REPORTS @SCM

AN ELECTRONIC JOURNAL OF THE SOCIETAT CATALANA DE MATEMÀTIQUES

Volume 4, num. 1, 2018

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 4, number 1

December 2018

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

Societat
Catalana de
Matemàtiques

Institut
d'Estudis
Catalans

Editorial Team

Editors-in-chief

Núria Fagella, Universitat de Barcelona (holomorphic dynamical systems, geometric function theory)
Enric Ventura, Universitat Politècnica de Catalunya (algebra, group theory)

Associate Editors

Marta Casanellas, Universitat Politècnica de Catalunya (algebraic geometry, phylogenetics)
Pedro Delicado, Universitat Politècnica de Catalunya (statistics and operations research)
Alex Haro, Universitat de Barcelona (dynamical systems)
David Marín, Universitat Autònoma de Barcelona (complex and differential geometry, foliations)
Xavier Massaneda, Universitat de Barcelona (complex analysis)
Eulàlia Nualart, Universitat Pompeu Fabra (probability)
Joaquim Ortega-Cerdà, Universitat de Barcelona (analysis)
Francesc Perera, Universitat Autònoma de Barcelona (non commutative algebra, operator algebras)
Julian Pfeifle, Universitat Politècnica de Catalunya (discrete geometry, combinatorics, optimization)
Albert Ruiz, Universitat Autònoma de Barcelona (topology)
Gil Solanes, Universitat Autònoma de Barcelona (differential geometry)

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.

This work is subject to a Recognition - Non Commercial - Without derivative works Creative Commons 3.0 Spain license, unless the text, pictures or other illustrations indicate the contrary. License's full text can be read at http://creativecommons.org/licenses/by-nc-nd/3.0/es/deed.ca. Readers can reproduce, distribute and communicate the work as long as its authorship and publishing institution are recognized and also if this does not entail commercial use or derivative work.
©The authors of the articles

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 4, number 1, 2018

Table of Contents

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

A generalization of Pascal's mystic hexagram

***Sergi Baena Miret**
Universitat de Barcelona
sergibaena@ub.edu

∗Corresponding author

Resum *(CAT)*
El teorema clàssic de Pascal afirma que si un hexàgon a $\mathbb{P}^2(\mathbb{C})$ està inscrit en una cònica llavors els costats oposats de l'hexàgon es troben en tres punts que s'ubiquen sobre una recta, anomenada recta de Pascal. Zhongxuan Luo va donar el 2007 una generalització del teorema de Pascal per a corbes de grau arbitrari. En el present article es donen dues demostracions d'aquesta generalització. La primera és autocontiguda i fa ús del teorema de Carnot, mentre que la segona es basa en el teorema fonamental de Max Noether.

Abstract *(ENG)*
Pascal's classical theorem asserts that if a hexagon in $\mathbb{P}^2(\mathbb{C})$ is inscribed in a conic, then the opposite sides of the hexagon lie on a straight line, called Pascal line. Zhongxuan Luo gave in 2007 a generalization of Pascal's theorem for curves of arbitrary degree. In the present article, two proofs of this generalization are given. The first one is self-contained and makes use of Carnot's theorem, while the second proof is based on Max Noether's Fundamental theorem.

Keywords: *Pascal's Theorem, Characteristic Ratio, Carnot's Theorem, Pascal Mapping, Max Noether's Fundamental Theorem.*
MSC (2010): *Primary: 14N05, 14N15, 14H50. Secondary: 51N35.*
Received: *July 9, 2017.*
Accepted: *October 25, 2018.*

Acknowledgement
The author would like to thank Dr. J.C. Naranjo for his many insightful suggestions. He would like to thank also C. Cruz, A. Rojas and C. Torres for their many advices. Finally, he wishes to thank the anonymous referee whose comments have greatly improved the paper.

Societat Catalana de Matemàtiques
Institut d'Estudis Catalans

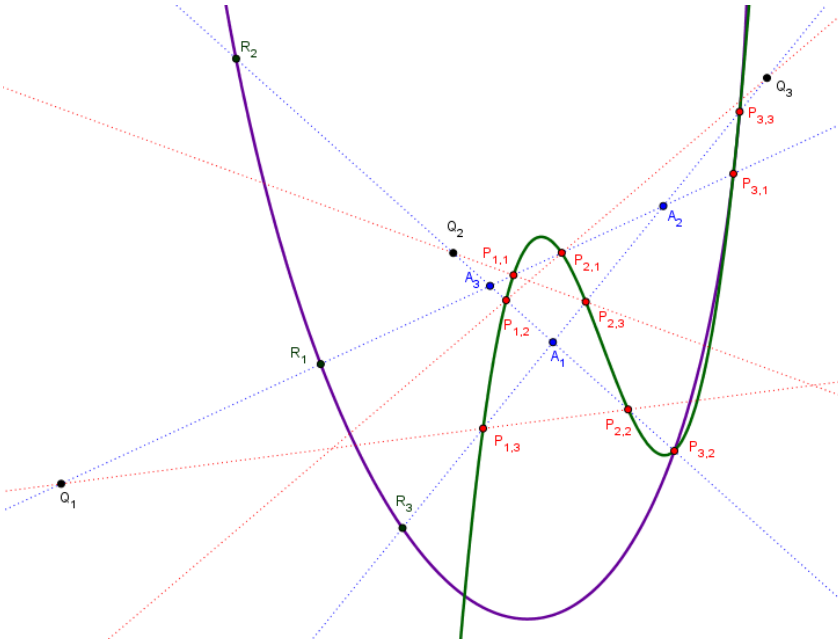

Figure 1: Pascal Type theorem for n = 3.

1. Introduction

One of the classical results in projective geometry is Pascal's theorem, also known as Pascal's Mystic Hexagram. This theorem, obtained by Blaise Pascal in 1640, asserts that if a hexagon in $\mathbb{P}^2(\mathbb{C})$ is inscribed in a conic, then the opposite sides of the hexagon lie on a straight line; c.f. [6, § 5.6, Cor. 1].

There are many known generalizations of Pascal's theorem. For example, Chasles' theorem (c.f. [3]) or the Cayley-Bacharach theorem (c.f. [5]) are generalizations of Pascal's theorem. In [8], Zhongxuan Luo presents another generalization of Pascal's theorem (see Fig. 1): *Let l_1, l_2, l_3 be three non-concurrent lines and take a collection of $n \geq 2$ points $S_i \subset l_i$ on each line, such that $S_i \cap l_j = \emptyset$ for $j \neq i$. Choose two points $P_{1,i}, P_{2,i} \in S_i$ on each collection and let R_1, R_2, R_3 be the triple of points given by the Pascal mapping (see Definition 2.11) applied to the six chosen points. Then the $3n$ points $S_1 \cup S_2 \cup S_3$ lie on an algebraic curve of degree n that contains none of the lines l_1, l_2, l_3 if and only if there exists an algebraic curve of degree $n-1$ intersecting each line l_i in $\{R_i\} \cup S_i \setminus \{P_{1,i}, P_{2,i}\}$.*

The aim of this article is to present two different proofs of Zhongxuan Luo's extension of Pascal's theorem. The first proof is elementary and makes use of a version of Carnot's theorem; see Section 3. The approach is similar to [8], but we do not use spline theory. The second proof is based on Max Noether's Fundamental theorem; see Section 4.

Throughout this paper we work in the complex projective plane $\mathbb{P}^2(\mathbb{C})$ and we set a projective reference $R = \{A_1, A_2, A_3; O\}$, so that $A_1 = (1:0:0)$, $A_2 = (0:1:0)$, $A_3 = (0:0:1)$ and $O = (1:1:1)$. Observe that then $A_{2,3} = OA_1 \cap A_2A_3 = (0:1:1)$, $A_{3,1} = OA_2 \cap A_3A_1 = (1:0:1)$, and $A_{1,2} = OA_3 \cap A_1A_2 = (1:1:0)$, where for points $A, B \in \mathbb{P}^2(\mathbb{C})$, we mean AB to be the projective line that joins A and B. Moreover, we set $l_1 = A_2A_3$, $l_2 = A_3A_1$ and $l_3 = A_1A_2$ to be the sides of the projective triangle

$A_1A_2A_3$ with vertices at points A_1, A_2, A_3.

2. An extension of Carnot's theorem

In this section we review a generalization of Carnot's and Menelaus's theorems, which allows to determine whether a certain configuration of points lies on an algebraic curve of a given degree. We also introduce some constructions from [8] that appear in the generalization of Pascal's theorem.

Definition 2.1. Given points $P_1, \dots, P_r \in A_iA_j \backslash \{A_i, A_j\}$, where $i, j \in \{1, 2, 3\}$, $i \neq j$. We define the characteristic ratio of $P_1, \dots, P_r$ with respect to the reference R to be $[A_i, A_j; P_1, \dots, P_r]_R = \prod_{k=1}^{r}(A_i, A_j, A_{i,j}, P_k)$, where $(A_i, A_j, A_{i,j}, P_k)$ denotes the cross ratio; c.f. [4, § 5.2].

The notion of characteristic ratio defined in [8] and the one defined in Definition 2.1 are inverse to each other.

Example 2.2. Let $P_k = (0 : \lambda_k : 1) \in A_2A_3$ with $\lambda_k \in \mathbb{C} \setminus \{0\}$, $k = 1, \dots, r$. Then, $[A_2, A_3; P_1, \dots, P_r]_R = \prod_{k=1}^{r}(A_2, A_3, A_{2,3}, P_k) = \prod_{k=1}^{r} \lambda_k$.

With this notation at hand, Menelaus's theorem (c.f. [7]) and Carnot's theorem (c.f. [2]) can be stated as follows.

Theorem 2.3 (Menelaus's Theorem)**.** *Let $P_i \in l_i$, $i = 1, 2, 3$, be points different from A_1, A_2 and A_3. Then, P_1, P_2 and P_3 are collinear if and only if $[A_2, A_3; P_1]_R[A_3, A_1; P_2]_R[A_1, A_2; P_3]_R = -1$.*

Theorem 2.4 (Carnot's Theorem)**.** *Let $P_1, P_2 \in l_1$, $P_3, P_4 \in l_2$, and $P_5, P_6 \in l_3$ be six distinct points different from A_1, A_2 and A_3. Then, $P_1, P_2, \dots, P_6$ lie on a conic disjoint with $\{A_1, A_2, A_3\}$ if and only if $[A_2, A_3; P_1, P_2]_R[A_3, A_1; P_3, P_4]_R[A_1, A_2; P_5, P_6]_R = 1$.*

The next theorem is a natural generalization of Menelaus's and Carnot's theorems to curves of arbitrary degree. It is called Carnot's theorem in [1] and is equivalent to [8, Thm. 4.4]. For completeness we provide a proof here.

Theorem 2.5. *Let $S_i = \{P_{1,i}, \dots, P_{n,i}\}$ be a collection of n different points of $l_i \setminus \{A_1, A_2, A_3\}$, $i = 1, 2, 3$. Then, $S_1 \cup S_2 \cup S_3$ lie on an algebraic curve of degree n disjoint with $\{A_1, A_2, A_3\}$ if and only if*

$$[A_2, A_3; P_{1,1}, \dots, P_{n,1}]_R[A_3, A_1; P_{1,2}, \dots, P_{n,2}]_R[A_1, A_2; P_{1,3}, \dots, P_{n,3}]_R = (-1)^n.$$

Proof. Recall that the cases $n = 1$ and $n = 2$ are Menelaus's theorem and Carnot's theorem respectively. Then we can assume that $n \geq 3$. We denote

$$\mathbb{C}[X, Y, Z]_n = \{F \in \mathbb{C}[X, Y, Z];\ F \text{ homogeneous polynomial of degree } n\}.$$

With this in hand, we define the map $\varphi \colon \mathbb{C}[X, Y, Z]_n \big/ (XYZ) \to \mathbb{C}[Y, Z]_n \times \mathbb{C}[X, Z]_n \times \mathbb{C}[X, Y]_n$ such that $\varphi([F(X, Y, Z)]) = (F(0, Y, Z), F(X, 0, Z), F(X, Y, 0))$.

Clearly, φ is well defined and linear. Let us see that it is also injective: if $[F] \in ker(\varphi)$, then $\varphi([F(X, Y, Z)]) = (F(0, Y, Z), F(X, 0, Z), F(X, Y, 0)) = (0, 0, 0)$. Thus, X, Y and Z divide F. Therefore, $[F] = [0]$.

Hence, φ is an isomorphism over its image. We claim that the image of φ is exactly the set $M_n \subset \mathbb{C}[Y,Z]_n \times \mathbb{C}[X,Z]_n \times \mathbb{C}[X,Y]_n$ defined as

$$M_n := \left\{ \left(\sum_{i+j=n} B_{i,j} Y^i Z^j, \sum_{i+j=n} C_{i,j} X^i Z^j, \sum_{i+j=n} D_{i,j} X^i Y^j \right) ; B_{n,0} = D_{0,n}, B_{0,n} = C_{0,n}, C_{n,0} = D_{n,0} \right\}.$$

Clearly, $Im(\varphi) \subseteq M_n$. Moreover, by computing dimensions, we find

$$\begin{aligned} \dim_{\mathbb{C}} \left(\mathbb{C}[X,Y,Z]_n \big/ (XYZ) \right) = 3n &= \dim_{\mathbb{C}}(\mathbb{C}[Y,Z]_n) + \dim_{\mathbb{C}}(\mathbb{C}[X,Z]_n) \\ &+ \dim_{\mathbb{C}}(\mathbb{C}[X,Y]_n) - 3 = \dim_{\mathbb{C}}(M_n). \end{aligned}$$

It follows that $Im(\varphi) = M_n$.

Let $P_{i,1} = (0 : a_i : 1)$, $P_{i,2} = (1 : 0 : b_i)$, and $P_{i,3} = (c_i : 1 : 0)$ with $a_i, b_i, c_i \in \mathbb{C} \setminus \{0\}$, $i = 1, 2, \dots, n$. Note that, by degree reasons, any curve of degree n containing $S_1 \cup S_2 \cup S_3$ and not containing l_1, l_2 nor l_3 must be disjoint with $\{A_1, A_2, A_3\}$. Such a curve exists if and only if

$$\left(\lambda_1 \prod_{i=1}^{n} (a_i Z - Y),\ \lambda_2 \prod_{i=1}^{n} (b_i X - Z),\ \lambda_3 \prod_{i=1}^{n} (c_i Y - X) \right) \in M_n, \tag{1}$$

for some $\lambda_1, \lambda_2, \lambda_3 \in \mathbb{C} \setminus \{0\}$. According to the definition of M_n, a necessary and sufficient condition for (1) to be true is that the following system

$$\begin{cases} \lambda_1 \prod_{i=1}^{n} a_i &= (-1)^n \lambda_2, \\ (-1)^n \lambda_1 &= \lambda_3 \prod_{i=1}^{n} c_i, \\ \lambda_2 \prod_{i=1}^{n} b_i &= (-1)^n \lambda_3, \end{cases} \tag{2}$$

has a non-trivial solution for $\lambda_1, \lambda_2, \lambda_3$. However, the system (2) has a non-trivial solution if and only if $(-1)^n = \prod_{i=1}^{n} a_i b_i c_i = [A_2, A_3; P_{1,1}, \dots, P_{n,1}]_R [A_3, A_1; P_{1,2}, \dots, P_{n,2}]_R [A_1, A_2; P_{1,3}, \dots, P_{n,3}]_R$. This completes the proof. □

Next, we introduce some notions from [8] that will be needed in the generalization of Pascal's theorem.

Definition 2.6. The characteristic map $\sigma_{i,j} : A_i A_j \to A_i A_j$ relative to $A_i, A_j, A_{i,j}$ is the projective involution that satisfies $\sigma_{i,j}(A_i) = A_j$, $\sigma_{i,j}(A_j) = A_i$, and $\sigma_{i,j}(A_{i,j}) = A_{i,j}$, where $i, j = 1, 2, 3$, $i \neq j$.

Observation 2.7. If $P = \sigma_{i,j}(Q)$ is the image of Q under the characteristic map relative to $A_i, A_j, A_{i,j}$, then $[A_i, A_j; P, Q]_R = (A_i, A_j, A_{i,j}, P)(A_i, A_j, A_{i,j}, Q) = 1$.

An interesting fact is that if we take a point in each side of a triangle, and we consider their respective images of the characteristic map relative to each side, then these six points lie on a conic. This property will take an important role in Section 4.

Proposition 2.8. *Let $P_1 \in l_1$, $P_2 \in l_2$ and $P_3 \in l_3$ be three points different from A_1, A_2, A_3. Then, P_1, P_2, P_3, $\sigma_{2,3}(P_1)$, $\sigma_{3,1}(P_2)$ and $\sigma_{1,2}(P_3)$ lie on a conic.*

Proof. By Observation 2.7, $[A_2, A_3; \sigma_{2,3}(P_1), P_1]_R = [A_3, A_1; \sigma_{3,1}(P_2), P_2]_R = [A_1, A_2; \sigma_{1,2}(P_3), P_3]_R = 1$. Then, the result follows by Carnot's theorem. □

The following results contain basic properties of the characteristic map; c.f. [8, Cor. 3.7, 3.9].

Proposition 2.9. *Any three distinct points $P_1 \in l_1$, $P_2 \in l_2$, and $P_3 \in l_3$ different from A_1, A_2, A_3 are collinear if and only if $\sigma_{2,3}(P_1)$, $\sigma_{3,1}(P_2)$, and $\sigma_{1,2}(P_3)$ are collinear.*

Proof. Let $P_1 = (0 : a : 1)$, $P_2 = (1 : 0 : b)$, and $P_3 = (c : 1 : 0)$, with $a, b, c \in \mathbb{C} \setminus \{0\}$. Then, $\sigma_{2,3}(P_1) = (0 : 1 : a)$, $\sigma_{3,1}(P_2) = (b : 0 : 1)$, and $\sigma_{1,2}(P_3) = (1 : c : 0)$. Menelaus's theorem asserts that a necessary and sufficient condition of P_1, P_2, P_3 to be collinear is that

$$-1 = [A_2, A_3; P_1]_R[A_3, A_1; P_2]_R[A_1, A_2; P_3]_R = abc. \tag{3}$$

Similarly, their images under their corresponding characteristic map are collinear if and only if

$$-1 = [A_2, A_3; \sigma_{2,3}(P_1)]_R[A_3, A_1; \sigma_{3,1}(P_2)]_R[A_1, A_2; \sigma_{1,2}(P_3)]_R = \frac{1}{abc}. \tag{4}$$

Since both equalities (3) and (4) are equivalent, this completes the proof. □

Proposition 2.10. *Let $P_1, P_2 \in l_1$, $P_3, P_4 \in l_2$, and $P_5, P_6 \in l_3$ be any six distinct points different from A_1, A_2, A_3. Then, $P_1, P_2, \dots, P_6$ lie on a conic if and only if their images by the corresponding characteristic map lie on a conic as well.*

Proof. Let $P_1 = (0 : a_1 : 1)$, $P_2 = (0 : a_2 : 1)$, $P_3 = (1 : 0 : b_1)$, $P_4 = (1 : 0 : b_2)$, $P_5 = (c_1 : 1 : 0)$, and $P_6 = (c_2 : 1 : 0)$, with $a_i, b_i, c_i \in \mathbb{C} \setminus \{0\}$, $i = 1, 2$. Then, $\sigma_{2,3}(P_1) = (0 : 1 : a_1)$, $\sigma_{2,3}(P_2) = (0 : 1 : a_2)$, $\sigma_{3,1}(P_3) = (b_1 : 0 : 1)$, $\sigma_{3,1}(P_4) = (b_2 : 0 : 1)$, $\sigma_{1,2}(P_5) = (1 : c_1 : 0)$, and $\sigma_{1,2}(P_6) = (1 : c_2 : 0)$. By Carnot's theorem, the six points $P_1, P_2, \dots, P_6$ lie on a conic if and only if

$$1 = [A_2, A_3; P_1, P_2]_R[A_3, A_1; P_3, P_4]_R[A_1, A_2; P_5, P_6]_R = a_1a_2b_1b_2c_1c_2. \tag{5}$$

Similarly, their images under their corresponding characteristic map lie on a conic if and only if

$$\begin{aligned} 1 = [A_2, A_3; &\sigma_{2,3}(P_1), \sigma_{2,3}(P_2)]_R[A_3, A_1; \sigma_{3,1}(P_3), \sigma_{3,1}(P_4)]_R \\ &\times [A_1, A_2; \sigma_{1,2}(P_5), \sigma_{1,2}(P_6)]_R = \frac{1}{a_1a_2b_1b_2c_1c_2}. \end{aligned} \tag{6}$$

Since both equalities (5) and (6) are equivalent, this completes the proof. □

The following construction from [8] plays a crucial role in Zhongxuan Luo's generalization of Pascal's theorem.

