REPORTS@SCM

AN ELECTRONIC JOURNAL OF THE SOCIETAT CATALANA DE MATEMÀTIQUES

Volume 5, num. 1, 2020

ISSN (electronic edition): 2385 - 4227

http://reportsascm.iec.cat

Institut d'Estudis Catalans

Image: Arnaud Chéritat

REPORTS@SCM

AN ELECTRONIC JOURNAL OF THE
SOCIETAT CATALANA DE MATEMÀTIQUES

Volume 5, number 1

December 2020

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

Societat Catalana de Matemàtiques

Institut d'Estudis Catalans

Focus and Scope

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

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

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

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

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

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

Style revision by Enric Ventura.

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

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

Reports@SCM
Volume 5, number 1, 2020

Table of Contents

REPORTS@SCM

AN ELECTRONIC JOURNAL OF THE
SOCIETAT CATALANA DE MATEMÀTIQUES

K-theory for C*-algebras: The hexagonal exact sequence

*Eduard Vilalta Vila
Universitat Autònoma de Barcelona
evilalta@mat.uab.cat

*Corresponding author

Resum *(CAT)*
Aquest treball té com a objectiu introduir el lector a la teoria K per C^*-àlgebres demostrant-ne dos dels seus resultats centrals coneguts: La periodicitat de Bott i la successió exacta cíclica de sis termes. Aquests dos resultats constitueixen una eina essencial de cara al càlcul explícit dels K-grups d'una C^*-àlgebra, i han estat utilitzats amb èxit en l'estudi de diverses famílies. De cara a enunciar-los, ens desviem lleugerament de la literatura estàndard i introduïm la notació K', que permet simplificar els resultats i definicions necessàries per entendre les seves demostracions.

Abstract *(ENG)*
The aim of this work is to introduce the reader to C^*-algebraic K-theory whilst proving two of its main known results: Bott periodicity and the hexagonal exact sequence. These constitute a determinant tool for the explicit computation of the K-groups of a C^*-algebra, and have been used succesfully to study a variety of families. In order to state them, we deviate slightly from the standard literature and introduce the notation K', which allows us to simplify the results and definitions needed to understand their proofs.

Keywords: *Operator Algebras, Non-commutative Topology, K-Theory.*
MSC (2010): *46L05, 46L35, 46L80, 46L85.*
Received: *October 13, 2019.*
Accepted: *November 15, 2019.*

Acknowledgement
The author was partially supported by grant 2017/COLAB/00487 of the Spanish Ministerio de Educación, Cultura y Deporte.

1. Introduction

The development of C^*-algebraic K-theory was initiated in the early 1970s, when G.A. Elliott classified the so-called approximately finite dimensional C^*-algebras by using their ordered K_0 groups, see [2]. Since then, C^*-algebraic K-theory has become an important tool in the treatment of operator algebras, and has been used succesfully to classify a considerably large family of separable and simple C^*-algebras.

In analogy to the topological K-theory developed by Aityah–Hirzebruch, in C^*-algebraic K-theory one defines a family of functors K_n from the category of C^*-algebras to that of abelian groups, thus assigning to every C^*-algebra A a family of groups $K_n(A)$. The computation of these groups, usually known as the K-groups of the algebra, provides useful information on the structure of the sets of projections and unitaries of A.

Towards this computation, and in contrast to algebraic K-theory, there exist a number of tools that make the treatment of the K-groups of a C^*-algebra manageable. Amongst them, there are two that are of particular importance: The first one, known as Bott periodicity, is the C^*-algebraic equivalent to the periodicity obtained in topological K-theory, and states that all K-groups of even and odd subscripts are isomorphic to K_0 and K_1, respectively; see [1, Ch. 9].

The second result, which is a consequence of the first one, allows us to construct a hexagonal exact sequence from any exact sequence of C^*-algebras. In particular, the existence of such a sequence implies that one can compute the K-groups of a C^*-algebra by studying the K-groups of one of its ideals and its corresponding quotient; see [1, § 9.3].

Therefore, the aim of this work is to introduce the reader to C^*-algebraic K-theory whilst proving these two results. More explicitly, for any exact sequence of C^*-algebras

$$0 \longrightarrow I \xrightarrow{\varphi} A \xrightarrow{\phi} B \longrightarrow 0, \tag{1}$$

we wish to obtain the associated hexagonal exact sequence

$$\begin{array}{ccccc} K_1(I) & \xrightarrow{K_1(\varphi)} & K_1(A) & \xrightarrow{K_1(\phi)} & K_1(B) \\ \delta_0 \uparrow & & & & \downarrow \delta_1 \\ K_0(B) & \xleftarrow{K_0(\phi)} & K_0(A) & \xleftarrow{K_0(\varphi)} & K_0(I). \end{array} \tag{2}$$

To this end, we will assume without loss of generality (see, for example, [4, § 1.1.5]) that (1) is of the form

$$0 \longrightarrow I \xrightarrow{i} A \xrightarrow{\pi} A/I \longrightarrow 0, \tag{3}$$

where I is an ideal of A and i, π are the usual inclusion and quotient mappings.

The remainder of this paper has been divided into three parts, where some familiarity with C^*-algebras is assumed. All the C^*-algebraic background needed for these sections can be found in many textbooks; see, for example, [3].

In Section 2 we recall the definitions of the K-functors as well as some of their properties. We also introduce the notation K_n', which will allow us to shorten the definitions of this section and the proofs of Section 3. As the aim of Section 2 is for the reader to get acquainted with the basics of K-theory, we omit all the proofs.

The index map, denoted by δ_1, is then defined in Section 3, where we first prove that the K'-functors are exact and invariant under homotopy. As these functors are equivalent to the K-functors, the construction of the index map together with this fact gives us a six-term exact sequence which is not yet cyclic.

Finally, in Section 4 we prove Bott periodicity and construct the hexagonal exact sequence by closing the sequence obtained in Section 3. Since the proofs of some of the preliminary lemmas in this section are rather long and arduous, we only provide a reference for them.

2. An overview of C*-algebraic K-theory

In this first section we briefly review the concepts and results of C^*-algebraic K-theory that will be used in the subsequent sections. All the proofs can be found in [4, 5]. Throughout this paper, A and B will denote C^*-algebras and, given any two square matrices a, b over A, we will refer to the matrix $\mathrm{diag(a,b)}$ by $a \oplus b$.

2.1 The projection group K$_0$ and the unitary group K$_1$

We begin our overview defining the first two K-groups. As we will later prove, these are the only ones up to isomorphism.

Proposition 2.1. *Let $P_n(A)$ be the sets of projections in $M_n(A)$ and denote by $P_\infty(A)$ their disjoint union. Then, by writting $p \sim_0 q$ if and only if $p = vv^*$ and $q = v^*v$ for some rectangular matrix v, one gets that $(P_\infty(A)/\sim_0, \oplus)$ is a commutative monoid with the class of 0 as its unit.*

Definition 2.2. For any unital C^*-algebra A, the group $K_0(A)$ is defined to be the Grothendieck group of the monoid above, where we denote the class of an element $p \in P_\infty(A)$ by $[p]_0$. If A does not have a unit, we define the group $K_0(A)$ as the kernel of the map $K_0(\pi)\colon K_0(\tilde{A}) \to K_0(\mathbb{C})$, where $K_0(\pi)([p]_0) = [\pi(p)]_0$ and π is the usual projection map from $\tilde{A}$ to $\mathbb{C}$ applied entry-wise (here, we use $\tilde{A}$ to denote the unitification of A).

Remark 2.3. It can be shown that every element in $K_0(A)$ is of the form $[p]_0 - [1_n \oplus 0_n]_0$ for some projection $p \in P_\infty(\tilde{A})$ whose scalar part $s(p)$ is $1_n \oplus 0_n$. Moreover, if A is unital, it follows from its construction that every element in $K_0(A)$ is of the form $[p]_0 - [q]_0$ for some $p, q \in P_\infty(A)$, where we can assume that both projections are of the same size.

As A is equipped with a norm, we can study its induced topology. In particular, we say that two elements a and b are homotopic in a subset $S \subset A$, in symbols $a \sim_h b$, if there exists a continuous path in S going from a to b. For example, given two projections p, q homotopic in $P_n(A)$, one can see that $[p]_0 = [q]_0$. Conversely, if $[p]_0 = [q]_0$, then $p \oplus 0_s \sim_h q \oplus 0_t$ for some positive integers s and t.

We will also say that two $*$-homomorphisms φ_0 and φ_1 from A to B are homotopic if there is a continuous map $t \mapsto \varphi_t$ from $[0, 1]$ to the $*$-homomorphisms from A to B such that $t \mapsto \varphi_t(a)$ is a homotopy for each $a \in A$. Moreover, two C^*-algebras A and B are said to be homotopic if there exist two $*$-homomorphisms ϕ and φ such that $\phi \circ \varphi \sim_h \mathrm{id_B}$ and $\varphi \circ \phi \sim_h \mathrm{id_A}$.

Proposition 2.4. *Let A be a unital C^*-algebra and consider the set $U_\infty(A) = \cup_n U_n(A)$, where $U_n(A)$ denotes the set of all unitary $n \times n$ matrices over A. Then, the equivalence relation "$u \sim_1 v$ if and only if $u \oplus 1_n \sim_h v \oplus 1_m$ in $U_N(A)$ for some suitable integers n, m, and N" makes $(U_\infty(A)/\sim_1, \oplus)$ into a commutative group with the class of 1 as its unit.*

Definition 2.5. Given a unital C^*-algebra A, the group $K_1(A)$ is the commutative group defined above, where we refer to the class of a unitary $u \in U_\infty(A)$ as $[u]_1$. If A does not have a unit, we define $K_1(A) := K_1(\tilde{A})$.

Remark 2.6. It can be proven that every element in $K_1(A)$ is of the form $[u]_1$ with $u \in U_\infty^+(\tilde{A})$, where $U_\infty^+(\tilde{A})$ is the set of unitaries whose scalar part is of norm 1.

Example 2.7. It is easy to see that two elements $p, q \in P_\infty(\mathbb{C})$ are equivalent under $\sim_0$ if and only if $\dim(\mathrm{Im(p)}) = \dim(\mathrm{Im(q)})$. Thus, it follows that $K_0(\mathbb{C}) \cong \mathbb{Z}$. Moreover, recall that a unitary u in a unital C^*-algebra is homotopic to 1 in $U(A)$ if and only if its spectrum is not $\mathbb{T}$; see [4, Lem. 2.1.3(ii)]. Therefore, as all unitaries in $U_\infty(\mathbb{C})$ have finite spectrum, they must be equivalent to 1 under $\sim_1$. This implies that $K_1(\mathbb{C}) = 0$. One can also adapt these arguments to see that $K_0(B(H)) = K_1(B(H)) = 0$ for any separable infinite dimensional Hilbert space H.

2.2 Suspension functor and higher index K-groups

Once the K_0 and K_1 groups have been defined, one can make use of the suspension functor S to define two families of groups: the higher index K-groups and the K'-groups. Even though these two families turn out to be the same, the introduction of the K'-groups allows us to simplify both the definitions and proofs regarding the properties of the higher index K-groups.

Recall that the suspension functor S is an exact covariant functor mapping a C^*-algebra A to $SA := \{f \in C(\mathbb{T}, A) \mid f(1) = 0\}$, and a $*$-homomorphism $\phi\colon A \to B$ to the $*$-homomorphism $S\phi$ from SA to SB defined as $S\phi(f) = \phi \circ f$.

Definition 2.8. By using the notation $S^0 = \mathrm{id}$ and $S^n = S^{n-1} \circ S$, we define the higher index K-groups $K_n(A) = K_1(S^{n-1}A)$ and the K'-groups $K_n'(A) = K_0(S^n(A))$.

Now let $\phi\colon A \to B$ be a $*$-homomorphism. We denote by $K_n'(\phi)\colon K_n'(A) \to K_n'(B)$ the group homomorphism $K_n'(\phi)([p]_0 - [s(p)]_0) = [S^n\tilde{\phi}(p)]_0 - [S^n\tilde{\phi}(s(p))]_0$. One can check that this definition makes K_n' into functors. A proof of the theorem below can be found in [5, Thm. 7.2.5].

Theorem 2.9. *Given any C^*-algebra A, consider the map $\theta_{A,n}\colon K_n(A) \to K_n'(A)$ defined as*

$$\theta_{A,n}([u]_1) = [w(1_m \oplus 0_m)w^*]_0 - [1_m \oplus 0_m]_0\,, \qquad u \in U_m^+(\widetilde{S^{n-1}A}),$$

where w is a homotopy between 1_{2m} and $u \oplus u^*$ in $U_{2m}(S^{n-1}A)$. Then, $\theta_{A,n}$ is an isomorphism for every integer $n \geq 1$.

Definition 2.10. Given a $*$-homomorphism $\phi\colon A \to B$, we define $K_n(\phi)\colon K_n(A) \to K_n(B)$ as $K_n(\phi) = \theta_{B,n}^{-1} \circ K_n'(\phi) \circ \theta_{A,n}$. Together with this definition, the K-groups also become functors.

3. Homotopy invariance and the index map

The goal of this section is to define the map δ_1 from (2) and prove that the two rows together with the right column of (2) form an exact sequence. However, we will first show that the K_n' functors are invariant under homotopy, as this is one of the main tools used in the explicit computation of the K groups of a C^*-algebra. Note that, by their definition and Theorem 2.9, this will imply that the functors K_n are also invariant under homotopy.

Theorem 3.1. *Given two homotopic $*$-homomorphisms φ_0 and φ_1 from A to B, we have that $K_n'(\varphi_0) = K_n'(\varphi_1)$, for every $n \geq 0$.*

Proof. Fix $n \in \mathbb{N}$ and let q be an element in $P_k(\widetilde{S^nA})$ for some k. Then, write q as the sum $q = p + \alpha 1_{\widetilde{S^nA},k}$ with $p \in M_k(S^nA)$ and $\alpha \in M_k(\mathbb{C})$. For every $t \in [0,1]$, define the elements $p_t = S^n\varphi_t(p)$ and $q_t = \widetilde{S^n\varphi_t}(q)$, where $t \mapsto \varphi_t$ is the homotopy from φ_0 to φ_1.

As $\widetilde{S^n\varphi_t}$ is a $*$-homomorphism, it follows that q_t is a projection for every t. Moreover, one gets that $q_t = p_t + \alpha 1_{\widetilde{S^nB},k}$ and, consequently, that $t \mapsto q_t$ is continuous if and only if $t \mapsto p_t$ is continuous.

Now let $\delta_p \colon [0,1] \times \mathbb{T}^n \to A$ be the map defined as $\delta(p)(t,(z_1,\dots,z_n)) = p_t(z_1)(z_2)\cdots(z_n)$, and note that, for any two pairs $(t_1,\xi_1),(t_2,\xi_2) \in [0,1] \times \mathbb{T}^n$, one gets

$$\|\delta_p(t_1,\xi_1) - \delta_p(t_2,\xi_2)\| \leq \|\varphi_{t_1}(p(\xi_1)) - \varphi_{t_2}(p(\xi_1))\| + \|p(\xi_1) - p(\xi_2)\|.$$

Thus, since p is continuous and $t \mapsto \varphi_t$ is a homotopy, we have that δ_p is also continuous. Furthermore, as δ_p has compact support, the map is uniformly continuous.

It then follows that $t \mapsto p_t$ is continuous and that $t \mapsto q_t$ is a homotopy of projections. Therefore, one gets $[q_0]_0 = [q_1]_0$ for any $q \in P_\infty(\widetilde{S^nA})$, which implies the equality $K_0(\widetilde{S^n\varphi_0}) = K_0(\widetilde{S^n\varphi_1})$, from which the desired result follows. □

Example 3.2. Let X be a compact, Hausdorff, and contractible topological space. Then, the K_0 and K_1 groups of the C^*-algebra $C(X,\mathbb{C})$ are isomorphic to $\mathbb{Z}$ and 0, respectively, as $C(X,\mathbb{C})$ is homotopic to $\mathbb{C}$. Recall that X is contractible if there exists a point x_0 and a continuous map $c\colon X \times [0,1] \to X$ such that $c(x,0) = x$ and $c(x,1) = x_0$ for every $x \in X$. Then, a pair of functions giving the homotopy between $C(X,\mathbb{C})$ and $\mathbb{C}$ are $z \mapsto z1_{C(X,\mathbb{C})}$ and $f \mapsto f(x_0)$. For more details, see [4, Ex. 3.3.6].

Proposition 3.3. *For any exact sequence of the form (3), the induced sequence*

$$K_n(I) \xrightarrow{K_n(i)} K_n(A) \xrightarrow{K_n(\pi)} K_n(A/I) \tag{4}$$

is exact for every n.

Proof. As we have previously noted, it follows from their definition and Theorem 2.9 that proving the result for K_n' is equivalent to proving it for K_n. Moreover, by using the functoriality and exactness of S, one can see that the diagram

$$\begin{array}{ccccc} K_n'(I) & \longrightarrow & K_n'(A) & \longrightarrow & K_n'(A/I) \\ \downarrow & & \downarrow & & \downarrow \\ K_0(\mathrm{Im}(S^n i)) & \longrightarrow & K_0(S^nA) & \longrightarrow & K_0(S^nA/Im(S^ni)) \end{array}$$

is commutative and has isomorphisms as columns, for every $n \geq 0$. Thus, we only need to prove that $\ker(K_0(\pi)) \subset \mathrm{Im}(K_0(i))$, as we can restrict ourselves to $n = 0$ and the other inclusion is clear.

Now, given an element $[p]_0 - [s(p)]_0$ in $\ker(K_0(\pi))$, find a unitary $u \in U_N(\widetilde{A/I})$ such that, for suitable integers n, k, N, $u(\tilde{\pi} \oplus 1_n \oplus 0_k) = s(p) \oplus 1_n \oplus 0_k$. Then, by taking a unitary w homotopic to 1_{2N} in $U_{2N}(A)$ such that $\tilde{\pi}(w) = u \oplus u^*$, we can define the projection

$$r = w(p \oplus 1_n \oplus 0_{k+N})w^*.$$

As $\tilde{\pi}(r) \in M_\infty(\mathbb{C}1_A)$ by construction, it follows that $r \in M_\infty(\tilde{I})$. In particular, we have that

$$[p]_0 - [s(p)]_0 = [r]_0 - [s(r)]_0 \in \mathrm{Im}(\mathrm{K}_0(\mathrm{i})),$$

as required. □

Theorem 3.4. *For any exact sequence of the form (3), there exists a group homomorphism δ_1 such that the following sequence is exact:*

$$\begin{array}{ccccc} K_1(I) & \xrightarrow{K_1(i)} & K_1(A) & \xrightarrow{K_1(\pi)} & K_1(A/I) \\ & & & & \downarrow \delta_1 \\ K_0(A/I) & \xleftarrow{K_0(\pi)} & K_0(A) & \xleftarrow{K_0(i)} & K_0(I). \end{array} \tag{5}$$

Proof. Given an element $[u]_1 \in K_1(A/I)$ with $u \in U_k^+(\widetilde{A/I})$, we define its image through the index map δ_1 as

$$\delta_1([u]_1) = [w(1_n \oplus 0_k)w^*]_0 + [1_n \oplus 0_k]_0,$$

where $v \in U_k^+(\widetilde{A/I})$ is such that $u \oplus v \sim_h 1_{n+k}$ in $U_k^+(\widetilde{A/I})$, and w is a unitary lift of $u \oplus v$. It can be proven that δ_1 is indeed a well defined group homomorphism; see, for example, [5, Prop. 8.1.3].

Then, it follows from Proposition 3.3 that we only need to prove the equalities $\mathrm{Im}(\delta_1) = \ker(\mathrm{K}_0(\mathrm{i}))$ and $\ker(\delta_1) = \mathrm{Im}(\mathrm{K}_1(\pi))$. Moreover, note that the inclusions $\mathrm{Im}(\delta_1) \subseteq \ker(\mathrm{K}_0(\mathrm{i}))$ and $\mathrm{Im}(\mathrm{K}_1(\pi)) \subseteq \ker(\delta_1)$ are clear.

Thus, let $[u]_1 \in \ker(\delta_1)$ with $u \in U_m^+(\widetilde{A/I})$ for some m, and let w be a unitary lift of $u \oplus u^*$. As $\delta_1([u]_1) = 0$, we can find an integer k and a matrix $v \in M_{2(k+2m)}(\tilde{I})$ such that $vv^* = 1_{2n} - q \oplus 1_k \oplus 0_n$ and $v^*v = 1_{2n} - (1_m \oplus 0_m) \oplus (1_k \oplus 0_n)$, where $q = w(1_m \oplus 0_m)w^*$ and $n = k + 2m$.

By using the previous two equalities together with $vv^*v = v$, it is easy to check that $\tilde{\pi}(v) = 0_m \oplus X$ for some $X \in M_{2n-m}(\mathbb{C}1_{\tilde{A}})$. Therefore, there exists a complex $(2n+m) \times (2n+m)$ matrix U such that $\tilde{\pi}(qw \oplus v) = u \oplus U$. As $U \sim_h 1_{2n+m}$, it follows that

$$[u]_1 = [u]_1 + [U]_1 = K_1(\pi)([qw \oplus v]_1) \in \mathrm{Im}(\mathrm{K}_1(\pi)).$$

Now let $x = [p]_0 - [1_n \oplus 0_n]_0 \in \ker(K_0(i))$ with $p \in P_{2n}(\tilde{I})$. Then, there exists an integer $k \in \mathbb{N}$ for which $p \oplus 1_k \oplus 0_m \sim_0 (1_n \oplus 0_n) \oplus 1_k \oplus 0_m =: s_k$, with $m = 3(2n+k)$. Moreover, we can find a complex matrix P such that $Ps_kP^* = 1_{n+k} \oplus 0_{n+m}$ and, consequently, we have

$$q := P(p \oplus 1_k \oplus 0_m)P^* \sim_0 1_{n+k} \oplus 0_{n+m} =: d.$$

Take $w \in U_\infty^+(\tilde{A})$ such that $wqw^* = d$ and note that $\tilde{\pi}(w)$ commutes with d. It follows that $\tilde{\pi}(w) = a \oplus b$ for some $a \in U_{n+k}^+(\tilde{A})$. Finally, as w is a unitary lift of $a \oplus b$, we get

$$x = [q]_0 - [d]_0 = [wdw^*]_0 - [d]_0 = \delta_1(\tilde{\pi}(a)) \in \mathrm{Im}(\delta_1),$$

as desired. □

4. Bott periodicity and the hexagonal exact sequence

We begin this section by noting that $M_n(\widetilde{SA})$ can be identified with the set of continuous functions $f \in C(\mathbb{T}, M_n(\tilde{A}))$ such that $f(1) \in M_n(\mathbb{C}1_{\tilde{A}})$. In particular, we can write

$$U_n(\widetilde{SA}) = \{f \in C(\mathbb{T}, U_n(\tilde{A})) \mid f(1) \in M_n(\mathbb{C}1_{\tilde{A}})\}.$$

By using this identification, we can now define the Bott map. What follows is a combination of [4, Ch. 11] and [5, Ch. 9]:

Definition 4.1. Let A be a unital C^*-algebra and take $p \in P_n(A)$. We define $f_p \in U_n(\widetilde{SA})$ as the map from $\mathbb{T}$ to A such that $f_p(z) = zp + (1_n - p)$. The Bott map $\beta_A \colon K_0(A) \to K_1(SA)$ is then defined as $\beta_A([p]_0 - [q]_0) = [f_p f_q^*]_1$ for any element $[p]_0 - [q]_0$ in $K_0(A)$.

Remark 4.2. It can be proven that β_A is a well defined homomorphism; see, e.g., [4, § 11.1].

If A is not unital, we define the Bott map of A to be the only homomorphism for which the following diagram is commutative:

$$\begin{array}{ccccccccc} 0 & \longrightarrow & K_0(A) & \longrightarrow & K_0(\tilde{A}) & \rightleftarrows & K_0(\mathbb{C}) & \longrightarrow & 0 \\ & & \downarrow{\scriptstyle \beta_A} & & \downarrow{\scriptstyle \beta_{\tilde{A}}} & & \downarrow{\scriptstyle \beta_{\mathbb{C}}} & & \\ 0 & \longrightarrow & K_1(SA) & \longrightarrow & K_1(S\tilde{A}) & \rightleftarrows & K_1(S\mathbb{C}) & \longrightarrow & 0. \end{array}$$

In particular, it follows that we only need to prove that the Bott map is an isomorphism in the unital case. Thus, we will assume from now on that A has a unit.

