ϕ respectively, for which we already know the values u_i and ϕ_i. Substituting into the weak form $v = N_i$ (for $i \notin \mathcal{B}_u$) and $w = N_i$ (for $i \notin \mathcal{B}_\phi$), and $u(x,t) \simeq u^h(x,t)$, $\phi(x,t)\simeq\phi^h(x,t)$ we obtain the following system of equations

$$\begin{cases} \displaystyle\sum_{j=1}^n\int_\Omega N_iN_j\frac{\partial u_j}{\partial t} = -c\sum_{j=1}^n\int_\Omega\nabla N_i\cdot\nabla N_j u_j - \frac{l}{2}\sum_{j=1}^n\int_\Omega N_iN_j\frac{\partial\phi_j}{\partial t}, & i\notin\mathcal{B}_u, \\ \displaystyle\alpha\xi^2\sum_{j=1}^n\int_\Omega N_iN_j\frac{\partial\phi_j}{\partial t} = -\xi^2\sum_{j=1}^n\int_\Omega\nabla N_i\cdot\nabla N_j\phi_j + \sum_{j=1}^n\int_\Omega N_iN_jG(u_j,\phi_j), & i\notin\mathcal{B}_\phi, \end{cases}$$

where instead of computing the integral $\int_\Omega N_jG(u,\phi)$, which would be computationally costly, the approximation $G(u,\phi)\approx\sum_j G(u_j,\phi_j)N_j$ is used. This introduces the same error as a spline interpolation, which approximates the calculation with an error h^{p+1}, with p being the order of the polynomial employed. Given that in all cases, we use linear basis functions, which translates into an h^2 error. This coincides with the error due to the FEM method, and so the convergence remains unaffected.

If we use

$$M_{ij} = \int_\Omega N_iN_j\,d\Omega, \quad K_{ij} = \int_\Omega\nabla N_i\cdot\nabla N_j\,d\Omega, \quad i\notin\mathcal{B}, \quad j = 1,\dots,n.$$

We can rewrite the whole system as

$$\begin{cases} M\dot{u} = -cKu - \frac{l}{2}M\dot{\phi}, \\ \alpha\xi^2 M\dot{\phi} = -\xi^2 K\phi + MG(u,\phi), \end{cases}$$

where $G(u,\phi)$ denotes the vector with components $G(u_i,\phi_i)$.

At the time of implementation, we divide the vectors u, ϕ and matrix M in their components on the Dirichlet boundary and in the rest of the domain. Noting them by the subscripts $_F$ and$_I$, respectively, we can write:

$$M = \left[\begin{array}{c|c} M_{II} & M_{IF} \end{array}\right], \quad u = \left[\begin{array}{c} u_I \\ \hline u_F \end{array}\right], \quad \phi = \left[\begin{array}{c} \phi_I \\ \hline \phi_F \end{array}\right].$$

With this decomposition, the system can be written as

$$\begin{cases} M_{II}\dot{u}_I = -M_{IF}\dot{u}_F - cKu - \dfrac{l}{2}M\dot{\phi}, \\ \alpha\xi^2 M_{II}\dot{\phi}_I = -\alpha\xi^2 M_{IF}\dot{\phi}_F - \xi^2 K\phi + MG(u,\phi). \end{cases} \tag{7}$$

We solve (7) at each time step using Euler's method (an explicit, 1st order method):

$$\begin{cases} u_I^{n+1} = u_I^n + \Delta t M_{II}^{l,-1}\left(-M_{IF}\dot{u}_F - cKu - \dfrac{l}{2}M\dot{\phi}\right), & \text{(8a)} \\ \phi_I^{n+1} = \phi_I^n + \dfrac{\Delta t}{\alpha\xi^2}M_{II}^{l,-1}(-\alpha\xi^2 M_{IF}\dot{\phi}_F - \xi^2 K\phi + MG(u,\phi)). & \text{(8b)} \end{cases}$$

This is the final system that can be found in the code. For each time step, for efficiency purposes, $\dot{\phi}$ is first computed and used in (8a) and (8b). The reader should note the l superindex in the matrix M, which indicates that we use the lumped mass matrix – a diagonal matrix obtained by summing over rows. Throughout this work, this correction has proven to not affect the convergence of the results, whereas it presents an important computational advantage when it comes to solving a system with the matrix M.