Definition 2.11. The Pascal mapping is the map $\Psi := (\sigma_{2,3} \times \sigma_{3,1} \times \sigma_{1,2}) \circ \Phi$, where $\Phi\colon (l_1 \setminus \{A_2, A_3\})^2 \times (l_2 \setminus \{A_3, A_1\})^2 \times (l_3 \setminus \{A_1, A_2\})^2 \to l_1 \times l_2 \times l_3$ satisfies

$$\Phi((P_1, P_2), (P_3, P_4), (P_5, P_6)) = \{P_1P_2 \cap P_4P_5, P_3P_4 \cap P_6P_1, P_5P_6 \cap P_2P_3\}.$$

If we denote $Q_1 = P_1P_2 \cap P_4P_5$, $Q_2 = P_3P_4 \cap P_6P_1$, and $Q_3 = P_5P_6 \cap P_2P_3$ then, $\Psi((P_1, P_2), (P_3, P_4), (P_5, P_6)) = \{\sigma_{2,3}(Q_1), \sigma_{3,1}(Q_2), \sigma_{1,2}(Q_3)\}$.

3. A Pascal type theorem

In this section we present the generalization of Pascal's theorem given in [8]. We give an elementary proof based on the results of the previous section. In Section 4 we will give a second proof using Max Noether's Fundamental theorem. First let us recall the complete version of Pascal's original theorem. For completeness, we provide a proof here based on Menelaus's and Carnot's theorems.

Theorem 3.1. *Let $P_1, P_2 \in l_1$, $P_3, P_4 \in l_2$, and $P_5, P_6 \in l_3$ be six distinct points, all of them different from A_1, A_2, A_3, and let $Q_1 = P_1P_2 \cap P_4P_5$, $Q_2 = P_3P_4 \cap P_6P_1$, and $Q_3 = P_5P_6 \cap P_2P_3$. Then, $P_1, P_2, \dots, P_6$ lie on a conic if and only if Q_1, Q_2, Q_3 are collinear.*

Proof. Let $P_1 = (0 : a_1 : 1)$, $P_2 = (0 : a_2 : 1)$, $P_3 = (1 : 0 : b_1)$, $P_4 = (1 : 0 : b_2)$, $P_5 = (c_1 : 1 : 0)$, and $P_6 = (c_2 : 1 : 0)$, with $a_i, b_i, c_i \neq 0$, $i = 1, 2$; it follows that $Q_1 = (0 : -1 : b_2c_1)$, $Q_2 = (a_1c_2 : 0 : -1)$, and $Q_3 = (-1 : b_1a_2 : 0)$. By Carnot's theorem, $P_1, P_2, \dots, P_6$ lie on a conic disjoint from $\{A_1, A_2, A_3\}$ if and only if

$$1 = [A_2, A_3; P_1, P_2]_R[A_3, A_1; P_3, P_4]_R[A_1, A_2; P_5, P_6]_R = a_1a_2b_1b_2c_1c_2. \tag{7}$$

Similarly, by Menelaus's theorem, we have that a necessary and sufficient condition for Q_1, Q_2, Q_3 to be collinear is that

$$-1 = [A_2, A_3; Q_1]_R[A_3, A_1; Q_2]_R[A_1, A_2; Q_3]_R = \frac{-1}{b_2c_1}\frac{-1}{a_1c_2}\frac{-1}{b_1a_2}. \tag{8}$$

Since both equalities (7) and (8) are equivalent, this proves the theorem. □

Notice that by Proposition 2.9, if Q_1, Q_2, Q_3 lie on the same line, then the points in the Pascal mapping $\Psi((P_1, P_2), (P_3, P_4), (P_5, P_6))$ are also collinear points. It is precisely this version of Pascal's theorem that was generalized by Zhongxuan Luo to higher degrees. The precise statement is the following.

Theorem 3.2 (Pascal Type Theorem). *Let $S_j = \{P_{i,j}\}_{i=1}^n$ be a collection of $n \geq 2$ distinct points on the set $l_j \setminus \{A_1, A_2, A_3\}$, $j = 1, 2, 3$. Let us choose two points on each collection S_j, and let R_1, R_2, R_3 be the triple given by the Pascal mapping applied to the six chosen points. Then, the $3n$ points $S_1 \cup S_2 \cup S_3$ lie on an algebraic curve of degree n disjoint with $\{A_1, A_2, A_3\}$ if and only if there exists an algebraic curve of degree $n-1$ disjoint with $\{A_1, A_2, A_3\}$ which contains R_1, R_2, R_3 and the $3(n-2)$ points from $S_1 \cup S_2 \cup S_3$ that have not been chosen.*

Proof. Let us take $a_i, b_i, c_i \in \mathbb{C} \setminus \{0\}$ and $P_{i,1} = (0 : a_i : 1)$, $P_{i,2} = (1 : 0 : b_i)$, and $P_{i,3} = (c_i : 1 : 0)$ for every $i = 1, \dots, n$. Without loss of generality, let us choose the points $P_{1,1}$, $P_{2,1}$, $P_{1,2}$, $P_{2,2}$, $P_{1,3}$ and $P_{2,3}$, to apply the Pascal mapping; see Fig. 1 above. Then,

$$\Psi\left((P_{1,1}, P_{2,1}), (P_{1,2}, P_{2,2}), (P_{1,3}, P_{2,3})\right) = \{R_1, R_2, R_3\}, \tag{9}$$

where $R_1 = \sigma_{2,3}(Q_1) = (0 : b_2c_1 : -1)$, $R_2 = \sigma_{3,1}(Q_2) = (-1 : 0 : a_1c_2)$, and $R_3 = \sigma_{1,2}(Q_3) = (b_1a_2 : -1 : 0)$, with $Q_1 = P_{1,1}P_{2,1} \cap P_{2,2}P_{1,3}$, $Q_2 = P_{1,2}P_{2,2} \cap P_{2,3}P_{1,1}$, and $Q_3 = P_{1,3}P_{2,3} \cap P_{2,1}P_{1,2}$.

By Theorem 2.5, the $3n$ points $S_1 \cup S_2 \cup S_3$ lie on an algebraic curve of degree n disjoint with $\{A_1, A_2, A_3\}$ if and only if

$$\begin{aligned}(-1)^n &= [A_2, A_3; P_{1,1}, \dots, P_{n,1}]_R[A_3, A_1; P_{1,2}, \dots, P_{n,2}]_R[A_1, A_2; P_{1,3}, \dots, P_{n,3}]_R \\ &= \prod_{i=1}^n (A_2, A_3, A_{2,3}, P_{i,1})(A_3, A_1, A_{3,1}, P_{i,2})(A_1, A_2, A_{1,2}, P_{i,3}) = \prod_{i=1}^n a_ib_ic_i.\end{aligned} \tag{10}$$

Similarly, there exists an algebraic curve of degree $n-1$ disjoint with $\{A_1, A_2, A_3\}$ which contains R_1, R_2, R_3 and the $3(n-2)$ points from $S_1 \cup S_2 \cup S_3$ that have not been chosen if and only if

$$\begin{aligned}(-1)^{n-1} &= [A_2, A_3; P_{3,1}, \dots, P_{n,1}, R_1]_R[A_3, A_1; P_{3,2}, \dots, P_{n,2}, R_2]_R[A_1, A_2; P_{3,3}, \dots, P_{n,3}, R_3]_R \\ &= \left[(A_2, A_3, A_{2,3}, R_1)\prod_{i=3}^{n}(A_2, A_3, A_{2,3}, P_{i,1})\right]\left[(A_3, A_1, A_{3,1}, R_2)\prod_{i=3}^{n}(A_3, A_1, A_{3,1}, P_{i,2})\right] \\ &\times \left[(A_1, A_2, A_{1,2}, R_3)\prod_{i=3}^{n}(A_1, A_2, A_{1,2}, P_{i,3})\right] = -\prod_{i=1}^{n} a_i b_i c_i.\end{aligned} \tag{11}$$

Since both equalities (10) and (11) are equivalent, this completes the proof. □

4. A Pascal type theorem and Max Noether's fundamental theorem

We give a new proof of Theorem 3.2 based on Max Noether's Fundamental theorem; in particular, we will make us of a corollary of it. To do so, we need a few basic notions about algebraic curves in $\mathbb{P}^2(\mathbb{C})$; for more details, see [6].

Max Noether's Fundamental theorem is concerned with the following question (c.f. [6, § 5.5]): suppose C, C' are two projective plane curves with no common factors, and C'' is another curve satisfying $C \cap C' \subset C \cap C''$, when counted with multiplicity. So, when is there a curve that intersects C in the points of $C \cap C''$ that are not in $C \cap C'$?

For our purpose, we do not use directly Max Noether's Fundamental theorem, but we use a corollary of it. First, if C, C' are projective plane curves with no common components, the intersection cycle $C \cdot C'$ is defined as the formal sum

$$C \cdot C' = \sum_{P \in C \cap C'} m_P(C, C')P,$$

where $m_P(C, C')$ is the multiplicity of the point P in $C \cap C'$; c.f. [6, § 5.5]. In particular, $m_P(C, C') = 0$ if and only if $P \notin C \cap C'$.

If C, C' are projective plane curves of degree n and m respectively, CC' is the projective plane curve of degree $n + m$ consisting on the union of C and C'.

With this notation at hand, we are in conditions to state the corollary of Max Noether's Fundamental theorem; c.f. [6, § 5.5, Cor. 2].

Theorem 4.1. *Let C, C', C'' be projective plane curves such that C' and C'' have no common component with C. If all the points of $C \cap C'$ are simple points of C and $C \cdot C'' \geq C \cdot C'$ (i.e., $m_P(C, C'') \geq m_P(C, C')$ for every $P \in C$), then there is a curve Γ of degree $deg(\Gamma) = deg(C'') - deg(C')$ such that $C \cdot \Gamma = C \cdot C'' - C \cdot C'$.*

Now we are in conditions to give the new proof of Theorem 3.2.

Proof. (Pascal Type Theorem) Let us take the same notation as in (9). Here is when Proposition 2.8 becomes crucial, since it asserts that the points $Q_1, Q_2, Q_3, R_1, R_2, R_3$ lie on a conic; let Γ_2 be that conic.

Let $C = (P_{2,1}P_{1,2})(P_{2,2}P_{1,3})(P_{2,3}P_{1,1})$ be the cubic generated by the three opposites sides, disjoint with $\{A_1, A_2, A_3\}$, of the hexagon with vertices $P_{1,1}, P_{2,1}, P_{1,2}, P_{2,2}, P_{1,3}, P_{2,3}$.

First assume that the $3n$ points $S_1 \cup S_2 \cup S_3$ lie on an algebraic curve Γ_n of degree n disjoint with $\{A_1, A_2, A_3\}$ and consider the algebraic curve $\Gamma_{n+2} = \Gamma_2\Gamma_n$ of degree $n+2$. Then, we have that

$$\Gamma_{n+2} \cdot l_1 l_2 l_3 = \sum_{i=1}^{n} (P_{i,1} + P_{i,2} + P_{i,3}) + Q_1 + Q_2 + Q_3 + R_1 + R_2 + R_3,$$

and $C \cdot l_1 l_2 l_3 = P_{1,1} + P_{2,1} + P_{1,2} + P_{2,2} + P_{1,3} + P_{2,3} + Q_1 + Q_2 + Q_3$. Therefore, $\Gamma_{n+2} \cdot l_1 l_2 l_3 - C \cdot l_1 l_2 l_3 = \sum_{i=3}^{n} (P_{i,1} + P_{i,2} + P_{i,3}) + R_1 + R_2 + R_3$.

By Theorem 4.1, there exists a curve Γ of degree $\deg(\Gamma) = \deg(\Gamma_{n+2}) - \deg(C) = n-1$ such that $\Gamma \cdot l_1 l_2 l_3 = \sum_{i=3}^{n} (P_{i,1} + P_{i,2} + P_{i,3}) + R_1 + R_2 + R_3$. So, Γ is an algebraic curve of degree $n-1$ that passes through the $3(n-1)$ points $P_{3,1}, \dots, P_{n,1}, P_{3,2}, \dots, P_{n,2}, P_{3,3}, \dots, P_{n,3}, R_1, R_2, R_3$.

Reciprocally, suppose that there exists an algebraic curve Γ'_{n-1} of degree $n-1$ disjoint with $\{A_1, A_2, A_3\}$ that contains R_1, R_2, R_3 and the $3(n-2)$ points from $S_1 \cup S_2 \cup S_3$ that have not been chosen. Consider the algebraic curve $\Gamma'_{n+2} = C\Gamma'_{n-1}$ of degree $n+2$. Then, we have that

$$\Gamma'_{n+2} \cdot l_1 l_2 l_3 = \sum_{i=1}^{n} (P_{i,1} + P_{i,2} + P_{i,3}) + Q_1 + Q_2 + Q_3 + R_1 + R_2 + R_3,$$

and $\Gamma_2 \cdot l_1 l_2 l_3 = Q_1 + Q_2 + Q_3 + R_1 + R_2 + R_3$. Therefore, $\Gamma'_{n+2} \cdot l_1 l_2 l_3 - \Gamma_2 \cdot l_1 l_2 l_3 = \sum_{i=1}^{n} (P_{i,1} + P_{i,2} + P_{i,3})$.

By Theorem 4.1, there exists a curve Γ' of degree $\deg(\Gamma') = \deg(\Gamma'_{n+2}) - \deg(\Gamma_2) = n$ such that $\Gamma' \cdot l_1 l_2 l_3 = \sum_{i=1}^{n} (P_{i,1} + P_{i,2} + P_{i,3})$. So, Γ' is an algebraic curve of degree n that passes through the $3n$ points $P_{1,1}, \dots, P_{n,1}, P_{1,2}, \dots, P_{n,2}, P_{1,3}, \dots, P_{n,3}$. This completes the proof. □

References

[1] Retrieved October 10, 2018, from `https://www.encyclopediaofmath.org/index.php/Carnot_theorem`.

[2] L. Carnot, "Géométrie de position", Paris, (1803).

[3] M. Chasles, "Traité des sections coniques faisant suite au Traité de géométrie supérieure", Paris, Gauthier-Villars, imprimeur-libraire du bureau des longitudes, de l'école impériale polytechnique successeur de Mallet-Bachelier, Primière Partie, pp. 19, (1865).

[4] H.S M. Coxeter and S.L. Greitzer, "Geometry Revisited", *Math. Assoc. Amer.* (1967), 107.

[5] D. Eisenbud, M. Green, and J. Harris, "Cayley–Bacharach theorems and conjectures", *American Mathematical Society* **33**(3) (1996), 295–323.

[6] W. Fulton, "Algebraic curves: an introduction to algebraic geometry", *Reverté, s.a.* (1971).

[7] B. Grünbaum and G.C. Shepard, "Ceva, Menelaus, and the area principle", *Math. Mag.* **68** (1995), 254–268.

[8] Z. Luo, "An invariant of algebraic curves from the Pascal theorem", *School of Software, School of Mathematical Sciences, Dalian University of Technology, Dalian, 116024, China*, arXiv: 1201.1344v1, 2007.

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

Birch and Swinnerton-Dyer conjecture: old and new

*Óscar Rivero-Salgado
Universitat Politècnica de Catalunya
oscar.rivero@upc.edu

*Corresponding author

Resum *(CAT)*
La conjectura de Birch i Swinnerton-Dyer (BSD) és un dels sis problemes del mil·lenni que encara no s'ha resolt. Tot i que es va formular després de diferents experiments numèrics, hi ha diverses raons teòriques i analogies amb objectes matemàtics més senzills que ens fan pensar que l'enunciat és cert. Repassarem primer algunes d'aquestes motivacions i explicarem els resultats i generalitzacions més rellevants que es coneixen. A la darrera part ens apropem al mon en el que el rang analític és dos, una situació poc treballada, i ens trobem així amb la conjectura el·líptica de Stark, molt relacionada amb BSD.

Abstract *(ENG)*
The Birch and Swinnerton-Dyer (BSD) conjecture is one of the millennium problems that has not been solved yet. Although it was formulated after different numerical experiments, there are several theoretical reasons and analogies with simpler mathematical objects that lead us to believe that it is true. We go through some of these analogies, and at the same time, we explain the most relevant results and generalizations that are currently known. At the end, we move to the rank two situation, recovering the elliptic Stark conjecture, closely related to BSD.

Keywords: *BSD, elliptic curve, L-series, modular forms, Gross–Zagier Stark conjecture.*
MSC (2010): *11G18, 14G35.*
Received: *April 17, 2017.*
Accepted: *October 28, 2018.*

Acknowledgement
The author thanks V. Rotger for his advice during the realization of this project, which is part of his bachelor thesis. He also thanks the referee for comments that contributed to improve the quality of this presentation.

1. Introduction

During the first half of the 1960s, mathematicians Bryan Birch and Peter Swinnerton-Dyer formulated, after having performed different numerical computations, a conjecture relating the rank of an elliptic curve over a number field with the order of vanishing of the corresponding L-series at the point $s = 1$, lying outside the region of convergence of the defining series. The conjecture asserts that the rank of the elliptic curve (also called the algebraic rank) agrees with the order of vanishing of the L-series at $s = 1$ (the analytic rank), and moreover, it gives a formula for the first non-zero coefficient in the Taylor development of this L-series. At that time, not much knowledge concerning the theory of elliptic curves was available. In the thirties, Mordell had proved the finiteness of the rank for an elliptic curve defined over $\mathbb{Q}$, and then Weil generalized the proof to the case of number fields. It was also Weil who did a more detailed study of the L-series attached to a projective variety, which allowed to have a broaden perspective of the real meaning of this analytic object. There was a lot of progress along the 20-th century, that culminated with the proof of the modularity theorem, first by Taylor and Taylor–Wiles for the case of elliptic curves over $\mathbb{Q}$ with semistable reduction, and then in the general case (but again only over $\mathbb{Q}$) by Breuil, Conrad, Diamond and Taylor. This theorem allows to attach to an elliptic curve a normalized modular form of weight two with the same L-series, which turns out to be useful in many different settings (for instance, to prove the analytic continuation and the functional equation of the L-series of the elliptic curve).