We will first prove that β_A is surjective. Let $GL_0(M_n(A))$ be the set of invertible $n \times n$ matrices that are homotopic to the identity. Then, define the following sets:

$$\begin{aligned} \mathrm{Inv}_0^n &:= C(\mathbb{T}, GL_0(M_n(A))), \\ \mathrm{Pol}_m^n &:= \{f \in \mathrm{Inv}_0^n \mid f(z) = \sum_{i=0}^{m} a_i z^i,\ a_i \in M_n(A)\}, \\ \mathrm{Trig}_m^n &:= \{f \in \mathrm{Inv}_0^n \mid f(z) = \sum_{i=-m}^{m} a_i z^i,\ a_i \in M_n(A)\}. \end{aligned}$$

Remark 4.3. One can check that $U_n(\widetilde{SA})$ is a subset of Inv_0^n for every n, and that, if two unitaries are homotopic in Inv_0^n, then they are also homotopic in $U_n(\widetilde{SA})$; see [4, § 11.2].

As we have already mentioned in the introduction, we omit the proof of the next lemma.

Lemma 4.4 ([5, Lem. 9.2.3–9.2.7])**.** *For every unital C^*-algebra A and every integer $n \in \mathbb{N}$, we have:*

(i) for every $f \in \mathrm{Inv}_0^n$, there exists an integer m and an element $h \in \mathrm{Trig}_m^n$ such that $f \sim_h h$ in Inv_0^n;

(ii) for every integer m there exists a continuous map μ_m^n from Pol_m^n to Pol_1^{mn+n} such that $\mu_m^n(f)$ is homotopic to $f \oplus 1_{mn}$ in Inv_0^n, for every f;

(iii) for any degree one polynomial $f \in \mathrm{Pol}_1^n$ there exists an element $\gamma(f)$ of the form f_p such that $f \sim_h \gamma(f)$ in Pol_1^n; moreover, the map $f \mapsto \gamma(f)$ is continuous.

Proposition 4.5. *The Bott map is surjective.*

Proof. Let $[f]_1 \in K_1(SA)$ with $f \in U_n(\widetilde{SA})$. By Lemma 4.4(1), we can find an element $h \in \mathrm{Trig}_m^n$ such that $f \sim_h h$ in Inv_0^n. As $z^{-N} \oplus 1_{M-1} \sim_h f^*_{1_N \oplus 0_{M-N}}$ in $U_M(\widetilde{SA})$ for every pair N, M such that $N \leq M$, we get $(hz^N)z^{-N} \oplus 1_M \sim_h (hz^N \oplus 1_M)f^*_{p_N}$ for every $N \leq M$, and where $p_N = 1_N \oplus 0_{M-N}$.

In particular, if N is large enough, hz^N is polynomial. Thus, for any such N, Lemma 4.4(2) ensures that we can find a degree one polynomial r such that $hz^N \oplus 1_t \sim_h r$ for some t. Moreover, it follows from Lemma 4.4(3) that there exists some element f_p homotopic to r.

By now adding some extra 1's in the diagonal, we have that $f \oplus 1_{M+nt} \sim_h f_{p \oplus 0_M} f^*_{p_N}$ and, consequently, that $\beta_A([p]_0 - [p_N]_0) = [f]_1$. □

We now prove that β_A is injective. Once again, we will omit the proof of the following lemma.

Lemma 4.6 ([5, § 9.1.2 & Lem. 9.2.10])**.** *For every unital C^*-algebra A and every integer $n \in \mathbb{N}$, we have:*

(i) the map $\pi \colon \{f_p \mid p \in P_n(A)\} \to P_n(A)$ sending an element f_p to p is continuous;

(ii) for any homotopy $f \mapsto f_t$ in Inv_0^n, there exists a positive integer N such that $f \mapsto f_t$ can be uniformly approximated by a homotopy $c \mapsto c_t$ in Trig_N^n that is piecewise linear. In particular, if $f_0, f_1 \in \mathrm{Trig}_N^n$, one can set $c_0 = f_0$ and $c_1 = f_1$.

Proposition 4.7. *The Bott map is injective.*

Proof. Let $[p]_0 - [q]_0 \in K_0(A)$ be such that $\beta_A([p]_0 - [q]_0) = 0$ or, equivalently, such that $[f_p f_q^*]_1 = 0$. Then, by possibly adding some zeros diagonally, we have that $f_p \sim_h f_q$. By Lemma 4.6(2), we can find a polynomial homotopy between $z^N f_p$ and $z^N f_q$ for some integer N. Moreover, once again adding some zeros and ones diagonally, and using Lemma 4.4(2,3), we obtain a homotopy $t \mapsto f_{p_t}$ such that $p_0 = p$ and $p_1 = q$. Finally, Lemma 4.6(1) gives a homotopy between p and q, from which we get that $[p]_0 = [q]_0$ and, therefore $[p]_0 - [q]_0 \in \ker(\beta_A)$. □

Combining Propositions 4.5 and 4.7 above, one gets the desired result:

Theorem 4.8. *For any C^*-algebra A, the Bott map β_A is an isomorphism between the groups $K_0(A)$ and $K_2(A)$. Consequently, all the K-groups $K_n(A)$ of even subindexes are isomorphic to $K_0(A)$, and those with odd subindexes are isomorphic to $K_1(A)$.*

Other proofs of this result are indeed possible, such as the recent one in [6]. In it, Voiculescu's almost commuting matrices are used to define a homomorphism $\alpha_A \colon K_1(SA) \to K_0(A)$, which is then shown, with the help of Atiyah's rotation trick, to be the mutual inverse of β_A (the author thanks the anonymous referee for this reference).

With Bott periodicity at hand, we can now construct the hexagonal exact sequence.

Theorem 4.9. *For any exact sequence of the form* (3), *there exists a group homomorphism* δ_0 *such that the following hexagonal sequence is exact*

$$\begin{array}{ccccc} K_1(I) & \xrightarrow{K_1(i)} & K_1(A) & \xrightarrow{K_1(\pi)} & K_1(A/I) \\ \delta_0 \uparrow & & & & \downarrow \delta_1 \\ K_0(A/I) & \xleftarrow{K_0(\pi)} & K_0(A) & \xleftarrow{K_0(i)} & K_0(I). \end{array}$$

Proof. Given an exact sequence $0 \to I \to A \to A/I \to 0$, consider the suspended sequence

$$0 \to SI \to SA \to S(A/I) \to 0$$

and its corresponding index δ_1' from Theorem 3.4. Then, define δ_0 as the composition $\theta_{I,1}^{-1} \circ \delta_1' \circ \beta_{A/I}$, where $\theta_{I,1}^{-1}$ is the isomorphism from Theorem 2.9.

$$\begin{array}{cccccc} & K_2(A/I) & & & & \\ & \downarrow \theta_{I,1}^{-1}\circ\delta_1' & & & & \\ \beta_{A/I} \nearrow & K_1(I) & \xrightarrow{K_1(i)} & K_1(A) & \xrightarrow{K_1(\pi)} & K_1(A/I) \\ & & & & & \downarrow \delta_1 \\ & K_0(A/I) & \xleftarrow{K_0(\pi)} & K_0(A) & \xleftarrow{K_0(i)} & K_0(I) \end{array}$$

By Theorem 3.4, we only need to prove that the sequence is exact at $K_1(I)$ and $K_0(A/I)$. To do this, simply note that the following diagram is commutative

$$\begin{array}{ccccccc} K_0(A) & \xrightarrow{K_0(\pi)} & K_0(A/I) & \xrightarrow{\delta_0} & K_1(I) & \xrightarrow{K_1(i)} & K_1(A) \\ \beta_A \updownarrow & & \beta_{A/I} \updownarrow & & \updownarrow \theta_I & & \updownarrow \theta_A \\ K_1(SA) & \xrightarrow{K_1(S\pi)} & K_1(S(A/I)) & \xrightarrow{\delta_1} & K_0(SI) & \xrightarrow{K_0(Si)} & K_0(SA), \end{array}$$

and that all of its columns are isomorphisms. As the second row is exact by Theorem 3.4, so is the first one. □

Example 4.10. Let H be an infinite dimensional separable Hilbert space and consider the Calkin algebra $Q(H) = B(H)/K(H)$, where $K(H)$ is the algebra of compact operators on H. Then, the exact sequence

$$0 \longrightarrow K(H) \xrightarrow{i} B(H) \xrightarrow{\pi} Q(H) \longrightarrow 0$$

induces, by Theorem 4.9, the hexagonal exact sequence

$$\begin{array}{ccccc} K_1(K(H)) & \xrightarrow{K_1(i)} & K_1(B(H)) & \xrightarrow{K_1(\pi)} & K_1(Q(H)) \\ \delta_0 \uparrow & & & & \downarrow \delta_1 \\ K_0(Q(H)) & \xleftarrow{K_0(\pi)} & K_0(B(H)) & \xleftarrow{K_0(i)} & K_0(K(H)). \end{array}$$

Moreover, recall from Example 2.7 that $K_0(B(H)) = K_1(B(H)) = 0$. Thus, δ_0 and δ_1 are isomorphisms.

By using that the K-groups are stable (see [4, Prop. 6.4.1 & Prop. 8.2.8]), one can also see that $K_0(K(H)) \cong K_0(\mathbb{C}) \cong \mathbb{Z}$ and that $K_1(K(H)) \cong K_1(\mathbb{C}) = 0$. Therefore, the K_0 and K_1 groups of the Calkin algebra are isomorphic to 0 and $\mathbb{Z}$ respectively.

References

[1] B. Blackadar, "K-Theory for operator algebras", Springer-Verlag, New York, 1986.

[2] G.A. Elliott, "On the classification of inductive limits of sequences of semisimple finite-dimensional algebras", *Journal of Algebra* **38** (1976), 29–44.

[3] G.J. Murphy, "C^*-Algebras and operator theory", Academic Press, 1990.

[4] M. Rørdam, F. Larsen, and N. Laustsen, "An introduction to K-theory for C^*-algebras", Cambridge University Press, 2000.

[5] N.E. Wegge-Olsen, "K-Theory and C^*-algebras: a friendly approach", Oxford University Press, 1993.

[6] R. Willett, "Bott periodicity and almost commuting matrices", *preprint*, 2019.

REPORTS @SCM

AN ELECTRONIC JOURNAL OF THE
SOCIETAT CATALANA DE MATEMÀTIQUES

A $\mathcal{C}^0$ Interior Penalty Method for 4^{th} order PDEs

Dani Fojo
Universitat Politècnica de Catalunya
daniel.fojo@estudiant.upc.edu

David Codony
Universitat Politècnica de Catalunya
david.codony@upc.edu

*__Sonia Fernández-Méndez__
Universitat Politècnica de Catalunya
sonia.fernandez@upc.edu

∗Corresponding author

Resum *(CAT)*
En aquest treball desenvolupem i estudiem el comportament d'un mètode per a la solució d'EDPs de 4^{rt} ordre amb aproximacions d'Elements Finits $\mathcal{C}^0$ estàndar. El mètode es basa en una forma feble que introdueix integrals entre elements per imposar continuïtat $\mathcal{C}^1$ en forma feble. El mètode es desenvolupa per les equacions que modelitzen una placa de Kirchoff, però es preveu que l'extensió a altres EDPs de 4^{rt} ordre sigui natural. La convergència i aplicabilitat del mètode s'estudia amb exemples numèrics.

Abstract *(ENG)*
A method to solve 4^{th}-order PDEs using the Finite Element Method (FEM) with standard $\mathcal{C}^0$ elements is derived and studied. It is based on a special weak form accounting for the discontinuous derivatives of the approximation and imposing their normal continuity across element sides in weak form. The method is developed for the equations of the deformation of a Kirchoff plate, but its extension to other 4^{th}-order PDEs is expected to be straightforward. The accuracy and convergence of the resulting numerical approximation is studied with numerical experiments.

Keywords: *4^{th}-order PDE, $\mathcal{C}^0$ Finite Elements, Nitsche's method, Interior Penalty Method.*
MSC (2010): *35Q74, 74R10, 65M60.*
Received: *November 21, 2019.*
Accepted: *May 16, 2020.*

Acknowledgement
This work was supported by the European Research Council (StG-679451 to Irene Arias) and the Generalitat de Catalunya (2017-SGR-1278).

Societat Catalana de Matemàtiques

Institut d'Estudis Catalans

1. Introduction

There are two main strategies for the numerical solution of 4^{th} order PDEs. The first one consists on considering an approximation space with $\mathcal{C}^1$ continuity to discretize a weak form involving 2^{nd} order derivatives. The main drawback of this approach is that the definition of $\mathcal{C}^1$-continuous approximations on non-cartesian meshes, such as the ones necessary to adapt to non-rectangular domains, is really cumbersome. Thus, these approximations are limited to the solution of problems in rectangular domains or in combination with a technique for embedded domains; see for instance [3].

An alternative is splitting the 4^{th} order PDE in two 2^{nd} order PDEs, allowing the use of $\mathcal{C}^0$ Finite Element (FE) approximations. However, the approximation spaces for the primal unknown and the additional unknown must fulfil some conditions for stability that, again, lead to approximation spaces with cumbersome definitions, and difficult extension to high-order approximations; see for instance [4].

A not so common approach is considering a modified weak form suitable for standard $\mathcal{C}^0$ FE approximations, imposing continuity of the derivative in weak form. This is the strategy considered in this work.

The developed formulation is based on the ideas of the Interior Penalty Method (IPM) [1], which considers discontinuous approximations and imposes $\mathcal{C}^0$ continuity in weak form, in the context of 2^{nd} order PDEs. Here, the same ideas are applied for 4^{th} order PDEs, but considering $\mathcal{C}^0$ approximations and imposing the continuity of the derivative in weak form. The resulting weak form involves second order derivatives, two different types of Dirichlet and Neumann boundary conditions and punctual forces on corners of the boundary. It coincides with the one proposed and analyzed in [2], but without some limitations for the boundary conditions. The derivation here is based on the use of the surface divergence theorem, instead of considering a limit from rounded corners to sharp corners, leading to a more natural understanding of the contribution of interior and boundary corners (vertices). In addition, a convergence study based on numerical experiments, assessing the real applicability of the method, is included here, and an strategy based on an eigenvalue problem is also proposed for the estimate of the value for the stabilization parameter to ensure coercivity.

Einstein notation (repeated indexes sum over) is assumed in the whole text.

2. A $\mathcal{C}^0$ Interior Penalty Method for Kirchoff plates

The equations modelling the behaviour of a plate with the Kirchoff model are

$$\frac{\partial^2 \sigma_{ij}(u)}{\partial x_i \partial x_j} = f \text{ on} \Omega \tag{1a}$$
$$u = g_1 \qquad \text{on} \Gamma_D^1 \tag{1b}$$
$$\frac{\partial u}{\partial \mathbf{n}} = g_2 \qquad \text{on} \Gamma_D^2 \tag{1c}$$
$$t(u) = t_n \qquad \text{on} \Gamma_N^1 \tag{1d}$$
$$r(u) = r_n \qquad \text{on} \Gamma_N^2 \tag{1e}$$
$$j_k(u) = j_k^{\text{ext}} \quad \text{on } V_k \in V_N, \tag{1f}$$

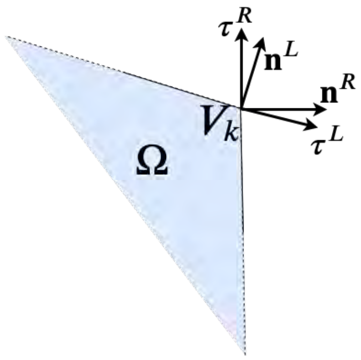

Figure 1: Left and right tangent and normal vectors at a corner of the domain.

where

$$
\begin{aligned}
\sigma_{ij}(u) &= C_{ijkl}\frac{\partial^2 u}{\partial x_k \partial x_l} \\
t(u) &= \left(\frac{\partial \sigma_{ij}(u)}{\partial x_i} - \nabla^\tau \cdot (\mathbf{n}) n_i \sigma_{ij}(u)\right) n_j + \nabla^\tau \cdot (\boldsymbol{\sigma}(u)\cdot \mathbf{n}) \\
r(u) &= \mathbf{n}\cdot\boldsymbol{\sigma}(u)\cdot\mathbf{n} \\
j_k(u) &= \boldsymbol{\tau}^{\mathbf{L}}_{\mathbf{k}}\cdot\boldsymbol{\sigma}(u)\cdot\mathbf{n}^{\mathbf{L}}_{\mathbf{k}} + \boldsymbol{\tau}^{\mathbf{R}}_{\mathbf{k}}\cdot\boldsymbol{\sigma}(u)\cdot\mathbf{n}^{\mathbf{R}}_{\mathbf{k}},
\end{aligned}
\tag{2}
$$

$\Gamma^1_D \cup \Gamma^1_N = \Gamma^2_D \cup \Gamma^2_N = \partial\Omega$, V_N are the vertices in the boundary inΓ^1_N, $\mathbf{n}$ is the exterior unitary normal vector, $\boldsymbol{\tau}$ is the tangent vector, and $\nabla^\tau \cdot \mathbf{f} := \tau_k \partial f_k/\partial\boldsymbol{\tau}$. At each vertex, superscripts L and R refer to the left and right sides that meet there; see Fig. 1

In these equations u is the vertical displacement on the plate, the 4th order tensor $\mathbf{C}$ depends on the material, equation (1a) is the 4th-order PDE stating equilibrium with the vertical applied load f, equations (1b) and (1c) are the first and second Dirichlet conditions, equations (1d) and (1e) are the first and second Neumann conditions, and equation (1f) imposes punctual forces on the exterior vertices where the displacement is unknown. The boundary conditions (1c), (1e), and (1f), that may be not intuitive, can be justified from mechanical reasonings, or can be derived from the weak form of (1a).

Let us consider now a partition of Ω in subdomainsΩ_e, that will in fact be the elements, and definitions for broken domain and boundaries, such as $\widehat{\Omega} = \bigcup_e \Omega_e$. Then, the problem can be stated as

$$\frac{\partial^2 \sigma_{ij}(u)}{\partial x_i \partial x_j} = f \text{ on } \widehat{\Omega} \tag{3a}$$

$$u = g_1 \quad \text{on } \widehat{\Gamma^1_D} \tag{3b}$$

$$\frac{\partial u}{\partial \mathbf{n}} = g_2 \quad \text{on } \widehat{\Gamma^2_D} \tag{3c}$$

$$t(u) = t_n \quad \text{on } \widehat{\Gamma^1_N} \tag{3d}$$

$$r(u) = r_n \quad \text{on } \widehat{\Gamma^2_N} \tag{3e}$$

$$j_k(u) = j_k^{\text{ext}} \quad \text{on } V_k \in V_N \tag{3f}$$

$$[\![u\mathbf{n}]\!] = \mathbf{0} \quad \text{on} \Gamma \tag{3g}$$

$$\left[\!\!\left[\frac{\partial u}{\partial \mathbf{n}}\right]\!\!\right] = 0 \quad \text{on}\Gamma \tag{3h}$$
$$[\![t(u)]\!] = 0 \quad \text{on}\Gamma \tag{3i}$$
$$[\![\mathbf{n}r(u)]\!] = \mathbf{0} \quad \text{on}\Gamma \tag{3j}$$
$$\sum_{e\in E_k} j_k^e(u) = 0 \text{ on } V_k \in V_{int}, \tag{3k}$$

where Γ is the union of all interior sidesΓ_f,

$$\Gamma = \left[\bigcup_e \partial\Omega_e\right] \setminus \partial\Omega = \bigcup_f \Gamma_f, \tag{4}$$

E_k is the set of elements touching the vertex V_k, V_{int} is the set of all interior vertices (i.e., vertices in Γ), and the jump operator is defined on each sideΓ_f as $[\![a]\!] = a^L + a^R$, with a^L and a^R being the values from the elementsΩ^L andΩ^R sharing the side. Note that the jump operator is always used including the normal vector, for instance, $[\![u\mathbf{n}]\!] = u^L\mathbf{n}^L + u^R\mathbf{n}^R = (u^L - u^R)\mathbf{n}^L$, thus, it is zero for a continuous function.

Equations (3g) and (3h) impose continuity of the displacement and its normal derivative. And equations (3i), (3j), and (3k) impose equilibrium of internal forces across sides between elements and on internal vertices.

Now, multiplying equation (3a) by an arbitrary function v, integrating over any elementΩ_e and using twice integration by parts leads to

$$\int_{\Omega_e} vf\, d\Omega = \int_{\Omega_e} \frac{\partial^2 v}{\partial x_i \partial x_j}\sigma_{ij}(u)\, d\Omega - \int_{\partial\Omega_e} \frac{\partial v}{\partial x_i}\sigma_{ij}(u) n_j\, dS + \int_{\partial\Omega_e} v\frac{\partial \sigma_{ij}(u)}{\partial x_j} n_i\, dS \tag{5}$$

Now, the derivative in the first boundary integral can be split in normal and tangential derivative as $\frac{\partial v}{\partial x_i} = \tau_i \frac{\partial v}{\partial \boldsymbol{\tau}} + n_i \frac{\partial v}{\partial \mathbf{n}}$, and the integral for the tangential derivative can be expressed as

$$\int_{\partial\Omega_e} \tau_i \frac{\partial v}{\partial \boldsymbol{\tau}}\sigma_{ij}(u) n_j\, dS = \int_{\partial\Omega_e} \nabla^\tau \cdot (v\boldsymbol{\sigma}(u)\cdot\mathbf{n})\, dS - \int_{\partial\Omega_e} v\nabla^\tau \cdot (\boldsymbol{\sigma}(u)\cdot\mathbf{n})\, dS,$$

or, using the surface diverge theorem,

$$\int_{\partial\Omega_e} \tau_i \frac{\partial v}{\partial \boldsymbol{\tau}}\sigma_{ij}(u) n_j\, dS = \int_{\partial\Omega_e} \nabla^\tau \cdot (\mathbf{n}) v\, \mathbf{n}\cdot\boldsymbol{\sigma}(u)\cdot\mathbf{n}\, dS + \sum_{s=1}^{\#sides\,\Omega_e} v\left[\boldsymbol{\tau}\cdot\boldsymbol{\sigma}(u)\cdot\mathbf{n}\right]_0^{end} - \int_{\partial\Omega_e} v\nabla^\tau \cdot (\boldsymbol{\sigma}(u)\cdot\mathbf{n})\, dS,$$

where $[\cdot]_0^{end}$ denotes the value at the end minus the value at the beginning of the side, for each side of the element. Thus, equation (5) can now be written as

$$\begin{aligned}
\int_{\Omega_e} vf\, d\Omega = &\int_{\Omega_e} \frac{\partial^2 v}{\partial x_i \partial x_j}\sigma_{ij}(u)\, d\Omega \\
&+ \int_{\partial\Omega_e} v\left[\left(\frac{\partial \sigma_{ij}(u)}{\partial x_i} - \nabla^\tau\cdot(\mathbf{n}) n_i \sigma_{ij}(u)\right) n_j + \nabla^\tau\cdot(\boldsymbol{\sigma}(u)\cdot\mathbf{n})\right] dS \\
&- \int_{\partial\Omega_e} \frac{\partial v}{\partial \mathbf{n}}\left[n_i\sigma_{ij}(u)n_j\right] dS - \sum_{k=1}^{\#vertices\,\partial\Omega_e} v\left(\boldsymbol{\tau}_k^L \sigma(V_k)\mathbf{n}_k^L + \boldsymbol{\tau}_k^R \sigma(V_k)\mathbf{n}_k^R\right),
\end{aligned} \tag{6}$$

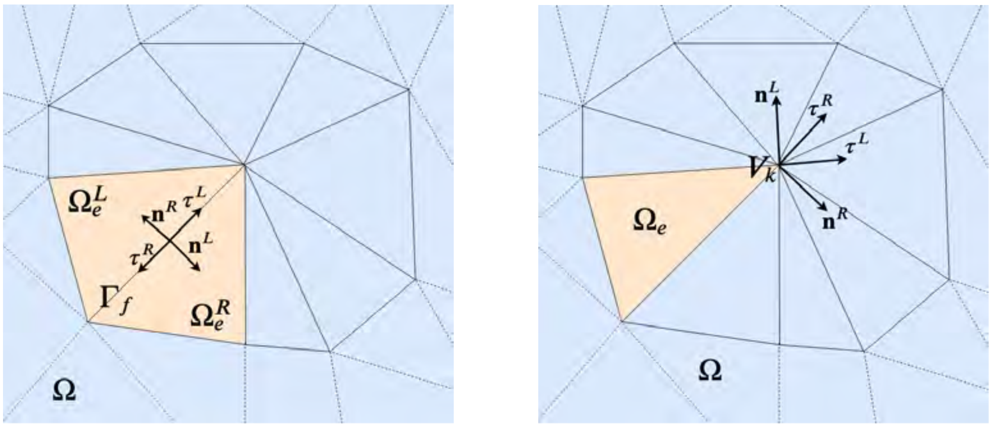

Figure 2: Left and right tangent and normal vectors for a side Γ_f inside the mesh, and left and right tangent and normal vectors for an interior vertex of an arbitrary element Ω_e inside the mesh.

