# Boletim

## da Sociedade Portuguesa de Matemática

Matemáticos Portugueses pelo Mundo

77
*Número*
DEZEMBRO 2019

Universidade de Lisboa, 2017
Universidade do Porto, 2019

*4 euros*
[semestral]

*Boletim*
*da Sociedade Portuguesa de Matemática*

**Propriedade e Edição**
Sociedade Portuguesa de Matemática, Av. da República, 45–3º E, 1050–187 Lisboa
☎ 217939785 FAX 217952349
Internet: http://www.spm.pt



**Impressão e acabamento:** Dossier | Comunicação e Imagem • Alameda dos Oceanos, lote 3.10.06 B • Parque das Nações • 1990–186 Lisboa

**Tiragem:** 1500 exemplares

**Depósito Legal:** nº 63134/93

**ISSN:** 0872–3672

**ICS:** 106437

ISSN 0872–3672

## SUMÁRIO

**Artigos**

# Prefácio

Desde o seu primeiro volume, em 1978, o Boletim da SPM tem sido uma publicação dedicada à comunidade matemática portuguesa, constituindo um espaço diversificado de informação e promovendo a circulação de ideias e a troca de experiências sobre investigação e sobre o ensino superior.

Tanto o Boletim, como a comunidade matemática ligada ao nosso país, têm evoluído ao longo dos tempos, e têm-se adaptado a diferentes realidades.

Enquanto que, até aos anos 90, a esmagadora maioria desta comunidade formava parte dos quadros das nossas instituições de ensino superior, desde então, e ainda com mais relevância neste novo século, há cada vez mais doutorados portugueses em universidades estrangeiras, sendo a internacionalização da nossa investigação matemática uma realidade incontornável.

A conferência "Matemáticos Portugueses pelo Mundo", organizada por pela primeira vez em 2017, no Instituto Superior Técnico, da Universidade de Lisboa, tendo como oradores 10 matemáticos portugueses na diáspora, veio demonstrar a pertinência de preservar os elos de ligação desta comunidade.

Tanto na conferência de 2017, como na segunda edição em 2019, na Faculdade de Ciências da Universidade do Porto, foi notória a grande qualidade dos temas abordados pelos oradores convidados, tendo havido, em geral, o cuidado de apresentar os temas de forma acessível a uma audiência não especializada.

Desta forma, é com grande naturalidade e sentido de oportunidade que o Boletim da SPM se associa a esta iniciativa, registando materialmente uma parte significativa dos trabalhos apresentados nestas duas edições.

Esperamos, desta forma, contribuir para a visibilidade interna e externa da investigação matemática feita fora de Portugal, e para a criação de novas oportunidades de interacção científica no contexto académico português.

Carlos Florentino
Ana Jacinta Soares

*Dezembro de 2019*

# Introdução

"Matemáticos Portugueses pelo Mundo" (MPM) é uma conferência bienal que visa reunir matemáticos portugueses dos quatro cantos do planeta. Entre os seus objetivos, contam-se a disseminação de investigação de ponta e o estímulo para novas colaborações científicas, resultando num maior sentido de unidade e de interligação na comunidade matemática lusófona.

A última edição da conferência MPM teve lugar de 24 a 26 de junho de 2019 no Departamento de Matemática da Universidade do Porto, e contou com quinze oradores oriundos de um vasto leque de áreas, incluindo a álgebra, a análise, a estatística, a geometria, os sistemas dinâmicos, e a teoria dos números. As palestras geraram inúmeras perguntas e uma animada discussão, que se estendeu para além dos coffee breaks. Houve ainda uma sessão de encerramento com representantes das universidades do Porto, Lisboa e Coimbra, onde se discutiu a relevância de atividades como a conferência MPM, e a situação dos matemáticos profissionais em Portugal.

O programa científico foi complementado por uma agenda social que capitalizou a data estratégica do início da conferência (São João), e incluiu uma visita à exposição "Suite Vollard" de Pablo Picasso no Palácio das Artes e um jantar na Cooperativa Árvore. O São Pedro apoiou incondicionalmente todo o evento.

Está prevista uma próxima edição dos MPM, a decorrer na Universidade de Coimbra algures durante o Verão de 2021. Até lá!

Jorge Freitas
Samuel Lopes
Diogo Oliveira e Silva

*Editores convidados*

*Novembro de 2019*

# Matemáticos Portugueses pelo Mundo

*Departamento de Matemática, IST*
*Universidade de Lisboa*

*13 e 14 de Junho de 2017*

https://mpm.math.tecnico.ulisboa.pt

# Matemáticos Portugueses pelo Mundo

*Departamento de Matemática, Faculdade de Ciências*
*Universidade do Porto*

*24 a 26 de Junho de 2019*

https://cmup.fc.up.pt/matematicos/

# A Price Model with Finitely Many Agents

*AbdulRahman Alharbi*, *Tigran Bakaryan*, *Rafael Cabral*, *Sara Campi*, *Nicholas Christoffersen*, *Paolo Colusso*, *Odylo Costa*, *Serikbolsyn Duisembay*, *Rita Ferreira*, *Diogo Gomes*, *Shibei Guo*, *Julian Gutierrezpineda*, *Phebe Havor*, *Michele Mascherpa*, *Simone Portaro*, *Ricardo Ribeiro*, *Fernando Rodriguez*, *Johan Ruiz*, *Fatimah Saleh*, *Calum Strange*, *Teruo Tada*, *Xianjin Yang*, *Zofia Wróblewska*