However, in spite of all this great progress in number theory, the conjecture of Birch and Swinnerton-Dyer remains unsolved and only some special cases have been proved. The most remarkable result was obtained by Gross–Zagier and Kolyvagin, who proved the conjecture in analytic rank at most one. For that, they made use of what is known as an Euler system, a compatible collection of cohomology classes along a tower of fields. In this case, it is the Euler system of Heegner points. However, Heegner points are futile in analytic rank greater than one since they are torsion. In those settings, new tools based on p-adic methods have recently been introduced and the interested reader is referred to [1, 2, 3, 7, 9] for a wider perspective.

The organization of this note is as follows. First of all, Section 2 gives a motivation for the conjecture based on the parallelism with the finiteness results available for the group of units of a number field. Our last aim in this section is to present the statement of the conjecture. Then, in Section 3 we state some of the most interesting results and generalizations of the BSD conjecture, particularly the so-called equivariant BSD. Finally, Section 4 explores some of the new insights when the analytic rank is greater than one, and in particular we recover the elliptic Stark conjecture, where the parallelism between units in number fields and points in elliptic curves is again present. For some results about elliptic curves, L-series and modular curves that we freely use along the exposition, we refer to the excellent book [4].

2. Motivation for the BSD conjecture

2.1 First analogies

There are two main reason that make the BSD conjecture specially appealing at first sight: it can be seen as a local-global principle, and at the same time, it makes a link between the algebraic side (the rank of the elliptic curve) and the analytic side (the L-function). We go more carefully through each of these points:

(i) The BSD conjecture can be understood as a local-global principle. The L-function is an analytic object obtained by gluing different pieces that are constructed just counting points over finite fields. From it, we expect to derive some properties about the global behavior (the rank of the elliptic curve over $\mathbb{Q}$ or more generally over a number field). This follows the spirit of Hasse–Minkowski theorem, which states that a homogeneous quadratic form represents 0 over $\mathbb{Q}$ if and only if it represents 0 in all the completions of the rational numbers (the real numbers $\mathbb{R}$ and the p-adic fields $\mathbb{Q}_p$, for the different rational primes p).

However, we know that the Hasse principle is not true for cubic curves, as Selmer showed with his celebrated example $3x^3 + 4y^3 + 5z^3 = 0$, so in general we cannot expect a generalization of this result. This would be something similar to expect that one can prove that a polynomial with integer coefficients is irreducible over $\mathbb{Q}$ just by knowing that is irreducible over all the finite fields $\mathbb{F}_p$. Some examples showing that this is false are the polynomials $x^4 + 1$ or $x^4 - 10x^2 + 1$ (in particular, any irreducible degree four polynomial whose Galois group over $\mathbb{Q}$ is $\mathbb{Z}/2\mathbb{Z} \times \mathbb{Z}/2\mathbb{Z}$, or more generally any irreducible degree n polynomial whose Galois group does not contain an n-cycle). The point is that there is an ubiquitous group, the Tate–Shafarevich group, that measures the failure to the Hasse principle, and we expect that for elliptic curves this group is finite.

(ii) The BSD conjecture gives a relation between the algebraic or geometric side (the rank of the elliptic curve) and the analytic side (the L-series). This gives a connection with another remarkable conjecture, the Bloch–Kato conjecture, that also establishes a link between the rank of vanishing of an L-series and the dimension of appropriate cohomology groups. In particular, we will see that when trying to generalize the Gross–Zagier formula to analytic rank greater than one we necessarily pass through some construction of families of compatible cohomology classes, for which some *explicit reciprocity laws* connecting these classes with appropriate L-functions are available. It turns out that proving this kind of equalities is easier in the p-adic world than in the complex one.

As we have suggested, elliptic curves over a number field are not as easy to understand as one may expect at first sight, so it is natural to look for *simpler* analogies, such as the *ring of integers of a number field*. There are two main remarkable results that come from Minkowski's theorem, that asserts that a set in $\mathbb{R}^n$ which is big enough must contain a rational point: these two results are the finiteness of the class number and the finite generation of the group of units. We would like to look for analogues in the case of an elliptic curve:

(i) The analogue of the rank of the group of units is the rank of the elliptic curve. Both of them are known to be finite and in fact the proof of the Mordell–Weil theorem makes use of the classical result for number fields. This last analogy is specially relevant. We will see how the L-series of the number field encodes information about the rank of the group of units, and we expect the same for elliptic curves via the BSD conjecture.

(ii) One may think that the natural analogue of the class group is the Picard group, that can be defined for any ringed space X as the first cohomology group $H^1(X, \mathcal{O}_X^\times)$. In the case of curves, it turns out to be isomorphic to the jacobian of the curve, that for an elliptic curve is the elliptic curve itself. However, there is another object that we later present, the *Tate–Shafarevich group*, which turns out to be a more appropriate analogue. However, proving its finiteness is equally hard and again, results are limited to situation of analytic rank at most one.

2.2 The L-function of an elliptic curve

L-series are analytic functions attached to motives, which are essentially pieces of the cohomology of a variety over a field. In particular, when M is a motive over $\mathbb{Q}$ with coefficients in a field $E \subset \mathbb{C}$, we may attach to it the complex L-function $L(M,s) = \prod_p P_{M,p}(p^{-s})^{-1}$, where the product runs over all rational primes and $P_{M,p}$ is the characteristic polynomial (suitably normalized) of the Frobenius at p, Frob_p, acting on the motive M of weight w. It converges on $\Re(s) > 1 + w/2$.

The easiest example is concerned with a Dirichlet character $\chi\colon (\mathbb{Z}/N\mathbb{Z})^\times \to \mathbb{C}^\times$. In this case, the identification $(\mathbb{Z}/N\mathbb{Z})^\times \simeq \mathrm{Gal}(\mathbb{Q}(e^{2\pi i/N})/\mathbb{Q})$ allows us to work from the perspective of Galois representations. In a natural way, one can construct $\rho_\chi : G_\mathbb{Q} \to \mathbb{C}^\times \simeq \mathrm{GL}_1(\mathbb{C})$ as the composition of the projection $G_\mathbb{Q} \to \mathrm{Gal}(\mathbb{Q}(e^{2\pi i/N})/\mathbb{Q})$ with the character χ. With this approach, the local factor $P_{M,p}(p^{-s})$ is nothing but the characteristic polynomial of $\rho_\chi(\mathrm{Frob}_p)$, for a choice of Frob_p. Alternatively, this local factor corresponds to the Frobenius acting on the p-torsion of $\mathbb{C}^\times$ (the multiplicative group generated by ζ_p), and it is just $1/(1-\chi(p)p^{-s})$ (as a Galois character, χ acts as $\zeta_p \mapsto \zeta_p^{\chi(p)}$). Then, we obtain the L-function

$$L(\chi, s) = \prod_{p \nmid N} \frac{1}{1-\chi(p)p^{-s}},$$

that in the particular case that χ is identically one agrees with the usual zeta-function. This function can be analytically continued to the whole complex plane (this is a classical result in complex analysis).

This same study can be done for a general number field, and then the corresponding L-function encodes information about all the primes of that number field. In particular, given a number field K and $\rho\colon G_K \to \mathrm{GL}(V)$,

$$L(\rho, s) = \prod_{\mathfrak{p} \subset \mathcal{O}_K} \frac{1}{P_{\rho,\mathfrak{p}}(\mathbb{N}_{K/\mathbb{Q}}(\mathfrak{p})^{-s})}.$$

When ρ is the trivial representation, we obtain the usual Dedekind zeta function of the number field K, ζ_K. It converges absolutely for $\Re(s) > 1$ and it extends to a meromorphic function defined for all complex numbers s and with a simple pole at $s = 1$. The following result is a wonderful analogy for BSD, since we can see how in some sense the value at $s = 1$ allows us to recover arithmetic information. This is the so-called class number formula:

$$\lim_{s\to 1}(s-1)\cdot\zeta_K(s) = \frac{2^{r_1}\cdot(2\pi)^{r_2}\cdot h_K\cdot \mathrm{Reg}_K}{w_K\cdot\sqrt{|D_K|}},$$

where we have made use of the usual conventions of writing r_1, r_2 for the number of real embeddings and half of the complex embeddings of K, respectively; h_K for the class number; Reg_K for the regulator of the number field; w_K for the number of roots of unity in K and D_K for the discriminant of $K/\mathbb{Q}$. Many cases of this result can be more deeply analyzed in the realm of class field theory.

In the case of an elliptic curve (say over $\mathbb{Q}$), the way to introduce the L-function is to consider the Galois action over the so-called Tate module. The construction we describe may seem ad-hoc but it really works in a more general framework and can be extended to more general algebraic varieties. Recall that in the number field case we have used the p-torsion of $\mathbb{C}^\times$ (the group generated by ζ_p), so now we can consider the p-torsion of an elliptic curve. Take $E/\mathbb{Q}$ an elliptic curve. When p is a prime of good reduction (the cubic curve modulo p is still non-singular), it turns out that the p-torsion over $\bar{\mathbb{Q}}$ (denoted by $E[p]$) is a finite group isomorphic to $\mathbb{Z}/p\mathbb{Z} \times \mathbb{Z}/p\mathbb{Z}$ which respects the action of the Galois group, so we have a morphism $\rho_E : G_{\mathbb{Q}} \to \mathrm{GL}_2(\mathbb{Z}/p\mathbb{Z})$. Since this can also be done for $E[p^n]$, taking projective limits we get a representation (that we denote with the same letter) $\rho_E : G_{\mathbb{Q}} \to \mathrm{GL}_2(\mathbb{Z}_p)$. This projective limit is called the Tate module.

It is quite easy to prove (see [13, Ch. 5] for more details) that the characteristic polynomial of Frob_p acting on the p-torsion evaluated at p^{-s} is $1 - a_p(E)p^{-s} + p^{1-2s}$ for the primes of good reduction. Then,

$$L(E, s) = \prod_{p \text{ good}} \frac{1}{1 - a_p p^{-s} + p^{1-2s}} \prod_{p \text{ bad}} \frac{1}{1 - a_p p^{-s}},$$

where the extra factors correspond to the primes of bad reduction. A priori, it may not be clear that this function could be analytically continued, and in fact this is a consequence of the work of Wiles and others towards the proof of the modularity theorem. The idea is based on the introduction of the so-called modular forms, that are functions on the upper half-plane satisfying certain transformation properties, the so-called *modular forms*. The symbol $S_2(N)$ arises for the *weight two* modular forms of level N. vanishing at infinity.

Theorem 2.1 (Modularity)**.** *Let E be an elliptic curve over $\mathbb{Q}$ of conductor N. Then, there exists a modular form $f \in S_2(N)$ such that $L(E, s) = L(f, s)$.*

This means that the coefficients $a_p(E) = p + 1 - \sharp E(\mathbb{F}_p)$ agree with the Fourier coefficients of the modular form f.

Corollary 2.2. *The L-function $L(E, s)$ has an analytic continuation and an integral representation of the form*

$$(2\pi)^{-s}\Gamma(s)L(E, s) = \int_0^\infty f(it)t^{s-1}dt.$$

2.3 Curves of genus zero

Before giving our first formulation of BSD, we can try to study another analogy for curves of genus zero. In particular, let us investigate some local-global properties for conics in a different setting. Consider for instance the circle $x^2 + y^2 = 1$ and count solutions modulo a certain prime. These solutions can be parameterized following the usual method of considering lines through a fixed point of the conic. That way, we get that all the solutions are given in terms of a variable t in the form

$$(x, y) = \Big(\frac{t^2 - 1}{t^2 + 1}, \frac{2t}{t^2 + 1}\Big),$$

so the number of solutions is $p-1$ or $p+1$, since t^2+1 has either zero or two solutions modulo p depending on the residue of p modulo 4 (this works for odd p). Then, we can consider the proportion between the number of points modulo a certain prime and the size of the prime, and multiplying all the quotients we directly get

$$\prod_p \frac{p}{N_p} = \sum_{n=1}^{\infty} \frac{(-1)^{n+1}}{2n - 1} = \frac{\pi}{4},$$

where the last equality follows from Wallis formula. If we denote by $N_{\mathbb{R}}$ the measure of the "real solutions" (the length of the unit circle), we get a very curious result:

$$\prod_p \frac{N_p}{p} \cdot N_{\mathbb{R}} = \frac{4}{\pi} \cdot 2\pi = 8,$$

that is precisely twice the number of integer solutions of $x^2 + y^2 = 1$.

Observe that for counting the number of points over $\mathbb{F}_p$ in the case of elliptic curves, an heuristic argument (that may seem naïve at first sight) is the following. Taking the Weierstrass equation $y^2 = x^3 + Ax + B =: f(x)$, for each value of x modulo p we can obtain either that $f(x) = 0$ (one solution), that $f(x)$ is a non-zero square (two solutions) or that it is a non-square (zero solutions). Since these two last cases occur the same number of times, we expect in average $p + 1$ solutions (considering the point at infinity). The truth is that Hasse's bound is precisely $a_p = |p + 1 - \#E(\mathbb{F}_p)| < 2\sqrt{p}$, which can be seen as some kind of Riemann hypothesis for elliptic curves, since from here we may define a certain zeta function, and Hasse's bound implies that the zeros of this zeta function has real part 1/2. Then, it makes sense again to consider the quantity

$$f(T) = \prod_{p \leq T} \frac{N_p}{p}.$$

One of the first versions preceding BSD was the following one:

Conjecture 2.3. *For each elliptic curve E over $\mathbb{Q}$, there exists a constant C such that $\lim_{T\to+\infty} f(T) = C \cdot \log(T)^r$, where r is the rank of the elliptic curve. Roughly speaking, "many points over the different $\mathbb{F}_p$ force many points over $\mathbb{Q}$".*

2.4 The BSD conjecture

At this point of the discussion, we are in conditions of presenting the extended version of the BSD conjecture. However, the reader should note that it is enough to prove that the rank of $E/\mathbb{Q}$ equals the order of vanishing at $s = 1$ of $L(E, s)$ in order to receive the prize of the Clay Mathematical Institute.

Conjecture 2.4. *Let r be the rank of $E(\mathbb{Q})$ and $P_1, \dots, P_r$ be linearly independent elements of $E(\mathbb{Q})$. Then,*

$$\lim_{s\to 1} \frac{L(E, s)}{(s-1)^r} = \Big(\Omega \prod_{p\,\mathrm{bad}} c_p\Big) \frac{\mathrm{Sha}(E/\mathbb{Q}) \det(\langle P_i, P_j\rangle)}{(\#E_{\mathrm{tors}})^2},$$

where $\Omega = \int_{E(\mathbb{R})} |\omega|$ is the integral of the canonical differential; c_p corresponds to the bad reduction prime p raised to some explicit power and $\mathrm{Sha}(E/\mathbb{Q})$ is the order of the Tate–Shafarevich group.

This Tate–Shafarevich group is an important actor in the different versions of BSD and measures the failure to the Hasse principle. To properly introduce it, let us define it together with the n-th Selmer group, $S^{(n)}(E/\mathbb{Q})$.

$$\begin{aligned} S^{(n)}(E/\mathbb{Q}) &= \{\gamma \in H^1(\mathbb{Q}, E[n]) \mid \text{ for all } p, \gamma_p \text{ comes from } E(\mathbb{Q}_p)\} \\ &= \ker\Big(H^1(\mathbb{Q}, E[n]) \to \prod_{p=2,3,\dots,\infty} H^1(\mathbb{Q}_p, E)\Big), \end{aligned}$$

$$\mathrm{Sha}(E/\mathbb{Q}) = \ker\Big(H^1(\mathbb{Q}, E) \to \prod_{p=2,3,\dots,\infty} H^1(\mathbb{Q}_p, E)\Big).$$

These two groups are related via the short exact sequence

$$0 \to E(\mathbb{Q})/nE(\mathbb{Q}) \to S^{(n)}(E/\mathbb{Q}) \to \mathrm{Sha}(E/\mathbb{Q})[n] \to 0.$$

It is conjectured that $\mathrm{Sha}(E/\mathbb{Q})$ is finite. This group turns out to appear in many other situations and in general Selmer groups are a powerful tool very related with the Euler systems we will briefly present at the end of our discussion. They also arise in the formulation of the Iwasawa main conjecture, another milestone in number theory.

3. Results and generalizations

3.1 The Gross–Zagier, Kolyvagin theorem

Our next aim is to explore some of the known results around BSD conjecture. In 1976, Coates and Wiles proved that the conjecture was true in analytic rank zero (that is, when $L(E,1) \neq 0$, the algebraic rank is zero) for elliptic curves with complex multiplication (informally, curves with many endomorphisms).

The most remarkable result about BSD was proved by Gross–Zagier and Kolyvagin.

Theorem 3.1 (Gross–Zagier, Kolyvagin)**.** *Let E be an elliptic curve over $\mathbb{Q}$. Then,*

(i) if $L(E,1) \neq 0$, then $\#E(\mathbb{Q}) < \infty$ (the algebraic rank is zero);

(ii) if $L(E,1) = 0$ and $L'(E,1) \neq 0$, then the algebraic rank of the elliptic curve is one and there is an efficient method for calculating $E(\mathbb{Q})$.

In both cases $\mathrm{Sha}(E/\mathbb{Q})$ is finite.

The proof of this result requires the introduction of an extremely powerful tool, the so-called Heegner points. In general, for an imaginary quadratic extension K of $\mathbb{Q}$, we write H_n for the ring class field of K of conductor n. A Heegner system attached to (E,K) is a collection of points $P_n \in E(H_n)$ indexed by integers n prime to N satisfying certain (explicit) norm compatibility properties. When (E,K) satisfies the Heegner hypothesis (that is, all primes dividing the conductor of E split in $K/\mathbb{Q}$), there is a non-trivial Heegner system attached to (E,K). Let $\{P_n\}_n$ be a Heegner system and let $P_K = \mathrm{Trace}_{H_1/K}(P_1) \in E(K)$. More generally, consider $\chi : \mathrm{Gal}(H_n/K) \to \mathbb{C}^\times$ a primitive character of a ring class field extension of K of conductor n and let

$$P_n^\chi = \sum_{\sigma \in \mathrm{Gal}(H_n/K)} \bar{\chi}(\sigma) P_n^\sigma \in E(H_n) \otimes \mathbb{C}.$$

The following formula provides the relation between the Heegner system $\{P_n\}$ and the special values of the complex L-series $L(E/K,s)$ and its twists.

Theorem 3.2. *Let $\langle , \rangle_n$ be the canonical Néron–Tate height on $E(H_n)$ extended by linearity to a Hermitian pairing on $E(H_n) \otimes \mathbb{C}$. Then,*

(i) $\langle P_K, P_K \rangle = *L'(E/K,1)$;

(ii) $\langle P_n^\chi, P_n^{\bar{\chi}} \rangle = *L'(E/K,\chi,1)$.

Here, $*$ means equality up to a non-zero factor that can be explicitly described.

The remarkable fact in this story is that a non-trivial Heegner system, beyond yielding lower bounds on the size of the Mordell–Weil group of E over ring class fields of K, also leads to upper bounds on the Mordell–Weil group and the Shafarevich–Tate group of E/K.

Theorem 3.3. *Let $\{P_n\}$ be a Heegner system attached to (E, K). If P_K is non-torsion, then*

(i) the Mordell–Weil group $E(K)$ is of rank one, so that P_K generates a finite-index subgroup of $E(K)$;

(ii) the Shafarevich–Tate group of E/K is finite.