where the L and R indices refer to the values at the vertices from the left and right sides, and the tangent vectors $\boldsymbol{\tau}_k^{L,R}$ point outward on the vertex for each side; see Fig. 2, right. Now, applying Definitions (2), equation (6) becomes

$$\int_{\Omega_e} vf \, d\Omega = \int_{\Omega_e} \frac{\partial^2 v}{\partial x_i \partial x_j} \sigma_{ij}(u) \, d\Omega + \int_{\partial\Omega_e} vt^e(u) \, dS - \int_{\partial\Omega_e} \frac{\partial v}{\partial \mathbf{n}} r^e(u) \, dS - \sum_{k=1}^{\#\text{vertices } \partial\Omega_e} vj_k^e(u), \quad (7)$$

where a superscript e remarks that the value is taken from element Ω_e. Summing (7) for all elements,

$$\begin{aligned} \int_{\widehat{\Omega}} vf \, d\Omega = & \int_{\widehat{\Omega}} \frac{\partial^2 v}{\partial x_i \partial x_j} \sigma_{ij}(u) \, d\Omega + \int_{\partial\Omega} vt(u) \, dS - \int_{\partial\Omega} \frac{\partial v}{\partial \mathbf{n}} r(u) \, dS \\ & + \int_{\Gamma} v \llbracket t(u) \rrbracket \, dS - \int_{\Gamma} \left[\frac{\partial v^L}{\partial \mathbf{n}^L} r^L(u) + \frac{\partial v^R}{\partial \mathbf{n}^R} r^R(u) \right] dS \\ & - \sum_{V_k \in V_{\text{int}}} v \sum_{e \in E_k} j_k^e(u) - \sum_{V_k \in V_{\text{ext}}} v \sum_{e \in E_k} j_k^e(u), \end{aligned}$$

where V_{int}, V_{ext} are the set of interior and exterior vertices, respectively. Now, using the identity

$$\frac{\partial v^L}{\partial \mathbf{n}^L} r^L(u) + \frac{\partial v^R}{\partial \mathbf{n}^R} r^R(u) = \left\llbracket \frac{\partial v}{\partial \mathbf{n}} \right\rrbracket \{r(u)\} + \{\nabla v\} \cdot \llbracket \mathbf{n} r(u) \rrbracket ,$$

with the mean operator $\{a\} := \frac{1}{2}\left(a^L + a^R\right)$, and the equilibrium equations (3i), (3j), and (3k), leads to

$$\int_{\widehat{\Omega}} vf \, d\Omega = \int_{\widehat{\Omega}} \frac{\partial^2 v}{\partial x_i \partial x_j} \sigma_{ij}(u) \, d\Omega + \int_{\partial\Omega} vt(u) \, dS - \int_{\partial\Omega} \frac{\partial v}{\partial \mathbf{n}} r(u) \, dS - \int_{\Gamma} \left\llbracket \frac{\partial v}{\partial \mathbf{n}} \right\rrbracket \{r(u)\} \, dS - \sum_{V_k \in V_{\text{ext}}} v \sum_{e \in E_k} j_k^e(u).$$

Finally, replacing the Neumann boundary conditions, (1d), (1e), and (1f), imposing the first Dirichlet boundary condition in strong form (that is, (1b) and, consequently $v = 0$ on Γ_D^1), and adding some integrals with null sum (as a consequence of the $\mathcal{C}^1$ continuity of the solution (3h) and the second Dirichlet boundary condition (1c)), we get the final weak form: find $u \in \mathcal{H}^2(\widehat{\Omega}) \cap \mathcal{C}^0(\Omega)$ such that $u = g_1$ on Γ_D^1 and $a(u, v) = \ell(v)$, for all $v \in \mathcal{H}^2(\widehat{\Omega}) \cap \mathcal{C}^0(\Omega)$ such that $v = 0$ on Γ_D^1, where

$$
\begin{aligned}
a(u, v) = & \int_{\widehat{\Omega}} \frac{\partial^2 v}{\partial x_i \partial x_j} \sigma_{ij}(u)\, d\Omega \\
& - \int_{\Gamma} \left[\!\left[\frac{\partial v}{\partial \mathbf{n}} \right]\!\right] \{r(u)\}\, dS - \int_{\Gamma} \{r(v)\} \left[\!\left[\frac{\partial u}{\partial \mathbf{n}} \right]\!\right] dS + \beta \int_{\Gamma} \left[\!\left[\frac{\partial v}{\partial \mathbf{n}} \right]\!\right] \left[\!\left[\frac{\partial u}{\partial \mathbf{n}} \right]\!\right] dS \\
& - \int_{\Gamma_D^2} \frac{\partial v}{\partial \mathbf{n}} r(u)\, dS - \int_{\Gamma_D^2} r(v) \frac{\partial u}{\partial \mathbf{n}}\, dS + \alpha \int_{\Gamma_D^2} \frac{\partial v}{\partial \mathbf{n}} \frac{\partial u}{\partial \mathbf{n}}\, dS,
\end{aligned} \tag{8a}
$$

$$
\begin{aligned}
\ell(v) = & \int_{\widehat{\Omega}} v f\, d\Omega - \int_{\Gamma_N^1} v t_n\, dS + \int_{\Gamma_N^2} \frac{\partial v}{\partial \mathbf{n}} r_n\, dS + \sum_{V_k \in V_N} v j_k^{\text{ext}} \\
& - \int_{\Gamma_D^2} r(v) g_2\, dS + \alpha \int_{\Gamma_D^2} \frac{\partial v}{\partial \mathbf{n}} g_2\, dS.
\end{aligned} \tag{8b}
$$

The terms added to the weak form recover symmetry and coercivity of the bilinear form, provided that parameters β and α are large enough. They also weakly enforce continuity of the normal derivative across elements interior sides (continuity along sides is given by the $\mathcal{C}^0$ continuity), and the second Dirichlet boundary condition. In fact, the parameters β and α act as penalty parameters, but differently to a non-consistent penalty formulation, moderate values of the parameters, of order $\mathcal{O}(h^{-1})$, provide convergence for any degree of approximation, avoiding the typical ill-conditioning problems of non-consistent penalty methods. Proper values for the parameters can be obtained solving an eigevalue problem, as commented in Section 3.

The methodology considered here for the weak imposition of interface conditions and boundary conditions is inspired by the Interior Penalty Method [1], developed in the context of discontinuous approximations to weakly impose $\mathcal{C}^0$ continuity, and on Nitsche's method [6], developed for Dirichlet boundary conditions, here applied for the second Dirichlet boundary condition. Both methods are well known in the context of second-order PDEs. The difficulties for its application with fourth-order PDEs have been overcome here thanks to the use of the surface divergence theorem.

The FE solution can now be obtained replacing the classical $\mathcal{C}^0$ FE approximations in the weak form and solving the resulting linear system of equations for the nodal values.

3. Analysis of the β parameter

A study of the value of parameter β ensuring the coercivity of the bilinear form (8a), which can be easily replicated for parameter α, is presented next. It is inspired in the analysis developed in [3] for embedded domains. We consider the problem with $\Gamma_N^2 = \partial\Omega$ (i.e., without second Dirichlet boundary conditions), and a FE space V_0^h, which discretizes the space of functions in $\mathcal{H}^2(\widehat{\Omega}) \cap \mathcal{C}^0(\Omega)$ with null value on Γ_D^1. The matrix resulting from the discretization will be positive definite if $a(v, v) > 0$ for all non-null $v \in V_0^h$.

Using Cauchy–Schwartz inequality, the bilinear form can be bounded as

$$a(v,v) = \int_{\widehat{\Omega}} \frac{\partial^2 v}{\partial x_i \partial x_j}\sigma_{ij}(v)\, d\Omega - 2\int_\Gamma \left[\!\left[\frac{\partial v}{\partial \mathbf{n}}\right]\!\right] \{r(v)\}\ dS + \beta \int_\Gamma \left[\!\left[\frac{\partial v}{\partial \mathbf{n}}\right]\!\right]^2 dS$$
$$\geq \int_{\widehat{\Omega}} \frac{\partial^2 v}{\partial x_i \partial x_j}\sigma_{ij}(v)\, d\Omega - 2\left\|\left[\!\left[\frac{\partial v}{\partial \mathbf{n}}\right]\!\right]\right\|_{L^2(\Gamma)} \|\{r(v)\}\|_{L^2(\Gamma)} + \beta\left\|\left[\!\left[\frac{\partial v}{\partial \mathbf{n}}\right]\!\right]\right\|^2_{L^2(\Gamma)}.$$

Now, let us consider a constant c (depending only on the considered FE discretization space) such that

$$\|\{r(v)\}\|^2_{L^2(\Gamma)} \leq c^2 \int_{\widehat{\Omega}} \frac{\partial^2 v}{\partial x_i \partial x_j}\sigma_{ij}(v)\, d\Omega \quad \forall v \in V_0^h. \tag{9}$$

Then, using Young's inequality $\left(ab \leq \frac{a^2}{2\varepsilon} + \frac{\varepsilon}{2}b^2\, \forall a, b, \varepsilon > 0\right)$, we have

$$a(v,v) \geq \left[1 - \frac{c^2}{\varepsilon}\right] \int_{\widehat{\Omega}} \frac{\partial^2 v}{\partial x_i \partial x_j}\sigma_{ij}(v)\, d\Omega + [\beta - \varepsilon]\left\|\left[\!\left[\frac{\partial v}{\partial \mathbf{n}}\right]\!\right]\right\|^2_{L^2(\Gamma)} \tag{10}$$

for any $\varepsilon > 0$. Thus, the matrix will be positive definite if $\beta > c^2$.

In practice, β can be taken slightly larger than the largest eigenvalue of the generalized eigenvalue problem $\mathbf{KV} = \lambda \mathbf{MV}$, being $\mathbf{M}$ and $\mathbf{K}$ the matrices corresponding to the discretization of $\int_{\widehat{\Omega}} \frac{\partial^2 v}{\partial x_i \partial x_j}\sigma_{ij}(u)\, d\Omega$ and $\int_\Gamma \{r(v)\}\{r(u)\}\ dS$, respectively. Moreover, under nested mesh refinement, with characteristic element size h, the matrices $\mathbf{M}$ and $\mathbf{K}$ scale as h^{-3} and h^{-2}, respectively, thus, the maximum eigenvalue (and parameter β) scales as h^{-1}.

4. Numerical Experiments

4.1 Convergence and sensitivity to β parameter

As commented in the introduction, [2] presents a theoretical convergence analysis of the formulation valid for smooth boundaries (without corners) or pure Dirichlet boundary conditions. The conclusion of the analysis is that the method is convergent for large enough parameter β, but too large values may lead to suboptimal convergence. To assess the applicability of the method in real computations, the accuracy and convergence of the numerical solution of (1a), with boundary conditions (1b) and (1e) in the whole boundary, is tested next with

$$\sigma_{ij}(u) = \frac{\tau^3}{12}\left(2\mu \frac{\partial^2 u}{\partial x_i \partial x_j} + \lambda \frac{\partial^2 u}{\partial x_k \partial x_k}\delta_{ij}\right),$$

and material parameters $\tau = \lambda = \mu = 1$, in a square domain $\Omega = [0,1]^2$. The body force f, the Dirichlet boundary value g_1 and the second Neumann boundary value r_n are chosen in accordance with the analytical solution $u(x,y) = x^4 y$. Note that with this boundary conditions, the method depends only on β, and not on the parameter associated to second Dirichlet boundary conditions α.

Fig. 3 shows the evolution of the L^2 error of the displacement under uniform refinement with characteristic element size h. Optimal convergence would lead to errors $||u - u^h||_{L_2} = \mathcal{O}(h^{p+1})$ for degree of

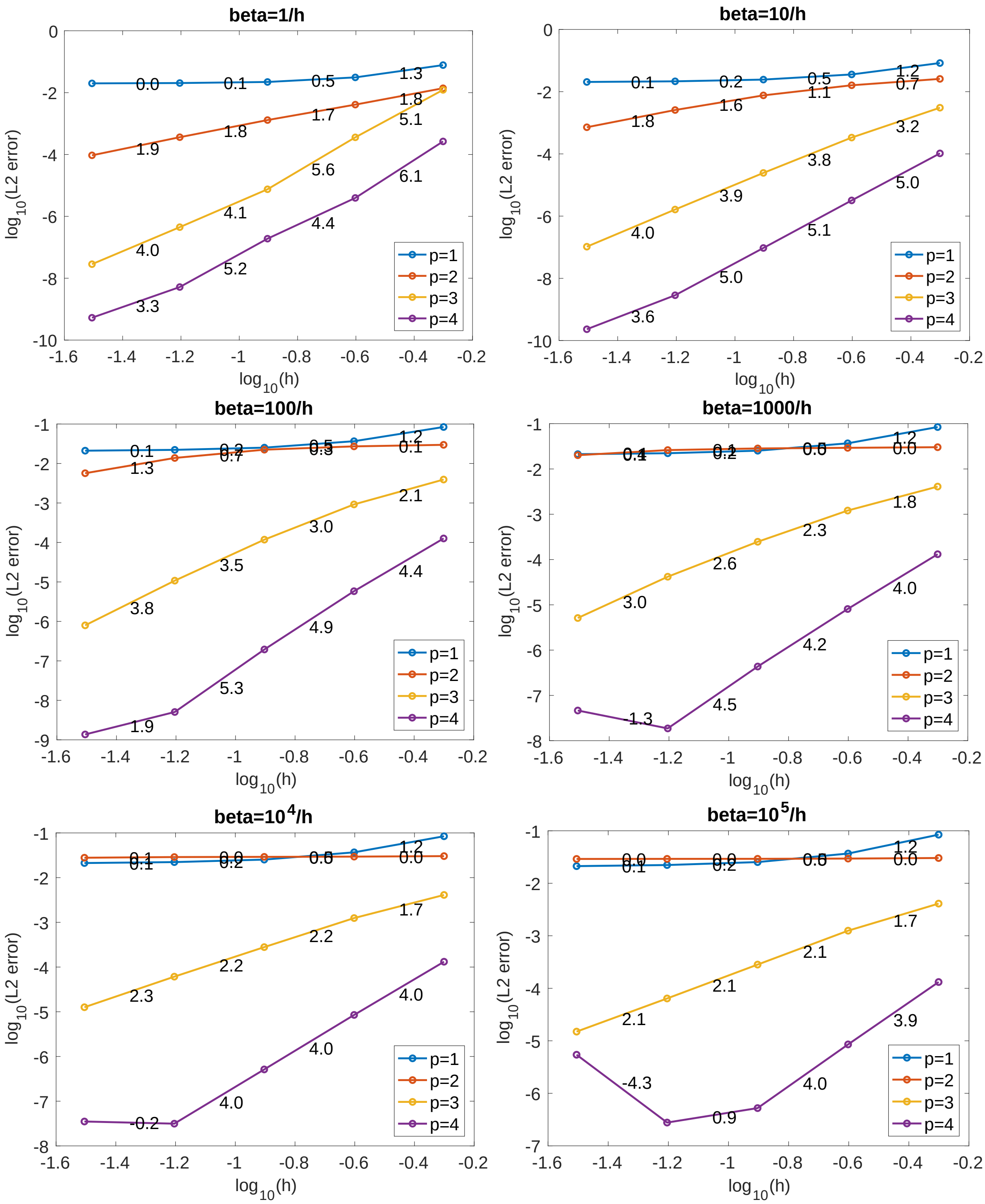

Figure 3: Convergence plots for different values of β and different degrees p. The numbers indicate the slopes of the segments.

approximation p, which would correspond to slope $p+1$ in the plots. As expected, for degree $p = 1, 2$, the approximation space is not rich enough to weakly impose continuity of the derivatives and at the same time properly approximate the solution. This kind of locking leads to poor accuracy and convergence. However, for degree $p \geq 3$, a reasonably large range for β ($(1-100)\tau^3/h$ for $p=3$) provides optimal convergence. Higher values of β lead to slightly suboptimal convergence, but still provide accurate results and good convergence. Much higher values of β are not recommended mainly because they may lead to a very ill-conditioned matrix, but also because we expect a continuous degradation in the convergence and accuracy due to the locking associated to a too strong imposition of the continuity of the derivative.

Thus, in practice, the recommendation is using a value of β slightly larger than the one corresponding to the maximum eigenvalue of the problem in Section 3. It is also worth noting that, assuming material parameters λ and μ constant or in a small range, $\mathbf{M}$ scales as $\tau^3\, h^{-2}$, $\mathbf{K}$ scales as $\tau^6\, h^{-3}$, and, therefore, the maximum eigenvalue scales as τ^3/h. Consequently, if the eigenvalue problem is solved for a particular mesh and a particular value of τ, the value for β for finer nested meshes, or other values of τ, can be estimated without solving the eigenvalue problem.

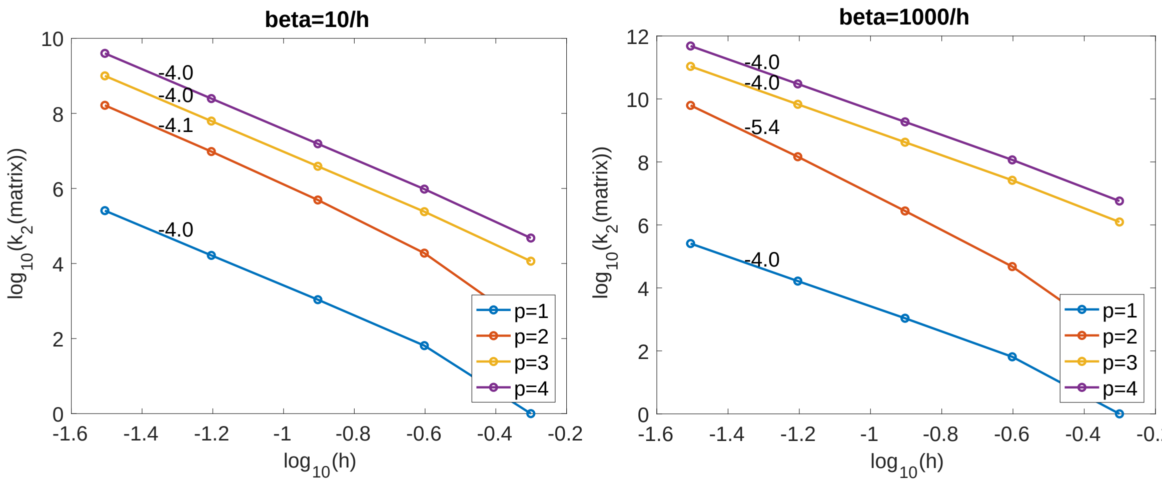

Figure 4: Condition number for degrees $p = 1, \ldots, 4$, and for two different values of β, varying the element size h.

Fig. 4 shows the condition number of the matrix corresponding to the discretization of the problem for two different values of β, both above the bound for positiveness of the matrix. As expected, the condition number increases when increasing β. In addition, we observe an increase in the condition number as $\mathcal{O}(h^{-4})$, that is the expected behaviour for the numerical solution of a fourth-order PDE, regardless of the discretization method.

4.2 Plate under uniform distributed load

A more realistic problem is solved in this section to demonstrate the applicability of the proposed method: a plate under a uniformly distributed applied load of $f = 100$Pa. In this case, the material parameters are $\mu = E/(2(1+\nu))$ and $\lambda = \nu E/(1-\nu^2)$, with Young's modulus $E = 200 \cdot 10^9$Pa, Poisson's ratio $\nu = 0.28$ and thickness of the plate $\tau = 0.001$m, corresponding to a thick steel plate taken from [5]. The problem is solved on the $p=4$ mesh depicted in Fig. 5, discretizing a plate of 1×1 meters. The considered penalty

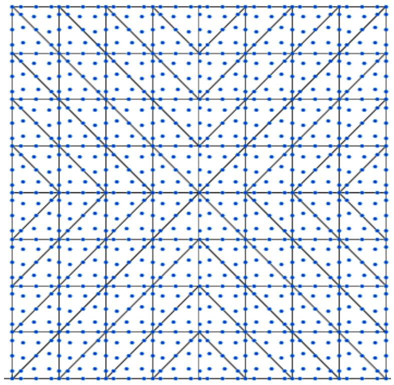

Figure 5: FE mesh. Blue dots are nodes for the $p = 4$ approximation.

parameters are $\beta = \alpha = 10\tau^3\mu/h = 781.25/h$ with element size $h = 0.125$m. Fig. 6 shows the solution for a simply supported plate (left) and a clamped plate (right). As expected, deformations are much larger for the simply supported plate.

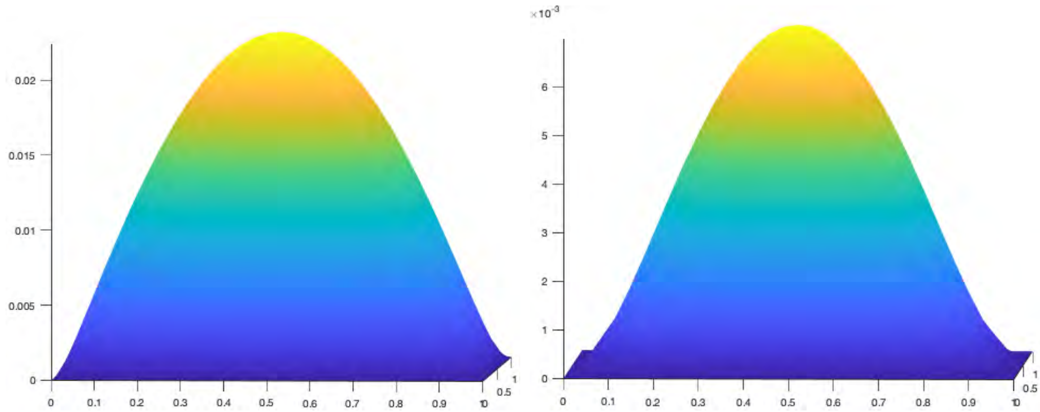

Figure 6: Solution of the problem with distributed load for a simply supported plate (left) and a clamped plate (right).

In both cases the first boundary condition is $u = 0$ on $\partial\Omega$. The second boundary condition is $r_n(u) = 0$ (Neumann) for the simply supported plate, and $\partial u/\partial n$ (Dirichlet) for the clamped plate. The null normal derivative on the boundary can be clearly observed in the right solution, corresponding to the clamped plate.

5. Conclusions and final remarks

A method for the solution of 4$^{\text{th}}$-order PDEs, with standard $\mathcal{C}^0$ FE approximations, has been proposed. The method has been developed and tested for the solution of the equations of Kirchoff plates. It is based on a formulation that weakly imposes $\mathcal{C}^1$ continuity across element sides. Numerical experiments are in agreement with the theoretical analysis in [2]: a large enough penalty parameter β is needed to ensure coercivity of the bilinear form, and convergence, but, on the other hand, too large β parameters may lead to suboptimal convergence. However, the numerical experiments show that a wide range of β parameter, within 3 orders of magnitude of difference, provides optimal convergence for degree $p \geq 3$, demonstrating the robustness of the method in practice. In fact, even for very large parameters, that should in practice not be considered to avoid ill-conditioning, the loss of optimal convergence is not catastrophic, since accurate results are still obtained.

The method is promising for the solution of other problems modelled by 4^{th}-order PDEs, such as the ones modelling strain-gradient elasticity or flexoelectricity, overcoming the inconveniences or limitations of other techniques. Differently to B-spline approximations or Hermite interpolants, the discretization with standard FE allows the use of non-cartesian meshes, fitting to the boundary of non-rectangular domains, without the need to use a technique for embedded domains in non-fitted meshes, and avoiding the typical ill-conditioning problems related to the so-called cut elements. On the other hand, standard $\mathcal{C}^0$ FEs are easy to define and implement for any degree, differently to mixed approximations whose definition is cumbersome and not developed for high-order approximations and, in addition, involve additional unknowns increasing the computational cost.

In the next future, we aim to apply the same methodology to other 4^{th}-order PDEs, and study its applicability and robustness in real applications of interest.

References

[1] D.N. Arnold, "An interior penalty finite element method with discontinuous elements", *SIAM J. Numer Anal* **4** (1982), 742–760.

[2] S.C. Brenner and L.Y Sung, "$\mathcal{C}^0$ Interior penalty methods for fourth order elliptic boundary value problems on polygonal domains", *Journal of Scientific Computing* **22–23** (2005), 83–118.

[3] D. Codony, O. Marco, S. Fernández-Méndez, and I. Arias, "An immersed boundary hierarchical B-spline method for flexoelectricity", *Computer Methods in Applied Mechanics and Engineering* **354** (2019), 750–782.

[4] F. Deng, Q. Deng, W. Yu, and S. Shen, "Mixed finite elements for flexoelectric solids", *Journal of Applied Mechanics* **84**(8) (2017).

[5] D.B. Miracle, S.L. Donaldson, D. Scott, et al., *ASM handbook* (2001) 21.

[6] J. Nitsche, "Über ein variationsprinzip zur lösung von Dirichlet-problemen bei verwendung von teilräumen, die keinen randbedingungen unterworfen sind", *Abhandlungen aus dem mathematischen Seminar der Universität Hamburg* **36** (1971), 9–15.