Applied Mathematics Summer School
King Abdullah University of Science and Technology
Thuwal, Saudi Arabia

**Resumo:** Neste trabalho, estudamos um modelo de formação de preços numa população com um número finito de agentes que compram e vendem uma mercadoria. A oferta desta mercadoria é exógena e os agentes são racionais uma vez que pretendem minimizar os custos de transacção. O problema em estudo é formulado como um jogo dinâmico entre $N$ jogadores com uma condição de equilíbrio de mercado. O limite deste problema de $N$ jogadores é um "mean field game". Posteriormente, mostramos como reformular o nosso jogo como um problema de optimização do custo total. Mostramos a existência de uma solução usando o método direto do cálculo das variações. Por fim, mostramos que o preço é o multiplicador de Lagrange para a condição de equilíbrio entre a oferta e a procura.

**Abstract** Here, we propose a price-formation model, with a population consisting of a finite number of agents storing and trading a commodity. The supply of this commodity is determined exogenously, and the agents are rational as they seek to minimize their trading costs. We formulate our problem as an $N$-player dynamic game with a market-clearing condition. The limit of this $N$-player problem is a mean-field game (MFG). Subsequently, we show how to recast our game as an optimization problem for the overall trading cost. We show the existence of a solution using the direct method in the calculus of variations. Finally, we show that the price is the Lagrange multiplier for the balance condition between supply and demand.

**palavras-chave:** Formação de preço, jogos dinâmicos, equilíbrio de mercado.

**keywords:** Price formation model, dynamic games, market equilibrium.

# 1 Prologue

This document is the result of the second KAUST Summer Camp in Applied Partial Differential Equations that took place from August 25 to September 8 of 2019. The purpose of this summer camp is to give an intense hands-on research experience in cutting edge topics to BS/BSc and MS students. Participants attended mini-courses that provide them with the tools to reach the results we present here. For the research project, the participants worked in small groups. These were coordinated by Professor Diogo Gomes, together with his Ph.D. Students and Postdocs and Research Scientist Rita Ferreira.

Participants also had the opportunity to get acquainted with a variety of research topics pursued by KAUST scholars as a means of broadening their mathematical perspectives and future opportunities at KAUST. On the weekends, there were cultural activities, such as sightseeing in the UNESCO Cultural Heritage neighborhood of Al Balad, a snorkeling trip, and a Hejazi Fish Dinner.

# 2 Introduction

Mean-field game (MFG) theory studies the behavior of large populations of identical rational agents in competition, where the behavior of each agent is determined by their state and by statistical information of the remaining players. In [9], Gomes and Saúde studied a price formation problem using an MFG approach. In this paper, we address a similar price formation problem (**Problem 1**) in a market with $N$ identical rational agents who trade continuously a commodity whose supply, $Q$, is a given exogenous variable and whose price, $\varpi$, is determined by the balance between supply and demand. The agents are rational, in the sense that they seek to minimize their trading cost. The collective behavior of the agents, coupled with the market clearing condition, determines the evolution of the price, $\varpi$. More precisely, we consider the following problem:

**Problem 1.** *Let $Q \in C^1([0,T])$ be the supply rate per agent. Let $L \in C^2(\mathbb{R}\times\mathbb{R})$, the Lagrangian, be a non-negative function, convex in the second component. Let $\Psi \in C^1(\mathbb{R})$ be a non-negative terminal cost. Let $N \in \mathbb{N}$ be the number of agents. At time $0$, each agent $i$ owns $x_0^i$ units of the commodity.*

*Find a price, $\varpi : [0,T] \to \mathbb{R}$, and trajectories, $\mathbf{x}_i : [0,T] \to \mathbb{R}$, with initial conditions $\mathbf{x}_i(0) = x_0^i$, such that for each $1 \leqslant i \leqslant N$, $\mathbf{x}_i$ minimizes the functional*

$$\int_0^T \Big(L(\mathbf{x}_i(s), \dot{\mathbf{x}}_i(s)) + \varpi(s)\dot{\mathbf{x}}_i(s)\Big)ds + \Psi(\mathbf{x}_i(T)) \tag{2.1}$$

*subjected to the balance condition*

$$\frac{1}{N}\sum_{i=1}^{N} \dot{\mathbf{x}}_i(t) = Q(t), \quad \forall t \in [0,T]. \tag{2.2}$$

In the preceding problem, $\mathbf{x}_i(t)$ is the amount of commodity held by the agent $i$ at time $t$; hence, $\dot{\mathbf{x}}_i(t)$ denotes the rate at which the agent $i$ trades. The functional in (2.1) represents the cost for each agent. The running cost is composed of the trading cost that comprises the instantaneous cost of the commodity $\varpi\dot{\mathbf{x}}_i$ and indirect costs such as storage or market impact encoded in the term $L(\mathbf{x}, \dot{\mathbf{x}})$. The preference of the agents at the final time, $T$, is encoded in the term $\Psi(\mathbf{x}_i(T))$, the terminal cost. The equation in (2.2) is the requirement that the market clears; that is, supply equals demand at all times.