With these ingredients, the proof of the theorem of Gross–Zagier and Kolyvagin is relatively easy; see [4, Ch. 10].

3.2 The equivariant BSD conjecture

As it occurs in many cases, the formulation of a stronger version of the problem can help to clarify some questions around it. Let us explain now the so-called equivariant BSD conjecture. Let $K/\mathbb{Q}$ be a finite Galois extension. Then, by the Mordell–Weil theorem $E(K)\otimes\mathbb{C}\cong\oplus V_i^{r_i}$, where $r=\sum r_i\dim(V_i)<\infty$. That is, we have decomposed a representation into the sum of irreducible representations. Consider for the sake of clarity $K=\mathbb{Q}(\sqrt{-D})$; the Galois group has just a non-trivial element, say χ. Then,

$$E(K)\otimes\mathbb{C}=V_1^{r_1}\oplus V_\chi^{r_\chi}=(E(\mathbb{Q})\otimes\mathbb{C})\oplus(E(K)^\chi\otimes\mathbb{C}),$$

where the last summand is the set of elements v in $E(K)\otimes\mathbb{C}$ such that $\bar{v}=-v$. Observe that if $P\in E(K)$, then $P+\chi(P)\in E(\mathbb{Q})$ and $P-\chi(P)\in E(K)^\chi$.

We can mimic this decomposition for the L-series and express

$$L(E/K,s)=\prod_i L(E/K,V_i,s)^{\dim(V_i)},$$

in such a way that $\operatorname{ord}_{s=1}L(E/K,s)=\sum\operatorname{ord}_{s=1}L(E/K,V_i,s)\dim(V_i)$ (this V_i in the L-function refers to the twist by a certain given representation).

Again, the simplest instance of this phenomenon is the quadratic case. There, $L(E/K,\chi,s)$ can be seen as the L-function of what is called a quadratic twist of E, an elliptic curve that is isomorphic to E not over $\mathbb{Q}$, but over the quadratic extension K. For instance, the elliptic curves $y^2=x^3-x$ and $2y^2=x^3-x$ are not isomorphic over $\mathbb{Q}$, but over $\mathbb{Q}(\sqrt{2})$. Then, if D is the discriminant of the extension, we denote by $E^D: Dy^2=x^3+Ax+B$. We give a brief explanation of why in this case $L(E/K,s)=L(E,s)L(E^D,s)$, by comparing the local factors at p, where p is a prime of good reduction.

Let n_p be the number of points of E and m_p the number of points of E^D; write $a_p=p+1-n_p$ and $b_p=p+1-m_p$. When p splits in K, $Dy^2=f(x)$ has the same number of solutions than $y^2=f(x)$. Then, $a_p=b_p$ and each of the primes contributes to the L-function of the curve over K with the same factor, that is present once both in the L-function of E and E^D.

In the inert case, it is easy to check that $n_p+m_p=2+2p$ and hence $a_p+b_p=0$. Then,

$$(1-a_pp^{-s}+p^{1-2s})(1-b_pp^{-s}+p^{1-2s})=1+2p^{1-2s}+p^{2-4s}+(-a_p^2)p^{-2s}.$$

Taking into account that when p is inert its norm is p^2, what we have is that the inverse of the local factor is $1-a_{p^2}p^{-2s}+p^{2-4s}$ and everything gets reduced to proving that $a_{p^2}=a_p^2-2p$, which can be deduced from standard properties of L-series of elliptic curves; see again [13, §5.2].

Conjecture 3.4 (Equivariant BSD)**.** *With the previous notations,* $\mathrm{ord}_{s=1} L(E/K, V_i, s) = \dim(V_i)$.

Let us briefly comment some of the cases in which this equivariant version of BSD has been proved in analytic rank zero:

(i) ρ is the odd self-dual two-dimensional Galois representation induced from a ring class (or dihedral) character of an imaginary quadratic field; this follows from the work of Gross–Zagier and Kolyvagin;

(ii) ρ is a Dirichlet character; this follows from the work of Kato;

(iii) ρ is an odd irreducible two-dimensional Galois representation satisfying mild restrictions;

(iv) ρ is an irreducible constituent of the tensor product of two odd irreducible two dimensional Galois representations which is self-dual and satisfies some other mild restrictions.

4. Rank two and beyond

One of the challenges when working with BSD it to produce tools that allow us to obtain new results in analytic rank 2 or greater. Moreover, given a Galois representation ρ with underlying vector space V_ρ defined over a number field L, when the analytic rank of $E \otimes \rho$ is positive we have the objective of constructing non-zero elements in $E(H)_L^\rho := \sum_\phi \phi(V_\rho)$, where ϕ runs over a basis of $\mathrm{Hom}_{G_\mathbb{Q}}(V_\rho, E(H) \otimes L)$. We are going to point out some of the new directions trying to emphasize new insights that can help to a better understanding of the problem.

A great progress came with the use of p-adic methods. In 1986, Mazur, Tate and Teitelbaum published "On p-adic analogues of the conjectures of Birch and Swinnerton-Dyer" [11]. They comment in the introduction that since the p-adic analogue of the Hasse–Weil L-function had been defined and also p-adic theories analogous to the theory of the canonical height had been introduced, "it seemed to us to be an appropriate time to embark on the project of formulating a p-adic analogue of the conjecture of Birch and Swinnerton-Dyer, and gathering numerical data in its support [...] The project has proved to be anything but routine".

The first surprising aspect is the appearance of a factor that they call p-adic multiplier, which is a simple local term not equal to any recognizable Euler factor. It can vanish at the central point and throw off the order of vanishing of the p-adic L-function at that point (exceptional case). We expect that in the exceptional case the order of vanishing of the p-adic L-function is one greater than the order of vanishing of the classical L-function. This agrees with the fact that when E has split multiplicative reduction at p, one can define in a natural way the extended Mordell–Weil group, the rank of which is one greater than the rank of the usual Mordell–Weil group. Further, we can define a p-adic height pairing on this extended Mordell–Weil group. In this case, the formulation of BSD involves the regulator of the extended Mordell–Weil group. In this setting, we can produce a conjectural relationship between the special value of the first derivative of the p-adic L-function of E and the algebraic part of the special value of the classical L-function of E. It turns out that the former quantity is the product of the latter and the factor

$$\mathcal{L}_p(E) = \log_p(q_p(E))/\,\mathrm{ord}_p(q_p(E)),$$

where $q_p(E)$ is the p-adic multiplicative period of E. The quantity $\mathcal{L}_p(E)$ is known as the L-invariant of E.

One of the most popular techniques to deal with these p-adic conjectures is the use of Euler systems. Roughly speaking, they are collection of compatible elements of Galois cohomology classes indexed by towers of fields. The most well-known examples are cyclotomic and elliptic units, but also Heegner points, that has been previously presented (more precisely, Heegner points are an example of *anticyclotomic* Euler system). For the study of the rank two setting, the Euler systems that are more useful are those envisaged by Kato, also known as systems of Rankin–Selberg–Garrett type. They consist on the image under étale or syntomic regulators of certain cycles occurring in the higher Chow groups (or K-groups) of modular curves. At the same time, the complex L-function is replaced by its p-adic counterpart. This p-adic L-functions are usually constructed via the interpolation of classical L-values, divided by suitable period, along a certain *interpolation region*.

There are several *explicit reciprocity laws* relating the Perrin–Riou big-logarithm (a map interpolating the dual exponential map and the Bloch–Kato logarithm) of the cohomology classes with special values of p-adic L-functions lying outside the region of non-classical interpolation. This kind of results put a link between the algebraic and analytic side, that is at the end what aims the BSD conjecture. The most *surprising* idea is that in the p-adic setting we may consider certain L-function where we do not only have the usual variable s, but a weight variable which makes the modular form f to vary in a continuous way (we move the modular form along what is called a Hida family).

We continue now by introducing the elliptic Stark conjecture. It is a more constructive alternative to BSD, allowing the efficient computation of p-adic logarithms of global points. In some sense, it can be seen as trying to unify two of the currently known constructions of global points on elliptic curves over $\mathbb{Q}$, Heegner points and the conjectural Stark–Heegner points attached to real quadratic cycles on $\mathbb{H}_p \times \mathbb{H}$. Stark's conjectures give complex analytic formulas for units in number fields (their logarithms) in term of the leading terms of Artin L-functions at $s = 0$. Let $g = \sum a_n(g)q^n$ be a cusp form of weight one, level N and odd character χ. Consider also H_g, the field cut out by an Artin representation ρ_g and $L \subset \mathbb{Q}(\zeta_n)$, the field generated by the Fourier coefficients of g. We denote by V_g the vector space underlying ρ_g. In this framework, Stark's conjecture states the following.

Conjecture 4.1 (Stark)**.** *Let g be a cuspidal newform of weight one with Fourier coefficient in L. Then, there is a modular unit $u_g \in (\mathcal{O}_{H_g}^\times \otimes L)^{\sigma_\infty=1}$ (where σ_∞ stands for the complex conjugation) such that $L'(g, 0) = \log(u_g)$.*

There are some cases (the reducible one, the imaginary dihedral case), where it has been proved. The general result is still unknown to be true.

In [5], the authors formulate some kind of analogue in the realm of points in elliptic curves. The motivation for all this work came for the previous results around Katz's p-adic L-function, the Mazur–Swinnerton-Dyer p-adic L-function and in general, the various types of p-adic Rankin L-functions. Let E be an elliptic curve attached to $f \in S_2(N)$. We introduce the following notations, where χ is a Dirichlet character modulo N, with N relatively prime with a fixed p; more details can be found in [5]:

(i) $M_k(Np, \chi)$ is the space of classical modular forms of weight k, level Np and character χ;

(ii) $M_k^{(p)}(N, \chi)$ is the corresponding space of p-adic modular forms;

(iii) $M_k^{\mathrm{oc}}(N, \chi)$ is the subspace of overconvergent modular forms, a p-adic Banach space where the Hecke operator U_p acts completely continuously. It satisfies $M_k(Np, \chi) \subset M_k^{\mathrm{oc}}(N, \chi) \subset M_k^{(p)}(N, \chi)$.

Coleman's theorem asserts that when h is overconvergent and ordinary of weight ≥ 2, then it is classical;

(iv) $d = q\,d/dq$ is the Atkin–Serre d operator on p-adic modular forms;

(v) when $f \in M_2^{\mathrm{oc}}(N)$, then $F := d^{-1}f \in M_0^{\mathrm{oc}}(N)$, where the d^{-1} refers to the limit of d^t when t tends p-adically to -1;

(vi) $e_{\mathrm{ord}} := \lim_n U_p^{n!}$ is Hida's ordinary projection.

Let $\gamma \in M_k(Np,\chi)^\vee$ and $h \in M_k(N,\chi)$. We define the so-called p-adic iterated integral of f and h along γ as $\int_\gamma f \cdot h := \gamma(e_{\mathrm{ord}}(F \times h)) \in \mathbb{C}_p$. Our aim would be to give an arithmetic interpretation for $\int_{\gamma_{g_\alpha}} f \cdot h$ as $\gamma_{g_\alpha} \in M_1(Np,\chi)^\vee[g_\alpha]$, where this notation refers to elements having the same system of Hecke eigenvalues as g_α. This *integral* can be recast as a special value of a *triple product* p-adic L-function. For the sake of simplicity we must do some assumptions:

(i) certain local signs in the functional equation for $L(E, V_{gh}, s)$ are 1, and in particular $\mathrm{ord}_{s=1} L(E, V_{gh}, s)$ is even;

(ii) $V_{gh} = V_1 \oplus V_2 \oplus W$, where $\mathrm{ord}_{s=1} L(E, V_1, s) = \mathrm{ord}_{s=1} L(E, V_2, s) = 1$ and $L(E, W, 1) \neq 0$. BSD predicts that V_1 and V_2 occur in $E(H_{gh}) \otimes L$ with multiplicity one;

(iii) the geometric Frobenius acts on V_1 (V_2) with eigenvalue $\alpha_g\alpha_h$ ($\alpha_g\beta_h$). Here, α_g and β_g (resp. α_h and β_h) stand for the roots of the p-th Hecke polynomial of the modular form.

Conjecture 4.2 (Elliptic Stark). *Under the above conditions,*

$$\int_{\gamma_{g_\alpha}} f \cdot h = \frac{\log_{E,p}(P_1)\log_{E,p}(P_2)}{\log_p u_{g_\alpha}},$$

where $P_j \in V_j$-isotypic component of $E(H_{gh}) \otimes L$ and $\sigma_p P_1 = \alpha_g\alpha_h \cdot P_1$, $\sigma_p P_2 = \alpha_g\beta_h \cdot P_2$. Further, u_{g_α} is a Stark unit in the $\mathrm{Ad}^0(V_g)$-isotypical part of $(\mathcal{O}_{H_g}^\times) \otimes L$ and $\sigma_p u_{g_\alpha} = (\alpha_g/\beta_g) \cdot u_{g_\alpha}$ (that is coherent with the fact that the Frobenius must act in the left hand side trivially).

The result has been proved by Darmon, Lauder and Rotger when g and h are theta series attached to the same imaginary quadratic field K and the prime p splits in K. In that setting, P_1 and P_2 are expressed in terms of Heegner points and u_{g_α} in terms of elliptic units. The assumption that p is split is crucial for the use of both Katz's p-adic Kronecker limit formula and also for the p-adic Gross–Zagier formula of Bertolini, Darmon and Prasanna.

This conjecture is adapted in [6] to express it in the setting of units in number fields, where some of the self-duality assumptions can be relaxed. In particular, the conjecture is rephrased in terms of the special value $L_p(g \otimes h, 1)$, where g and h are two weight one modular forms and $L_p(g \otimes h, s)$ is the Hida–Rankin p-adic L-function attached to the convolution of two modular forms. This value is expected to encode information about units and p-units in the field cut out by the Galois representation attached to $g \otimes h$. In [12], this conjecture is proved when g and h are self-dual, and there is a further study of the question via the Euler system of Beilinson–Flach elements constructed in [2, 3, 10]. The setting of points in elliptic curves is treated in [8] using the families of cohomology classes of [9].

As a way of finishing this survey about the conjecture of Birch and Swinnerton-Dyer, we would like to emphasize the idea that mathematicians have not still envisaged a successful approach to the problem useful for its general proof, but many interesting ideas not only for this area but for many others have emerged in the last years. In particular, those ideas concerning Euler systems and p-adic methods have been successfully applied to many other instances, such as the study of the Iwasawa main conjecture for elliptic curves.

References

[1] M. Bertolini, F. Castella, H. Darmon, S. Dasgupta, K. Prasanna, and V. Rotger. "p-adic L-functions and Euler systems: a tale in two trilogies", *Automorphic forms and Galois representations*, Vol. 1, LMS Lecture Notes **414** Cambridge University Press (2014) 52–102.

[2] M. Bertolini, H. Darmon, and V. Rotger, "Beilinson–Flach elements and Euler systems I: syntomic regulators and p-adic Rankin L-series", *J. Algebraic Geometry* **24** (2015), 355–378.

[3] M. Bertolini, H. Darmon, and V. Rotger, "Beilinson–Flach elements and Euler systems II: p-adic families and the Birch and Swinnerton-Dyer conjecture", *J. Algebraic Geometry* **24** (2015), 569–604.

[4] H. Darmon. "Rational points on modular elliptic curves", American Mathematical Society, 2004.

[5] H. Darmon, A. Lauder, and V. Rotger, "Stark points and p-adic iterated integrals attached to modular forms of weight one", *Forum of Mathematics, Pi* **3** (2015), 95 pages.

[6] H. Darmon, A. Lauder, and V. Rotger, "Gross-Stark units and p-adic iterated integrals attached to modular forms of weight one", *Ann. Math. Québec* **40** (2016), 325–354.

[7] H. Darmon and V. Rotger, "Diagonal cycles and Euler systems I: a p-adic Gross-Zagier formula", *Annales Scientifiques de l'Ecole Normale Supérieure* **47**(4) (2014), 779–832.

[8] H. Darmon and V. Rotger, "Elliptic curves of rank two and generalised Kato classes", accepted at *Research in Math. Sciences*, special issue in memory of Robert Coleman, (2016).

[9] H. Darmon and V. Rotger, "Diagonal cycles and Euler systems II: the Birch and Swinnerton-Dyer conjecture for Hasse-Weil-Artin L-series", *Journal of the American Mathematical Society* **30**(3) (2017), 601–672.

[10] G. Kings, D. Loeffler, and S.L. Zerbes, "Rankin–Eisenstein classes and explicit reciprocity laws", *Cambridge J. Math* **5**(1) (2017), 1–122.

[11] B. Mazur, J. Tate, and J. Teitelbaum, "On p-adic analogues of the conjecture of Birch and Swinnerton-Dyer", *Inventiones Matematicae* **84** (1986), 1–49.

[12] O. Rivero and V. Rotger, "Derived Beilinson–Flach elements and the arithmetic of the adjoint of a modular form", preprint available at Arxiv.

[13] J. Silvermann, "The arithmetic of elliptic curves", *Graduate Texts in Mathematics* (Springer), 1986.

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

Lefschetz properties in algebra and geometry

*Martí Salat Moltó
Universitat de Barcelona
msalatmo7@alumnes.ub.edu
∗Corresponding author

Resum *(CAT)*
La propietat dèbil de Lefschetz (WLP) té un paper important tant a l'àlgebra com a la geometria. A [3], Mezzetti, Miró-Roig i Ottaviani van demostrar que el fet que certs ideals Artinians de $\mathbb{K}[x_0, \dots, x_n]$ fallin la WLP té relació amb l'existència de projeccions de la varietat de Veronese que satisfan una equació de Laplace. Aquest vincle dóna lloc a la definició de sistema de Togliatti. En aquest article, enunciarem alguns resultats recents obtinguts sobre el tema. En particular exposem la classificació dels sistemes de Togliatti minimals i llisos generats per $2n + 3$ monomis de grau $d \geq 10$ obtinguts a [8].

Abstract *(ENG)*
The weak Lefschetz property (WLP) plays an important role both in algebra and geometry. In [3], Mezzetti, Miró-Roig and Ottaviani found that the failure of the WLP for some particular Artinian ideals in $\mathbb{K}[x_0, \dots, x_n]$ is related with the existence of projections of the Veronese variety satisfying one Laplace equation. This relation gives rise to the definition of Togliatti system. In this note, we state some recent results on this topic. In particular, we expose the classification of minimal smooth Togliatti systems generated by $2n + 3$ monomials of degree $d \geq 10$ obtained in [8].

Keywords: *weak Lefschetz property, Togliatti systems, Laplace equations.*
MSC (2010): *Primary 13E10, 14N15. Secondary 14M25, 53A20.*
Received: *September 27, 2018.*
Accepted: *March 19, 2019.*

Acknowledgement
The author would like to thank professor R.M. Miró-Roig for her advisoring task. Her suggestions and comments have been an invaluable gift. The comments of an anonymous referee have also been very useful.

Societat Catalana de Matemàtiques
Institut d'Estudis Catalans

1. Introduction

The weak and strong Lefschetz properties on graded Artinian algebras have been an object of study along the last few decades. We say that a graded Artinian algebra $A = \bigoplus_i A_i$ has the strong Lefschetz property (SLP) if the multiplication by a dth power of a general linear form has maximal rank (i.e., $\times L^d \colon A_i \to A_{i+d}$ is either injective or surjective for every i). In particular, we say that A has the weak Lefschetz property (WLP) if the same occurs for $d = 1$. These properties have connections among different areas such as algebraic geometry, commutative algebra and combinatorics. Sometimes quite surprisingly, these connections give new approaches to other problems, which are a priori are unrelated.