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

Scheme of pairs of matrices with vanishing commutator

*Bartomeu Llopis Vidal
Universitat Politècnica de Catalunya (UPC)
bartomeu.llopis.vidal@gmail.com
∗Corresponding author

Resum *(CAT)*
En aquest treball estudiarem l'esquema de parells de matrius $n \times n$ amb commutador nul, pel que es conjectura que és reduït, Cohen–Macaulay i normal. Demostrarem que és regular en codimensió 3 però no en codimensió 4. També aportem resultats similars per a altres esquemes relacionats amb el nostre esquema original. En una segona part del treball estudiem les singularitats de l'esquema de parelles de matrius que commuten a partir de l'estudi dels corresponents esquemes de jets i altres invariants de singularitats com el log-canonical threshold.

Abstract *(ENG)*
In this work we will study the scheme of $n \times n$ matrices with vanishing commutator, which is conjectured to be reduced, Cohen–Macaulay and normal. We will prove that it is regular in codimension 3 but not in codimension 4. We will also bring similar results for other schemes related to our original one. In a second part of the paper, we study the singularities of the scheme of pairs of commuting matrices from the study of the corresponding jet schemes and other singularity invariants such as the log-canonical threshold.

Keywords: *commuting variety, jet schemes, log-canonical threshold.*
MSC (2010): *Primary 13H10, 14B05, 15A27. Secondary 14-04.*
Received: *August 25, 2020.*
Accepted: *October 10, 2020.*

Acknowledgement
This work is part of my undergraduate thesis done at KULeuven under the supervision of Nero Budur and supported by an Erasmus+ KA103 grant, a MOBINT grant by AGAUR and a Mobilitat CFIS grant. I would like to thank Nero Budur, for the supervision, and Josep Àlvarez Montaner, for his immense support and help, specially in bringing this article to existence.

Societat Catalana de Matemàtiques
Institut d'Estudis Catalans

1. Introduction

The aim of this work is to study the scheme of pairs of $n \times n$ matrices over an algebraically closed field K with vanishing commutator.

Definition 1.1. Let K be an algebraically closed field. For any integer $n \geq 1$, consider the scheme associated to the following set with the natural scheme structure,

$$X_n = \{(A, B) \in \mathrm{Mat}(n, K)^{\times 2} \mid [A, B] = 0\},$$

where $[A, B] = AB - BA$, and we consider $\mathrm{Mat}(n, K)^{\times 2}$ as an affine $2n^2$-dimensional space, where A and B are generic matrices. Throughout the text, we refer to this scheme as the *commuting scheme*[1] which we will also denote as X_n. Its reduced associated scheme is usually referred to as the *commuting variety* (see [7], [10], [16]) or the *variety of commuting matrices*.

Equivalently, $X_n = \operatorname{Spec} R_n/I_n$ where $R_n = K[\{a_{i,j}, b_{i,j}\}_{1\leq i,j\leq n}]$, for the matrices $A = (a_{i,j})_{1\leq i,j\leq n}$, $B = (b_{i,j})_{1\leq i,j\leq n}$, and the ideal $I_n = (f_{i,j})_{1\leq i,j\leq n}$ is generated by

$$f_{i,j} = \begin{cases} \displaystyle\sum_{\substack{k=1\\ k\neq i}}^{n} (a_{i,k}b_{k,i} - a_{k,i}b_{i,k}) & \text{if } i = j,\\ \displaystyle\sum_{\substack{k=1\\ k\notin\{i,j\}}}^{n} (a_{i,k}b_{k,j} - a_{k,j}b_{i,k}) + a_{i,j}(b_{j,j} - b_{i,i}) - b_{i,j}(a_{j,j} - a_{i,i}) & \text{if } i \neq j. \end{cases}$$

Remark. $\{f_{i,j}\}_{i\neq j} \cup \{f_{i,i}\}_{i\neq k}$ is a generating set of I_n for any k and has a minimal number of generators.

An important property of X_n, first proven by Motzkin and Taussky [8] (as well as a bit later by Gerstenhaber [2]), is the following theorem:

Theorem 1.2. *X_n is irreducible and of dimension $n^2 + n$ for all $n \geq 1$.*

Moreover, there is a long standing conjecture atributed to M. Artin and M. Hochster[2] (cf. [6], [11], [7], [1], [12], [13]) on the properties of X_n:

Conjecture 1.3. *X_n is reduced, Cohen–Macaulay and normal for all $n \geq 1$.*

This conjecture is actually a specific case, for $\mathfrak{g} = \mathfrak{gl}_n$, of the following one:

Conjecture 1.4. *Let $\mathfrak{g}$ be a reductive Lie algebra. Then, the associated scheme to*

$$\mathcal{C}(\mathfrak{g}) = \{(a, b) \in \mathfrak{g} \mid [a, b] = 0\}$$

is reduced, irreducible, Cohen–Macaulay and normal.

[1] We use this nomenclature as a parallelism with the use of *commuting variety* for the reduced associated scheme.

[2] It is cited as being posed by M. Artin and M. Hochster in 1982 ([6], [11], [7]), but none of the references cites those two authors directly and we have not been able to find a direct source that supports it.

Even though we know of the existence of this wider conjecture, we will only focus on the specific case of X_n.

It is interesting to remember that the properties over the scheme can be checked over the associated ring and, for that, we can use Serre's conditions:

Definition 1.5. Given a Noetherian commutative ring A and an integer $k \geq 0$, A is said to fulfil Serre's condition if

(i) R_k if $A_\mathfrak{p}$ is a regular local ring for any prime ideal $\mathfrak{p} \subset A$ such that $\mathrm{height}(\mathfrak{p}) \leq k$.

(ii) S_k if $\mathrm{depth}\, A_\mathfrak{p} \geq \inf\{k, \mathrm{height}(\mathfrak{p})\}$ for any prime $\mathfrak{p}$.

Theorem 1.6 (Serre's criteria)**.** *Given a Noetherian commutative ring A, then*

(i) A is reduced i ffA satisfies R_0 and S_1;

(ii) A is normal i ffA satisfies R_1 and S_2;

(iii) A is Cohen–Macaulay iff A satisfies S_k for all $k \geq 0$;

(iv) A is regular iff A satisfies R_k for all $k \geq 0$.

Other questions that can be asked are related to the singularities of these schemes. In this sense, it is thought to have rational singularities (in characteristic 0)[3], though maybe the conjecture could be about whether they have log-canonical or log-terminal singularities, and the equivalents in characteristic $p > 0$, F-rational, F-pure or strongly F-regular. These properties, in characteristic 0, can be studied through the associated jet schemes, so we will take a look at them in the last section.

On another matter, those are not easy problems, so one ends up questioning oneself about similar schemes. In our case, we studied, among others, the pairs of matrices whose commutator's diagonal vanishes, that is, the scheme associated to:

$$X_{\mathrm{diag}} = \{(A, B) \in \mathrm{Mat}(K)^{\times 2} \mid \mathrm{diag}([A, B]) = 0\},$$

where $\mathrm{diag}(M)$ applied to a matrix M is the projection onto the diagonal elements, (i.e., $M = (m_{i,j})_{1\leq i,j\leq n} \mapsto \mathrm{diag}(M) = (m_{i,i})_{1\leq i\leq n}$).

2. Scheme of commuting matrices

In this section we present the results that we obtained on the *commuting scheme*. First of all, we point out that Conjecture 1.3 is known to be true for small n:

Proposition 2.1 (see [4], [5])**.** *X_n is reduced, irreducible and Cohen–Macaulay but not Gorenstein for $n \leq 4$.*

The proof of this result was obtained using the computational algebra system Macaulay2 ([3]). In that matter, we have redone the computations with a small improvement that might be helpful in attempting the proof for $n = 5$.

[3]The statement of rational singularities is not a published conjecture or open problem, but it would fit in the behaviour of a more general family of schemes that are closely related to it, studied in [1].

Proposition 2.2. *$\mathcal{O}_{X_n} := R_n/I_n$ is Cohen–Macaulay (respectively reduced) iff, for any $1 \leq i, j \leq n$, the quotient $\mathcal{O}_{X_n}/(a_{i,i}, b_{j,j})$ is Cohen–Macaulay (respectively reduced). Where $(a_{i,i}, b_{j,j})$ is the ideal (sheaf) generated by the (i,i)-th entry of the matrix A and the (j,j)-th entry of the matrix B.*

Furthermore, we have proven the following result:

Theorem 2.3. *X_n is regular in codimension 3 but not 4 for all $n \geq 1$. That is, it satisfies Serre's conditions R_0, R_1, R_2 and R_3 but not R_k for any $k \geq 4$.*

This result has the following implications:

Proposition 2.4. *The singular locus of X_n^{red}, the associated reduced scheme of X_n, has codimension at least 4. If X_n is reduced, then its singular locus has codimension 4.*

Proposition 2.5. *If X_n has any embedded component, it must have at most dimension $n^2 + n - 4$.*

In particular, Theorem 2.3 implies, through Serre's criteria (Theorem 1.6), the following proposition:

Proposition 2.6. *If X_n is Cohen–Macaulay, then it is reduced and normal.*

The implication of being reduced was known previously (cf. [4]), but the argumentation was different (see [15, Prob. 2.7.1]). The implication of being normal was also known as an implication of it being reduced and the following theorem:

Theorem 2.7 ([12])**.** *Given a connected non-commutative reductive lie algebra $\mathfrak{g}$ over an algebraically closed field K of characteristic 0, let $\mathcal{C}^{\text{red}}(\mathfrak{g}) = \{(a,b) \in \mathfrak{g} \mid [a,b] = 0\}$ be the reduced scheme of pairs of commuting elements. Then $\operatorname{codim}_{\mathfrak{g}\times\mathfrak{g}}(\mathcal{C}^{\text{red}}(\mathfrak{g}))^{\text{sing}} \geq 2$, where $(\mathcal{C}^{\text{red}}(\mathfrak{g}))^{\text{sing}}$ stands for the singular locus of $\mathcal{C}^{\text{red}}(\mathfrak{g})$.*

Even though Proposition 2.6 can be deduced from results that were already known, its implications to X_n for $n \leq 4$ do not seem to be recorded in the literature. In any case, we have:

Proposition 2.8. *X_n is reduced, irreducible, Cohen–Macaulay and normal, but not Gorenstein, for $n \leq 4$.*

The proof of Theorem 2.3 is too long to be included in its full extension, so we will just give the main ideas.

Sketch of Proof of Theorem 2.3. For ease of reading we have divided the proof in three parts. Throughout we will use the Jacobian smoothness criterion.

1. *R_0 and R_1 properties.*
   Let us consider B in Jordan canonical form. If we name J_k the nilpotent Jordan block of size k, then there exist $\lambda_1, \dots, \lambda_r \in K$ pairwise different elements and $a_1, \dots, a_r > 0$ integers satisfying $a_1 + \cdots + a_r = n$, such that B is a block diagonal matrix of the form $B = \operatorname{diag}(\lambda_1 I_{a_1} + J_{a_1}, \dots, \lambda_r I_{a_r} + J_{a_r}) = (b_{i,j})_{1 \leq i,j \leq n}$.
   In this case:

$$c_{i,j}^{r,s} := \frac{\partial f_{r,s}}{\partial a_{i,j}} = \begin{cases} 1 & \text{if } i = r,\ s = j+1 \leq n \text{ and } b_{j,j} = b_{j+1,j+1}, \\ -1 & \text{if } j = s,\ r = i-1 \geq 0 \text{ and } b_{i-1,i-1} = b_{i,i}, \\ b_{j,j} - b_{i,i} & \text{if } (i,j) = (r,s) \text{ and } b_{j,j} \neq b_{i,i}, \\ 0 & \text{otherwise.} \end{cases}$$

First, we will prove that $\det(c_{i,j}^{r,s})_{\substack{b_{r,r}\neq b_{s,s}\\ b_{i,i}\neq b_{j,j}}} \notin I_n$, where the columns of the matrix are indexed by the (i,j) and the rows by (r,s), both with the same ordering. We observe that the product of the diagonal elements is $\prod_{\{(i,j)|b_{i,i}\neq b_{j,j}\}} (b_{j,j} - b_{i,i}) \notin I_n$. We will prove that all the other products in the determinant vanish.

Let us pick the column (i,j) and assume that we have to pick a nonzero element outside the diagonal. If $j+1 \leq n$ and $b_{j,j} = b_{j+1,j+1}$, then $b_{i,i} \neq b_{j+1,j+1}$, so for the (i,j) column, we can get the entry of the $(i,j+1)$ row which has a value of 1. In this case, for the $(i,j+1)$ column we cannot get the diagonal element. If $i-1 \geq 0$ and $b_{i-1,i-1} = b_{i,i}$, then $b_{i-1,i-1} \neq b_{j,j}$ and for the (i,j) column we can get the entry of the $(i-1,j)$ row that has a value of -1. In this case, for the $(i-1,j)$ column we cannot get the diagonal element. Otherwise, the only nonzero element is the diagonal one.

A non-vanishing product would be equivalent to this process having a cycle, but either the i decreases or the j increases, so we can never have a cycle, and all products, apart from the diagonal one, vanish, as we wanted to show.

Now, we will reason by induction. Given (k,l) such that $b_{k,k} = b_{l,l}$, $l+1 \leq n$ and $b_{l,l} = b_{l+1,l+1}$, assume that all the columns with indexes in

$$\mathcal{S} = \{(i,j) \mid b_{i,i} \neq b_{j,j}\} \cup \{(i,j) \mid b_{i,i} = b_{j,j},\ j+1 \leq n,\ b_{j,j} = b_{j+1,j+1} \text{ and } (i,j) < (k,l)\},$$

where the ordering is the lexicographic order, are linearly independent. Then, $c_{k,l}^{k,l+1} = 1$ and for all $(i,j) \in \mathcal{S}$, $c_{i,j}^{k,l+1} = 0$, which proves that the columns with indexes in $\mathcal{S} \cup \{(k,l)\}$ are linearly independent. In this way, we have proven that the columns with indexes in

$$\mathcal{I} = \{(i,j) \mid b_{i,i} \neq b_{j,j}\} \cup \{(i,j) \mid b_{i,i} = b_{j,j},\ j+1 \leq n,\ b_{j,j} = b_{j+1,j+1}\}$$

are linearly independent.

Since the cardinality of $\mathcal{I}$ is $n^2 - n$, we get that this closed point is reduced.

Through the action of $\mathrm{GL}_n(K)$ we get that the open set that includes all closed points (A,B) where B is non-derogatory is regular.

Since the complementary of the set where A and B are non-derogatory can be checked to have codimension 2, this implies R_0 and R_1 for X_n.

2. *R_2 and R_3 properties.*

First of all, we notice:

$$X_n^{\text{red}} = Y \cup \bigcup Y^{r,s}_{\substack{(i_{1,1},\ldots,i_{1,t_1},i_{2,1},\ldots,i_{2,t_2},\ldots,i_{r,t_r})\\ (j_{1,1},\ldots j_{1,t_1'},j_{2,1},\ldots j_{2,t_2'},\ldots j_{r,t_r'})}},$$

where $Y = \{(A,B) \in X_n^{\text{red}} \mid A \text{ and } B \text{ are non-derogatory}\}$ and

$$Y^{r,s}_{\substack{(i_{1,1},\ldots,i_{1,t_1},i_{2,1},\ldots,i_{2,t_2},\ldots,i_{r,t_r})\\ (j_{1,1},\ldots j_{1,t_1'},j_{2,1},\ldots j_{2,t_2'},\ldots j_{r,t_r'})}}$$

is the set of pairs of commuting matrices (A,B) such that both are derogatory, A has r distinct generalised eigenvalues with Jordan decomposition in blocks of sizes $(i_{1,1},\ldots,i_{1,t_1},i_{2,1},\ldots,$

$i_{2,t_2}, \dots, i_{r,t_r})$, and B has s different generalised eigenvalues with Jordan decomposition in blocks of sizes $(j_{1,1}, \dots, j_{1,t'_1}, j_{2,1}, \dots, j_{2,t'_2}, \dots, j_{r,t'_r})$.

If $r \leq n-3$ or $s \leq n-3$, then

$$\dim\left(Y^{r,s}_{\substack{(i_{1,1},\dots,i_{1,t_1},i_{2,1},\dots,i_{2,t_2},\dots,i_{r,t_r})\\(j_{1,1},\dots,j_{1,t_1},j_{2,1},\dots,j_{2,t_2},\dots,j_{r,t_r})}}\right) \leq n-4,$$

for any $(i_{1,1}, \dots, i_{1,t_1}, i_{2,1}, \dots, i_{2,t_2}, \dots, i_{r,t_r})$ and any $(j_{1,1}, \dots, j_{1,t'_1}, j_{2,1}, \dots, j_{2,t'_2}, \dots, j_{r,t'_r})$, so we can ignore those sets.

Then, to prove the result, it is enough to check, for each one of the sets corresponding to $n-2 \leq r, s \leq n-1$, either that it has, at most, dimension $n-4$, or that it is composed of regular points. Checking for regularity is done by computing the rank of the Jacobian matrix.

3. *R_4 property failure.*

   Take the closed points of the form (A, B) where A and B are both diagonalisable and they both have $n-1$ distinct eigenvalues, such that, when simultaneously diagonalised, they have the form $gAg^{-1} = \operatorname{diag}(\lambda_2, \lambda_2, \lambda_3, \lambda_4, \dots, \lambda_n)$, $gBg^{-1} = \operatorname{diag}(\mu_2, \mu_2, \mu_3, \mu_4, \dots, \mu_n)$, for certain $g \in \mathrm{GL}_n(K)$ and certain $\lambda_i, \mu_j \in K$. It is immediate to check that the Jacobian matrix has rank at most $n^2 - n - 2$, so these are all non-regular points. On the other hand, the codimension is 4.

□

2.1 Related schemes

As we stated in the introduction, we have also worked with some similar schemes, which has lead to the solution of a small open problem posed by Hsu-Wen Young in his PhD dissertation [16]:

Theorem 2.9. *Given a field K, the scheme associated to $X = \{(A, B) \in \mathrm{Mat}(n, K)^{\times 2} \mid \operatorname{diag}([A, B]) = 0\}$, where $\operatorname{diag}(M)$ applied to a matrix M is the projection onto the diagonal elements (i.e., $M = (m_{i,j})_{1 \leq i,j \leq n} \mapsto \operatorname{diag}(M) = (m_{i,i})_{1 \leq i \leq n}$), is a reduced irreducible normal complete intersection scheme over K.*

Hsu-Wen Young proved it to be a reduced complete intersection for general n and checked it to be irreducible for $n \leq 3$. His motivation was mainly as a counterpart to the *diagonal commutator scheme*, which is the scheme:

$$D_n = \{(A, B) \in \mathrm{Mat}(n, F)^{\times 2} \mid [A, B] = \operatorname{diag}([A, B])\},$$

that is, the pairs of matrices whose commutator is diagonal.

The proof of Theorem 2.9 follows from an easy induction, the Jacobian smoothness criterion and the use of the following lemmas:

Lemma 2.10. *If R is a ring, and $a \in R$ is not a zero-divisor, then R is a domain (respectively reduced) if and only if R_a is a domain (respectively reduced).*

Remark. This implies that if we have an element $a \in R$ and an ideal such that $(I : (a)) = I$, I is prime (resp. radical) iff it is prime (resp. radical) in R_a (thanks to the localisation at a multiplicative set S being an exact functor from R-modules to $S^{-1}R$-modules).

Lemma 2.11. *Given a ring R, it is a domain (respectively reduced) iff the polynomial ring $R[X]$ is a domain (respectively reduced).*

3. Jet schemes

In this section we will study the jet schemes over X_n, which are known to be closely related to its singularities and which will allow us to get some results on the log-canonical threshold, another singularity invariant.

Definition 3.1. The *m-th jet scheme* associated to a scheme X over an algebraically closed field K is the set $X^{(m)}(K) = \mathrm{Hom}_K(\mathrm{Spec}(K[t]/t^{m+1}), X)$ with a natural scheme structure.

It is a well known result that the jet schemes over an affine scheme are again affine. Furthermore, there is the following result:

Theorem 3.2. *Given a field K and an affine scheme $X = \mathrm{Spec}(K[x_1, \dots, x_n]/I)$ over K, where $I = (f_1, \dots, f_r) \subset K[x_1, \dots, x_n]$ is an ideal, we have that the defining equations for the m-th jet scheme over the polynomial ring $K[\{x_{1,k}, \dots, x_{n,k}\}_{0 \le k \le m}]$ are*

$$\begin{gathered} f_1(\tilde{x}_1(t), \dots, \tilde{x}_n(t)) \cong 0 \mod t^{m+1}, \\ \vdots \\ f_r(\tilde{x}_1(t), \dots, \tilde{x}_n(t)) \cong 0 \mod t^{m+1}, \end{gathered}$$

where $\tilde{x}_i(t) = x_{i,0} + x_{i,1}t + \cdots + x_{i,m}t^m$.

Applied to our scheme, we get:

Proposition 3.3. *Over the ring $K\left[\{a_{i,j,k}, b_{i,j,k}\}_{\substack{0 \le k \le m \\ 1 \le i,j \le n}}\right]$, we define the matrices $A_k = (a_{i,j,k})_{1 \le i,j \le n}$, $B_k = (b_{i,j,k})_{1 \le i,j \le n}$. In this situation, the elements generating the ideal that defines the m-th jet scheme over X_n, which we name $X_n^{(m)}$, are the entries of the following matrices:*

$$\begin{gathered} [A_0, B_0] \\ [A_0, B_1] + [A_1, B_0] \\ [A_0, B_2] + [A_1, B_1] + [A_2, B_0] \\ \dots \\ [A_0, B_m] + [A_1, B_{m-1}] + \cdots + [A_{m-1}, B_1] + [A_m, B_0]. \end{gathered}$$

Remark. It is worth noticing that the group $\mathrm{GL}_n(K)$ acts on the scheme by simultaneous conjugation on all the matrices $X_0, \dots, X_m, Y_0, \dots, Y_m$.

The main results known about the jet schemes of X_n are:

Theorem 3.4 ([14])**.** *For $n \le 3$ and for all $m \ge 0$, the m-th jet scheme over X_n is irreducible and of dimension $(n^2 + n)(m + 1)$.*

Theorem 3.5 ([14])**.** *For all $m > 0$ exists an integer $N(m)$ such that for all $n \ge N(m)$, the m-th jet scheme over X_n is reducible.*

Even though it is not mentioned in that paper, the following proposition can be deduced from the proof of Theorem 3.5:

Proposition 3.6. *For all $m > 0$ there exists an integer $N(m)$ such that for all $n \geq N(m)$, the m-th jet scheme over X_n is not equidimensional and of dimension strictly greater than $(n^2 + n)(m + 1)$.*

Now, it is worth noticing the following result by Mustaţă:

Theorem 3.7 ([9])**.** *If X is a smooth variety over $\mathbb{C}$ and $Y \subset X$ is a closed sub-scheme, then the log-canonical threshold of the pair (X, Y) is given by*

$$\operatorname{lct}(X, Y) = \dim X - \sup_{m \geq 0} \frac{\dim Y^{(m)}}{m+1},$$

where $Y^{(m)}$ represents the m-th jet scheme over Y.

Joining these all, we obtain the following:

Proposition 3.8. *For $n \leq 3$, $\operatorname{lct}(\operatorname{Mat}(n, \mathbb{C})^{\times 2}, X_n) = n^2 - n = \operatorname{codim} X_n$.*

Proposition 3.9. *For $n \geq 30$, $\operatorname{lct}(\operatorname{Mat}(n, \mathbb{C})^{\times 2}, X_n) < n^2 - n = \operatorname{codim} X_n$.*

These results show the differences in the behaviour of the singularities of X_n depending on n.

The main results of Sethuraman and Šivic come from the existence of an irreducible open set of $X_N^{(m)}$ having dimension $(n^2+n)(m+1)$, which we denote $U_n^{(m)}$, formed by the set of closed points $(A(t), B(t)) = (A_0 + A_1 t + \cdots + A_m t^m, B_0 + B_1 t + \cdots + B_m t^m)$ where A_0 is non-derogatory, and the following lemmata:

Lemma 3.10. *Given a positive integer N, assume that $X_n^{(m)}$ is irreducible for all $n < N$. Then, for any point $(A, B) = (A(t), B(t)) \in X_N^{(m)}$ such that A_0 or B_0 have two distinct eigenvalues, we have that $(A, B) \in \overline{U}_N^{(m)}$, where $\overline{U}_N^{(m)}$ denotes the closure of $U_N^{(m)}$.*

And, if we define the corresponding open set where B_0 is non-derogatory as $U_n'^{(m)}$:

Lemma 3.11. *Let f be an automorphism of $X_n^{(m)}$ such that $f(U_n^{(m)}) = U_n^{(m)}$ or $f(U_n'^{(m)}) = U_n'^{(m)}$ or $f(U_n^{(m)} \cap U_n'^{(m)}) = U_n^{(m)} \cap U_n'^{(m)}$. Then, $(A, B) \in \overline{U}_n^{(m)}$ ifff $(A, B) \in \overline{U}_n^{(m)}$.*

Our method consists in proving that the closed subvariety where A_0 is in a specific nilpotent Jordan canonical form is irreducible. In this case, the set

$$S_{A_0} = \{(A'(t), B'(t)) \in X_n^{(m)} \mid \exists g \in \operatorname{GL}_n(F),\ \lambda \in F \text{ such that } A'_0 = gA_0g^{-1} + \lambda I\}$$

is irreducible. Finally, we have that there is a non-derogatory matrix B_0 commuting with A_0. Taking $A(t) = A_0 + 0t + \cdots + 0t^m$ and $B(t) = B_0 + 0t + \cdots + 0t^m$, we have that this pair belongs to $U_n^{(m)}$ and, therefore, $S_{A_0} \cap U_n^{(m)} \neq \varnothing$. Which, by the irreducibility of S_{A_0}, implies $S_{A_0} \subset \overline{U}_n^{(m)}$.