Formally, MFGs model the mean-field limit of $N$-player games as $N \to \infty$. However, the rigorous justification of this limit is unknown in the general case, despite recent substantial progress [1]. In our price formation problem, the $N$-player game is relatively tractable. The main goal of this work is to study this $N$-player problem, which is the first step towards the rigorous justification of the mean-field limit as $N$ goes to infinity. We expect our price to approximate the one presented [9] as the number of players increases. Also, each trajectory $\mathbf{x}_i$ should converge to the trajectory of the representative player of the continuum of agents model solved as an MFG. Notice that in this limiting process, the function $Q$ remains the same for both the finite and the continuous player models.

In their seminal paper, Lasry and Lions [15] presented three examples of mean-field modeling in economics. They were concerned with situations involving a large number of rational players with little individual effect on the game. Inspired by [15], Markowich et al. [18] discussed the existence and uniqueness of the solution for a one-dimensional parabolic evolution equation with a free boundary that models price formation. Caffarelli et al. [2] established the global existence and asymptotic behavior of a price formation model with free boundaries. Their results rely on a transformation, which takes the equation in their problem into the heat equation. Burger et al. [1] extended this problem to a Boltzman-type price formation model. Their solutions converge to the Lasry–Lions model as the transaction rate tends to infinity. The study of the behavior of rational agents in energy markets appeared in [16, 17] in the context of load-control problems. Switching space heaters on and off controls the load, for an MFG approach see [12, 13, 14]. Previous authors addressed the price issue by assuming that the demand is a given function of the price [11] or that the price is a given function of the demand, see [3, 4, 5, 6, 10].

An $N$-player version of an economic growth model was presented in [8]. In a more recent paper [9], Gomes and Saúde introduced a price-formation model where a large number of small players seek to store and trade electricity. This model was a constrained MFG where the price is a Lagrange multiplier for the supply vs. demand balance condition.

Here, we prove the following main theorem:

**Theorem 2.1.** *Assume that $\Psi$, the terminal cost, is non-negative and uniformly convex, and $L \in C^2(\mathbb{R} \times \mathbb{R})$, the Langrangian, is non-negative, uniformly convex in the second component, and satisfies the following inequality uniformly in $(z, v) \in \mathbb{R} \times \mathbb{R}$:*

$$L(z, v) \geqslant \alpha |v|^q - \beta, \quad q \in (1, \infty), \alpha > 0, \beta \geqslant 0. \tag{2.3}$$

*Then,* ***Problem 1*** *has a unique solution.*

The existence is established in **Proposition 4.1** and the uniqueness in **Proposition 4.3**.

The condition (2.3) means that high trading rates are expensive. The utility function in Economics is the negative of our value function. Convexity properties of the value function translate into concavity for the utility function. Therefore, our convexity assumptions are natural from the Economics point of view.

This work starts with the description of the single-agent control problem and derives the Euler–Lagrange equation. It then deals with the $N$-agent problem. For this, we first show the existence of the minimizers by applying the direct method in the calculus of variations. Then, we provide an interpretation of the price of the commodity as the Lagrange multiplier of the corresponding multi-agent problem. Subsequently, we find necessary conditions for the trajectories to be minimizers, via a slight variation on the Euler–Lagrange equation. We conclude the work by proving the existence of a unique solution for **Problem 1** under convexity assumptions on $L$ and $\Psi$.

Finally, we point out that **Problem 1** can be coupled with a control problem for $Q$ on the production side, where the producer seeks to maximize profits.

## 3 Single-agent control problem

To build an $N$-player model, we first analyze a single-agent control problem. Using optimal control theory and calculus of variations, we derive the Euler–Lagrange equation and the boundary conditions.

Let $\mathcal{B}_t^q = W^{1,q}(t,T)$ be the set of admissible functions with $q \in (1,\infty)$ as in (2.3). Each agent seeks to find an optimal trajectory, $\mathbf{x} \in \mathcal{B}_0^q$, minimizing the functional

$$I[\mathbf{x}] = \int_0^T \Big( L(\mathbf{x}(t), \dot{\mathbf{x}}(t)) + \varpi(t)\dot{\mathbf{x}}(t) \Big)\, dt + \Psi(\mathbf{x}(T))$$

with an initial position, $\mathbf{x}(0) = x_0$.