The study of the Lefschetz properties started in 1980 with the work Stanley [9], which reached the following result:

Proposition 1.1. *Let $R = \mathbb{K}[x_0, \dots, x_n]$ be the polynomial ring in n variables, and $I = (x_0^{a_0}, \dots, x_n^{a_n}) \subset R$ be an Artinian monomial complete intersection. Let $L \in R_1$ be a general linear form. Then, for any positive integers d and i, the homomorphism $\times L^d \colon [R/I]_i \to [R/I]_{i+d}$ (induced by multiplication by L^d) has maximal rank.*

Afterwards, Watanabe [12] continued this research connecting the Lefschetz properties to the Sperner theory in combinatorics. Later more connections between the Lefschetz properties and vector bundles, line arrangements on the plane or the Fröberg conjecture have been discovered (see, for instance, [6, 7]). In this note, however, we will focus on another connection based on the so-called Togliatti systems.

In Mezzetti–Miró-Roig–Ottaviani [3], the authors related the failure of the weak Lefschetz property of Artinian ideals to the existence of projective varieties satisfying at least one Laplace equation. This connection (see Proposition 2.10) between a pure algebraic notion and a differential geometry concept gives rise to Togliatti systems (see Definition 2.12), an important family of Artinian ideals generated by homogeneous forms of the same degree d failing the WLP in degree $d-1$.

Let $\mathbb{K}$ be an algebraically closed field of characteristic 0 and $R = \mathbb{K}[x_0, \dots, x_n]$. Given an ideal $I \subset R$ generated by homogeneous forms of the same degree d, there is no difference between $[R/I]_i$ and R_i whenever $i \le d-1$. Therefore, the lowest possible degree in which R/I can fail the WLP is $d-1$. Thus, Togliatti systems are precisely Artinian ideals I failing the WLP in the very first possible degree: from $d-1$ to d. Equivalently in [3, Thm. 3.2] it was proved that Togliatti systems give rise to projections of the Veronese variety $V(n, d)$ satisfying at least one Laplace equation of order $d-1$.

After reviewing this connection, the rest of this note is devoted to study the classification of minimal (smooth) monomial Togliatti systems (see 2.12) and present some recent results found in [8]. The complete classification of Togliatti systems is still an open problem. However, if we restrict our attention to *monomial* Togliatti systems a lot of combinatorial tools emerge and the picture becomes clearer (see, for instance, [2, 3]). In [4] the complete classification of minimal smooth monomial Togliatti systems of quadrics and cubics, was given. This classification uses graph theory and other combinatoric tools in its proof, and cannot be easily generalized to classify all minimal smooth monomial Togliatti systems of degree $d \ge 4$. In order to overcome this difficulty and address the classification problem for an arbitrary degree $d \ge 4$, a new strategy was proposed in [2]. First of all, upper and lower bounds on the number of generators $\mu(I)$ of a minimal monomial Togliatti system are given. Namely, if I is a minimal Togliatti system in R generated by monomials of degree $d \ge 4$, then $2n+1 \le \mu(I) \le \binom{n+d-1}{n}$ where $n \ge 2$ and $d \ge 4$. A second step consists of classifying all smooth monomial Togliatti systems reaching this lower bound or close to it; see Remark 3.3. In [2], a complete classification of all minimal smooth Togliatti systems generated by $2n+1 \leq \mu(I) \leq 2n+2$ monomials, was achieved. In particular, it was shown that, except a few cases when $n = 2$, these Togliatti systems have a very peculiar shape. However, this shape is not directly generalizable to Togliatti systems with $\mu(I) \geq 2n+3$. Actually, even when $\mu(I) = 2n+3$ very different situations occur. For instance, there is no *smooth* minimal Togliatti system generated by $2n+3$ monomials of degree $d \geq 4$ whenever $n \geq 3$; see Proposition 3.4. Therefore we focus the study on minimal Togliatti systems in three variables generated by 7 monomials of degree $d \geq 4$. In [8] a complete classification of all the minimal Togliatti systems in $\mathbb{K}[x, y, z]$ generated by 7 monomials (see Theorem 3.7) was given. As a corollary, a complete classification of *smooth* minimal Togliatti systems in R generated by $2n+3$ monomials was achieved. Section 3 is devoted to present and motivate these two recent results.

2. Preliminaries

This section is devoted to recall all the definitions related to the Lefschetz properties and Laplace equations, and a review of the connection between those two notions and the definition of Togliatti systems.

Definition 2.1. Let $I \subset R$ be an Artinian ideal and let us consider $A = R/I$ with the standard graduation $A = \bigoplus_{i=0}^{r} A_i$. Let $L \in R_1$ be a general linear form. Then:

(i) A has the strong Lefschetz property (SLP) if, for all positive integer d and for all $1 \leq i \leq r-d$, the homomorphism $\times L^d \colon [A]_i \to [A]_{i+d}$ has maximal rank;

(ii) A has the weak Lefschetz property (WLP) if, for all $1 \leq i \leq r-1$, the homomorphism $\times L \colon [A]_i \to [A]_{i+1}$ has maximal rank.

Remark 2.2. It is clear that having the SLP implies having the WLP, however, the converse is not true. For instance, it can be proved that $I = (x_0^2, x_1^3, x_2^5, x_0x_1, x_0x_2^2, x_1x_2^3, x_1^2x_2^2)$ has the WLP but fails the SLP in degrees 2 and 1, and also that $I = (x_0^3, x_1^3, x_2^3, (x_0+x_1+x_2)^3)$ has the WLP but fails the SLP in degrees 3 and 1.

In this note we will focus on the failure of the weak Lefschetz property for Artinian ideals. Let us first see some examples:

Example 2.3. (i) $I = (x^3, y^3, z^3, xyz)$ fails the WLP in degree 2.

(ii) $I = (x^4, y^4, z^4, t^4, xyzt)$ fails the WLP in degree 5.

(iii) By [5, Thm. 4.3], the ideals $I = (x_0^{n+1}, \dots, x_n^{n+1}, x_0 \dots, x_n)$ fail the WLP in degree $\binom{n+1}{2} - 1$.

Remark 2.4. Notice that the first ideal fails the WLP in the first non trivial place while the others fail later. This particularity will be studied in the sequel in more detail.

Even though an Artinian ideal I is expected to have the WLP, establishing this property for a concrete family of Artinian ideals can be a hard problem. For instance, Stanley [9] and Watanabe [12] proved that a *general* Artinian complete intersection has the WLP. However, to see wether *every* Artinian complete intersection with codimension ≥ 4 has the WLP remains an open problem.

In the last decades there have been established multiple connections between Lefschetz properties and other areas of mathematics, such as combinatorics, representation theory or geometry; see, for instance, [3, 9, 12]. In particular, Mezzetti–Miró-Roig–Ottaviani [3] connected the failure of the WLP with rational varieties satisfying Laplace equations. Let us recall some differential geometry definitions.

Definition 2.5. Let $X \subset \mathbb{P}^N$ be a rational variety of dimension n with parametrization $\Psi\colon \mathbb{P}^n \dashrightarrow X$, $(t_0 : \ldots : t_n) \mapsto (F_0(t_0, \ldots, t_n) : \ldots : F_N(t_0, \ldots, t_n))$. We call *s-th osculating vector space* on $x = \Psi(t_0 : \ldots : t_n)$ the vector space

$$T_x^{(s)} X := \left\langle \frac{\partial^s \Psi}{\partial_{t_0}^{k_0} \ldots \partial_{t_n}^{k_n}}(t_0 : \ldots : t_n) | k_0 + \cdots + k_n = s \right\rangle.$$

Finally, we call *s-th osculating projective space* on $x \in X$ the projectivization of the vector space above: $\mathbb{T}_x^{(s)} X := \mathbb{P}\left(T_x^{(s)} X\right)$.

Remark 2.6. There are $\binom{n+s}{s} - 1$ vectors $(k_0, \ldots, k_n)$ satisfying $k_0 + \cdots + k_n = s$. Then, in a general point $x \in X$, the *expected dimension* of $T_x^{(s)} X$ is $\binom{n+s}{s} - 1$. However, if there are linear dependencies among the partial derivatives of order s, this bound is not reached. In this case, Ψ satisfies a linear partial differential equation of order s. This motivates the following definition.

Definition 2.7. Let $X \subset \mathbb{P}^N$ be a rational projective variety of dimension n. We say that X *satisfy δ Laplace equations of order s* if

(i) for all smooth point $x \in X$ we have $\dim T_x^{(s)} X < \binom{n+s}{s} - 1$, and

(ii) for general $x \in X$, $\dim T_x^{(s)} X = \binom{n+s}{s} - 1 - \delta$.

Remark 2.8. If $N < \binom{n+s}{s} - 1$, then $T_x^{(s)} X$ is spanned by more vectors than the ambiance space. So, X trivially satisfies at least one Laplace equation of order s.

Let us now restrict our attention to Artinian ideals generated by homogeneous forms of the same degree d. To these ideals we can associate two different rational varieties:

Definition 2.9. Let $I = (F_1, \ldots, F_r) \subset R$ be an Artinian ideal generated by r forms of degree d. Let I^{-1} be the ideal generated by the inverse Macaulay system of I; see, for instance, [3, § 3]. Consider $\phi_{[I^{-1}]_d}\colon \mathbb{P}^n \dashrightarrow \mathbb{P}^{\binom{n+d}{d}-r-1}$ to be the rational map associated to $[I^{-1}]_d$ and $\phi_{I_d}\colon \mathbb{P}^n \to \mathbb{P}^{r-1}$ to be the morphism (I is Artinian) associated to I_d. We define $X_{n,[I^{-1}]_d} := \overline{\operatorname{Im}(\phi_{[I^{-1}]_d})}$, which is the projection of the d-th Veronese variety $V(n, d)$ from $\langle F_1, \ldots, F_r\rangle$, and $X_{n,I_d} := \operatorname{Im}(\phi_{I_d})$, which is the projection of $V(n, d)$ from $\langle [I^{-1}]_d\rangle$.

Finally, we can establish the following important relation.

Proposition 2.10 (Mezzetti–Miró-Roig–Ottaviani, [3, Thm. 3.2]). *Let $I \subset R$ be an Artinian ideal generated by r forms $F_1, \ldots, F_r$ of degree d. If $r \leq \binom{n+d-1}{n-1}$, then the following conditions are equivalent:*

(a) the ideal I fails the WLP in degree $d - 1$;

(b) the forms $F_1, \ldots, F_r$ become $k-$linearly dependent on a general hyperplane $H \subset \mathbb{P}^n$; and

(c) the n dimensional variety $X_{n,[I^{-1}]_d}$ satisfies at least one Laplace equation of order $d - 1$.

Remark 2.11. (i) The bound on the number of generators ensures the possible failure of the WLP is due to injectivity.

(ii) For $1 \leq i \leq d-2$, since $[R/I]_i \cong R_i$, the homomorphism $\times L\colon [R/I]_i \to [R/I]_{i+1}$ is injective. Hence, the first possible degree where $\times L$ can fail to have maximum rank is precisely $d-1$.

This result motivates the following definitions.

Definition 2.12. Let $I = (F_1, \dots, F_r) \subset R$ be an Artinian ideal generated by $r \leq \binom{n+d-1}{n-1}$ forms of degree d. We say that I is a *Togliatti system* if it satisfies any of the three equivalent conditions from Proposition 2.10. Moreover, we say that

(i) I is a *minimal Togliatti system* if for any $1 \leq s \leq r$ and $\{F_{i_1}, \dots, F_{i_s}\} \subset \{F_1, \dots, F_r\}$, $I' = (F_{i_1}, \dots, F_{i_s})$ is not a Togliatti system;

(ii) I is a *monomial Togliatti system* if I can be generated by monomials;

(iii) I is a *smooth Togliatti system* if $X_{n,[I^{-1}]_d}$ is a smooth variety.

Remark 2.13. (i) The name was given in honor to E. Togliatti who proved that the only smooth Togliatti system of cubics is $I = (x_0^3, x_1^3, x_2^3, x_0x_1x_2)$; see, for instance, [10, 11].

(ii) To address the classification of Togliatti systems it is crucial to investigate when they are minimal. In the next section we will focus on the minimality of a Togliatti system.

(iii) If I is generated by r monomials of degree d, then I^{-1} is generated by all the $\binom{n+d}{d} - r$ monomials of degree d which do not generate I. In particular, $X_{n,I}$ and $X_{n,[I^{-1}]_d}$ are two closely related toric varieties. These varieties carry a lot of combinatorial properties which ease their study as can be seen in [2, 3, 4, 8]. Henceforward, we will restrict our attention to (smooth) monomial Togliatti systems.

This preliminary section ends with a result that shows why dealing with monomial ideals simplifies the study of Lefschetz properties.

Proposition 2.14 (Migliore–Miró-Roig–Nagel, [5, Prop. 2.2])**.** *Let $I \subset R$ be an Artinian monomial ideal. Then R/I has the WLP if and only if $x_0 + \cdots + x_n$ is a Lefschetz element for R/I.*

Therefore, to check the WLP for monomial Artinian ideals there is no need to study the multiplication by a *general* linear form L, but only for the particular form $L = x_0 + \cdots + x_n$.

3. Classification of monomial Togliatti systems

In this section we will address the problem of classifying (smooth) monomial Togliatti systems. We will start by taking account of all the most recent results achieved, all of them, in the present decade. Finally we will establish two new results that enlarge the classification of (smooth) monomial Togliatti systems. Furthermore, these new results provide a huge amount of new examples of monomial Togliatti systems which are non trivial in the sense of Proposition 3.2.

In Michałek–Miró-Roig [4], the authors completely classified all smooth monomial Togliatti systems of quadrics and cubics using graph theory. However, the classification problem of (smooth) monomial

Togliatti systems generated by forms of degree $d \geq 4$ becomes much more involved and, up to now, a complete classification seems out of reach. Therefore, Mezzetti and Miró-Roig changed the strategy in [2] and focused on studying the number of generators of a (smooth) monomial Togliatti system. Moreover, they gave upper and lower bounds of this number of generators, and classified all the smooth monomial Togliatti systems near the lower bound. Before stating the main results of this note, let us fix some notation and review some motivational results from [2].

Definition 3.1. For every $n, d \in \mathbb{N}$, we denote by $\mathcal{T}(n,d)$ the set of all minimal monomial Togliatti systems, and by $\mathcal{T}^s(n,d)$ the set of all minimal smooth monomial Togliatti systems. Furthermore, we write

(i) $\mu(n,d) = \min\{\mu(I) | I \in \mathcal{T}(n,d)\}$,

(ii) $\mu^s(n,d) = \min\{\mu(I) | I \in \mathcal{T}^s(n,d)\}$,

where $\mu(I)$ stands for the minimal number of generators of an ideal $I \subset \mathbb{K}[x_0, \dots, x_n]$.

With this notation, we start establishing lower bounds for the quantities $\mu(n,d)$ and $\mu^s(n,d)$. The first result in this direction gives rise to a very important family of Togliatti systems:

Proposition 3.2. *Let $n, d \in \mathbb{N}$ and let m be a monomial of degree $d-1$. Then, the ideal $(x_0^d, \dots, x_n^d) + m(x_0, \dots, x_n)$ is a minimal monomial Togliatti system. These minimal monomial Togliatti systems are called* trivial Togliatti systems*.*

Proof. By Propositions 2.14 and 2.10, it is enough to restrict the generators of I to the hyperplane $x_0 + \dots + x_n$, and see that they become linearly dependent; see, for instance, [2, Rem. 3.8]. □

Remark 3.3. (i) Let I be trivial Togliatti system given by a monomial $m = x_0^{a_0} \cdots x_n^{a_n}$ of degree $d-1$. Then, the n dimensional variety $X_{n,[I^{-1}]_d}$ parametrized by $\Psi = \phi_{[I^{-1}]_d}$ (see Definition 2.9) satisfies a very simple Laplace equation: $\partial^{d-1}/\partial^{a_0} \cdots \partial^{a_n} \Psi = 0$.

(ii) Observe that a trivial Togliatti system I satisfies that $2n+1 \leq \mu(I) \leq 2n+2$. In fact, Mezzetti and Miró-Roig showed in [2] that for $n \geq 2$ and $d \geq 4$, $\mu(n,d) = \mu^s(n,d) = 2n+1$. Furthermore, they proved that, for $d \gg 4$, all minimal monomial smooth Togliatti systems of forms of degree d with $2n+1 \leq \mu(I) \leq 2n+2$ are trivial.

(iii) Of course, for minimal monomial Togliatti systems I with $\mu(I) \geq 2n+3$ we cannot expect them to be trivial. In particular, its study gives rise to geometrically more interesting examples.

The remaining of this section is devoted to study the classification of (smooth) monomial Togliatti systems generated by $2n+3$ forms of degree d. Let us start the discussion by stating the following result:

Proposition 3.4 (Mezzetti–Miró-Roig, [2, Prop. 4.4])**.** *Let $n \geq 3$ and $d \geq 4$. Then, there is no $I \in \mathcal{T}^s(n,d)$ with $\mu(I) = 2n+3$.*

This result significantly reduces the task of classifying all minimal smooth monomial Togliatti systems I generated by $2n+3$ forms of degree $d \geq 4$. Namely, it will be enough to consider the three variables case. Moreover, the hypothesis of Proposition 2.10 implies that a Togliatti system in $\mathbb{K}[x,y,z]$ generated by 7 forms of degree d must satisfy that $7 \leq \binom{2+d-1}{2-1} = d+1$.

In other words, our goal now is shifted to classify all minimal smooth monomial Togliatti systems in $\mathbb{K}[x,y,z]$ generated by forms of degree $d \geq 6$. To establish this classification we need to fix some notation.

Definition 3.5. Let us denote the ideal $T = (x^3, y^3, z^3, xyz)$ and the following sets of monomial ideals:

$$A = \{(y^3, y^2z, yz^2, z^3), (xy^2, xz^2, y^3, z^3), (x^2y, y^3, y^2z, z^3), (x^2z, y^3, y^2z, z^3), (xz^2, y^3, y^2z, z^3), (xz^2, y^3, y^2z, yz^2), (x^2z, y^3, y^2z, yz^2), (xyz, xz^2, y^3, yz^2), (xy^2, xz^2, y^3, yz^2), (xyz, xz^2, y^3, y^2z), (xy^2, xz^2, y^2z, yz^2), (x^2z, xy^2, y^2z, yz^2), (x^2z, xz^2, y^3, y^2z), (x^2z, xz^2, y^3, yz^2), (x^2y, xy^2, y^3, z^3), (x^2z, xy^2, y^3, z^3), (x^2z, xyz, y^3, y^2z), (x^2z, xyz, y^3, yz^2), (x^2y, xz^2, y^3, y^2z), (x^2y, xz^2, y^3, yz^2), (x^2z, xy^2, y^3, yz^2)\},$$

$$B = \{(x^3z, xy^2z, y^4, yz^3), (x^2yz, xz^3, y^4, y^3z), (x^2z^2, xy^2z, y^4, z^4), (x^2yz, y^4, y^2z^2, z^4)\},$$

and

$$C = \{(x^3yz, xy^2z^2, y^5, z^5), (x^2yz^2, xy^3z, y^5, z^5)\}.$$

Finally, for any $d \geq 1$ integer, let be $M(d) := \{x^ay^bz^c \mid d-1 \geq a, b, c \geq 0,\ a+b+c = d\}$.