We also used similar methods to set bounds on the dimension of the jet schemes.

Let us define

$$Y_{(r_1,\dots,r_s)} = \{(A(t), B(t)) \in X_n^{(1)} \mid A_0 = J_{(r_1,\dots,r_s)}\},$$
$$\widetilde{Y}_{(r_1,\dots,r_s)} = \{(A(t), B(t)) \in X_n^{(1)} \mid \exists g \in \mathrm{GL}_n(F),\ \exists \lambda \in K \text{ s.t. } gA_0g^{-1} + \lambda I = J_{(r_1,\dots,r_s)}\},$$

where $J_{(r_1,\dots,r_s)}$ refers to the nilpotent matrix in Jordan canonical form with s blocks of sizes $r_1, \dots, r_s$.

The results that we obtained using the described method and basic linear algebra are the following:

Proposition 3.12. *The reduced scheme associated to*

(i) $Y_{(1,\dots,1)}$ *is irreducible for all* $n \geq 1$*;*

(ii) $Y_{(n/r,\dots,n/r)}$*, for* $r|n$*, is irreducible if and only if* $X_{n/r}^{(r-1)}$ *is irreducible;*

(iii) $Y_{(n-r,1,\overset{r}{\dots},1)}$*, for* $r \geq 0$ *is irreducible for all* $n \geq r+2$ *if and only if it is for some* $n \geq r+2$*;*

(iv) $Y_{(n-2,1,1)}$ *is irreducible for all* $n \geq 4$*;*

(v) $Y_{(n-r,1,\overset{r}{\dots},1)}$*, for* $r \geq 0$*, has the same codimension for all* $n \geq r+2$*;*

(vi) $\widetilde{Y}_{(n-r,1,\overset{r}{\dots},1)}$*, for* $r \geq 0$*, has dimension at most* $2(n^2+n)$*;*

(vii) $\widetilde{Y}_{((n-1)/2,(n-1)/2,1)}$*, for* $n = 5$*, has dimension at most* $2(n^2+n)$*.*

All these results allowed us to prove the following:

Theorem 3.13. *The first jet scheme over* X_4 *is irreducible of dimension* $2(4^2+4) = (m+1)(n^2+n)$*.*

Theorem 3.14. *The first jet scheme over* X_5 *has dimension* $2(5^2+5) = (m+1)(n^2+n)$*.*

These results on the jet schemes have implications on another open problem (see [14]) that deals with the dimension of $K[A_1, \dots, A_m]$, the algebra generated by m square $n \times n$ commuting matrices over a field K. The question is whether it is bounded by n. The answer is positive for $m = 2$ and negative for $m \geq 4$ (cf. [14]).

Specifically, Sethuraman and Šivic introduced a relation between the jet schemes over X_n with algebras generated by three commuting matrices:

Proposition 3.15 ([14])**.** *Given* K *an algebraically closed field and* $k \geq 0$ *an integer, if* J_{k+1} *is the nilpotent Jordan block of dimension* $k+1$*,* C *is a block diagonal matrix in* $\mathrm{Mat}(n(k+1), K)$ *consisting of* n *copies of* J_{k+1} *along the diagonal up to addition of scalars and* A*,* B *two matrices commuting with* C*, then if* $X_n^{(k)}$ *is irreducible* $\dim K[A, B, C] \leq n(k+1)$*.*

In particular, if we combine this proposition with the results that we obtained on the first jet scheme over X_4, we obtain the following new result:

Corollary 3.16. *Let* K *be an algebraically closed field. If* J_2 *is the nilpotent Jordan block of dimension* 2*,* C *is a block diagonal matrix in* $\mathrm{Mat}(8, K)$ *consisting of* 4 *copies of* J_2 *along the diagonal up to addition of scalars and* A*,* B *two matrices commuting with* C*, then* $\dim K[A, B, C] \leq 8$*.*

References

[1] N. Budur. Rational singularities, quiver moment maps, and representations of surface groups, *Int. Math. Res. Not. IMRN* (2019), rnz236.

[2] M. Gerstenhaber. On dominance and varieties of commuting matrices, *Ann. of Math.* **73**(2) (1961), 324–348.

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

[4] F. Hreinsdóttir. A case where choosing a product order makes the calculations of a Groebner basis much faster, *J. Symbolic Comput.* **18**(4) (1994), 373–378.

[5] F. Hreinsdóttir. Miscellaneous results and conjectures on the ring of commuting matrices, *An. Ştiinţ. Univ. "Ovidius" Constanţa Ser. Mat.* **14**(2) (2006), 45–60.

[6] Z. Kadyrsizova. Nearly commuting matrices, *J. Algebra* **497** (2018), 199–218.

[7] A. Knutson. Some schemes related to the commuting variety, *J. Algebraic Geom.* **14**(2) (2003), 283–294.

[8] T. Motzkin, O. Taussky. Pairs of matrices with property L. II, *Trans. Amer. Math. Soc.* **80** (1955), 387–401.

[9] M. Mustată. Singularities of pairs via jet scheme, *J. Amer. Math. Soc.* **15**(3) (2002), 599–615.

[10] M. Majidi-Zolbanin, B. Snapp. A note on the variety of pairs of matrices whose product is symmetric, *Contemp. Math.* **555** (2011), 146–150.

[11] N.V. Ngo. Commuting varieties of r-tuples over Lie algebras, *J. Pure Appl. Algebra* **218**(8) (2014), 1400–1417.

[12] V.L. Popov. Irregular and singular loci of commuting varieties, *Transform. Groups* **13**(3–4) (2008), 819–837.

[13] A. Premet. Nilpotent commuting varieties of reductive lie algebras, *Invent. Math.* **154**(3) (2003), 653–683.

[14] Bharath Al Sethuraman, K. Šivic. Jet schemes of the commuting matrix pairs scheme, *Proc. Amer. Math. Soc.* **137**(12) (2009), 3953–3967.

[15] W.V. Vasconcelos. *Computational methods in commutative algebra and algebraic geometry*, Algorithms and Computation in Mathematics **2**, Springer-Verlag, Berlin-Heidelberg, 1998.

[16] H.-W. Young. Components of algebraic sets of commuting and nearly commuting matrices, Ph.D. thesis, University of Michigan, Horace H. Rackham School of Graduate Studies, 2010.

REPORTS @SCM

AN ELECTRONIC JOURNAL OF THE
SOCIETAT CATALANA DE MATEMÀTIQUES

A negative result for hearing the shape of a triangle: a computer-assisted proof

*Gerard Orriols Giménez
ETH Zürich.
gerard.orriols@math.ethz.ch
∗Corresponding author

Resum *(CAT)*
En aquest article demostrem que existeixen dos triangles diferents pels quals el primer, segon i quart valor propi del Laplacià amb condicions de Dirichlet coincideixen. Això resol una conjectura proposada per Antunes i Freitas i suggerida per la seva evidència numèrica. La prova és assistida per ordinador i utilitza una nova tècnica per tractar l'espectre d'un l'operador, que consisteix a combinar un Mètode d'Elements Finits per localitzar aproximadament els primers valors propis i controlar la seva posició a l'espectre, juntament amb el Mètode de Solucions Particulars per confinar aquests valors propis a un interval molt més precís.

Abstract *(ENG)*
We prove that there exist two distinct triangles for which the first, second and fourth eigenvalues of the Laplace operator with zero Dirichlet boundary conditions coincide. This solves a conjecture raised by Antunes and Freitas and suggested by their numerical evidence. We use a novel technique for a computer-assisted proof about the spectrum of an operator, which combines a Finite Element Method, to locate roughly the first eigenvalues keeping track of their position in the spectrum, and the Method of Particular Solutions, to get a much more precise bound of these eigenvalues.

Keywords: *computer-assisted proof, Laplace eigenvalues, spectral geometry, Finite Element Method, Method of Particular Solutions.*
MSC (2010): *35P05, 35R30, 58J53.*
Received: *September 2, 2020.*
Accepted: *October 30, 2020.*

Acknowledgement
The author would like to thank *Reports@SCM* for the opportunity to publish his Bachelor's thesis as a consequence of being awarded the Noether prize from the SCM.

Societat Catalana de Matemàtiques

Institut d'Estudis Catalans

1. Introduction

The main result of the thesis is the following theorem.

Theorem 1.1. *The first, second and fourth eigenvalues of the Laplace operator on an Euclidean triangle with null Dirichlet boundary conditions are not enough to determine it up to isometry.*

This is a conjecture proposed by Antunes and Freitas in [1], suggested by numerical evidence, but a rigorous proof was required. The Dirichlet eigenvalues of the Laplace operator for a triangle Ω are real numbers λ such that there is a nonzero smooth function u defined on Ω and continuous on $\overline{\Omega}$ such that

$$\begin{cases} -\Delta u = \lambda u & \text{in } \Omega, \\ u = 0 & \text{on } \partial\Omega. \end{cases}$$

It is a well known fact that the set of such λ forms an increasing sequence $0 < \lambda_1 < \lambda_2 \leq \lambda_3 \leq \cdots$ whose only limit point is ∞, and that the corresponding eigenfunctions u_j form an orthonormal basis of $L^2(\Omega)$. The eigenvalues of a domain are closely related with its geometric properties, constituting an active area of research called *spectral geometry*. A classical example of this relationship is Weyl's law, which relates the asymptotics of the eigenvalues to the volume of the domain, and a later result by McKean and Singer states that the perimeter is also determined by the eigenvalues [20]. More results of this kind can be found in [2] and [24]. Other results about how the geometry of a domain determines its spectrum can be found in Henrot's book [14].

The question of the determination of a domain given the set of its Laplace eigenvalues was posed by Mark Kac in his famous paper "Can one hear the shape of a drum?" [16]. Since then, the answer has been found to be negative in general; in particular, for euclidean polygons, the first example of a pair of non-isometric polygons with the same spectrum is due to Gordon, Webb and Wolpert [13]. However, there are positive results when we restrict the determination to a class of domains, the most successful of which was found by Zelditch [25], who proved spectral determination for analytic domains with two classes of symmetries.

Less is known about domains with more irregular boundaries, the simplest of which are polygons. In the case of triangles, it has been proven that the whole spectrum of the Laplace operator determines the shape of a triangle ([7], with a recent simple proof by [9]), and later Chang and DeTurck proved that only a finite amount of eigenvalues, which depends on λ_1 and λ_2, is enough [5]. It is natural to try to improve the result to only a finite and fixed amount of eigenvalues, answering the question "Can a *human* hear the shape of a *triangular* drum?".

Since the space of triangles up to isometries has dimension 3, we would expect that 3 eigenvalues should be enough, and, if so, it is not clear which ones. Antunes and Freitas ([1]) conjectured that indeed the three first eigenvalues λ_1, λ_2 and λ_3 do determine the shape of a triangle. Numerical evidence by themselves seems to indicate that this is not the case for λ_1, λ_2 and λ_4, and in this paper we will prove this fact (Theorem 1.1). This will give an example of an obstruction to determining the shape of a triangle from a finite portion of its spectrum.

The proof of the theorem is computer-assisted: this means that first some topological, analytic and geometric arguments are used to reduce the proof of the theorem to a finite but large number of computations, which are then verified by a computer. The computations are carried in a rigorous way, using the technique of interval arithmetic which keeps track of propagated error bounds for all the computations.

An expanded version of this work, written jointly with Javier Gómez-Serrano and including the codes for the computer verifications, will appear elsewhere and is available as a preprint [11].

2. Structure of the proof of Theorem 1.1

By the scaling of the problem, we reduce our search to the set of triangles with a fixed base length (together with additional conditions that ensure that we only consider one triangle for each similarity class); instead of looking for all three eigenvalues λ_1, λ_2, λ_4 to be equal, we just require the quotients $\xi_{21} = \lambda_2/\lambda_1$ and $\xi_{41} = \lambda_4/\lambda_1$ to take the same value. Since the eigenvalues scale by r^{-2} when the lengths of a triangle are scaled by r, if two non-congruent triangles are found with the same quotients, there will be a scaling that makes all three eigenvalues coincide.

Fixing the first two vertices of the triangle to be $(0, 0)$ and $(1, 0)$, we use the coordinates (c_x, c_y) of the third vertex to parametrize the search space. Our approach consists in using a topological argument to show that in each of two disjoint regions in this parameter space there is a triangle in which ξ_{21} and ξ_{41} take the same prescribed value. More precisely, we claim that there are two distinct triangles for which $\xi_{21} = \bar{\xi}_{21} := 1.67675$ and $\xi_{41} = \bar{\xi}_{41} := 2.99372$.

Since rigorous calculations with the computer are done using interval arithmetic, we need a topological technique to transform the closed condition into an open condition tolerates error intervals. For that purpose we will use the Poincaré–Miranda theorem (see [19]):

Theorem 2.1. *Given two continuous functions $f, g\colon [-1, 1]^2 \to \mathbb{R}$ such that $f(x, y)$ has the same sign as x when $x = \pm 1$ and $g(x, y)$ has the same sign as y when $y = \pm 1$, there exists a point $(x, y) \in [-1, 1]^2$ such that $f(x, y) = g(x, y) = 0$.*

The regions that we will consider are two parallelograms around the points $A = (0.63500, 0.27500)$ and $B = (0.84906, 0.31995)$, designed such that ξ_{21} and ξ_{41} have approximately a constant value each in a pair of opposite edges. Using the computer we will verify that the functions $\xi_{21} - \bar{\xi}_{21}$ and $\xi_{41} - \bar{\xi}_{41}$ each have a constant and opposite sign in opposite edges of the parallelogram, and hence by the theorem, together with the well known domain continuity of eigenvalues, we will conclude that such two distinct triangles exist.

The vectors defining the parallelogram are obtained from the inverse of an approximation of the differential of the $\mathbb{R}^2$-valued funcion (ξ_{21}, ξ_{41}) at the points A and B, scaled so as to minimize the computation time. This setup is displayed in Figure 1 together with a plot of non rigorous contour lines of the eigenvalue quotients.

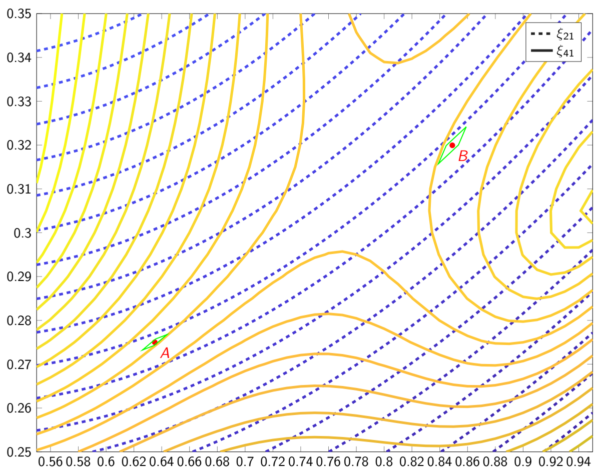

Figure 1: Numerical approximate plot of the quotients ξ_{21} (discontinuous lines) and ξ_{41} (continuous lines) around the region of interest. The validated parallelograms around A and B are shown in green.

The pointwise verification of the values ξ_{21} and ξ_{41} on the edges, which depends on an accurate calculation of λ_i for $i = 1, 2, 4$, consists of two steps. The first one, treated in Section 3, is about showing that the computed eigenvalues actually correspond to the ordered ones λ_1, λ_2 and λ_4; in order to do that, we will prove a lower bound for λ_5 combining techniques from the Finite Element Method with rigorous bounds linking the finite dimensional problem to the infinite dimensional one. The second step consists of finding accurate values of four eigenvalues that lie below the threshold obtained in the first part, which implies that they will indeed have to be the four lowest ones. This is done using the Method of Particular Solutions and recent rigorous bounds based on the L^2 norm of the boundary error of candidate approximate eigenfunctions, explained in Section 4.

We emphasize the difficulty of finding the order of an eigenvalue, which is a global problem, compared to the local easier task of refining its value. To the best knowledge of the author, this is the first computer-assisted proof in which these two distinct, local and global methods are used to verify eigenvalues of an operator.

The computer-verified enclosures of ξ_{21} and ξ_{41} explained above are only obtained for a finite set of points. In order to check the hypotheses of the Poincaré–Miranda theorem in all the edges of the parallelograms we will use an argument based on domain monotonicity for the Laplace eigenvalues to propagate the bounds to a neighborhood of the verified points. The details of this part are explained in Section 5. We now explain more about the implementation and execution of the automatic part of the proof.

2.1 Implementation of the computer-assisted proof

In the recent years, the application of calculations done by computers to mathematical proofs have become more popular due to the increment of computational resources, but in order to make sure that their results are rigorous, we need to control the errors that floating point arithmetic can accumulate. This is usually done by means of interval arithmetic, in which the data that a computer stores for a real number is an interval (two endpoints, or a midpoint and a radius) of real numbers, stored by two floating point numbers, instead of just one.

Operations between intervals are implemented to return intervals which are guaranteed to contain every possible result when the operands belong to the input intervals. For example, if $[x] = [\underline{x}, \overline{x}]$ and $[y] = [\underline{y}, \overline{y}]$ are two intervals, their sum will can be given by the interval $[x] + [y] = [\underline{x} + \underline{y}, \overline{x} + \overline{y}]$ and their product by $[x] \cdot [y] = [\min\{\underline{x}\underline{y}, \underline{x}\overline{y}, \overline{x}\underline{y}, \overline{x}\overline{y}\}, \max\{\underline{x}\underline{y}, \underline{x}\overline{y}, \overline{x}\underline{y}, \overline{x}\overline{y}\}]$. The same rule applies to function implementations: a function f evaluated on $[x]$ should return an interval containing every $f(x)$ for $x \in [x]$. We refer to the book [23] for an introduction to validated numerics, in which most of the techniques used here are explained, and to [10] for a more specific treatment of computer-assisted proofs in PDE.

The validated computations are performed using the rigorous arithmetic library Arb, developed by Fredrik Johansson [15], which can be found at `http://arblib.org`. Other non-rigorous computations are made using common libraries such as ALGLIB or Boost. The validation of one of the sides of a parallelogram can use approximately from 500 to 2000 points, and the total running time can take from 4 to 14 hours in 120 parallel machines, approximately. These benchmarks are greatly improved in [11] thanks to changes described throughout the text.

3. Separation of the first four eigenvalues

In order to find a rigorous lower bound for the fifth eigenvalue of a triangle we will use a recent bound found by Liu [17], which is similar to the one in [6] but simplifies the hypotheses and improves the constant. Both use the non-conforming Finite Element Method of Crouzeix–Raviart; other rigorous bounds with conforming finite elements were explored, like [18], but the bound is worse and the method is harder to implement with validated numerics because its mass matrix is not diagonal.

The Crouzeix–Raviart finite-element method uses a triangulation of the domain Ω, which in our case we will take to be the trivial triangulation given by N^2 triangles with sides equal to $1/N$ of the original one and similar to it. The basis functions are indexed by the interior edges of the triangulation: if E is a common edge of triangles T_1, T_2, the basis function ψ_E is the unique function supported on $T_1 \cup T_2$ such that restricted to each triangle is affine, takes the value 1 in the midpoint of E and the value 0 in the midpoints of the other edges of T_1 and T_2.

We define the coefficients of the stiffness and mass matrices $\mathbf{A} = (a_{EF})$, $\mathbf{B} = (b_{EF})$ by the bilinear forms

$$a_{EF} = \int_\Omega \nabla\psi_E \cdot \nabla\psi_F, \quad b_{EF} = \int_\Omega \psi_E\psi_F.$$

For our choice of triangulation, $\mathbf{B}$ is simply a multiple of the identity $2|\Omega|\mathbf{I}/(3N^2)$, whereas $\mathbf{A}$ is a sparse matrix. This will allow us to work with a matrix eigenvalue problem for the symmetric matrix $\mathbf{M} = \mathbf{B}^{-1}\mathbf{A}$ instead of a generalized one. The main result that we will use is the following ([17, Thm. 2.1 and Rmk. 2.2]):

Theorem 3.1. *Consider a polygonal domain Ω with a triangulation so that each triangle has diameter at most h. Let λ_k be the k-th eigenvalue of Ω and $\lambda_{k,h}$ the k-th eigenvalue of the Crouzeix–Raviart discretized problem for Ω. Then*

$$\frac{\lambda_{h,k}}{1 + C_h^2\lambda_{h,k}} \leq \lambda_k, \tag{1}$$

where $C_h \leq 0.1893h$ is a constant.

In order to be able to deal with approximate eigenvalues we will need in addition the following lemma from [21, Thm. 15.9.1].

Lemma 3.2. *Let $(\tilde{\lambda}_h, \tilde{\mathbf{u}}_h)$ be an approximate algebraic eigenpair such that $\tilde{\lambda}_h$ is closer to some λ_h than to any other discrete eigenvalue. Suppose that the coefficient vector $\tilde{\mathbf{u}}_h$ is normalised with respect to $\mathbf{B}$, $\|\mathbf{B}\tilde{\mathbf{u}}_h\|_{\mathbf{B}^{-1}} = \|\tilde{\mathbf{u}}_h\|_{\mathbf{B}} = 1$. Then the algebraic residual $\mathbf{r} := \mathbf{A}\tilde{\mathbf{u}}_h - \tilde{\lambda}_h\mathbf{B}\tilde{\mathbf{u}}_h$ satisfies*

$$|\lambda_h - \tilde{\lambda}_h| \leq \|\mathbf{r}\|_{\mathbf{B}^{-1}}.$$

Remark 3.3. We can combine Theorem 3.1 with Lemma 3.2 using the monotonicity of (1), using $\lambda_{h,k} - \|\mathbf{r}\|_{\mathbf{B}^{-1}}$ as a lower bound of $\lambda_{h,k}$ instead.

It is easy to obtain estimations $\tilde{\lambda}_h$ with a very small residual; the hardest part here before applying the theorem is to check that they have indeed the correct index, i.e., that they are closer to the appropriate λ_h than to any other discrete eigenvalue. Thus we need to control the whole spectrum of the discrete problem.

More precisely, in order to get a lower bound for λ_5 we need to separate the first 5 eigenvalues from the rest so that we can control them, and in order to do that we will perform Givens rotations until the intervals provided by Gershgorin's theorem can be separated in two disjoint components, one containing the 5 smallest eigenvalues and the other containing the rest. If this holds, then the strong version of

Gershgorin's theorem will guarantee an upper bound for $\lambda_{h,5}$. Therefore, provided that the residuals are all very small and that all approximate eigenvalues are different (which happens in our setting), Lemma 3.2 will guarantee that there are 5 distinct discrete eigenvalues below the upper bound and therefore they will be forced to have the correct indices.

This allows us to verify $\lambda_{h,5}$ with an error only depending on its residual, and using Remark 3.3, get a lower bound of λ_5. Therefore, the first four eigenvalues can be separated just by checking that they are distinct and smaller than this lower bound.

The Givens rotations must be applied in a rigorous way using interval arithmetic, although the angle of the rotation is chosen in a non rigorous way. The rotations are performed by steps: in each step, several iterations are performed to reduce the upper bound of the lowest 5 Gershgorin intervals below a fixed threshold and to increase the lower bound of the highest Gershgorin intervals above another threshold. These thresholds are improved at each step progressively (the former is reduced, the latter is increased).

The iterations consist in making a Givens rotation to set to zero each off-diagonal entry whose absolute value exceeds the maximum Gershgorin radius allowed (i.e. the difference between the diagonal value and the current threshold) divided by the number of off-diagonal entries. This heuristic tolerates small absolute values in off-diagonal entries and stops when the Gershgorin radius reached is small enough.

The execution of this algorithm for our data required a subdivision into N^2 triangles, for N between 18 and 21, and had a running time of between 15 and 45 minutes at a precision of 1024 bits. After the presentation of the thesis, with Gómez-Serrano, we simplified this part and reduced its computation time by showing that we can bound explicitly the difference between the exact diagonal form of the matrix and a nonrigorous diagonalization, provided that it is precise enough. The argument is based on stability bounds of an application of the Gram–Schmidt orthogonalization process to the proposed almost-orthogonal approximate eigenvector basis (see [11] for more details).