If $\mathbf{x}$ is a minimizer, then for any $\mathbf{y} \in C_c^\infty((0,T])$, and every $\epsilon \in \mathbb{R}$, we have

$$I[\mathbf{x}] \leqslant I[\mathbf{x} + \epsilon\mathbf{y}] \ .$$

Thus, the function $i : \mathbb{R} \to \mathbb{R}$ defined by $i(\epsilon) = I[\mathbf{x} + \epsilon\mathbf{y}]$ attains a local minimum at $\epsilon = 0$. Then, $i'(\epsilon)|_{\epsilon=0} = 0$. Accordingly, computing $i'(0)$ and using the fact that $\mathbf{y}$ is arbitrary, we obtain the Euler–Lagrange equation

$$D_x L(\mathbf{x}, \dot{\mathbf{x}}) - \frac{d}{dt}(D_v L(\mathbf{x}, \dot{\mathbf{x}}) + \varpi) = 0$$

and the *natural boundary condition*

$$D_v L(\mathbf{x}(T), \dot{\mathbf{x}}(T)) + \varpi(T) + \Psi'(\mathbf{x}(T)) = 0.$$

**Example 3.1.** Consider a Lagrangian of the form

$$L(x,v) = L(v).$$

Then, the Euler–Lagrange equation becomes

$$\frac{d}{dt}(D_v L(\dot{\mathbf{x}}) + \varpi) = 0 \Leftrightarrow D_v L(\dot{\mathbf{x}}) + \varpi = K,$$

where $K$ is some constant. Since $L$ is uniformly convex, $D_v L$ is strictly monotone and, thus, invertible. Therefore,

$$\dot{\mathbf{x}} = (D_v L)^{-1}(K - \varpi) \ .$$

So, if the price $\varpi(t)$ increases, the agents buy less or sell. In particular, if $L(v) = \frac{v^2}{2}$, then

$$D_v L(\dot{\mathbf{x}}) = \dot{\mathbf{x}}.$$

Hence, the Euler–Lagrange equation becomes

$$\dot{\mathbf{x}} = K - \varpi. \tag{3.1}$$

Equation (3.1) shows that as the price increases, $\dot{\mathbf{x}}$ decreases.

Let $\Psi(x) = \frac{x^2}{2}$, which means that agents seek to minimize $|\mathbf{x}(T)|^2$. This choice of $\Psi$ corresponds to the portfolio liquidation problem. The Euler–Lagrangian equation and corresponding natural boundary condition give

$$\begin{cases} \frac{d}{dt}\left(\dot{\mathbf{x}}(t) + \varpi(t)\right) = 0 \\ \dot{\mathbf{x}}(T) + \mathbf{x}(T) = -\varpi(T) \\ \mathbf{x}(0) = x_0 \ . \end{cases} \tag{3.2}$$

Thus, from (3.1) and (3.2), we get

$$K = \frac{1}{1+T}\left[\int_0^T \varpi(t)dt - x_0\right].$$

Define the average price

$$\hat{\varpi} = \frac{1}{T}\int_0^T \varpi(t)dt.$$

The agent buys when

$$\dot{\mathbf{x}}(t) > 0.$$

According to (3.1), the above inequality holds if

$$\varpi(t) < \frac{T\hat{\varpi} - x_0}{T+1}.$$

Thus, an agent buys when the price is below the threshold price on the right-hand side of the preceding inequality.

## 4 A constrained minimization problem for N agents

We use the single-agent control problem to formulate an $N$-agent minimization problem that includes the balance condition. We prove existence, uniqueness, and then we provide a characterization of such minimizer by showing that the price is the Lagrange multiplier of an equivalent minimization problem.

### 4.1 A variational problem

Notice that, for each agent, (2.1) is a functional that is independent of the dynamics of the other agents. Hence, **Problem 1** is equivalent to the following minimization problem

$$x \min_{\mathbf{x},\mathbf{x}(0)=x_0} \frac{1}{N}\sum_{i=1}^{N}\left(\int_0^T \Big(L(\mathbf{x}_i(s),\dot{\mathbf{x}}_i(s)) + \varpi(s)\dot{\mathbf{x}}_i(s)\Big)ds + \Psi(\mathbf{x}_i(T))\right) \tag{4.1}$$

$$\text{subject to} \quad \frac{1}{N}\sum_{i=1}^{N}\dot{\mathbf{x}}_i(t) = Q(t) \quad \forall t \in [0,T] \ . \tag{4.2}$$

Substituting (4.2) into (4.1) we get

$$\begin{aligned}&\frac{1}{N}\sum_{i=1}^{N}\left(\int_0^T \Big(L(\mathbf{x}_i(s),\dot{\mathbf{x}}_i(s)) + \varpi(s)\dot{\mathbf{x}}_i(s)\Big)ds + \Psi(\mathbf{x}_i(T))\right)\\ &= \frac{1}{N}\sum_{i=1}^{N}\left(\int_0^T L(\mathbf{x}_i(s),\dot{\mathbf{x}}_i(s))ds + \Psi(\mathbf{x}_i(T))\right) + \int_0^T \varpi(s)Q(s)ds \ ,\end{aligned}$$

and since $\varpi(s)Q(s)$ is independent of $\mathbf{x}$ at every $s$, the minimization problem is equivalent to

$$\min_{\mathbf{x},\mathbf{x}(0)=x_0} \frac{1}{N}\sum_{i=1}^{N}\int_0^T L(\mathbf{x}_i(s),\dot{\mathbf{x}}_i(s))ds + \Psi(\mathbf{x}_i(T)) \tag{4.3}$$

$$\text{subject to} \quad \frac{1}{N}\sum_{i=1}^{N}\dot{\mathbf{x}}_i(t) = Q(t) \quad \forall t \in [0,T] \ . \tag{4.4}$$

We now prove the existence of optimal trajectories.