5. Conclusions and final remarks

The present work introduces the nature and interpretation of Phase-Field models in the context of the Stefan problem, highlighting some of its advantages and limitations. One remarkable property, the approximation to the solution of the Stefan (sharp-interface) model for small ξ, is illustrated by a 2D analog of a 1D example with analytical solution. In addition, a further example with physical meaning is included.

Regarding the implementation, the code has been tested to be not only precise but also efficient, with a special mention to the use of the lumped mass matrix. Besides this, the limit $\xi \geq 1.25h$ has been seen to work in all cases.

An interesting extension of the work would be to surpass the assumptions made, by considering the generalized Stefan problem, as described in [4]. For example, different diffusivity constants c could be used for each phase (possibly with a linear approximation in the interface), or the latent heat could be modelled as a function of temperature $l(u)$. From a numerical perspective, the implementation of an adaptive mesh

– as done in [1] – would take full advantage of the FEM, while implicit time integration methods would allow larger time steps, but with the burden of having to solve a non-linear system of equations. The increased stability would allow the simulation of a wider class of phenomena, such as the coalescence of interfaces.

References

[1] G. Beckett, J.A. Mackenzie, M.L. Robertson, "A moving mesh finite element method for the solution of two-dimensional Stefan problems", *J. Comput. Phys.* **168(2)** (2001), 500–518.

[2] G. Caginalp, "An analysis of a phase field model of a free boundary", *Arch. Rational Mech. Anal.* **92(3)** (1986), 205–245.

[3] J.W. Cahn, J.E. Hilliard, "Free energy of a nonuniform system. I. Interfacial free energy", *J. Chem. Phys.* **28(2)** (1958), 258–267.

[4] J. Crank, *Free and Moving Boundary Problems*, Oxford Science Publications, The Clarendon Press, Oxford University Press, New York, 1984.

[5] M. Fabbri, V.R. Voller, "The phase-field method in the sharp-interface limit: A comparison between model potentials", *J. Comput. Phys.* **130(2)** (1997), 256–265.

[6] R. Kobayashi, "Modeling and numerical simulations of dendritic crystal growth", *Physica D: Nonlinear Phenomena* **63(3-4)** (1993), 410–423.

[7] L.D. Landau, E.M. Lifshitz, *Statistical Physics. Course of Theoretical Physics. Volume 5*, Third Edition, Part 1, Pergamon Press, London, 1980.

[8] J.T. Lin, "The numerical analysis of a phase field model in moving boundary problems", *SIAM J. Numer. Anal.* **25(6)** (1988), 1015–1031.

[9] A. Muixí, S. Fernández-Méndez, A. Rodríguez-Ferran, "Adaptive refinement for phase-field models of brittle fracture based on Nitsche's method", *Comput. Mech.* **66(1)** (2020), 69–85.

[10] A.C. Polycarpou, *Introduction to the Finite Element Method in Electromagnetics*, Synthesis Lectures on Computational Electromagnetics, Morgan & Claypool, 2006.

[11] N. Provatas, K. Elder, *Phase-Field Methods in Material Science and Engineering*, Wiley-VCH, 2010.

[12] J. Stefan, "Ueber die Theorie der Eisbildung, insbesondere über die Eisbildung im Polarmeere", *Ann. Phys.* **278(2)** (1891), 269–286.

[13] K.S. Surana, A.D. Joy, L.A. Quiros, J.N. Reddy, "Mathematical models and numerical solutions of liquid-solid and solid-liquid phase change", *Journal of Thermal Engineering* **1(2)** (2015), 61–98.

[14] A.A. Wheeler, W.J. Boettinger, G.B. McFadden, "Phase-field model for isothermal phase transitions in binary alloys", *Phys. Rev. A* **45(10)** (1992), 7424–7439.

Reports@SCM
Volume 7, number 1, 2022

Table of Contents