4. Rigorous eigenvalue bounds for individual triangles

Our approach to find tight bounds for the eigenvalues of triangles uses the Method of Particular Solutions (MPS), introduced by Fox, Henrici and Moler in [8] and more recently revived by Betcke and Trefethen [4]. In this method, a function u is written as a linear combination of functions ϕ_i ($1 \leq i \leq N$) that satisfy pointwise the equation $(\Delta + \lambda)\phi_i = 0$ for a fixed λ. The coefficients are chosen to optimize the proximity of the function to the eigenspace for the actual eigenvalue λ_j, in a sense made precise in [4], and this is measured by the least singular value of a certain matrix that involves the values of u at discrete points of the boundary $\partial\Omega$. This parameter is minimized with respect to λ by using a golden ratio search. This provides a candidate $\lambda \in \mathbb{R}$ and coefficients c_i for which $u(x) = \sum_{i=1}^{N} c_i \phi_i(x)$ can be computed with arbitrary precision.

The functions that we will use for the MPS consist of two types: the first ones are of the form $\phi(x) = Y_0(\sqrt{\lambda}|x - x_0|)$, for x_0 a point outside Ω. The second type of functions are parametrized by a vertex of the triangle and a positive integer j, and take the form $\psi_j(r,\theta) = J_{j\alpha}(\sqrt{\lambda}r)\sin(j\alpha\theta)$, where (r,θ) are the polar coordinates of the point with respect to a vertex in the triangle whose total angle is π/α, and θ is measured from an adjacent side. The first kind of functions allow us to approximate the function in the interior of the triangle and near the sides, while the second kind gives the correct asymptotic behavior of the solution near the vertices of the triangle.

We must mention that, shortly before the submission of the thesis, a remarkable paper by Gopal and Trefethen ([12]) introduced a new basis of functions that offers root-exponential convergence, meaning that with a lot fewer functions one could obtain a better fitting in much less time. In the updated version of this work [11], Gómez-Serrano and the author use this so-called *lightning Laplace solver* method to improve the total running time from around a thousand hours to just about 42.

The main tool that we will use to find rigorous bounds for eigenvalues is the L^2 bound given by Barnett and Hassell [3]. However, their method is optimized for high eigenvalues, so we will have to adapt some of the steps to our case of small eigenvalues. We summarize the main results that we will use. Let Ω be a triangle and $u \in C^2(\Omega)$ be nonzero such that $(\Delta + \lambda)u = 0$. Consider the tension

$$t[u] = \frac{\|u\|_{L^2(\partial\Omega)}}{\|u\|_{L^2(\Omega)}}.$$

Let λ_j, u_j be the sequence of eigenvalues and eigenfunctions of Ω, satisfying $(\Delta + \lambda_j)u_j = 0$ with Dirichlet null boundary conditions. Let v_j be the normal derivative of u_j, defined on $\partial\Omega$. We define the operator

$$A(\lambda) = \sum_{\lambda_j} \frac{v_j\langle v_j, \cdot\rangle}{(\lambda - \lambda_j)^2},$$

and its decomposition as a sum of three:

$$A_{\text{near}}(\lambda) = \sum_{|\lambda-\lambda_j|\leq\sqrt{\lambda}} \frac{v_j\langle v_j, \cdot\rangle}{(\lambda - \lambda_j)^2},$$

$$A_{\text{far}}(\lambda) = \sum_{\lambda/2\leq\lambda_j\leq 2\lambda,\, |\lambda-\lambda_j|>\sqrt{\lambda}} \frac{v_j\langle v_j, \cdot\rangle}{(\lambda - \lambda_j)^2},$$

$$A_{\text{tail}}(\lambda) = \sum_{\lambda_j<\lambda/2 \text{ or } \lambda_j>2\lambda} \frac{v_j\langle v_j, \cdot\rangle}{(\lambda - \lambda_j)^2},$$

where $\langle\cdot,\cdot\rangle$ is the standard inner product on $L^2(\partial\Omega)$. This operator is useful because its norm is controlled by the tension (see [3, §3]):

$$t[u]^{-2} \leq \|A(\lambda)\|. \tag{2}$$

Moreover we have the following explicit bounds from [3, Lem. 4.1] and [3, Lem. 4.2]:

$$\|A_{\text{far}}(\lambda)\| \leq C_1, \tag{3}$$

$$\|A_{\text{tail}}(\lambda)\| \leq C_2\lambda^{-1/2}, \tag{4}$$

with constants C_1, C_2 given below. For the near term, since we are working with very low eigenvalues, $\sqrt{\lambda}$ is actually small enough that only the summand with $\lambda_j = \lambda$ appears:

$$\|A_{\text{near}}(\lambda)\| = \frac{\|v_j\|^2_{L^2(\partial\Omega)}}{(\lambda - \lambda_j)^2}. \tag{5}$$

For convex domains, like in our case, Section 6 of [3] offers explicit bounds for the constants. By keeping track of all the constants used in their derivation, it is not hard to see that for a triangle with inradius ρ, one can take $C_1, C_2 < 28(1+\rho)/\rho$.

Putting (2)–(5) together we have

$$t[u]^{-2} \leq \frac{\|v_j\|^2_{L^2(\partial\Omega)}}{(\lambda-\lambda_j)^2} + 7C_\Omega(1+\lambda^{-1/2}). \tag{6}$$

Finally, recall Rellich's formula [22]:

$$\int_{\partial\Omega} (\partial_n u_j)^2 \mathbf{x}\cdot\mathbf{n}\, ds = 2\lambda_j.$$

For our choice of origin of coordinates, this just gives us $\|v_j\|^2_{L^2(\partial\Omega)} = 2\lambda_j/\rho$. Inserting this into (6), we have proved:

Proposition 4.1. *The distance d from λ to the spectrum of the Laplacian on Ω can be bounded above by*

$$t[u]^{-2} \leq \frac{2\tilde{\lambda}_j}{\rho d^2} + 7C_\Omega(1+\lambda^{-1/2}),$$

where $\tilde{\lambda}_j$ is an upper bound for λ_j.

Thus we must obtain a rigorous upper bound for the L^2 norm of the candidate eigenfunction on the boundary of Ω, and a rigorous lower bound for its interior L^2 norm, in order to use the previous proposition and deduce the existence of an actual eigenvalue near the candidate one.

4.1 Upper bound of the boundary norm

This computation is done by dividing the sides of the triangle into many small intervals, in positions given by Chebyshev nodes, and in each of them performing a validated computation using Taylor series: the Taylor polynomial of the function at the center point is evaluated in the whole interval, and to this value the remainder of Taylor's theorem is added. A validated enclosure for this remainder consists of the Taylor polynomial of the function at the whole interval evaluated in the whole interval.

When this computation exceeds a threshold in absolute value (in our case, 10^{-5}), the interval is split in half and the validating function is called recursively for the two halves. In the end, all contributions from all intervals are added up to get the L^2 bound. This calculation takes approximately 10 minutes per point at a precision of 128 bits, using a total of 317 charge basis functions and 15 vertex basis functions.

4.2 Lower bound of the interior norm

This bound is obtained by using a grid of 8×8 small triangles that occupy a smaller triangle of side 0.8 times the original one (hence the area of each triangle is $0.01|\Omega|$). Figure 2 displays the grid for the plot of the first and the fourth eigenfunction of triangle B. In each of the triangles a lower bound of the absolute value of u is obtained using the same method as above (Taylor series bounds and splitting recursively).

Whenever we obtain a validated estimate, say $u \geq a > 0$, on ∂T, where T is one of the small triangles in the grid, we can extend this inequality to the whole T by using the minimum principle. More precisely, it is enough to show that $-\Delta u \geq 0$ in T to get that the minimum of u is in ∂T and hence is at least a. If this does not hold, it means that $\lambda u = -\Delta u < 0$ at some point inside T. This means that the open set $U = \{u < 0\} \cap T \subset T$ has λ as a Dirichlet eigenvalue, and hence by the Faber–Krahn inequality, we get a lower bound for its area:

$$0.01|\Omega| = |T| \geq |U| \geq \frac{\pi j_{0,1}^2}{\lambda} > \frac{18.1684}{\lambda}.$$

This is a contradiction by orders of magnitude for our triangles ($|\Omega| \leq 0.25$ and $\lambda < 1000$).

The calculation of this part also takes approximately 10 minutes per point at a precision of 128 bits, using a total of 317 charge basis functions and 15 vertex basis functions.

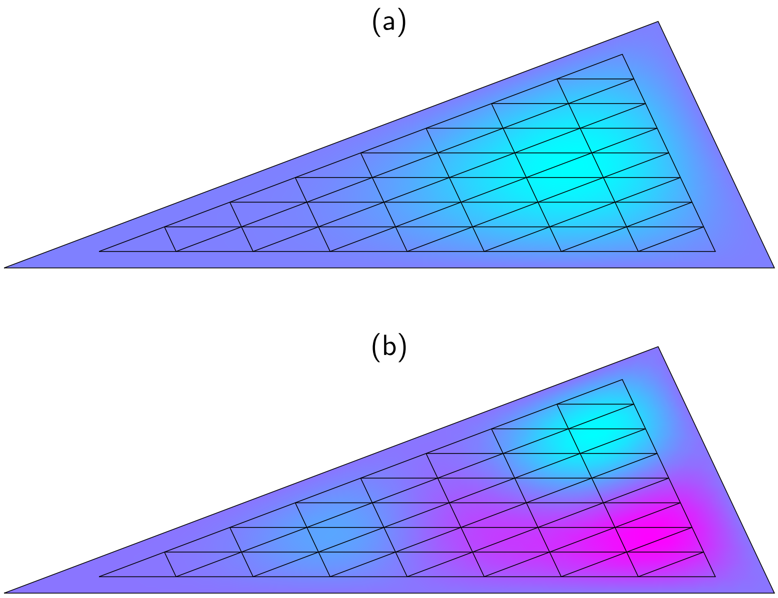

Figure 2: Grid used to validate a lower bound for $\|u\|_{L^2(\Omega)}$ for triangle B, with (a) the first eigenfunction and (b) the second eigenfunction plotted on top.

5. Extension of the bounds to a region of triangles

Our goal is to propagate the rigorous bounds of an eigenvalue of the Laplacian of a triangle with Dirichlet boundary conditions to a neighborhood of triangles. In the original version of the thesis, this was done using a continuity argument based on the fact that an operator norm bound of the difference of two compact operators (in this case, the inverse of the Laplacian and of a deformed version of it associated to a neighboring triangle) translates into a bound of the difference of all their respective eigenvalues. We later realized that obtaining an explicit bound for the difference of such operators was harder than we thought, and discovered another method which is conceptually much simpler and surprisingly propagates the bound into longer intervals. Therefore we will just sketch this simpler method, and refer to [11] for more details.

Lemma 5.1. *Let T and T' be two triangles, whose vertices are $A = (0,0)$, $B = (1,0)$, and $C = (c_x, c_y)$, $C' = (c'_x, c'_y)$ respectively ($c_y, c'_y > 0$). Consider the cross products $p = \overrightarrow{AC} \times \overrightarrow{AC'} = c_x c'_y - c_y c'_x$ and $q = \overrightarrow{BC} \times \overrightarrow{BC'} = (c_x - 1)c'_y - c_y(c'_x - 1)$. Then,*

(i) if both $p, q < 0$, there is a homothety of T' by a factor $1 - p/c'_y$ that contains T;

(ii) if both $p, q > 0$, there is a homothety of T' by a factor $1 + q/c'_y$ that contains T.

Proof. The proof is very similar in the two cases, so we will only do it for the first one. We want to find the homothety of scale $1 + r$ that keeps the vertex B of triangle T' fixed and such that the image of its opposite side contains vertex C of T. The condition becomes simpler once we apply an inverse homothety to T and T', so that it results in the points A, $C''' = (C + rB)/(1 + r)$, C' being aligned. The solution is $r = -p/c'_y$, which is positive by our condition. Moreover, triangle T lies below this homothety of T' because the vectors $\overrightarrow{BC}$ and $\overrightarrow{BC'}$ are in the correct orientation due to the condition $q < 0$. This suffices to check that T is contained in this homothety. □

Lemma 5.2. *With the same notation as in Lemma 5.1,*

(i) if $p > 0$ and $q < 0$, then $T \subset T'$;

(ii) if $p < 0$ and $q > 0$, there is a homothety of T' by a factor $c_y/c'_y > 1$ that contains T.

Proof. In the first case, the conditions on the signs of the cross products of the side vectors is equivalent to T being contained in T'. In the second case, the relative orientations of the sides guarantee that a homothetic triangle to T' of the same height as T whose top vertex coincides with C will contain T, and the ratio of this homothety is clearly c_y/c'_y. □

Using the reversed inclusions from the previous lemmas, an easy but tedious calculation which distinguishes the two cases above leads to the following result, that can be applied directly to propagate a bound on ξ_{21} or ξ_{41} to a neighborhood of a triangle.

Lemma 5.3. *Let T be a triangle as above, and consider perturbations of the third vertex of the form $C + tv$ defining triangles $T^{(t)}$, for $t \in [-\ell, \ell]$, where $v = (v_x, v_y)$. Let λ_n, $\lambda_n^{(t)}$ be the n-th Dirichlet eigenvalues of triangles T, $T^{(t)}$, respectively, and define $\xi_{n1}^{(t)}$ as the obvious eigenvalue quotient. Then we distinguish two cases depending on $p_v = \overrightarrow{AC} \times v$ and $q_v = \overrightarrow{BC} \times v$:*

(i) if p_v and q_v both have the same sign, then for all $t \in [-\ell, \ell]$

$$|\xi_{n1}^{(t)} - \xi_{n1}| \leq \xi_{n1} \left[\left(1 + \ell \frac{|p_v|}{c_y - \ell |v_y|}\right)^2 \left(1 + \ell \frac{|q_v|}{c_y - \ell |v_y|}\right)^2 - 1 \right];$$

(ii) if p_v and q_v have different signs, then for all $t \in [-\ell, \ell]$

$$|\xi_{n1}^{(t)} - \xi_{n1}| \leq \xi_{n1} \left[\left(\frac{c_y}{c_y - \ell |v_y|} \right)^2 - 1 \right].$$

Expression of gratitude

First of all, I would like to thank my supervisor Javier Gómez-Serrano for hosting me at Princeton University, suggesting me the problem, providing me references, giving me a lot of guidance and sharing with me many useful discussions. I am very grateful to the Mathematics Department of Princeton University, in particular the Graduate Program, for funding my university fee and allowing my stay as a visitor researcher. I would also like to thank CFIS and the administrators of the Mobility program for giving me the opportunity to do this research in another university and funding me. I also thank the MOBINT scholarship for providing partial financial support for my stay.

References

[1] P.R.S. Antunes, P. Freitas. On the inverse spectral problem for Euclidean triangles, *Proc. R. Soc. Lond. Ser. A Math. Phys. Eng. Sci.* **467**(2130) (2011), 1546–1562.

[2] T.P. Branson, P.B. Gilkey. The asymptotics of the Laplacian on a manifold with boundary, *Comm. Partial Differential Equations* **15**(2) (1990), 245–272.

[3] A.H. Barnett, A. Hassell. Boundary quasi-orthogonality and sharp inclusion bounds for large Dirichlet eigenvalues, *SIAM J. Numer. Anal.* **49**(3) (2011), 1046–1063.

[4] T. Betcke, L.N. Trefethen. Reviving the method of particular solutions, *SIAM Rev.* **47**(3) (2005), 469–491.

[5] P.-K. Chang, D. DeTurck. On hearing the shape of a triangle, *Proc. Amer. Math. Soc.* **105**(4) (1989), 1033–1038.

[6] C. Carstensen, J. Gedicke. Guaranteed lower bounds for eigenvalues, *Math. Comp.* **83**(290) (2014), 2605–2629.

[7] C. Durso. On the inverse spectral problem for polygonal domains, Ph.D. thesis, Massachusetts Institute of Technology, 1988.

[8] L. Fox, P. Henrici, C. Moler. Approximations and bounds for eigenvalues of elliptic operators, *SIAM J. Numer. Anal.* **4**(1) (1967), 89–102.

[9] D. Grieser, S. Maronna. Hearing the shape of a triangle, *Notices Amer. Math. Soc.* **60**(11) (2013), 1440–1447.

[10] J. Gómez-Serrano. Computer-assisted proofs in PDE: a survey, *SeMA J.* **76**(3) (2019), 459–484.

[11] J. Gómez-Serrano, G. Orriols. Any three eigenvalues do not determine a triangle, to appear in *J. Differential Equations*. Available at arXiv:1911.06758.

[12] A. Gopal, L.N. Trefethen. Solving Laplace problems with corner singularities via rational functions, *SIAM J. Numer. Anal.* **57**(5) (2019), 2074–2094.

[13] C. Gordon, D.L. Webb, S. Wolpert. One cannot hear the shape of a drum, *Bull. Amer. Math. Soc. (N.S.)* **27**(1) (1992), 134–138.

[14] A. Henrot. *Extremum problems for eigenvalues of elliptic operators*, Frontiers in Mathematics, Birkhäuser Verlag, Basel, 2006.

[15] F. Johansson. Arb: efficient arbitrary-precision midpoint-radius interval arithmetic, *IEEE Trans. Comput.* **66**(8) (2017), 1281–1292.

[16] M. Kac. Can one hear the shape of a drum? *Amer. Math. Monthly* **73**(4), part II (1966), 1–23.

[17] X. Liu. A framework of verified eigenvalue bounds for self-adjoint differential operators, *Appl. Math. Comput.* **267** (2015), 341–355.

[18] X. Liu, S. Oishi. Verified eigenvalue evaluation for the Laplacian over polygonal domains of arbitrary shape, *SIAM J. Numer. Anal.* **51**(3) (2013), 1634–1654.

[19] C. Miranda. Un'osservazione su un teorema di Brouwer, *Boll. Un. Mat. Ital. (2)* **3** (1940), 5–7.

[20] H.P. McKean, Jr., I.M. Singer. Curvature and the eigenvalues of the Laplacian, *J. Differential Geometry* **1**(1) (1967), 43–69.

[21] B.N. Parlett. *The symmetric eigenvalue problem*, Prentice-Hall Series in Computational Mathematics, Prentice-Hall, Inc., Englewood Cliffs, N.J., 1980.

[22] F. Rellich. Darstellung der Eigenwerte von$\Delta\ u+\lambda u = 0$ durch ein Randintegral, *Math. Z.* **46** (1940), 635–636.

[23] W. Tucker. *Validated numerics. A short introduction to rigorous computations*, Princeton University Press, Princeton, NJ, 2011.

[24] M. van den Berg, S. Srisatkunarajah. Heat equation for a region in $\mathbf{R}^2$ with a polygonal boundary, *J. London Math. Soc. (2)* **37**(1) (1988), 119–127.

[25] S. Zelditch. Inverse spectral problem for analytic domains. II. $\mathbb{Z}_2$-symmetric domains, *Ann. of Math. (2)* **170**(1) (2009), 205–269.

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

CM elliptic curves and the Coates–Wiles Theorem

*Martí Roset Julià
McGill University.
marti.rosetjulia@mail.mcgill.ca
*Corresponding author

Resum *(CAT)*
Descrivim un dels únics casos de la Conjectura de Birch i Swinnertor-Dyer que ha estat demostrat, l'anomenat teorema de Coates–Wiles. Sigui K un cos quadràtic imaginari amb anell d'enters $\mathcal{O}$ principal i sigui E una corba el·líptica definida sobre K amb multiplicació complexa per $\mathcal{O}$. El teorema de Coates–Wiles afirma que si la sèrie L associada a E/K no s'anul·la en 1, aleshores el conjunt de punts K-racionals de E és finit. La prova que explicarem, donada per Karl Rubin, utilitza la teoria de sistemes d'Euler.

Abstract *(ENG)*
We describe one of the few cases of the Birch and Swinnerton-Dyer Conjecture that has been already proved, the so called Coates–Wiles Theorem. Let K be an imaginary quadratic field with ring of integers $\mathcal{O}$ and class number 1 and let E be an elliptic curve defined over K with complex multiplication by $\mathcal{O}$. The Coates–Wiles Theorem states that if the L-series attached to E/K does not vanish at 1, then the set of K-rational points of E is finite. We explain a proof given by Karl Rubin, which uses the theory of Euler systems.

Keywords: *BSD Conjecture, Coates–Wiles Theorem, L-series, elliptic curves with CM, Euler systems, elliptic units.*
MSC (2010): *11G18, 14G35.*
Received: *September 4, 2020.*
Accepted: *November 7, 2020.*

Acknowledgement
I would like to express my gratitude to CFIS and to Fundació CELLEX for organizing and partially funding a mobility program to do part of this project at Princeton University. I would also like to thank the MOBINT Scholarship for partially funding this program.

Societat Catalana de Matemàtiques
Institut d'Estudis Catalans

1. Introduction

An elliptic curve E defined over a field F is a algebraic projective nonsingular curve of genus one with a distinguished F-rational point O. The Riemann–Roch Theorem shows that the set of affine F-rational points of E can be identified with the locus of solutions in $\mathbb{A}^2(F)$ of a Weierstrass equation

$$y^2 + a_1xy + a_3y = x^3 + a_2x^2 + a_4x + a_6, \tag{1}$$

with coefficients a_i in F. Then, O is the point at infinity. We will denote by $E(F)$ the set of points $P = (x, y)$ with $x, y \in F$ that satisfy (1) together with the point O.

Remarkably, $E(F)$ can be endowed with a natural group structure. It is given by the Chord-Tangent Method. Given two points $P, Q \in E(F)$, consider the point R of intersection of the line passing through P and Q with $E(F)$. Then, define $P + Q$ to be the intersection of the line through R and O with $E(F)$.

The endomorphisms of E are the morphisms $\phi\colon E \to E$ of algebraic curves that respect the group structure of E. The set $\mathrm{End}(E)$ of endomorphisms of E is a ring where the operations are addition and composition. Some examples of endomorphisms are the maps multiplication-by-m for some integer m, which are naturally defined by adding a point m times using the Chord-Tangent Method. For some curves these are all the possible endomorphisms. For others, $\mathrm{End}(E)$ can have more elements and in such a case we say that E has complex multiplication: the ring $\mathrm{End}(E)$ can be either an order in an imaginary quadratic field or a quaternion algebra, and this last option is not possible if F has characteristic 0. See [4, Chap. 2] for an outline of the main theorem of elliptic curves with complex multiplication over a field of characteristic 0.

From now on assume that F is a number field with ring of integers $\mathcal{O}_F$. It is natural to ask about the size of $E(F)$ and it turns out that we can use the group structure of $E(F)$ to say something about it. A very important example of that is the Mordell–Weil Theorem which states that $E(F)$ is a finitely generated group, i.e. $E(F) \cong \mathbb{Z}^r \oplus T$ where $r \geq 0$ is an integer and T is a finite group. We call $r = r_E$ the rank of E, a mysterious invariant that has been object of extensive study.

Based on computer calculations, a conjectural answer to find r_E was given by Birch and Swinnerton-Dyer in 1965, the so called BSD Conjecture. It connects the algebraic nature of r_E with an analytic object attached to E, the L-series. In order to define the latter suppose that every a_i lies in $\mathcal{O}_F$. Then, the L-series attached to E is defined by an infinite product over the prime ideals of $\mathcal{O}_F$

$$L(E_{/F}, s) = \prod_{\mathfrak{p}} \frac{1}{L_{\mathfrak{p}}(E_{/F}, \mathrm{N}\mathfrak{p}^{-s})},$$

where $L_{\mathfrak{p}}(E_{/F}, T)$ is a polynomial of degree ≤ 2 and it is called the local factor at $\mathfrak{p}$. To define it, consider a minimal Weierstrass equation of E (see [5, Chap. VII, §1]) and reduce it modulo $\mathfrak{p}$. It was proven by Hasse that whenever the reduced equation is an elliptic curve over the field $\mathbb{F}_{\mathrm{N}\mathfrak{p}}$, which we will denote by $\tilde{E}(\mathbb{F}_{\mathrm{N}\mathfrak{p}})$, we have $\#\tilde{E}(\mathbb{F}_{\mathrm{N}\mathfrak{p}}) = \mathrm{N}\mathfrak{p} - a_{\mathfrak{p}} + 1$, where $-2\sqrt{\mathrm{N}\mathfrak{p}} \leq a_{\mathfrak{p}} \leq 2\sqrt{\mathrm{N}\mathfrak{p}}$. In that case, we define $L_p(E_{/F}, T) = (1 - a_{\mathfrak{p}}T + \mathrm{N}\mathfrak{p}T^2)$. When the reduced curve is not smooth the definition for $L_{\mathfrak{p}}(E_{/F}, T)$ depends on the structure of the group of nonsingular points of $\tilde{E}(\mathbb{F}_{\mathrm{N}\mathfrak{p}})$ (see [5, App. C, §16]). Using the estimate of $a_{\mathfrak{p}}$ it is not hard to see that the Euler product converges on the right half plane $\{s \in \mathbb{C} : \mathrm{Re}(s) > 3/2\}$. Birch and Swinnerton-Dyer conjectured the following.