### 4.2 Existence of a solution

We use the direct method in the calculus of variations to obtain the existence of a minimizer of (4.3) and (4.4). For that, let

$$\mathcal{L}(\mathbf{x},\dot{\mathbf{x}}) = \frac{1}{N}\sum_{i=1}^{N}\left(L(\mathbf{x}_i,\dot{\mathbf{x}}_i) + \frac{1}{T}\Psi(\mathbf{x}_i(T))\right),$$

and

$$I_N[\mathbf{x}] = \int_0^T \mathcal{L}(\mathbf{x},\dot{\mathbf{x}})ds.$$

Then (4.3) and (4.4) becomes

$$\begin{aligned} &\min_{\mathbf{x},\mathbf{x}(0)=\mathbf{x}_0} \quad I_N[\mathbf{x}] \\ &\text{s.t.} \quad \langle \mathbf{x}(t)\rangle = Q(t). \end{aligned} \tag{4.5}$$

**Proposition 4.1.** *Let $L$ satisfy (2.3). Then **Problem 1** has a solution.*

*Proof.* We show that $\mathcal{L}$ is coercive and lower semicountinuous in $W^{1,q}$. It is enough to show that there exist $\bar{\alpha} > 0$, $\bar{\beta} \geqslant 0$, and $q > 1$ such that

$$\mathcal{L}(\mathbf{x}, \mathbf{p}) \geqslant \alpha |\mathbf{p}|^q - \beta$$

to obtain coercivity. The condition on the Lagrangian for each agent implies the coercivity on $\mathcal{L}$, since, by the non-negativity of $\Psi$, we have:

$$\begin{aligned} \mathcal{L}(\mathbf{x}, \mathbf{p}) = \frac{1}{N}\sum_{i=1}^{N}\left(L(\mathbf{x}_i, \mathbf{p}_i) + \frac{1}{T}\Psi(\mathbf{x}_i(T))\right) &\geqslant \frac{1}{N}\sum_{i=1}^{N}(\alpha|\mathbf{p}_i|^q - \beta) \\ &\geqslant \alpha \sum_{i=1}^{N} |\mathbf{p}_i|^q - \beta = \alpha ||\mathbf{p}||^q_{L^q} - \beta \\ &\geqslant \frac{\alpha C}{N}|\mathbf{p}|^q - \beta. \end{aligned}$$

The last inequality follows from the fact that in $\mathbb{R}^N$ all the $p$-norms are equivalent. The above establishes the coercivity of $\mathcal{L}$.

To show lower semicontinuity, we need to ensure the convexity of $\mathcal{L}$ on the second variable, and that $\mathcal{L}$ is bounded from below. Convexity follows from the convexity of $L$ in $\dot{\mathbf{x}}$. Boundedness from below follows from the coercivity condition.

We use the direct method in the calculus of variations to determine the existence of a minimizer for our problem. Define the admissible set

$$\mathcal{A}_t = \left\{ \mathbf{x} \in W^{1,q}(t,T) \mid \frac{\sum_{i=1}^N \dot{\mathbf{x}}_i(s)}{N} = Q(s), \mathbf{x}_i(0) = x_0^i, 1 \leqslant i \leqslant N,\ t \leqslant s \leqslant T \right\},$$

and set $\mathcal{A} = \mathcal{A}_0$. We notice that $\mathcal{A}$ is nonempty by taking $\dot{\mathbf{x}}_i = Q(t), \mathbf{x}_i(0) = x_0^i$. Since $\mathcal{L}$ is bounded from below, there exists a minimizing sequence, $(\mathbf{x}^n)_{n\in\mathbb{N}} \subset \mathcal{A}$ such that