This definition gives rise to the first examples of minimal Togliatti systems generated by 7 monomials of degree $d \geq 6$ in three variables.

Proposition 3.6. *Let $d \geq 6$. Then any of the following ideals is a minimal Togliatti system:*

(i) both (a) $I = (x^d, y^d, z^d) + m(x^2, y^2, xz, yz)$ and (b) $I = (x^d, y^d, z^d) + m(x^2, y^2, xy, z^2)$, for every $m \in M(d-2)$;

(ii) $I = (x^d, y^d, z^d) + x^{d-3}J$, for any $J \in A$;

(iii) $I = (x^d, y^d, z^d) + mT$, for all $m \in M(d-3)$;

(iv) $I = (x^d, y^d, z^d) + x^{d-4}J$, for every $J \in B$;

(v) $I = (x^d, y^d, z^d) + x^{d-5}J$, for any $J \in C$.

Proof. It is a strightforward computation to show that restricting the generators of each type of ideal to the linear form $x+y+z$ they become linearly dependent. Then, all of them are Togliatti systems according to Proposition 2.10. On the other hand, none of these ideals contain a Togliatti system generated by either 5 or 6 monomials. Hence, they are minimal. □

Moreover, as we will see in the next results, *almost* all minimal Togliatti systems in three variables generated by 7 monomials of degree $d \geq 6$ are of one of the types mentioned in Proposition 3.6. These results can be seen as a natural generalization of those obtained in Remark 3.3. The first result classifies all minimal Togliatti systems in three variables generated by 7 monomials of degree $d \geq 10$.

Theorem 3.7 (Miró-Roig–Salat, [8, Thm. 3.8])**.** *Let $I \subset \mathbb{K}[x, y, z]$ be a minimal Togliatti system generated by 7 monomials of degree $d \geq 10$. Then, up to a permutation of the variables, one of the following cases hold:*

(i) there is $m \in M(d-2)$ such that either (a) $I = (x^d, y^d, z^d) + m(x^2, y^2, xz, yz)$ or (b) $I = (x^d, y^d, z^d) + m(x^2, y^2, xy, z^2)$;

(ii) there is $J \in A$ such that $I = (x^d, y^d, z^d) + x^{d-3}J$;

(iii) *there is* $m \in M(d-3)$ *such that* $I = (x^d, y^d, z^d) + mT$*;*

(iv) *there is* $J \in B$ *such that* $I = (x^d, y^d, z^d) + x^{d-4}J$*;*

(v) *there is* $J \in C$ *such that* $I = (x^d, y^d, z^d) + x^{d-5}J$*.*

From this result, and using combinatorial properties of toric varieties, a complete classification of smooth minimal Togliatti systems in $n+1$ variables generated by $2n+3$ monomials of degree $d \geq 10$ can be given.

Set $M^0(d) = \{x_0^a x_1^b x_2^c \mid a+b+c = d \text{ and } a, b, c \geq 1\}$ for any integer $d \geq 3$. Then we can establish the following theorem regarding *smooth* Togliatti systems.

Theorem 3.8 (Miró-Roig–Salat, [8, Thm. 3.9])**.** *Let* $I \subset \mathbb{K}[x_0, \dots, x_n]$ *be a smooth minimal monomial Togliatti system of forms of degree* $d \geq 10$*. Assume that* $\mu(I) = 2n+3$*. Then* $n = 2$ *and, up to permutation of the coordinates, one of the following cases holds:*

(i) $I = (x_0^d, x_1^d, x_2^d) + m(x_0^2, x_1^2, x_0x_2, x_1x_2)$ *with* $m \in M^0(d-2)$*;*

(ii) $I = (x_0^d, x_1^d, x_2^d) + m(x_0^2, x_1^2, x_0x_1, x_2^2)$ *with* $m \in M^0(d-2)$*;*

(iii) $I = (x_0^d, x_1^d, x_2^d) + m(x_0^3, x_1^3, x_2^3, x_0x_1, x_2)$ *with* $m \in M^0(d-3)$*.*

Remark 3.9. For $6 \leq d \leq 9$ there are other examples of minimal monomial Togliatti systems $I = (x^d, y^d, z^d) + J \subset \mathbb{K}[x, y, z]$, generated by 7 monomials, which are not covered by Theorem 3.7. Using Macaulay2 software [1], we computed all of these additional ideals J, up to permutation of the variables: for $d = 6$,

$$
\begin{gathered}
(x^5y, x^3z^3, x^2y^3z, y^5z), (x^5z, x^3y^3, x^2y^2z^2, y^5z), (x^3z^3, x^2y^4, x^2y^2z^2, y^5z), (x^5z, x^3y^3, xyz^4, y^5z),\\
(x^4z^2, x^3y^3, x^2y^2z^2, y^4z^2), (x^3z^3, x^2y^4, x^2y^2z^2, y^4z^2), (x^4z^2, x^3y^3, xyz^4, y^4z^2), (x^3y^3, x^3z^3, x^2y^2z^2, y^3z^3),\\
xy(x^4, x^2y^2, xyz^2, y^4), xy(x^3z, x^2y^2, xyz^2, y^3z), xy(x^2y^2, x^2z^2, xyz^2, y^2z^2), xy(x^2y^2, x^2z^2, xz^3, y^2z^2),\\
xy(x^4, xz^3, y^4, y^2z^2), xy(x^4, x^2y^2, y^4, z^4), xy(x^4, xyz^2, y^4, z^4), xy(x^3z, x^2y^2, y^3z, z^4),\\
xz(x^3z, x^2z^2, xyz^2, y^4), xz(x^2yz, x^2z^2, xyz^2, y^4), xz(x^3z, xy^2z, xyz^2, y^4), xz(x^3z, x^2yz, xz^3, y^4),\\
xz(x^3z, x^2z^2, xz^3, y^4), xz(x^3z, xy^2z, xz^3, y^4), xz(x^2y^2, x^2z^2, xy^3, y^3z), xz(x^2y^2, x^2z^2, xy^2z, y^3z),\\
xz(x^2z^2, xy^3, xy^2z, y^3z), xz(x^2y^2, x^2z^2, xy^3, y^4), xz(x^2y^2, x^2z^2, y^4, y^3z), xz(x^2y^2, x^2z^2, y^4, y^2z^2),\\
xz(x^3z, x^2yz, xy^2z, y^4), xz(x^3z, x^2yz, xyz^2, y^4), x(xy^4, xyz^3, xz^4, y^3z^2), x(x^4z, x^2y^3, xy^2z^2, y^5),\\
x(x^4z, xyz^3, y^5, y^3z^2), x(x^2z^3, xy^4, xy^2z^2, y^5), x(x^4z, x^2yz^2, y^5, y^2z^3), x(x^2z^3, xy^4, xyz^3, y^3z^2),\\
x(x^4z, x^2z^3, xy^3z, y^5), x(x^4z, x^2y^3, y^5, yz^4), x(x^3z^2, x^2y^3, xz^4, y^3z^2), x(x^4z, xy^2z^2, y^5, yz^4),\\
x(x^2z^3, xy^4, xz^4, y^3z^2), x(x^2y^3, x^2z^3, y^4z, yz^4), x(x^4y, x^2z^3, xy^3z, y^5), x(x^2yz^2, xy^3z, y^5, z^5);
\end{gathered}
$$

for $d = 7$,

$$
\begin{gathered}
xy(x^2z^3, xy^4, xy^2z^2, y^5), xy(x^5, x^2y^2z, xyz^3, y^5), xy(x^4y, x^3y^2, xz^4, y^3z^2), xy(x^3y^2, x^2y^3, x^2z^3, y^2z^3),\\
xy(x^5, x^2y^2z, y^5, z^5), xy(x^4z, xy^4, y^5, z^5), xy(x^5, xyz^3, y^5, z^5), xz(x^3z^2, x^2z^3, xy^3z, y^5),\\
xz(x^4z, x^2yz^2, xz^4, y^5), xz(x^4z, xy^3z, xz^4, y^5), x(x^5z, x^2y^3z, xy^2z^3, y^6), x(xy^5, xy^2z^3, xz^5, y^4z^2),\\
x(x^5z, x^4y^2, x^2y^2z^2, y^3z^3), x(x^4y^2, x^4z^2, x^2y^2z^2, y^3z^3), x(x^3y^3, x^3z^3, x^2y^2z^2, y^3z^3),\\
x(x^4yz, x^2y^4, x^2z^4, y^3z^3), x(x^2y^4, x^2y^2z^2, x^2z^4, y^3z^3), x(x^4yz, xy^5, xz^5, y^3z^3), x(x^2y^2z^2, xy^5, xz^5, y^3z^3),\\
x(x^5z, x^2y^3z, y^6, yz^5), x(x^5z, xy^2z^3, y^6, yz^5), x(x^5z, x^4y^2, y^5z, yz^5), x(x^4y^2, x^4z^2, y^5z, yz^5),\\
x(x^3y^3, x^3z^3, y^5z, yz^5), x(x^4yz, x^2y^2z^2, y^6, z^6), x(x^4yz, y^6, y^3z^3, z^6), x(x^2y^2z^2, y^6, y^3z^3, z^6),\\
xyz(x^2y^2, x^2z^2, xy^3, y^4), xyz(x^3z, x^2yz, xy^2z, y^4), xyz(x^4, x^2y^2, xyz^2, y^4), xyz(x^3z, x^2yz, xyz^2, y^4),\\
xyz(x^3z, x^2z^2, xyz^2, y^4), xyz(x^2yz, x^2z^2, xyz^2, y^4), xyz(x^3z, xy^2z, xyz^2, y^4), xyz(x^3z, x^2yz, xz^3, y^4),
\end{gathered}
$$

$$xyz(x^3z, x^2z^2, xz^3, y^4), xyz(x^3y, xy^3, xz^3, y^4), xyz(x^3z, xy^2z, xz^3, y^4), xyz(x^2y^2, x^2z^2, xy^3, y^3z),$$
$$xyz(x^2y^2, x^2z^2, xy^2z, y^3z), xyz(x^2z^2, xy^3, xy^2z, y^3z), xyz(x^3z, x^2y^2, xyz^2, y^3z), xyz(x^3y, x^2z^2, xyz^2, y^3z),$$
$$xyz(x^2y^2, x^2z^2, y^4, y^3z), xyz(x^4, xy^3, xz^3, y^2z^2), xyz(x^4, xz^3, y^4, yz^3);$$

for $d = 8$,

$$xy(x^4z^2, x^3y^3, xyz^4, y^4z^2), xz(x^3z^3, x^2y^2z^2, xy^4z, y^6);$$

and for $d = 9$,

$$xyz(x^3z^3, x^2y^2z^2, xy^4z, y^6), xyz(x^3y^3, x^3z^3, x^2y^2z^2, y^3z^3), xyz(x^6, x^2y^2z^2, y^6, z^6).$$

This remark added to Proposition 3.7 completes the classification problem, for any degree $d \geq 6$, of all minimal monomial Togliatti systems in three variables with 7 generators.

References

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

[2] E. Mezzetti and R.R. Miró-Roig, "The minimal number of generators of a Togliatti system", *Annali di Mathematica Pura ed Applicata* **195** (2016), 2077–2098.

[3] E. Mezzetti, R.M. Miró-Roig, and G. Ottaviani, "Laplace equations and the weak Lefschetz property", *Canadian Journal of Mathematics* **65** (2013), 634–654.

[4] E. Michałek and R.M. Miró-Roig, "Smooth monomial Togliatti system of cubics", *J. Combinatorial Theory, Series A* **143** (2016), 66–87.

[5] J. Migliore, R.M. Miró-Roig, and U. Nagel, "Monomial ideals, almost complete intersections and the weak Lefschetz property", *Trans. Amer. Math. Soc.* **363**(1) (2011), 229–257.

[6] J. Migliore and U. Nagel, "A tour of the weak and strong Lefschetz properties".

[7] R.M. Miró-Roig, "Globally generated vector bundles and the weak Lefschetz property", *Arch. Math.* (2014) 103–255.

[8] R.M. Miró-Roig and M. Salat, "On the classification of Togliatti systems", *Comm. in Alg.* **46**(6), 2459–2475.

[9] R.P. Stanley, "Weyl groups, the hard Lefschetz theorem, and the Sperner property", *SIAM J. Algebr. Discrete Methods* **1** (1980), 168–184.

[10] E. Togliatti, "Alcuni esempi di superfici algebriche degli iperspazi che rappresentano un'equazione di Laplace", *Comm. Math. Helvetici* **1** (1929), 255–272.

[11] E. Togliatti, "Alcune osservazzioni sulle superfici razionali che rappresentano equazioni di Laplace", *Ann. Mat. Pura Appl.* **25**(4) (1946), 325–339.

[12] J. Watanabe, "The Dilworth number of Artinian rings and finite posets with rank function, commutative algebra and combinatorics", *Advance Studies in Pure Math* **11**, Kinokuniya Co., Amsterdam (1987).

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

A hybridizable discontinuous Galerkin phase-field model for brittle fracture

***Alba Muixí**
Universitat Politècnica de Catalunya
alba.muixi@upc.edu

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

Antonio Rodríguez-Ferran
Universitat Politècnica de Catalunya
antonio.rodriguez-ferran@upc.edu

∗Corresponding author

Resum *(CAT)*
Els models *phase-field* per a fractura fràgil descriuen les fractures com a zones danyades mitjançant un camp continu que varia abruptament entre els estats intacte i trencat. Per tal de capturar bé la solució, les malles han de ser fines a prop de la fractura. Presentem una formulació *Hybridizable Discontinuous Galerkin* (HDG) per a un model *phase-field* quasi-estàtic, basada en un esquema alternat a l'hora de resoldre el sistema. La motivació per a utilitzar HDG és que permet implementar estratègies d'adaptabilitat de manera senzilla.

Abstract *(ENG)*
Phase-field models for brittle fracture consider smeared representations of cracks, which are described by a continuous field that varies abruptly in the transition zone between unbroken and broken states. Computationally, meshes have to be fine locally near the crack to capture the solution. We present a Hybridizable Discontinuous Galerkin (HDG) formulation for a quasi-static phase-field model, based on a staggered approach to solve the system. The use of HDG for this model is motivated by the suitability of the method for adaptivity.

Keywords: *phase-field, staggered scheme, brittle fracture, hybridizable discontinuous Galerkin (HDG).*
MSC (2010): *35Q74, 74R10, 65M60.*
Received: *January 25, 2019.*
Accepted: *April 30, 2019.*

Acknowledgement
This work was supported by the AGAUR training grant FI-DGR 2017, the DAFOH2 project (Ministerio de Economía y Competitividad, MTM2013-46313-R) and the Generalitat de Catalunya (2017-SGR-1278).

Societat Catalana de Matemàtiques
Institut d'Estudis Catalans

1. Introduction

Models of fracture in brittle materials can be based on discontinuous and continuous descriptions of cracks.

Discontinuous models describe cracks as sharp entities by means of discontinuous displacement fields. The main disadvantage of these models is the lack of a rigorous strategy to determine initiation and propagation of cracks. Numerically, they are usually tackled by the eXtended Finite Element Method (X-FEM), which enables to solve the problem with meshes unfitted to the crack geometry [2, 14]. However, dealing with discontinuities in a X-FEM setting may be cumbersome in cases with complex patterns [18].

Alternatively, phase-field models for fracture represent cracks as damaged regions that have lost their load-carrying capacity, with continuous displacement fields in all the domain [3]. These models introduce an auxiliar field d, called the *phase-field variable* or *damage field*, which differentiates between the broken and unbroken states of the material and varies smoothly between them. The evolution of the phase-field variable as a result of the loading conditions handles naturally the initiation, propagation, branching and coalescence of cracks. Incorporating the crack evolution into the equations is the main advantage of phase-field models over the discontinuous ones.

The phase-field approach introduces a regularization length parameter ℓ, which comes from the smeared representation of the crack and can be related to its width. Since the goal is to approximate a sharp crack, the parameter ℓ is to be chosen small and the phase-field variable d will vary sharply in the damaged zone. Therefore, high spatial resolution is a key requirement to approximate properly the solution. The usual strategy is to refine the computational mesh locally where the crack is expected to propagate: a priori in the cases in which the crack path is known and by remeshing as the phase-field value evolves when it is not. Obviously, this implies a high computational cost. A reasonable approach to reduce the cost is defining an adaptive refinement method. The different strategies proposed in the literature offer an alternative to remeshing, though they are non-trivial; see [15] and the references therein. We refer to [1, 20] for an exhaustive review of existing phase-field models and the numerical challenges they present.

In this work, we use the Hybridizable Discontinuous Galerkin method (HDG) as an alternative to standard FEM to solve the phase-field equations. HDG was first proposed in [5] for second order elliptic problems and, due to its promising properties, has already been formulated for multiple problems, see for example [10, 16, 17].

As any other Discontinuous Galerkin (DG) method, HDG is based on the use of element-by-element discontinuous basis functions in a finite element setting. Continuity of the solution is imposed in weak form by means of numerical fluxes on element boundaries. DG methods are appealing to solve the equations of the phase-field model because of the possiblity of using different approximation bases in neighbouring elements, which will enable the straightforward definition of an adaptive refinement strategy. Among all DG methods, we choose HDG because it involves less degrees of freedom, with a computational efficiency close to standard continuous FE and better convergence properties [11, 21].

In Section 2 we provide a brief overview of the chosen phase-field model for brittle fracture and the staggered scheme to solve it. In Section 3, we present the HDG formulation of the equations. Finally, in Section 4, we compare the results obtained with the HDG formulation with the ones obtained with standard FEM for a benchmark problem, for both low and high-order degrees of approximation. All computations have been done with Matlab.

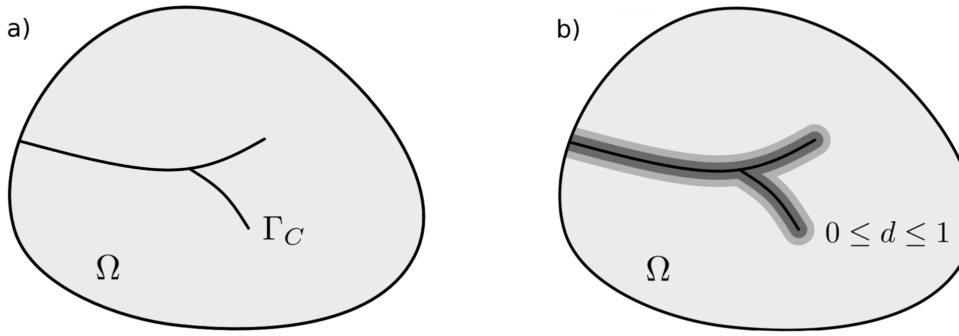

Figure 1: (a) Body with a sharp crack Γ_C. (b) Smeared crack representation.

2. Phase-field modelling of brittle fracture

In this work, we consider the quasi-static phase-field model for fracture proposed by Bourdin–Francfort–Marigo in [3].