Conjecture 1.1 (BSD Conjecture)**.** *The series $L(E_{/F}, s)$ admits an analytic continuation to the entire complex plane. Moreover $r_E = \mathrm{ord}_{s=1} L(E_{/F}, s)$.*

At this point it is worth mentioning the local global principle. The definition of the L-series attached to E has information of the curve E defined over the residue fields $\mathbb{F}_{\mathrm{N}\mathfrak{p}}$, which we can call local information, and the BSD Conjecture states that it is possible to deduce results of E over the global field F from it.

For the case where $F = \mathbb{Q}$ the work of Wiles et al. on the Shimura–Tanyiama–Weil Conjecture implies that $L(E_{/\mathbb{Q}}, s)$ has analytic continuation. The analytic continuation for the particular case where E has complex multiplication is known since the work of Deuring, who gave an expression of $L(E_{/F}, s)$ in terms of the so called Hecke L-series and Hecke who proved the analytic continuation of the latter. In this project we outline the proof of the following particular case of the BSD Conjecture. Let K be an imaginary quadratic field with ring of integers $\mathcal{O}$ and class number 1.

Theorem 1.2 (Coates–Wiles)**.** *Suppose E is defined over K and it has complex multiplication by $\mathcal{O}$. If $L(E_{/K}, 1) \neq 0$, then $E(K)$ is finite.*

We will expose a proof of this theorem given by Rubin in [1]. As we said, the analytic continuation of the L-series for our particular case was already known at this time so we will focus on proving that $E(K)$ is a finite group. Our exposition is organized in the following manner.

Section 2 provides an expression of the Selmer group of certain endomorphisms which will allow us to determine when they are trivial. Section 3 covers the theory of the Euler system of elliptic units. We introduce this system and explain how it is used to bound certain ideal class groups. Section 4 explains the connection between elliptic units and the L-series of E and combines the previous work to prove the theorem. It shows that if $L(E_{/K}, 1) \neq 0$, we can produce a concrete system of elliptic units. Applying the theory of Euler systems to it we will be able to give a sharp bound of the ideal class group studied in Section 3. This is precisely one of the conditions to show that certain Selmer group is trivial and with some additional work we will be able to conclude the proof.

2. The Selmer group

Let K be an imaginary quadratic field with ring of integers $\mathcal{O}$. Assume here and from now on that K has class number 1. Let $\mathfrak{p}$ be a prime ideal in $\mathcal{O}$ above a rational prime $p > 3$ that splits in K and let $\pi \in K$ be such that $(\pi) = \mathfrak{p}$. Let E be an elliptic curve defined over K with complex multiplication by $\mathcal{O}$. Fix $\bar{K}$ an algebraic closure of K, let $\mathrm{End}(E)$ be the ring of endomorphisms of E defined over $\bar{K}$ and fix the unique isomorphism $[\cdot]\colon \mathcal{O} \xrightarrow{\sim} \mathrm{End}(E)$ such that $[\alpha]^*\omega = \alpha\omega$ for every $\alpha \in \mathcal{O}$ and ω any invariant differential of E. When it is clear from the context, we will write α for the endomorphism $[\alpha]$. The goal of this section is to define the π-Selmer group of E over K, that will be denoted by $S_\pi(E_{/K})$ and characterize when it is trivial.

We begin by recalling the definition of $S_\alpha(E_{/F})$ for a number field $F \supset K$ and $\alpha \in \mathcal{O}$ and explaining why it will be relevant to prove the Coates–Wiles Theorem. First suppose that F is any field containing K and view E as an elliptic curve defined over F. Let $\bar{F}$ be an algebraic closure of F. If $\alpha \in \mathcal{O}$, denote by $E[\alpha]$ the kernel of $[\alpha]\colon E(\bar{F}) \to E(\bar{F})$ and if L is an extension of F contained in $\bar{F}$, let $E[\alpha](L)$ be the set of points of $E[\alpha]$ defined over L. Let $G_F = \mathrm{Gal}(\bar{F}/F)$. Consider the following exact sequence of G_F-modules

$$0 \to E[\alpha] \to E(\bar{F}) \xrightarrow{\alpha} E(\bar{F}) \to 0.$$

Taking G_F-cohomology leads to a long exact sequence, where we only write the first terms

$$0 \to E[\alpha](F) \to E(F) \xrightarrow{\alpha} E(F) \xrightarrow{\delta} H^1(F, E[\alpha]) \to H^1(F, E(\bar{F})) \xrightarrow{\alpha} H^1(F, E(\bar{F})),$$

where we are considering continuous morphisms and δ is the connecting morphism

$$\delta\colon E(F) \to H^1(F, E[\alpha]), \quad P \mapsto [\sigma \mapsto Q^\sigma - Q] \text{ for some } Q \text{ satisfying } \alpha Q = P.$$

From this sequence we can obtain the following short exact sequence

$$0 \to E(F)/\alpha E(F) \xrightarrow{\delta} H^1(F, E[\alpha]) \to H^1(F, E(\bar{F}))[\alpha] \to 0$$

(note that $H^1(F, E(\bar{F}))$ is an $\mathrm{End}(E)$-module and $H^1(F, E(\bar{F}))[\alpha]$ denotes the α-torsion of it).

We will study $E(F)/\alpha E(F)$ by studying its image by δ in $H^1(F, E[\alpha])$ for F a number field containing K. As we will see, this is easier if F is a local field containing K. This motivates the following: suppose that F is a number field containing K, fix a prime $\mathfrak{Q}$ (finite or infinite) of F and regard E as defined over the completion of F at $\mathfrak{Q}$, that from now on will be denoted by $F_\mathfrak{Q}$ (we will use similar notations to denote completions). Viewing E as an elliptic curve defined over $F_{F_\mathfrak{Q}}$ and repeating the process described above we obtain the short exact sequence

$$0 \to E(F_\mathfrak{Q})/\alpha E(F_\mathfrak{Q}) \xrightarrow{\delta} H^1(F_\mathfrak{Q}, E[\alpha]) \to H^1(F_\mathfrak{Q}, E)[\alpha] \to 0. \tag{2}$$

Using that $F \subset F_\mathfrak{Q}$, and $G_F \supset G_{F_\mathfrak{Q}}$, we have the natural map $E(F)/\alpha E(F) \to E(F_\mathfrak{Q})/\alpha E(F_\mathfrak{Q})$ and the restriction maps $H^1(F, E[\alpha]) \xrightarrow{\mathrm{res}_\mathfrak{Q}} H^1(F_\mathfrak{Q}, E[\alpha])$, $H^1(F, E(F)) \xrightarrow{\mathrm{res}_\mathfrak{Q}} H^1(F_\mathfrak{Q}, E(F_\mathfrak{Q}))$. We can consider these maps for every prime $\mathfrak{Q}$ of F to obtain the following commutative diagram

$$\begin{array}{ccccccccc}
0 & \longrightarrow & E(F)/\alpha E(F) & \xrightarrow{\ \delta\ } & H^1(F, E[\alpha]) & \longrightarrow & H^1(F, E)[\alpha] & \longrightarrow & 0 \\
 & & \downarrow & & \downarrow & & \downarrow & & \\
0 & \longrightarrow & \prod_\mathfrak{Q} E(F_\mathfrak{Q})/\alpha E(F_\mathfrak{Q}) & \xrightarrow{\ \delta\ } & \prod_\mathfrak{Q} H^1(F_\mathfrak{Q}, E[\alpha]) & \longrightarrow & \prod_\mathfrak{Q} H^1(F_\mathfrak{Q}, E)[\alpha] & \longrightarrow & 0.
\end{array}$$

Instead of studying the image of $E(F)/\alpha E(F)$ by δ, we will consider a larger group that is easier to characterize.

Definition 2.1. Let F be a number field containing K and let $\alpha \in \mathcal{O}$. Define the α-Selmer group of E/F as

$$S_\alpha(E_{/F}) = \{c \in H^1(F, E[\alpha]) : \mathrm{res}_\mathfrak{Q}(c) \in \delta(E(F_\mathfrak{Q})/\alpha E(F_\mathfrak{Q})) \text{ for all } \mathfrak{Q}\}.$$

Remark 2.2. One can think of the Selmer group $S_\alpha(E_{/F})$ as the smallest group defined by natural local conditions containing $\delta(E(F)/\alpha E(F))$.

The following proposition explains the relevance of the Selmer group of an elliptic curve.

Proposition 2.3. *Let $\alpha \in \mathcal{O}$. Suppose $S_\alpha(E_{/F}) = 0$, then $E(F)$ is finite.*

Proof. By definition of $S_\alpha(E_{/F})$, we have the injection $E(F)/\alpha E(F) \hookrightarrow S_\alpha(E_{/F})$. Thus, $E(F)/\alpha E(F) = 0$. Now the result follows from the Mordell–Weil Theorem (see [5, Ch. VIII, Thm. 4.1] for the statement and proof of Mordell–Weil Theorem). □

Here and from now on let $\alpha = \pi^n$. We proceed to study $S_\alpha(E_{/F})$. The main point in the following calculations is noting that the local conditions that appear in the definition of the Selmer group behave differently depending on whether the prime ideal $\mathfrak{Q}$ of F divides α or not. We begin studying the primes $\mathfrak{Q}$ such that $\mathfrak{Q} \nmid \alpha$.

Definition 2.4. Define the enlarged Selmer group of α as

$$S'_\alpha(E_{/F}) = \{c \in H^1(F, E[\alpha]) : \mathrm{res}_{\mathfrak{Q}}(c) \in \delta(E(F_{\mathfrak{Q}})/\alpha E(F_{\mathfrak{Q}})) \text{ for all } \mathfrak{Q} \nmid \alpha\}.$$

Clearly, $S_\alpha(E_{/F}) \subset S'_\alpha(E_{/F})$.

Theorem 2.5. *Suppose E is defined over K and let $K_n = K(E[\mathfrak{p}^n])$. Then,*

$$S'_\alpha(E_{/K}) \cong \mathrm{Hom}(M_n/K_n, E[\mathfrak{p}^n])^{\mathrm{Gal}(K_n/K)},$$

where M_n is the maximal abelian extension of K_n unramified outside primes above $\mathfrak{p}$.

Proof. This is done in two steps. First we compute

$$S'_\alpha(E_{/K_n}) = \{c \in \mathrm{Hom}(G_{K_n}, E[\mathfrak{p}^n]) : \mathrm{res}_{\mathfrak{Q}}(c) \in \delta(E(K_{n,\mathfrak{Q}})/\alpha E(K_{n,\mathfrak{Q}}) \text{ for all } \mathfrak{Q} \mid \alpha\},$$

where we used that G_{K_n} fixes $E[\mathfrak{p}^n]$ and $K_{n,\mathfrak{Q}}$ denotes the completion of K_n at the prime $\mathfrak{Q}$. Since E has good reduction at $\mathfrak{Q}$ (see [1, Thm. 5.7]), the inertia subgroup $I_{\mathfrak{Q}} \subset G_{K_{n,\mathfrak{Q}}}$ acts trivially on $E[\mathfrak{p}^m]$ for every $m \geq 1$ (see [1, Coroll. 3.17]). Therefore, the connecting morphism factors trough

$$E(K_{n,\mathfrak{Q}})/\alpha E(K_{n,\mathfrak{Q}}) \to \mathrm{Hom}(G_{K_{n,\mathfrak{Q}}}/I_{\mathfrak{Q}}, E[\mathfrak{p}^n]). \tag{3}$$

By (2) this map is injective and it can be seen that it is an isomorphism by showing that both groups are isomorphic to $\mathcal{O}/\mathfrak{p}^n$, see [1, Lem. 6.4]. From there it follows that $S'_\alpha(E_{/K_n}) \cong \mathrm{Hom}(M_n/K_n, E[\mathfrak{p}^n])$ by class field theory. The second step of the proof consists on applying [1, Lem. 6.2] to see that the inflation restriction exact sequence induces the isomorphism $S'_\alpha(E_{/K}) \simeq S'_\alpha(E_{/K_n})^{\mathrm{Gal}(K_n/K)}$. □

We are left with studying the local condition at $\mathfrak{p}$. Since $\mathfrak{p}$ is coprime to $\mathfrak{f}$, E has good reduction at $\mathfrak{p}$. Since $\mathrm{ord}_{\mathfrak{p}}(p) \leq 2 < p-1$, the logarithm induces an isomorphism $\log_E \colon E_1(K_{\mathfrak{p}}) \xrightarrow{\sim} \mathfrak{p}\mathcal{O}_{\mathfrak{p}}$, where $E_1(K_{\mathfrak{p}})$ is the set of points of $E(K_{\mathfrak{p}})$ that reduce to 0 modulo $\mathfrak{p}$. Moreover, since the reduction of E at $\mathfrak{p}$ has no $\mathfrak{p}$-torsion, $E(K_{\mathfrak{p}}) = E_1(K_{\mathfrak{p}}) \times \tilde{E}(k)$ and $\log_E$ can be extended to a map $\log_E \colon E(K_{\mathfrak{p}}) \to \mathcal{O}_{\mathfrak{p}}$ (see [1, Lem. 6.6]). By [1, Coroll. 5.20 (iv)], K_n/K is totally ramified at $\mathfrak{p}$. For every $n \geq 1$, denote $K_{n,\mathfrak{p}}$ the completion of K at the unique prime above $\mathfrak{p}$.

Definition 2.6. Define the following Kummer pairing

$$\langle\,,\,\rangle_{\pi^n} \colon E(K_{\mathfrak{p}}) \times K_{n,\mathfrak{p}}^\times \to E[\mathfrak{p}^n], \quad P, x \mapsto \langle P, x\rangle_{\pi^n} = Q^{[x, K_{n,\mathfrak{p}}]} - Q,$$

where $Q \in E(\bar{K}_{\mathfrak{p}})$ is such that $\pi^n Q = P$ and $[\cdot, K_{n,\mathfrak{p}}]$ is the local Artin map.

Definition 2.7. For every $n \geq 1$, define $\delta_n \colon K_{n,\mathfrak{p}}^\times \to E[\mathfrak{p}^n]$ by $\delta_n(x) = \langle R, x\rangle_{\pi^n}$.

Lemma 2.8 ([1, Lem. 6.8])**.** *For every n, the map δ_n is characterized by the fact that if $P \in E(K_{\mathfrak{p}})$ and $x \in K_{n,\mathfrak{p}}^\times$, we have $\langle P, x\rangle_{\pi^n} = (\pi^{-1}\log_E(P))\delta_n(x)$. Moreover, if $\mathcal{O}_{n,\mathfrak{p}}$ is the ring of integers of $K_{n,\mathfrak{p}}$, we have $\delta_n(\mathcal{O}_{n,\mathfrak{p}}^\times) = E[\mathfrak{p}^n]$.* □

The previous lemma shows that δ_n is essentially the connecting morphism given in (2). Combining this lemma with Theorem 2.5 and class field theory yields the following description of $S_{\pi^n}(E_{/K})$. For every number field F let $\mathbb{A}_F^\times$ denote the idele group of F.

Theorem 2.9 ([1, Thm. 6.9])**.** *Let $K_n = K(E[\mathfrak{p}^n])$ with idele group $\mathbb{A}_{K_n}^\times$. Define*

$$W_n = K_n^\times \prod_{v\mid\infty} K_{n,v}^\times \prod_{v\nmid\mathfrak{p}\infty} \mathcal{O}_{n,v}^\times \cdot \ker \delta_n.$$

Then, $S_{\pi^n}(E_{/K}) \cong \mathrm{Hom}(\mathbb{A}_{K_n}^\times / W_n, E[\mathfrak{p}^n])^{\mathrm{Gal}(K_n/K)}$. □

Let $\Delta = \mathrm{Gal}(K(E[\mathfrak{p}])/K)$. Then, Δ acts naturally on the $\mathcal{O}/\mathfrak{p}$-vector space $E[\mathfrak{p}]$. Let $\chi_E \colon \Delta \to \mathbb{F}_p^\times$ be the character of this representation. Let A be the p-part of the ideal class group of K_1. Note thatΔ acts on A in a natural way. For every character $\chi\colon \Delta \to \mathbb{F}_p^\times$ consider the composition, also denoted by χ, $\chi\colon \Delta \to \mathbb{F}_p \hookrightarrow \mathbb{Z}_p^\times$, where the last morphism is given by Hensel's Lemma. For a given $\mathbb{Z}[\Delta]$-module M, let $M^{(p)} = M \otimes_{\mathbb{Z}} \mathbb{Z}_p$, which is a $\mathbb{Z}_p[\Delta]$-module and let M^χ be the χ-isotypical component of $M^{(p)}$. Another application of class field theory gives the following result.

Corollary 2.10. *Consider the same notation as above and suppose that p splits in K. Then, $S_\pi(E_{/K}) = 0$ if and only if $A^{\chi_E} = 0$ and $\delta_1(\mathcal{O}_{K_1}^\times) \neq 0$.*

This characterizes when $S_\pi(E_{/K}) = 0$ which is the key point to prove the Coates–Wiles Theorem since, as we explained, $S_\pi(E_{/K}) = 0$ implies that $E(K)$ is finite.

3. The Euler system of elliptic units

Let E be an elliptic curve defined over K with complex multiplication by $\mathcal{O}$. Let ψ be the Hecke character attached to E with conductor $\mathfrak{f}$ (see [4, Chap. 2, §9]), viewed as a character on ideals. Choose a prime $\mathfrak{p}$ of K not dividing $6\mathfrak{f}$, let p be the rational prime below it and suppose that p splits in K. Fix an ideal $\mathfrak{a}$ of $\mathcal{O}$ coprime to $6\mathfrak{p}\mathfrak{f}$. Let $\mathcal{R}$ be the set of square free ideals of $\mathcal{O}$ coprime to $6\mathfrak{f}\mathfrak{p}\mathfrak{a}$. Finally, for $n \geq 0$ denote by $K_n = K(E[\mathfrak{p}^n])$, if $\mathfrak{r} \in \mathcal{R}$ denote by $K_n(\mathfrak{r}) = K(E[\mathfrak{p}^n\mathfrak{r}])$ and let $G_\mathfrak{r} = \mathrm{Gal}(K_n(\mathfrak{r})/K_n)$. In this section we introduce the Euler system of elliptic units and we explain how it can be used to bound the size of A^{χ_E} defined above. We will work with the following definition of Euler system.

Definition 3.1. An Euler system is a set of global units $\{\eta(n, \mathfrak{r}) \in K_n(\mathfrak{r})^\times \mid n \geq 1 \text{ and } \mathfrak{r} \in \mathcal{R}\}$ satisfying:

(i) if $\mathfrak{r}\mathfrak{q} \in \mathcal{R}$, where $\mathfrak{q}$ is a prime ideal of $\mathcal{O}$, $\mathrm{N}_{K_n(\mathfrak{r}\mathfrak{q})/K_n(\mathfrak{r})}\eta(n, \mathfrak{r}\mathfrak{q}) = \eta(n, \mathfrak{r})^{(1-\mathrm{Frob}_\mathfrak{q}^{-1})}$, and

(ii) if $\mathfrak{r} \in \mathcal{R}$ and $n \geq 1$, $\mathrm{N}_{K_{n+1}(\mathfrak{r})/K_n(\mathfrak{r})}\eta(n+1, \mathfrak{r}) = \eta(n, \mathfrak{r})$.

We now construct the so called Euler system of elliptic units. For that we need to introduce the following rational functions. Fix here and from now on an analytic isomorphism $\xi\colon \mathbb{C}/L \xrightarrow{\sim} E(\mathbb{C})$ where $L = \Omega\mathcal{O}$ and$\Omega \in \mathbb{C}$.

Definition 3.2. Choose a Weierstrass equation for E and denote by $\Delta(E)$ its discriminant. Let $\gamma \in \mathcal{O}$ be a generator of the ideal $\mathfrak{a}$. Define

$$\Theta_{E,\mathfrak{a}} = \gamma^{-12}\Delta(E)^{\mathrm{N}\mathfrak{a}-1} \prod_{P\in E[\mathfrak{a}]-O} (x - x(P))^{-6}.$$

Suppose that E is defined over K. Let $S \in E$ be an $\mathcal{O}$-generator of $E[\mathfrak{f}]$. Define

$$\Lambda_{E,\mathfrak{a}} = \prod_{\sigma\in \mathrm{Gal}(K(\mathfrak{f})/K)} \Theta_{E,\mathfrak{a}} \circ \tau_{S^\sigma},$$

where $\tau_{S^\sigma}(P) = P+S^\sigma$ for every $P \in E$ and $K(\mathfrak{f})$ is the ray class field of K modulo $\mathfrak{f}$. Define $\Theta_{L,\mathfrak{a}} = \Theta_{E,\mathfrak{a}}\circ\xi$ and$\Lambda_{L,\mathfrak{a}} = \Lambda_{E,\mathfrak{a}} \circ \xi$.

The system of elliptic units is obtained by evaluating$\Lambda_{L,\mathfrak{a}}$ at certain torsion points of E in the following way.

Definition 3.3. Given $n \geq 0$ and an integral ideal $\mathfrak{r} \in \mathcal{R}$ define $\eta_n^{(\mathfrak{a})}(\mathfrak{r}) = \Lambda_{E,\mathfrak{a}}(\xi(\psi(\mathfrak{p}^n\mathfrak{r})^{-1}\Omega))$. The set $\{\eta_n^{(\mathfrak{a})}(\mathfrak{r})\}$ for $n \geq 1$ and $\mathfrak{r} \in \mathcal{R}$ is the set of elliptic units.

Proposition 3.4 ([1, Prop. 8.2])**.** *The set* $\{\eta_n^{(\mathfrak{a})}(\mathfrak{r})\}$ *for* $n \geq 1$ *and* $\mathfrak{r} \in \mathcal{R}$ *is an Euler system.* □

Here and for the rest of this section we write $\eta(n,\mathfrak{r}) := \eta_n^{(\mathfrak{a})}(\mathfrak{r})$. Fix M a power of p and $n \geq 1$. We now explain how to construct a principal ideal of K_n starting from the unit $\eta(n,\mathfrak{r}) \in K_n(\mathfrak{r})$. This construction will be done only for $\mathfrak{r}$ in the following subgroup of $\mathcal{R}$.

Definition 3.5. Define $\mathcal{R}_{n,M}$ to be the subset of $\mathcal{R}$ with elements $\mathfrak{r} \in \mathcal{R}$ such that every prime $\mathfrak{q} \mid \mathfrak{r}$ satisfies:

(i) $\mathfrak{q}$ splits completely in K_n/K, and

(ii) $M \mid (\mathrm{N}\mathfrak{q} - 1)$.

In order to do the construction we will use Kolyvagin's derivative operator. For every $\mathfrak{q} \in \mathcal{R}$ prime ideal, fix $\sigma_\mathfrak{q} \in G_\mathfrak{q}$ a generator of the cyclic group $G_\mathfrak{q}$.

Definition 3.6. If $\mathfrak{q} \in \mathcal{R}$ prime, define $D_\mathfrak{q} = \sum_{i=1}^{\mathrm{N}\mathfrak{q}-2} i\sigma_\mathfrak{q}^i \in \mathbb{Z}[G_\mathfrak{q}]$. For an arbitrary ideal $\mathfrak{r} \in \mathcal{R}$, define $D_\mathfrak{r} := \prod_{\mathfrak{q}|\mathfrak{r}} D_\mathfrak{q} \in \mathbb{Z}[G_\mathfrak{r}]$.

Proposition 3.7. *Let* $n \geq 1$, $\mathfrak{r} \in \mathcal{R}_{n,M}$ *and* $\sigma \in G_\mathfrak{r}$. *Then,* $\eta(n,\mathfrak{r})^{(\sigma-1)D_\mathfrak{r}} \in (K_n(\mathfrak{r})^\times)^M$. *Moreover, there is a natural choice of Mth root of unity, that we will denote by* $(\eta(n,\mathfrak{r})^{(\sigma-1)D_\mathfrak{r}})^{1/M}$.