$$\lim_{n\to+\infty} I_N[\mathbf{x}^n] = \inf_{\mathbf{x}} I_N[\mathbf{x}].$$

By the coercivity of $\mathcal{L}$, we have

$$I_N[\mathbf{x}^n] \geqslant \alpha \|\dot{\mathbf{x}}^n\|_{L^q}^q - \beta T.$$

Thus, by Poincaré's inequality, $(\mathbf{x}^n)_{n\in\mathbb{N}}$ is bounded in $W^{1,q}(0,T)$. Then, there exists $\mathbf{x}^* \in W^{1,q}(0,T)$ such that, up to a subsequence, $\mathbf{x}^n$ converges weakly to $\mathbf{x}^*$. We notice that $\mathcal{A}$ is convex. Since $q > 1$, Morrey's theorem (see [7]) gives that $\mathcal{A}$ is closed. Thus, by Mazur's theorem, see ([7] Appendix D.4), $\mathcal{A}$ is weakly closed in $W^{1,q}(0,T)$, which implies that $\mathbf{x}^* \in \mathcal{A}$. Then, since $\mathcal{L}$ is bounded from below and convex in $p$, $I$ is sequentially weakly lower semicontinuous in $W^{1,q}(0,T)$. Thus, $\mathbf{x}^*$ is the minimizer of $I$ since

$$\inf_{\mathbf{x}} I_N[\mathbf{x}] = \lim_{n\to+\infty} I_N[\mathbf{x}^n] \geqslant I_N[\mathbf{x}^*] \geqslant \inf_{\mathbf{x}} I_N[\mathbf{x}]. \qquad \square$$

## 4.3 Uniqueness of solutions

Assume that

**Assumption 4.2.**

1. the map $(x,v) \mapsto L(x,v)$ is convex and for each $x \in \mathbb{R}$, the map $v \mapsto L(x,v)$ is uniformly convex; that is, there exists $\theta > 0$ such that for all $x, y, v, w \in \mathbb{R}$, we have

$$L(\lambda x+(1-\lambda)y, \lambda v+(1-\lambda)w) \leqslant \lambda L(x,v)+(1-\lambda)L(y,w)-\theta\lambda(1-\lambda)|v-w|^2.$$

2. $\Psi$ is uniformly convex.

We notice that the term $\varpi\dot{x}$ is linear in the velocity, thus convex.

**Proposition 4.3.** *Let $x \in \mathbb{R}^N$. Under Assumptions 1. and 2., the solution of the problem*

$$\min_{\mathbf{x}\in\mathcal{A}_t, \mathbf{x}(t)=x} I_N[\mathbf{x}]$$

*is unique.*

*Proof.* We prove the statement via contradiction. Assume that there exist two different minimizers, $\mathbf{x}$, $\mathbf{y} \in \mathcal{A}_t$ with $\mathbf{x}(t) = x$. Then, taking the middle point, $\dfrac{\mathbf{x}+\mathbf{y}}{2}$, we obtain

$$
\begin{aligned}
I_N\left[\frac{\mathbf{x}+\mathbf{y}}{2}\right] &= \frac{1}{N}\sum_{i=1}^{N}\left[\int_0^T L\left(\frac{\mathbf{x}_i+\mathbf{y}_i}{2},\frac{\dot{\mathbf{x}}_\mathbf{i}+\dot{\mathbf{y}}_\mathbf{i}}{2}\right)dt+\Psi\left(\frac{\mathbf{x}_\mathbf{i}(T)+\mathbf{y}_\mathbf{i}(T)}{2}\right)\right]\\
&\leqslant \frac{1}{2N}\sum_{i=1}^{N}\Bigg[\int_0^T L(\mathbf{x}_\mathbf{i},\dot{\mathbf{x}}_\mathbf{i})dt+\int_0^T L(\mathbf{y}_\mathbf{i},\dot{\mathbf{y}}_\mathbf{i})dt-\frac{\theta}{2}\int_0^T|\dot{\mathbf{x}}_\mathbf{i}-\dot{\mathbf{y}}_\mathbf{i}|^2dt\\
&\qquad\qquad +\Psi(\mathbf{x}_\mathbf{i}(T))+\Psi(\mathbf{y}_\mathbf{i}(T))\Bigg]\\
&= \frac{1}{2}I_N[\mathbf{x}]+\frac{1}{2}I_N[\mathbf{y}]-\sum_{i=1}^{N}\frac{\theta}{4N}||\dot{\mathbf{x}}_\mathbf{i}-\dot{\mathbf{y}}_\mathbf{i}||^2_{L^2(0,T)}\\
&= \min_{\mathbf{z}\in\mathcal{A}_t,\mathbf{z}(t)=x} I_N[\mathbf{z}]-\sum_{i=1}^{N}\frac{\theta}{4N}||\dot{\mathbf{x}}_\mathbf{i}-\dot{\mathbf{y}}_\mathbf{i}||^2_{L^2(0,T)}.
\end{aligned}
$$

Because $\theta>0$, the preceding inequality can hold only if $||\dot{\mathbf{x}}_\mathbf{i}-\dot{\mathbf{y}}_\mathbf{i}||^2_{L^2(0,T)}=0$ for all $i=1,\dots,N$. Consequently, there exists a constant, $c\in\mathbb{R}^N$, such that $\mathbf{x}(s)-\mathbf{y}(s)=c$ for all $s\in[t,T]$. Using the initial condition $\mathbf{x}(t)=\mathbf{y}(t)=x$, we get $c=0$, which contradicts the fact that $\mathbf{x}$ and $\mathbf{y}$ are distinct. Thus, $\mathbf{x}=\mathbf{y}$ and this concludes the proof. □

## 4.4 Price as a Lagrange Multiplier

For $F=(f_1,\dots,f_N)\in\mathbb{R}^N$, we denote its entry-wise average by

$$
\langle F\rangle := \frac{1}{N}\sum_{k=1}^{N} f_k.
$$

Before deriving the necessary optimality conditions, we introduce the following auxiliary result.