In [6], Francfort and Marigo state that the fracture process acts to minimize the total energy of a body, which can be expressed as the sum of its bulk elastic energy and the crack surface energy, that is

$$E(\mathbf{u}, \Gamma_C) = \int_\Omega \Psi_0(\varepsilon) \, \mathrm{d}V + G_c \int_{\Gamma_C} \mathrm{d}s, \tag{1}$$

with Ψ_0 the elastic energy density in the domain Ω and G_C the critical energy release rate for a crack Γ_C; see Fig. 1(a). We restrict ourselves to the case of linear elastic isotropic materials, for which the elastic energy density is given by $\Psi_0(\varepsilon) = (\varepsilon : \mathbf{C} : \varepsilon)/2$, where ε is the standard infinitesimal strain tensor, defined from the displacement $\mathbf{u}$ as $\varepsilon(\mathbf{u}) = \left(\boldsymbol{\nabla}\mathbf{u} + (\boldsymbol{\nabla}\mathbf{u})^T\right)/2$, and $\mathbf{C}$ is the fourth-order elasticity tensor depending on the Lamé parameters. The symbol : denotes the double contraction operator, that is, for instance, $(\mathbf{C} : \varepsilon)_{ij} = \sum_{k,l} C_{ijkl}\varepsilon_{kl}$.

To enable a numerical treatment of (1), Bourdin–Francfort–Marigo [3] introduced a regularized formulation by considering a smeared representation of the sharp crack Γ_C; see Fig. 1(b). The crack is defined by a new field $d(\mathbf{x}, t)$ which varies smoothly between two values representing the unbroken and broken states of the material, 0 and 1 respectively, and is therefore called the *phase-field* or *damage parameter*. The energy functional (1) is then approximated by

$$E_\ell(\mathbf{u}, d) = \int_\Omega \left((1-d)^2 + \eta\right) \Psi_0(\varepsilon) \, \mathrm{d}V + G_c \int_\Omega \left(\frac{d^2}{2\ell} + \frac{\ell}{2}|\nabla d|^2\right) \mathrm{d}V, \tag{2}$$

where ℓ regulates the width of the diffuse crack and η is a small dimensionless parameter to avoid a complete loss of stiffness in broken regions. It has been proved in Bourdin–Francfort–Marigo [4] that with $\ell \to 0$, the regularized functional (2) Γ-converges to (1). This implies that the set $\{d = 1\}$ tends to the sharp crack Γ_C as the width of the smeared representation tends to 0.

Minimizing the energy functional (2) we obtain the system

$$\begin{cases} \nabla \cdot \boldsymbol{\sigma} = \mathbf{0}, \\ -\ell^2 \Delta d + d = \dfrac{2\ell}{G_c}(1-d)\Psi_0, \end{cases} \tag{3}$$

with the stress tensor $\boldsymbol{\sigma}$ defined as

$$\boldsymbol{\sigma}(\mathbf{u}, d) = \left((1-d)^2 + \eta\right) \frac{\partial \Psi_0(\varepsilon)}{\partial \varepsilon} = \left((1-d)^2 + \eta\right) \mathbf{C} : \varepsilon(\mathbf{u}). \tag{4}$$

The resulting system of governing equations is to be solved using an incremental procedure for the loading process. Assuming the solution at load step n is known, in equilibrium with prescribed values $\mathbf{t}^n$ and $\mathbf{u}_D^n$, the system is solved for the load step $n+1$ using the corresponding boundary conditions

$$\begin{cases} \boldsymbol{\sigma} \cdot \mathbf{n} = \mathbf{t}^{n+1} & \text{on } \Gamma_N, \\ \mathbf{u} = \mathbf{u}_D^{n+1} & \text{on } \Gamma_D, \\ \nabla d \cdot \mathbf{n} = 0 & \text{on } \partial\Omega, \end{cases} \tag{5}$$

where $\mathbf{t}^{n+1}$ are prescribed tractions on Γ_N, $\mathbf{u}_D^{n+1}$ are prescribed displacements on Γ_D, with $\Gamma_D \cup \Gamma_N = \partial\Omega$ and $\Gamma_D \cap \Gamma_N = \emptyset$, and $\mathbf{n}$ stands for the outward unit normal vector.

Following Miehe–Hofacker–Welschinger in [12, 13], we replace Ψ_0 in the second equation in (3) by a history field variable $\mathcal{H}$ defined as $\mathcal{H}^{n+1}(\mathbf{x}) = \max_{\tau \in [1, n+1]} \Psi_0^\tau$, to enforce irreversibility of the crack evolution.

2.1 Staggered approach

The total energy (2) is convex with respect to $\mathbf{u}$ and d separately, but not with respect to both of them. This motivates the solution of the system by means of a staggered scheme: at each load step, we compute the displacement field $\mathbf{u}$ and the damage field d alternately until convergence. Given the solution $(\mathbf{u}^n, d^n)$ at load step n, the solution at load step $n+1$ is computed by iterating over i in the following scheme:

(i) Compute $\left[\mathbf{u}^{n+1}\right]^{i+1}$ by solving the equation

$$\nabla \cdot \boldsymbol{\sigma}\left(\left[\mathbf{u}^{n+1}\right]^{i+1}, \left[d^{n+1}\right]^{i}\right) = 0 \text{ in } \Omega, \tag{6}$$

with $\boldsymbol{\sigma}$ given by (4) and boundary conditions $\boldsymbol{\sigma} \cdot \mathbf{n} = \mathbf{t}^{n+1}$ on Γ_N, $\left[\mathbf{u}^{n+1}\right]^{i+1} = \mathbf{u}_D^{n+1}$ on Γ_D.

(ii) Update the history field $\left[\mathcal{H}^{n+1}\right]^{i+1} = \max\left(\mathcal{H}^n, \left[\Psi_0^{n+1}\right]^{i+1}\right)$.

(iii) Compute $\left[d^{n+1}\right]^{i+1}$ by solving

$$-\ell^2 \Delta \left[d^{n+1}\right]^{i+1} + \left[d^{n+1}\right]^{i+1} = \frac{2\ell}{G_c}\left(1 - \left[d^{n+1}\right]^{i+1}\right)\left[\mathcal{H}^{n+1}\right]^{i+1} \text{ in } \Omega, \tag{7}$$

with boundary condition $\left(\nabla \left[d^{n+1}\right]^{i+1}\right) \cdot \mathbf{n} = 0$ on $\partial\Omega$.

We take $(\mathbf{u}^0, d^0)(\mathbf{x}) = (\mathbf{0}, 0)$ for all $\mathbf{x}$ in Ω and $\left[d^{n+1}\right]^0 = d^n$ for $n > 0$. We keep iterating over i until some stopping criterion indicating convergence is satisfied. An alternative is to take sufficiently small load increments and use the staggered approach without iterating, see for example [1, 8], but the speed of propagation of the crack might be underestimated. Here, we iterate until the relative Euclidean norm of the difference of two consecutive iterates is smaller than a fixed tolerance, for both the displacement and damage fields.

Remark 2.1 (Efficiency). The staggered algorithm is simple and has been proved to be robust [12], but many iterations are needed to reach convergence even for simple benchmark problems. A monolithic scheme computing simultaneously both fields would be more efficient, but then one has to deal with the non-convexity of the functional (2) and the Jacobian matrix of the system being indefinite [9, 19].

3. HDG formulation

We aim to use HDG to solve the governing equations of the phase-field model. The adopted staggered approach enables an independent numerical treatment for each of the equations, so we can focus on the HDG formulations for the linear elasticity equilibrium equation (6) and the damage field equation (7). For the former equation, there are various options in the literature. Here, we consider the HDG formulation for linear elasticity by Fu–Cockburn–Stolarski [7] and Soon–Cockburn–Stolarski [17]. For the latter, we add the reaction term to the standard HDG formulation for diffusion by Cockburn–Gopalakrishnan–Lazarov [5]. Both formulations are recalled in this section.

Throughout the section, we assume the domain Ω covered by a finite element mesh with $\mathsf{n_{el}}$ disjoint elements K_i satisfying $\bar{\Omega} \subset \bigcup_{i=1}^{\mathsf{n_{el}}} \bar{K}_i$, $K_i \cap K_j = \emptyset$, for $i \neq j$, and denote the union of the $\mathsf{n_{fc}}$ faces Γ_f of the mesh as $\Gamma = \bigcup_{i=1}^{\mathsf{n_{el}}} \partial K_i = \bigcup_{f=1}^{\mathsf{n_{fc}}} \Gamma_f$.

3.1 HDG for the equilibrium equation

Let us consider the linear elasticity problem defined in (6), to be solved for a frozen damaged field d. The problem can be written in the broken space of elements as a set of local element-by-element equations and some global equations. Local problems impose the linear elasticity equation at each element K_i with Dirichlet boundary conditions, namely

$$\begin{cases} \nabla \cdot \boldsymbol{\sigma}(\mathbf{J}, d) = \mathbf{0} & \text{in } K_i, \\ \mathbf{J} - \boldsymbol{\nabla}\mathbf{u} = \mathbf{0} & \text{in } K_i, \\ \mathbf{u} = \hat{\mathbf{u}} & \text{on } \partial K_i, \end{cases} \tag{8}$$

for $i = 1 \dots \mathsf{n_{el}}$. The variable $\mathbf{J}$ is introduced to split the problem into a system of first order PDE and $\hat{\mathbf{u}}$ is a new trace variable defined on the skeleton of the mesh, Γ, which is single-valued; see Fig. 2. Note that, given $\hat{\mathbf{u}}$, the local problems (8) can be solved to determine $\mathbf{u}$ and $\mathbf{J}$ at each element.

The global problem is stated to determine the trace variable $\hat{\mathbf{u}}$. It imposes equilibrium of tractions on faces and also the boundary conditions, that is,

$$\begin{cases} [\![\boldsymbol{\sigma} \cdot \mathbf{n}]\!] = \mathbf{0} & \text{in } \Gamma \setminus \partial\Omega, \\ \boldsymbol{\sigma} \cdot \mathbf{n} = \mathbf{t}_N & \text{on } \Gamma_N, \\ \hat{\mathbf{u}} = \mathbf{u}_D & \text{on } \Gamma_D, \end{cases} \tag{9}$$

where $[\![\cdot]\!]$ denotes the jump operator defined at a face Γ_f as $[\![\odot]\!] = \odot_{L_f} + \odot_{R_f}$, where L_f and R_f denote the left and right elements sharing the face and $\odot_i$ denotes the value of $\odot$ from element K_i. Note that the continuity of $\mathbf{u}$ across Γ is satisfied due to the boundary condition $\mathbf{u} = \hat{\mathbf{u}}$ in the local problems and the fact that $\hat{\mathbf{u}}$ is single-valued.

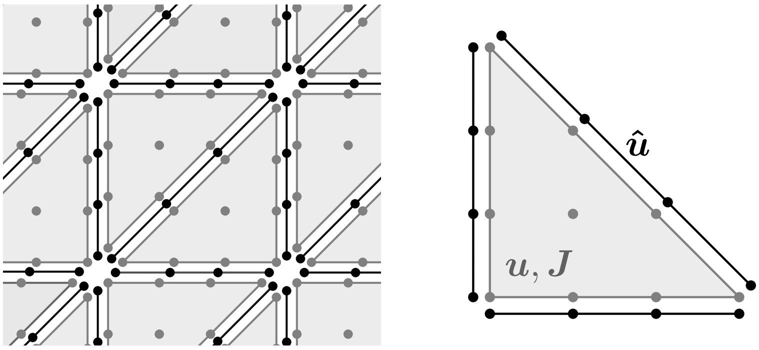

Figure 2: Left: HDG discretization of the domain with the mesh skeleton Γ. Right: detail of the HDG discretization for the local problem in one element.

The HDG formulation of the problem is obtained by the discretization of the local and global equations. To approximate the elemental variables, $\mathbf{u}$ and $\mathbf{J}$, and the trace variable, $\hat{\mathbf{u}}$, the discrete spaces considered are

$$\mathcal{V}^h(\Omega) = \{v \in L^2(\Omega) : v|_{K_i} \in \mathcal{P}_p(K_i) \text{ for } i = 1 \dots \mathrm{n_{el}}\},$$
$$\Lambda^h(\Gamma) = \{\hat{v} \in L^2(\Gamma) : \hat{v}|_{\Gamma_f} \in \mathcal{P}_p(\Gamma_f) \text{ for } f = 1 \dots \mathrm{n_{fc}}\},$$

where $\mathcal{P}_p$ denotes the space of polynomials of degree less or equal to p. To simplify the notation, we use $\mathbf{u}$, $\mathbf{J}$, $\hat{\mathbf{u}}$ to denote both the solutions and their approximations.

For an element K_i, the weak form for the local problem (8) is: given $\hat{\mathbf{u}} \in [\Lambda^h(\Gamma)]^n$, find $\mathbf{u} \in [\mathcal{V}^h(K_i)]^n$, $\mathbf{J} \in [\mathcal{V}^h(K_i)]^{n\times n}$ such that

$$\begin{aligned} &\int_{K_i} \mathbf{v}\cdot(\boldsymbol{\nabla}\cdot\boldsymbol{\sigma})\ \mathrm{d}V + \int_{\partial K_i} \mathbf{v}\cdot((\mathbf{C} : \tau(\hat{\mathbf{u}} - \mathbf{u})\otimes\mathbf{n})\cdot\mathbf{n})\ \mathrm{d}s = 0, \\ &\int_{K_i} \mathbf{Q} : \mathbf{J}\,\mathrm{d}V + \int_{K_i} (\boldsymbol{\nabla}\cdot\mathbf{Q})\cdot\mathbf{u}\,\mathrm{d}V - \int_{\partial K_i} (\mathbf{Q}\cdot\mathbf{n})\cdot\hat{\mathbf{u}}\,\mathrm{d}s = 0, \end{aligned} \tag{10}$$

for all $\mathbf{v} \in [\mathcal{V}^h(K_i)]^n$, for all $\mathbf{Q} \in [\mathcal{V}^h(K_i)]^{n\times n}$. The first equation in (10) is obtained from the first equation in (8) by applying integration by parts, replacing the numerical flux $\hat{\boldsymbol{\sigma}} := \boldsymbol{\sigma} + \mathbf{C} : (\tau(\hat{\mathbf{u}} - \mathbf{u})\otimes\mathbf{n})$ on the boundary and undoing the integration by parts. τ is a nonnegative stabilization parameter, which here is taken as a positive constant on all faces.

The discretization of the local problem (10) leads to the so-called *local solver* for each element K_i, which expresses $\mathbf{u}$ and $\mathbf{J}$ in terms of $\hat{\mathbf{u}}$, namely

$$\mathbf{u}_e = \mathbf{U}^{K_i}\boldsymbol{\Lambda}^i,\ \ \mathbf{J}_e = \mathbf{Q}^{K_i}\boldsymbol{\Lambda}^i, \tag{11}$$

with matrices $\mathbf{U}^{K_i}, \mathbf{Q}^{K_i}$. $\boldsymbol{\Lambda}^i$ is a vector containing the unknown nodal values of $\hat{\mathbf{u}}$ for all the faces of K_i, this is, $\boldsymbol{\Lambda}^i := \left[\hat{\mathbf{u}}^{F_{i,1},T}, \dots, \hat{\mathbf{u}}^{F_{i,m},T}\right]^T$.

For the global problem (9), the weak form is stated replacing $\boldsymbol{\sigma}$ by the numerical flux $\hat{\boldsymbol{\sigma}}$. The weak form is: find $\hat{\mathbf{u}} \in [\Lambda^h(\Gamma)]^n$ such that $\hat{\mathbf{u}} = \mathbb{P}^2(\mathbf{u}_D)$ on Γ_D and

$$\int_\Gamma \hat{\mathbf{v}}\cdot[\![\hat{\boldsymbol{\sigma}}\cdot\mathbf{n}]\!]\,\mathrm{d}s + \int_{\Gamma_N} \hat{\mathbf{v}}\cdot(\hat{\boldsymbol{\sigma}}\cdot\mathbf{n})\ \mathrm{d}s = \int_{\Gamma_N} \hat{\mathbf{v}}\cdot\mathbf{t}_N\,\mathrm{d}s, \tag{12}$$

for all $\hat{\mathbf{v}} \in [\Lambda^h(\Gamma)]^n$ such that $\hat{\mathbf{v}} = \mathbf{0}$ on Γ_D. The function $\mathbb{P}^2$ is the L^2 projection onto the discrete space on Γ_D. Discretizing the global weak form and replacing $\mathbf{u}$ and $\mathbf{J}$ in terms of $\hat{\mathbf{u}}$ by the local solver (11), we get a system for $\hat{\mathbf{u}}$. Once $\hat{\mathbf{u}}$ is determined, using the local solvers (11), we compute $\mathbf{u}$ and $\mathbf{J}$ in every element.

For this formulation, $\mathbf{u}$ converges with order $p+1$ in L^2 norm and $\mathbf{J}$ with order $p+1/2$; see Fu–Cockburn–Stolarski [7].

3.2 HDG for the damage field equation

The HDG formulation for the damage field equation (7) is obtained analogously to the formulation for linear elasticity. Introducing a new variable $\mathbf{q}$ to be the gradient of d, the local problems impose the equation in every element K_i with Dirichlet boundary conditions, and their weak form reads: given $\hat{d} \in \Lambda^h(\Gamma)$, find $d \in \mathcal{V}^h(K_i)$, $\mathbf{q} \in [\mathcal{V}^h(K_i)]^n$ such that

$$
\begin{aligned}
&-\int_{K_i} G_C \ell \, v \cdot \boldsymbol{\nabla} \cdot \mathbf{q} \, \mathrm{d}V - \int_{\partial K_i} G_C \ell \tau \, (\hat{d} - d) v \, \mathrm{d}s + \int_{K_i} \left(\frac{G_C}{\ell} + 2\mathcal{H} \right) v d \, \mathrm{d}V = \int_{K_i} v 2\mathcal{H}, \\
&\int_{K_i} \mathbf{w} \cdot \mathbf{q} \, \mathrm{d}V + \int_{K_i} (\boldsymbol{\nabla} \cdot \mathbf{w}) d \, \mathrm{d}V - \int_{\partial K_i} \mathbf{w} \cdot \mathbf{n} \, \hat{d} \, \mathrm{d}s = 0,
\end{aligned}
$$

for all $v \in \mathcal{V}^h(K_i)$, for all $\mathbf{w} \in [\mathcal{V}^h(K_i)]^n$. The numerical flux prescribed on the boundary of every element is $\hat{\mathbf{q}} := \mathbf{q} + \tau(\hat{d} - d)\mathbf{n}$, with τ the stabilization parameter.

The weak form of the global problem is: find $\hat{d} \in \Lambda^h(\Gamma)$ such that $\int_{\Gamma \setminus \partial\Omega} \hat{v} \cdot [\![\hat{\mathbf{q}} \cdot \mathbf{n}]\!] \, ds = 0$, for all $\hat{v} \in \Lambda^h(\Gamma)$. In this case, both d and $\mathbf{q}$ are proved to converge with order $p+1$ in L^2 norm; see Cockburn–Gopalakrishnan–Lazarov [5].

Remark 3.1 ($\mathcal{H}$ is evaluated at integration points). To solve the damage field equation we need the value of $\mathcal{H}$ at integration points. From the staggered scheme, $\mathcal{H}$ can be computed using the nodal values of $\mathbf{J}$ obtained by solving the equilibrium equation. Evaluating $\mathcal{H}$ at nodes may result in negative values when it is interpolated to integration points if we use approximation functions of degree greater than 1, even though it is a nonnegative function by definition. This leads to unphysical solutions. Also, it may cause the non-convergence of the staggered scheme if consecutive iterates alternate negative and positive values at some points. We will illustrate this behavior with a numerical example in next section. To avoid non-physical negative values $\mathbf{J}$ is interpolated to integration points and then these values are directly used to evaluate $\mathcal{H}$.