Proof. See [1, Prop. 8.4]. Note that both K_n and $K_n(\mathfrak{r})$ may contain Mth roots of unity. This is the reason why we need [1, Prop. 8.4 (i)] to specify a choice of an Mth root of $\eta(n,\mathfrak{r})^{(\sigma-1)D_\mathfrak{r}}$. For that, the so called universal Euler system is used (see [2, Chap. IV, §2] for more details). □

Definition 3.8. Let $n \geq 1$, $\mathfrak{r} \in \mathcal{R}_{n,M}$. Define the 1-cocycle $c \in H^1(G_\mathfrak{r}, K_n(\mathfrak{r})^\times)$ as

$$G_\mathfrak{r} \to K_n(\mathfrak{r})^\times, \quad c(\sigma) = (\eta(n,\mathfrak{r})^{(\sigma-1)D_\mathfrak{r}})^{1/M}.$$

By Hilbert's Theorem 90 we have that $H^1(G_{\mathfrak{r}}, K_n(\mathfrak{r})) = 0$, hence there exists $\beta \in K_n(\mathfrak{r})^\times$ such that $c(\sigma) = \beta^{\sigma-1}$. Raising this equality to the Mth power yields

$$z = \frac{\eta(x_{n,\mathfrak{r}})^{D_{\mathfrak{r}}}}{\beta^M} \in K_n^\times.$$

The element β is well defined up to multiplication by an element of K_n. Hence, z is well defined in $K_n^\times/(K_n^\times)^M$.

Definition 3.9. With the same notation used in the previous definition, define

$$\kappa_{n,M}(\mathfrak{r}) = \frac{\eta(x_{n,\mathfrak{r}})^{D_{\mathfrak{r}}}}{\beta^M} \in K_n^\times/(K_n^\times)^M.$$

Fix $n \geq 1$. In order to simplify the notation denote $F = K_n$ and let $\mathcal{O}_F$ be its ring of integers. We proceed to write the factorization of the ideal generated by $\kappa_{n,M}(\mathfrak{r}) \in F$ modulo Mth powers in terms of $\kappa_{n,M}(\mathfrak{s})$ for ideals $\mathfrak{s} \mid \mathfrak{r}$.

Definition 3.10. Denote the group of ideals of F additively as $\mathcal{I} = \bigoplus_{\mathfrak{Q}} \mathbb{Z}\mathfrak{Q}$, where the sum is over all prime ideals $\mathfrak{Q}$ of F. If $\mathfrak{q}$ is a prime ideal of K, we define $\mathcal{I}_{\mathfrak{q}} = \bigoplus_{\mathfrak{Q}|\mathfrak{q}} \mathbb{Z}\mathfrak{Q}$. For a given $y \in F$, denote by (y) the principal ideal generated by y, $(y)_{\mathfrak{q}}$ its projection to $\mathcal{I}_{\mathfrak{q}}$, $[y] \in \mathcal{I}/M\mathcal{I}$ the reduction modulo M and $[y]_{\mathfrak{q}}$ the respective projection.

Fix $\mathfrak{q} \in \mathcal{R}_{n,M}$ a prime of K. We will construct a function, $\phi_{\mathfrak{q}}$, that will allow us to relate $[\kappa_{n,M}(\mathfrak{r})]_{\mathfrak{q}}$ with the element $\kappa_{n,M}(\mathfrak{r}\mathfrak{q}^{-1})$. We start by defining a map

$$\phi'_{\mathfrak{q}} \colon (\mathcal{O}_F/\mathfrak{q}\mathcal{O}_F)^\times \to \mathcal{I}_{\mathfrak{q}}/M\mathcal{I}_{\mathfrak{q}}$$

that after a small modification will become the desired map. Note that $\mathfrak{q}$ splits completely in F. Therefore, we have

$$(\mathcal{O}_F/\mathfrak{q}\mathcal{O}_F)^\times \cong \prod_{\mathfrak{Q}|\mathfrak{q}} (\mathcal{O}_F/\mathfrak{Q})^\times,$$

where each of the terms in the right hand side is a cyclic group of order $\mathrm{N}\mathfrak{q} - 1$. On the other hand

$$\mathcal{I}_{\mathfrak{q}}/M\mathcal{I}_{\mathfrak{q}} \cong \oplus_{\mathfrak{Q}|\mathfrak{q}} (\mathbb{Z}/M\mathbb{Z}).$$

Since $M \mid (\mathrm{N}\mathfrak{q} - 1)$, in order to define a map $(\mathcal{O}_F/\mathfrak{q}\mathcal{O}_F)^\times \to \mathcal{I}_{\mathfrak{q}}/M\mathcal{I}_{\mathfrak{q}}$ it is enough to choose a generator of the cyclic group $(\mathcal{O}_F/\mathfrak{Q})^\times$ for every $\mathfrak{Q} \mid \mathfrak{q}$ and map it to $1 \in \mathbb{Z}/M\mathbb{Z}$. Now we explain how we choose these generators. For $\mathfrak{Q}$ dividing $\mathfrak{q}$ choose a prime $\mathfrak{Q}'$ of $F(\mathfrak{q})$ above it and consider $\pi_{\mathfrak{Q}}$ a local parameter at the prime $\mathfrak{Q}'$. Since the local field extension $F(\mathfrak{q})_{\mathfrak{Q}'}/F_{\mathfrak{Q}}$ is totally tamely ramified we have that the map

$$\mathrm{Gal}(F(\mathfrak{q})/F) \to \mathcal{O}_{F(\mathfrak{q}),\mathfrak{Q}'}{}^\times \to (\mathcal{O}_F/\mathfrak{Q})^\times, \quad \sigma \mapsto \pi_{\mathfrak{Q}}^{(1-\sigma)} \mapsto [\pi_{\mathfrak{Q}}^{(1-\sigma)}] \tag{4}$$

is a group isomorphism ([3, Chap. IV, Prop. 5]).

Definition 3.11. For $\mathfrak{Q}$ as above, define $\gamma_{\mathfrak{Q}} \in (\mathcal{O}_F/\mathfrak{Q})^\times$ to be the image of the fixed generator $\sigma_{\mathfrak{q}} \in G_{\mathfrak{q}}$ by the map in (4). It is a generator of $(\mathcal{O}_F/\mathfrak{Q})^\times$.

Definition 3.12. Define a map $\phi'_{\mathfrak{q}} \colon (\mathcal{O}_F/\mathfrak{q})^\times \to \mathcal{I}_{\mathfrak{q}}/M\mathcal{I}_{\mathfrak{q}}$ as follows. Given $\alpha \in (\mathcal{O}_F/\mathfrak{q})^\times$ and $\mathfrak{Q} \mid \mathfrak{q}$, let $a_{\mathfrak{Q}}(\alpha) \in \mathbb{Z}$ be such that $\alpha \equiv \gamma_{\mathfrak{Q}}^{a_{\mathfrak{Q}}(\alpha)} \mod \mathfrak{Q}$. Then define

$$\phi'_{\mathfrak{q}}(\alpha) = \sum_{\mathfrak{Q}|\mathfrak{q}} (a_{\mathfrak{Q}}(\alpha) \mod M)\mathfrak{Q}.$$

Finally, note that $\phi_{\mathfrak{q}}$ factors trough $(\mathcal{O}_F/\mathfrak{q})^\times/((\mathcal{O}_F/\mathfrak{q})^\times)^M$. Define $\phi_{\mathfrak{q}} = \phi'_q \circ j_{\mathfrak{q}}$, where $j_{\mathfrak{q}}$ is the natural map $\{\kappa \in F^\times/(F^\times)^M : [\kappa]_{\mathfrak{q}} = 0\} \to (\mathcal{O}_F/\mathfrak{q})^\times/((\mathcal{O}_F/\mathfrak{q})^\times)^M$. It is plain to see that $\phi_{\mathfrak{q}}$ is an isomorphism.

Theorem 3.13 (Factorization Theorem; [1, Prop. 8.10])**.** *Consider $\kappa_{n,M}(\mathfrak{r})$ and $\mathfrak{q}$ a prime ideal of K. Then,*

(i) if $\mathfrak{q} \nmid \mathfrak{r}$, $[\kappa_{n,M}(\mathfrak{r})]_{\mathfrak{q}} = 0$, and

(ii) if $\mathfrak{q} \mid \mathfrak{r}$: $[\kappa_{n,M}(\mathfrak{r})]_{\mathfrak{q}} = \phi_{\mathfrak{q}}(\kappa_{n,M}(\mathfrak{r}\mathfrak{q}^{-1}))$. □

Here and from now on suppose that $F = K_1$, i.e. $n = 1$. Note that $\eta(1, \mathcal{O}) \in \mathcal{O}_F^\times$ and denote by μ_F the subgroup of roots of unity of $\mathcal{O}_F^\times$. Let $\mathcal{C}$ be the $\mathbb{Z}[\Delta]$-submodule of $\mathcal{O}_F^\times$ generated by $\eta(1, \mathcal{O})$ and μ_F. The Factorization Theorem gives the factorization of principal ideals of the form $(\kappa_{1,M}(\mathfrak{r}))$ modulo M-th powers. If M is large enough, these factorizations give relations between the classes of the prime ideals generating A. This allows to give the following bound of the χ-isotypical component of A for every irreducible representation χ of Δ.

Theorem 3.14 ([1, Thm. 9.5])**.** *For every irreducible $\mathbb{Z}_p$-representation of Δ we have $\#A^\chi \leq \#(\mathcal{O}_F^\times/\mathcal{C})^\chi$.* □

Corollary 3.15. *Consider the same notation as above. Suppose that $\eta(1, \mathcal{O})^\chi \notin \mu_F^\chi((\mathcal{O}_F^\times)^\chi)^p$. Then, $A^\chi = 0$.*

4. Complex L-function of E and proof of Coates–Wiles Theorem

Let E be an elliptic curve defined over K with complex multiplication by $\mathcal{O}$. Let ψ be the Hecke character attached to E, viewed as a character on ideals, with conductor $\mathfrak{f}$ and denote by $\bar{\psi}$ its conjugate. Let $L(E_{/K}, s)$ be the complex L-function attached to E viewed as an elliptic curve over K. For a given ideal $\mathfrak{m}$ such that $\mathfrak{f} \mid \mathfrak{m}$ and $k \geq 1$ define $L_{\mathfrak{m}}(\psi^k, s) = \sum \psi^k(\mathfrak{b})/\mathrm{N}\mathfrak{b}^s$, where the sum is restricted to the ideals $\mathfrak{b}$ coprime to $\mathfrak{m}$. We similarly define $L_{\mathfrak{m}}(\bar{\psi}^k, s)$. The following theorem is due to Deuring.

Theorem 4.1 (Deuring; [4, Thm. 10.5 (a)])**.** *We have $L(E_{/K}, s) = L_{\mathfrak{f}}(\psi, s)L_{\mathfrak{f}}(\bar{\psi}, s)$.* □

Now we proceed to relate elliptic units with $L_{\mathfrak{f}}(\bar{\psi}^k, s)$ for $k \geq 1$.

Theorem 4.2. *For every $k \geq 1$,*

$$\frac{d^k}{dz^k} \log \Lambda_{L,\mathfrak{a}}(z)|_{z=0} = 12(-1)^k(k-1)!f^k(\mathrm{N}\mathfrak{a} - \psi(\mathfrak{a})^k)\Omega^{-k}L_{\mathfrak{f}}(\bar{\psi}^k, k).$$

Proof. This proof is done in several steps. First it is possible to relate the kth derivative of $\log\Theta_{L,\mathfrak{a}}(z)$ with respect to z with the Eisenstein series $E_k(z, L) = \lim_{s\to k} \sum_{\omega\in L'} (\bar{z} + \bar{\omega})^k / |z + \omega|^{2s}$, where $\lim_{s\to k}$ denotes evaluation at the analytic continuation. This is done in [1, Thm. 7.13]. Then, [1, Prop. 7.15] shows how to relate $E_k(z, L)$ with partial sums of $L_{\mathfrak{f}}(\bar{\psi}, k)$. Finally, since $\log\Lambda_{L,\mathfrak{a}}(z)$ is a sum of translates of $\log\Theta_{L,\mathfrak{a}}(z)$ (see Definition 3.2), it is possible to add all partial sums of $L_{\mathfrak{f}}(\bar{\psi}, k)$ to obtain the desired theorem (see [1, Thm. 7.17]). □

Let $\mathfrak{p}$ be a prime of K above p where E has good reduction and $\mathfrak{p} \nmid 6\mathfrak{f}$. Fix a Weierstrass model for E with coordinate functions x, y that has good reduction at $\mathfrak{p}$ and fix $\mathfrak{a}$ an ideal coprime to $6\mathfrak{f}\mathfrak{p}$. Let $\hat{E}$ be the formal group attached to E and let $x(Z) \in z^{-2}\mathcal{O}_{\mathfrak{p}}[[Z]], y(Z) \in z^{-3}\mathcal{O}_{\mathfrak{p}}[[Z]]$ be the power series corresponding to x and y as in [5, Chap. IV, §1]. Let $\lambda_{\hat{E}}(Z) \in Z + Z^2 K_{\mathfrak{p}}[[Z]]$ be the logarithm map of the formal group $\hat{E}$ (see [5, Chap. IV, §1]) and consider the operator $D = \frac{1}{\lambda'_{\hat{E}}(Z)} \frac{d}{dZ}$. Denote by $K(E)$ the function field of E and by identifying the coordinates (x, y) with $(x(Z), y(Z))$ and with $(\wp(z), \wp'(z)/2)$, where $\wp$ is the Weierstrass $\wp$-function. We have the following commutative diagram (see [1, Prop. 7.20]).

$$\begin{array}{ccccccc} K(\wp(z),\wp'(z)) & \longleftarrow & K(E) & \longrightarrow & K(x(Z),y(Z)) & \longrightarrow & K_{\mathfrak{p}}((Z)) \\ \downarrow{\scriptstyle \frac{d}{dz}} & & \downarrow & & \downarrow{\scriptstyle D} & & \downarrow{\scriptstyle D} \\ K(\wp(z),\wp'(z)) & \longleftarrow & K(E) & \longrightarrow & K(x(Z),y(Z)) & \longrightarrow & K_{\mathfrak{p}}((Z)). \end{array} \quad (5)$$

Theorem 4.3. *Denote by $\Lambda_{\mathfrak{p},\mathfrak{a}}(Z) \in K_{\mathfrak{p}}((Z))$ the image of $\Lambda_{E,\mathfrak{a}} \in K((E))$ by the map given in (5). Then, $\Lambda_{\mathfrak{p},\mathfrak{a}} \in \mathcal{O}_{\mathfrak{p}}[[Z]]^\times$ and for every $k \geq 1$*

$$D^k \log\Lambda_{\mathfrak{p},\mathfrak{a}}(Z)|_{Z=0} = 12(-1)^{k-1}(k-1)! f^k (\mathrm{N}\mathfrak{a} - \psi(\mathfrak{a})^k)\Omega^{-k} L_{\mathfrak{f}}(\bar{\psi}^k, 1).$$

Proof. The first statement is proven in [1, Thm. 7.22] while the second one follows from the fact that (5) is commutative and Theorem 4.2. □

Suppose here and from now on that $p > 7$ and p splits in K. Then, we can suppose that $\mathrm{N}\mathfrak{a} \neq \psi(\mathfrak{a})$ modulo $\mathfrak{p}$ (for every $\mathfrak{p}$ such that $p > 7$ such an $\mathfrak{a}$ exists by [1, Lem. 10.2]). Let $F = K_1 = K(E[\mathfrak{p}])$, which is totally ramified at $\mathfrak{p}$ and let $\mathfrak{P}$ be the unique prime above $\mathfrak{p}$. Consider $\eta(1, \mathcal{O}) = \Lambda_{L,\mathfrak{a}}(\psi(\mathfrak{p})^{-1}\Omega) = \Lambda_{\mathfrak{p},\mathfrak{a}}(z) \in \mathcal{O}_F^\times$. Let $\delta\colon \mathcal{O}_{F_{\mathfrak{P}}}^\times \to (1 + \mathfrak{P}\mathcal{O}_{F_{\mathfrak{P}}})/(1 + \mathfrak{P}^2\mathcal{O}_{F_{\mathfrak{P}}})$ be the natural projection, which is Δ-equivariant.

Proposition 4.4. *$L_{\mathfrak{f}}(E, 1)/\Omega$ is integral at $\mathfrak{p}$. Moreover, $\delta(\eta(1, \mathcal{O})) = 1$ if and only if $L(\bar{\psi}, 1)/\Omega \equiv 0 \mod \mathfrak{p}$. In particular, $L(\bar{\psi}, 1)/\Omega \not\equiv 0 \mod \mathfrak{p}$ implies that $\eta(1, \mathcal{O})^{\chi_E} \notin ((\mathcal{O}_{F,\mathfrak{P}}^\times)^{\chi_E})^p$.*

Proof. Let $P = \xi(\psi(\mathfrak{p})^{-1}\Omega) = (x, y)$ and $z = -x/y$. It follows from [1, Lem. 7.3] that $z = -y/x \in \mathcal{O}_{F,\mathfrak{P}}$ has valuation 1 at the prime $\mathfrak{P}$. Theorem 4.3 allows to write $\eta(1, \mathcal{O}) = \Lambda_{\mathfrak{p},\mathfrak{a}}(z)$ as a power series on z. The first terms are

$$\Lambda_{\mathfrak{p},\mathfrak{a}}(z) = \Lambda_{\mathfrak{p},\mathfrak{a}}(0) + \Lambda_{\mathfrak{a},\mathfrak{a}}(0) 12f(\mathrm{N}\mathfrak{a} - \psi(\mathfrak{a})) \frac{L_{\mathfrak{f}}(\bar{\psi}, 1)}{\Omega} z + O(z^2). \quad (6)$$

Since $\Lambda_{\mathfrak{p},\mathfrak{a}}(Z) \in \mathcal{O}_{\mathfrak{p}}[[Z]]^\times$ we have that $\Lambda_{\mathfrak{p},\mathfrak{a}}(0) \in \mathcal{O}_{\mathfrak{p}}^\times$. Since $\mathfrak{a}$ is chosen so that $\mathrm{N}\mathfrak{a} \neq \psi(\mathfrak{a})$ modulo $\mathfrak{p}$ we have $\Lambda_{\mathfrak{p},\mathfrak{a}}(0) 12f(\mathrm{N}\mathfrak{a} - \psi(\mathfrak{a})) \in \mathcal{O}_{\mathfrak{p}}^\times$ which shows that $L_{\mathfrak{f}}(\bar{\psi}, 1)/\Omega$ is integral at $\mathfrak{p}$.

To prove the second part of the statement we need to compute the projection of $\Lambda_{\mathfrak{p},\mathfrak{a}}(z)$ in $(1+\mathfrak{P}\mathcal{O}_{F,\mathfrak{P}})/(1+\mathfrak{P}^2\mathcal{O}_{F,\mathfrak{P}})$. Since $\mathrm{ord}_{\mathfrak{P}}(z)=1$, (6) reduces to

$$\eta(1,\mathcal{O}) \equiv \Lambda_{\mathfrak{p},\mathfrak{a}}(0)\left(1+12f(\mathrm{N}\mathfrak{a}-\psi(\mathfrak{a}))\frac{L_{\mathfrak{f}}(\bar{\psi},1)}{\Omega}z\right) \mod \mathfrak{P}^2.$$

Using again that $\Lambda_{\mathfrak{p},\mathfrak{a}}(0)\in\mathcal{O}_{\mathfrak{p}}^{\times}$ and that $\mathfrak{p}$ is totally ramified in F it follows that $\delta(\Lambda_{\mathfrak{p},\mathfrak{a}}(0))=1$. Hence, $\delta(\eta(1,\mathcal{O}))=1+12f(\mathrm{N}\mathfrak{a}-\psi(\mathfrak{a}))\frac{L(\bar{\psi},1)}{\Omega}z$ and the second result follows. Finally, the study of the formal group $\hat{E}$ gives a Δ-equivariant isomorphism $(1+\mathfrak{P}\mathcal{O}_{F_{\mathfrak{P}}})/(1+\mathfrak{P}^2\mathcal{O}_{F_{\mathfrak{P}}})\simeq E[\mathfrak{p}]$ (see [1, Lem. 10.4]). From there we see that $\delta(\eta(1,\mathcal{O})^{\chi_E})=\delta(\eta(1,\mathcal{O}))^{\chi_E}=\delta(\eta(1,\mathcal{O}))\neq 0$. Thus $\eta(1,\mathcal{O})^{\chi_E}\notin((\mathcal{O}_{F,\mathfrak{P}}^{\times})^{\chi_E})^p$, since otherwise its image by δ would be 1 (because $(1+\mathfrak{P}\mathcal{O}_{F_{\mathfrak{P}}})/(1+\mathfrak{P}^2\mathcal{O}_{F_{\mathfrak{P}}})$ is killed by $\mathrm{N}\mathfrak{P}=\mathrm{N}\mathfrak{p}\mid p$). ☐

Theorem 4.5. *Suppose that $L(\bar{\psi},1)/\Omega\not\equiv 0 \mod \mathfrak{p}$ and that $\mathrm{Tr}_{K/\mathbb{Q}}\psi(\mathfrak{p})\neq 1$. Let δ_1 be as in Definition 2.7. Then, $\delta_1(\mathcal{O}_F^{\times})\neq 0$.*

Proof. By Lemma 2.8 it is enough to see that $(\mathcal{O}_F^{\times})^{\chi_E}\twoheadrightarrow(\mathcal{O}_{F,\mathfrak{P}}^{\times})^{\chi_E}$. For that we make the following observation. Using the p-adic logarithm we see that $(\mathcal{O}_{F,\mathfrak{P}}^{\times}\otimes_{\mathbb{Z}_p}\mathbb{Q}_p)^{\chi_E}$ is 1-dimensional (recall that χ_E is 1-dimensional). Moreover, since $\mathrm{Tr}_{K/\mathbb{Q}}\psi(\mathfrak{p})\neq 1$ it can be seen that $\mu_p\not\subset\mathcal{O}_{F,\mathfrak{P}}^{\times}$ (see [1, Lem. 10.9 (i)]). Therefore, $(\mathcal{O}_{F,\mathfrak{P}}^{\times})^{\chi_E}$ is free of rank 1 over $\mathbb{Z}_p$. Since $\eta(1,\mathcal{O})^{\chi_E}\notin((\mathcal{O}_{F,\mathfrak{P}}^{\times})^{\chi_E})^p$ by Proposition 4.4, $\eta(1,\mathcal{O})^{\chi_E}\in\mathcal{O}_F^{\times}$ is a generator of $(\mathcal{O}_{F,\mathfrak{P}})^{\chi_E}$ giving the desired surjectivity. ☐

We can finally give the proof of the Coates–Wiles Theorem.

Theorem 4.6 (Coates–Wiles)**.** *Suppose that $L(E_{/K},1)\neq 0$. Then $E(K)$ is finite.*

Proof. Theorem 4.1 shows that $L_{\mathfrak{f}}(\bar{\psi},1)\neq 0$. By the Chebotarev Theorem there are infinite primes $\mathfrak{p}$ of K above a rational prime p such that p splits in K and $\mathrm{Tr}_{K/\mathbb{Q}}\psi(\mathfrak{p})\neq 1$. We can choose one such that $p>7$, $\mathfrak{p}$ coprime to $6\mathfrak{f}$ and $L_{\mathfrak{f}}(\bar{\psi},1)/\Omega$ is a unit at $\mathfrak{p}$.

Therefore we can apply the previous results of this section to $\mathfrak{p}$. Since $\mu_p\not\in F_{\mathfrak{P}}$ by [1, Lem. 10.9 (i)], by Proposition 4.4 and Corollary 3.15, $A^{\chi_E}=0$. In addition, Theorem 4.5 shows that $\delta_1(\mathcal{O}_F^{\times})\neq 0$. The conditions of Corollary 2.10 are satisfied so we can affirm $S_{\pi}(E_{/K})=0$, where $\pi\in\mathcal{O}$ such that $\mathfrak{p}=\pi\mathcal{O}$. Therefore $E(K)/\mathfrak{p}E(K)=0$, by the Mordell–Weil Theorem $E(K)$ has to be finite (see Proposition 2.3) and we are done. ☐

Expression of gratitude

I would like to start expressing my very great appreciation to Francesc Fité. He has been extremely generous with his time as well as patient when giving me advice about the project. I would also like to thank Christopher Skinner for giving me the opportunity of visiting Princeton University for 6 months and for being available whenever I needed. My thanks are also extended to Victor Rotger for suggesting me to work on this project. I would also like to thank the referee for a careful reading of the manuscript and for giving useful corrections and suggestions.

References

[1] K. Rubin. Elliptic curves with complex multiplication and the conjecture of Birch and Swinnerton-Dyer, in: *Arithmetic theory of elliptic curves* (Cetraro, 1997), Lecture Notes in Math. **1716**, Springer, Berlin, 1999, 167–234.

[2] K. Rubin. *Euler Systems*, Princeton University Press, 2000.

[3] J.-P. Serre. *Local fields*, Graduate Texts in Mathematics **67**, Springer Science & Business Media, 2013.

[4] J.H. Silverman. *Advanced topics in the arithmetic of elliptic curves*, Graduate Texts in Mathematics **151**, Springer Science & Business Media, 1994.

[5] J.H. Silverman. *The arithmetic of elliptic curves*, Graduate Texts in Mathematics **106**, Springer Science & Business Media, 2009.

Reports@SCM
Volume 5, number 1, 2020

Table of Contents