**Lemma 4.4.** *Let $F=(f_1,\dots,f_N)\in C((0,T);\mathbb{R}^N)$ be such that for all $P\in C_c^\infty([0,T];\mathbb{R}^N)$ with $\langle P(s)\rangle=0$, for all $s\in[0,T]$, $F$ satisfies*

$$
\int_0^T F(s)\cdot P(s)\,ds=0.
$$

*Then, there exists $c\in C(0,T)$ such that, for all $k=1,\dots,N$, we have*

$$
f_k(t)=c(t).
$$

*Proof.* Fix $R \in C_c^\infty([0,T];\mathbb{R}^N)$. Set $P$ by

$$P = R - \langle R \rangle \mathbb{1}.$$

Because $\langle P(s) \rangle = 0$, we have

$$0 = \int_0^T F \cdot P \, ds = \int_0^T (F \cdot R - N\langle R \rangle \langle F \rangle) \, ds = \int_0^T (F - \langle F \rangle \mathbb{1}) \cdot R \, ds.$$

Since $R$ is arbitrary, by the fundamental theorem of the calculus of variations, for all $k = 1, \ldots, N$ and $t \in (0,T)$, we have

$$f_k(s) - \langle F(s) \rangle = 0.$$

Hence, we obtain $c(\cdot) = \langle F(\cdot) \rangle \in C(0,T)$. □

In the next proposition, we derive the necessary conditions (Euler–Lagrange equations) for solutions of (4.5). Let $X$ be

$$X := \left\{ \mathbf{x} \in C^2([0,T], \mathbb{R}^N) \mid \langle \dot{\mathbf{x}}(s) \rangle = Q(s) \text{ for all } s \in [0,T] \right\}.$$

**Proposition 4.5.** *Assume that $L \in C^2(\mathbb{R}^2)$. Then, there exist $c \in C(0,T)$ and $\tilde{c} \in \mathbb{R}$ such that, if $\bar{\mathbf{x}} \in X \cap C^2([0,T];\mathbb{R}^N)$ is a minimizer of (4.5), then it solves*

$$\frac{\partial}{\partial x} L(\bar{\mathbf{x}}_k(t), \dot{\bar{\mathbf{x}}}_k(t)) - \frac{d}{dt}\left( \frac{\partial}{\partial v} L(\bar{\mathbf{x}}_k(t), \dot{\bar{\mathbf{x}}}_k(t)) \right) = c(t)$$

*and*

$$\frac{\partial}{\partial v} L(\bar{\mathbf{x}}_k(T), \dot{\bar{\mathbf{x}}}_k(T)) + \Psi'(\bar{\mathbf{x}}_k(T)) = \tilde{c}$$

*for all $t \in (0,T)$ and for all $k = 1, \ldots, N$.*

*Proof.* Let $\mathbf{y} \in C^\infty([0,T], \mathbb{R}^N)$ be such that $\mathbf{y}(0) = 0$ and $\langle \mathbf{y}(s) \rangle = 0$ for every $s \in [0,T]$. For $\epsilon \in \mathbb{R}$, we define $i : \mathbb{R} \to \mathbb{R}$ as

$$i(\epsilon) = \frac{1}{N} \sum_{k=1}^N \int_0^T \Big( L(\bar{\mathbf{x}}_k + \epsilon \mathbf{y}_k, \dot{\bar{\mathbf{x}}}_k + \epsilon \dot{\mathbf{y}}_k) + \varpi \cdot (\dot{\bar{\mathbf{x}}}_k + \epsilon \dot{\mathbf{y}}_k) \Big) \, ds + \Psi(\bar{\mathbf{x}}_k(T) + \epsilon \mathbf{y}_k(T)).$$

Since $\langle \mathbf{y} \rangle = 0$ and $\bar{\mathbf{x}} \in X$, we have

$$\langle \dot{\bar{\mathbf{x}}} + \epsilon \dot{\mathbf{y}} \rangle = Q(s).$$

Thus, we obtain

$$i(\epsilon) = \frac{1}{N}\sum_{k=1}^{N}\left[\int_0^T L(\bar{\mathbf{x}}_k + \epsilon\mathbf{y}_k, \dot{\bar{\mathbf{x}}}_k + \epsilon\dot{\mathbf{y}}_k)\, ds + \Psi(\bar{\mathbf{x}}_k(T) + \epsilon\mathbf{y}_k(T))\right] + \int_0^T \varpi\cdot Q\, ds.$$