4. Numerical example: L-shaped panel test

One of the typical benchmark problems in computational fracture is the L-shaped panel test. Consider the specimen in Fig. 3(left), which is fixed on the bottom and has imposed vertical displacement at a 30 mm distance to the right edge. Following Ambati–Gerasimov–De-Lorenzis [1], the material parameters are $\lambda = 6.16$ kN/mm^2, $\mu = 10.95$ kN/mm^2 and $G_c = 8.9 \cdot 10^{-5}$ kN/mm^2. The regularization length in the phase-field model is taken to be $\ell = 3$ mm and the residual stiffness is $\eta = 10^{-5}$. The stabilization parameter appearing in the HDG formulation of the equations is taken $\tau = 1$ for both the equilibrium and the damage field equations.

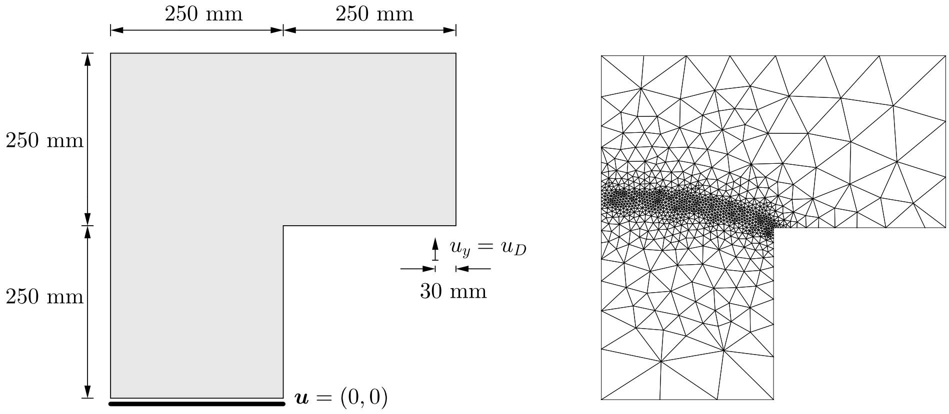

Figure 3: Left: geometry and boundary conditions of the test. Right: computational mesh.

We consider a triangular mesh with 1842 elements, pre-refined along the expected crack path with a mesh size of $h_{ref} = 3.5$ mm, see Fig. 3(right), and four nested meshes to this one obtained by dividing the mesh size by two for each level of refinement. The problem is solved with increments in the prescribed vertical displacement of $\Delta u_D = 10^{-3}$ mm and we iterate in the staggered scheme for each load step until convergence is reached with a tolerance of 10^{-6}.

Remark 4.1 (Boundary conditions). Imposing the vertical displacement at just one point causes unphysical damage near the point. To cancel this out and impose properly the boundary conditions, we set the damage to zero in the region after every iteration of the staggered scheme. Another strategy would be to assign a higher value of G_C where needed; see Yakovlev–Moxey–Kirby–Sherwin [20].

Comparison of FEM and HDG. We start by considering linear approximation functions. As expected, the solution tends to converge when refining the mesh. This can be observed in the load-displacement curves in Fig. 4, that show the evolution of the reaction force for an increasing imposed displacement u_D. We obtain similar results for both FEM and HDG, with slightly better accuracy in HDG. Recall that HDG has a better order of convergence for the gradient of the displacement field $\mathbf{J}$.

Spatial resolution. Using degree of approximation $p = 1$, the primary mesh with $h_{ref} = 3.5$ mm is not fine enough to approximate properly the smeared crack with $\ell = 3$ mm. The smeared crack becomes mesh-dependent and has a width of one element; see Fig. 5. In Fig. 6, we plot the damage field for different imposed displacements of the loading process with the 2-nested level mesh. The crack path obtained in this case is comparable with the results in the literature; see [1, 8].

Computation with high-order approximations. To increase the accuracy in space needed to capture the profile of the solution, one can take higher degree p of the approximation basis functions. With $p = 5$, we expect to obtain more accurate results than with $p = 1$ for the same mesh. Indeed, in Figure 7(left), we compare the load-displacement curve obtained with degree $p = 1$ and the 4-nested level mesh with the curves obtained for $p = 5$ and coarser meshes. In this case, using a higher-order degree of approximation gives us the same order of accuracy in the solution and with less degrees of freedom. In Fig. 7(right), we

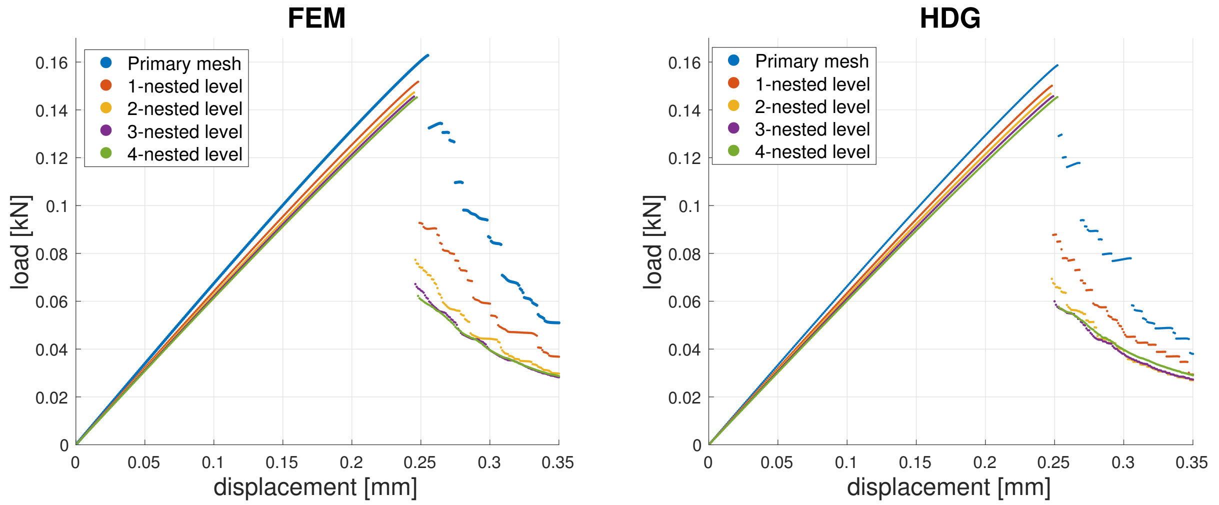

Figure 4: Load-displacement curves for the L-shaped panel test when using $p = 1$ for both FEM and HDG.

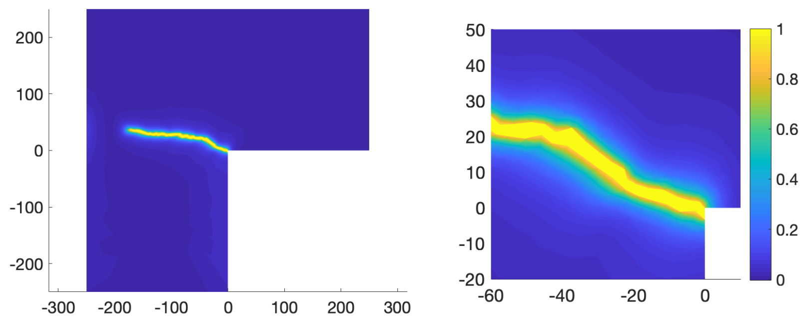

Figure 5: Damage field obtained with HDG at an imposed displacement of $u_D = 0.45$ mm. Degree of approximation $p = 1$, primary mesh and $\ell = 3$ mm.

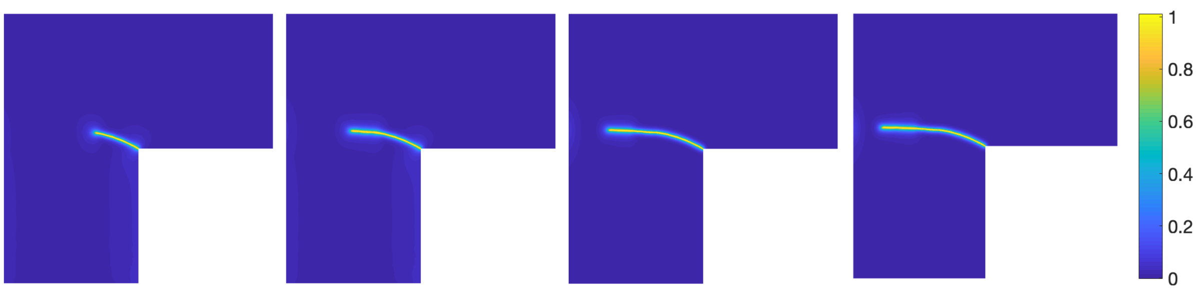

Figure 6: Damage field at displacements (a) $u_D = 0.25$ mm, (b) $u_D = 0.3$ mm, (c) $u_D = 0.4$ mm, (d) $u_D = 0.5$ mm. Degree of approximation $p = 1$, 2-nested level mesh and $\ell = 3$ mm.

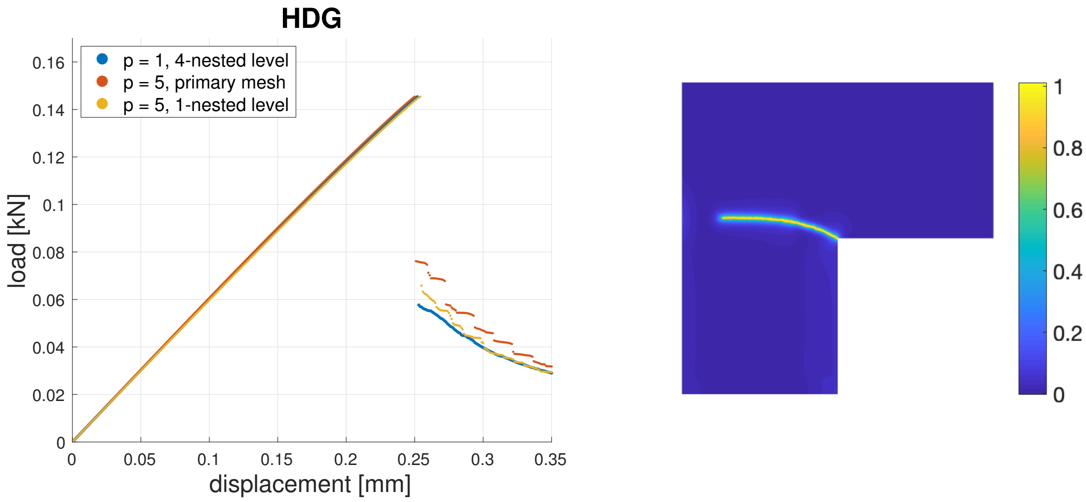

Figure 7: Left: load-displacement curves obtained with $p = 5$. Right: damage field at $u_D = 0.45$ mm for $p = 5$ and the primary mesh.

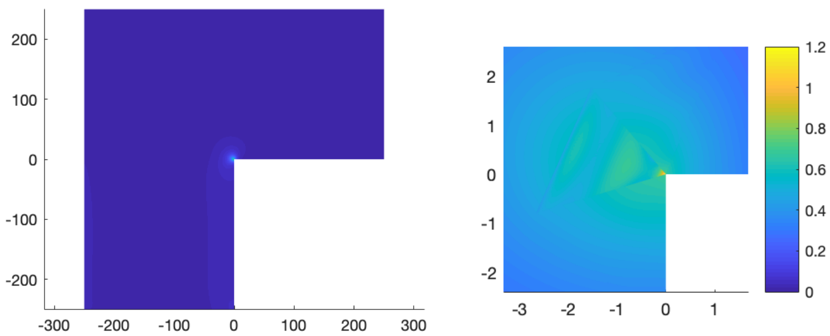

Figure 8: Evaluating $\mathcal{H}$ at nodes. Damage field for $u_D = 0.247$ mm. Whole body on the left, zoom on the right. Degree of approximation $p = 5$, primary mesh. The solution obtained is unphysical.

note that solving for $p = 5$ with the primary mesh we no longer observe the mesh dependence we have for $p = 1$ due to low spatial resolution.

Importance of evaluating $\mathcal{H}$ at integration points. As commented in Remark 3.1, if $\mathcal{H}$ is evaluated at nodes and then interpolated to Gauss points, it can reach negative values when using shape functions of degree $p > 1$. To illustrate this phenomenon, consider the L-shaped panel test with the primary mesh and degree of approximation $p = 5$. If we evaluate $\mathcal{H}$ at nodes, the damage field d is no longer in the interval $[0, 1]$. In Fig. 8, we can see the damage field obtained with this formulation for imposed vertical displacement $u_D = 0.247$ mm. Both the values of d and the pattern obtained are not a proper solution of the problem: the damage field presents oscillations and gets a value of 1.2 at the corner.

For the next load step, corresponding to imposed displacement $u_D = 0.248$ mm, the staggered scheme does not converge. In Fig. 9, we plot the relative Euclidean norm of the difference of consecutive iterates for d and the maximum and minimum values of damage obtained. Notice that the absolute value of the damage field gets arbitrarily large.

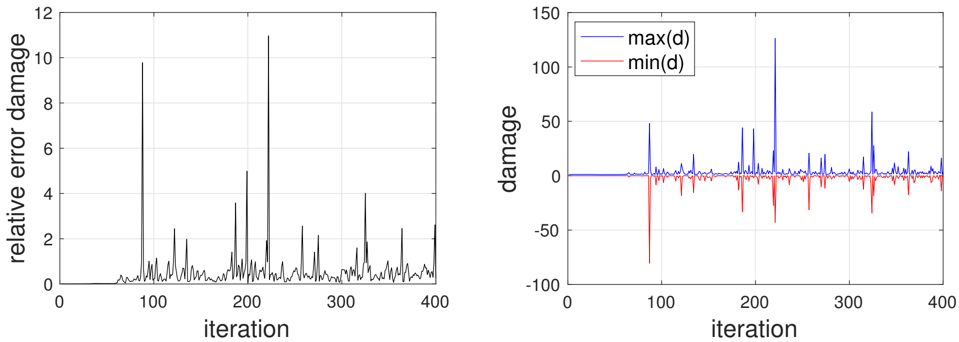

Figure 9: Evaluating $\mathcal{H}$ at nodes. For imposed displacement $u_D = 0.248$ mm, relative error of d (left) and maximum/minimum values of d (right) for number of iteration. The staggered scheme does not converge in this case.

5. Conclusions

We have proposed an HDG approach for phase-field models of brittle fracture using a staggered scheme that enables to uncouple the system. We have compared this formulation with the classical FEM formulation in a numerical example and both of them present the same behavior. As expected, the solution is more accurate when refining the mesh or increasing the degree of approximation. With HDG we obtain better accuracy than with FEM for the same mesh and degree of approximation, but at the price of a higher computational cost; see [11, 21].

The main drawback of phase-field models is their inefficiency coming from the remeshing needed if the crack path is not known. The HDG formulation is interesting for this problem because of the suitability of the method for adaptivity. The implementation of p-adaptivity and h-adaptivity for this formulation is subject of ongoing work.

References

[1] M. Ambati, T. Gerasimov, and L. De Lorenzis, "A review on phase-field models of brittle fracture and a new fast hybrid formulation", *Comput. Mech.* **55** (2015), 383–405.

[2] T. Belytschko and T. Red, "Elastic crack grow in finite element with minimal remeshing", *Int. J. Numer. Methods Eng.* **45**(5) (1999), 601–620.

[3] B. Bourdin, G.A. Francfort, and J.J. Marigo, "Numerical experiments in revisited brittle fracture", *J. Mech. Phys. Solids* **48** (2000), 797–826.

[4] B. Bourdin, G.A. Francfort, and J.J. Marigo, "The variational approach to fracture", *Journal of Elasticity* **91**(1-3) (2008), 5–148.

[5] B. Cockburn, J. Gopalakrishnan, and R. Lazarov, "Unified hybridization of discontinuous Galerkin, mixed and continuous Galerkin methods for second order elliptic problems", *SIAM J. Numer. Anal.* **47** (2009), 1319–1365.

[6] G.A. Francfort and J.J. Marigo, "Revisiting brittle fracture as an energy minimization problem",

Journal of Mechanics and Physics of Solids **46**(8) (1998), 1319–1342.

[7] G. Fu, B. Cockburn, and H. Stolarski, "Analysis of an HDG method for linear elasticity", *Int. J. Numer. Methods Eng.* **102** (2015), 51–575.

[8] R. Geelen, Y. Liu, J.E. Dolbow, and A. Rodríguez-Ferran, "An optimization-based phase-field method for continuous-discontinuous crack propagation", *Int. J. Numer. Methods Eng.* **116** (2018), 1–20.

[9] T. Gerasimov and L. De Lorenzis, "A line search monolithic approach for phase-field computing of brittle fracture", *Comput. Methods Appl. Mech. Engrg.* **312** (2016), 276–303.

[10] G. Giorgiani, D. Modesto, S. Fernández-Méndez, and A. Huerta, "High-order continuous and discontinuous Galerkin methods for wave problems", *Int. J. Numer. Methods Fluids* **73**(10) (2013), 883–903.

[11] R.M. Kirby, S.J. Sherwin, and B. Cockburn, "To CG or to HDG: A comparative study", *J. Sci. Comput.* **51** (2012), 183–212.

[12] C. Miehe, M. Hofacker, and F. Welschinger, "A phase-field model for rate-independent crack propagation: robust algorithmic implementation based on operator splits", *Comput. Methods Appl. Mech. Eng.* **199** (2010), 2765–2778.

[13] C. Miehe, F. Welschinger, and M. Hofacker, "Thermodynamically consistent phase-field models of fracture: variational principles and multi-field FE implementations", *Int. J. Numer. Methods Eng.* **83** (2010), 1273–1311.

[14] N. Moës, J. Dolbow, and T. Belytschko, "A finite element method for crack grow without remeshing", *Int. J. Numer. Methods Eng.* **46** (1999), 131–150.

[15] S. Nagaraja, M. Elhaddad, M. Ambati, S. Kollmannsberger, L. De Lorenzis, and E. Rank, "Phase-field modeling of brittle fracture with multi-level *hp*-FEM and the finite cell method", *Comput. Mech.* (2018).

[16] N.C. Nguyen and J. Peraire, "Hybridizable discontinuous Galerkin methods for partial differential equations in continuum mechanics", *Journal of Computational Physics* **231** (2012), 5955–5988.

[17] S.C. Soon, B. Cockburn, and H.K. Stolarski, "A hybridizable discontinuous Galerkin method for linear elasticity", *Int. J. Numer. Methods Eng.* **80** (2009), 1058–1092.

[18] N. Sukumar, J. Dolbow, and N. Moës, "Extended finite element method in computational fracture mechanics: a retrospective examination", *Int. J. Fract.* **196** (2015), 189–206.

[19] T. Wick, "Modified Newton methods for solving fully monolithic phase-field quasi-static brittle fracture propagation", *Comput. Methods Appl. Mech. Engrg.* **325** (2017), 577–611.

[20] J.Y. Wu, V.P. Nguyen, C.T. Nguyen, D. Sutulas, S. Bordas, and S. Sinaie, "Phase field modelling for fracture", *Advances in Applied Mathematics* **53** (2018).

[21] S. Yakovlev, D. Moxey, R.M. Kirby, and S.J. Sherwin, "To CG or to HDG: A comparative study in 3D", *J. Sci. Comput.* **67**(1) (2015), 192–220.

Reports@SCM
Volume 4, number 1, 2018

Table of Contents