We have that $i \in C^1(\mathbb{R})$ because $L \in C^2(\mathbb{R}^2)$. Thus, because $\bar{x}_k$ is a minimizer for all $k = 1, \ldots, N$, we have $i'(0) = 0$; that is

$$\frac{1}{N}\sum_{k=1}^{N}\left[\int_0^T \left(\frac{\partial}{\partial x}L(\bar{\mathbf{x}}_k, \dot{\bar{\mathbf{x}}}_k)\mathbf{y}_k + \frac{\partial}{\partial v}L(\bar{\mathbf{x}}_k, \dot{\bar{\mathbf{x}}}_k)\dot{\mathbf{y}}_k\right) ds + \Psi'(\bar{\mathbf{x}}_k(T))\mathbf{y}_k(T)\right] = 0.$$

Integrating by parts, we obtain

$$\frac{1}{N}\sum_{k=1}^{N}\Big[\int_0^T \Big(\frac{\partial}{\partial x}L(\bar{\mathbf{x}}_k, \dot{\bar{\mathbf{x}}}_k) - \frac{d}{dt}(\frac{\partial}{\partial v}L(\bar{\mathbf{x}}_k, \dot{\bar{\mathbf{x}}}_k))\Big)\mathbf{y}_k\, ds$$
$$+ \Big(\frac{\partial}{\partial v}L(\bar{\mathbf{x}}_k(T), \dot{\bar{\mathbf{x}}}_k(T)) + \Psi'(\bar{\mathbf{x}}_k(T))\Big)\mathbf{y}_k(T)\Big] = 0.$$

If we select $\mathbf{y}$ such that $\mathbf{y}(T) = 0$, by Lemma 4.4, we conclude that there exists $c \in C(0, T)$ such that, for all $k = 1, \ldots, N$, we have

$$\frac{\partial}{\partial x}L(\bar{\mathbf{x}}_k, \dot{\bar{\mathbf{x}}}_k) + \frac{d}{dt}\Big(\frac{\partial}{\partial v}L(\bar{\mathbf{x}}_k, \dot{\bar{\mathbf{x}}}_k)\Big) = c(t). \tag{4.6}$$

Define $\tilde{f}_k(T)$ by

$$\tilde{f}_k(T) = \frac{\partial}{\partial v}L(\bar{\mathbf{x}}_k(T), \dot{\bar{\mathbf{x}}}_k(T)) + \Psi'(\bar{\mathbf{x}}_k(T)).$$

For all $t \in [0, T]$, set $\tilde{F}(t) := (\tilde{f}_1(T), ..., \tilde{f}_n(T))$. Since $\tilde{f}_k$ is constant, applying Lemma 4.4 for $\tilde{F}$, there exists $\tilde{c} \in \mathbb{R}$ such that we have $\tilde{f}_k(t) = \tilde{f}_k(T) = \tilde{c}$, from which we conclude that

$$\frac{\partial}{\partial v}L(\bar{\mathbf{x}}_k(T), \dot{\bar{\mathbf{x}}}_k(T)) + \Psi'(\bar{\mathbf{x}}_k(T)) = \tilde{c}. \qquad \square$$

Let $c$ and $\tilde{c}$ be as in the statement of the preceding proposition, and let $\varpi \in C^1(0, T)$ solve

$$\dot{\varpi}(t) = -c(t), \quad \varpi(T) = -\tilde{c}.$$

Then, the necessary optimality conditions for $\mathbf{x}_k$ become

$$\frac{\partial}{\partial x}L(\bar{\mathbf{x}}_k, \dot{\bar{\mathbf{x}}}_k) + \frac{d}{dt}\Big(\frac{\partial}{\partial v}L(\bar{\mathbf{x}}_k, \dot{\bar{\mathbf{x}}}_k) + \varpi\Big) = 0$$

and
$$\frac{\partial}{\partial v} L(\bar{\mathbf{x}}_k(T), \dot{\bar{\mathbf{x}}}_k(T)) + \varpi(T) + \Psi'(\bar{\mathbf{x}}_k(T)) = 0$$
for all $k = 1, ..., N$.

The preceding equations are the optimality conditions for the functional

$$\frac{1}{N} \sum_{i=1}^{N} \int_0^T L(\mathbf{x}_i(s), \dot{\mathbf{x}}_i(s)) ds + \Psi(\mathbf{x}_i(T)) + \int_0^T \varpi(s) \left( \frac{1}{N} \sum_{i=1}^{N} \dot{\mathbf{x}}_i(s) - Q(s) \right) ds,$$

and the solution constructed in **Proposition 4.5** satisfy the constraint (4.2). Thus, we can regard $\varpi$ as a Lagrange multiplier for (4.2).

## References


[1] M. Burger, L. A. Caffarelli, P. A. Markowich, and Marie-Therese Wolfram. On a Boltzmann-type price formation model. *Proc. R. Soc. Lond. Ser. A Math. Phys. Eng. Sci.*, 469(2157):20130126, 20, 2013.

[2] L. A. Caffarelli, P. A. Markowich, and J.-F. Pietschmann. On a price formation free boundary model by Lasry and Lions. *C. R. Math. Acad. Sci. Paris*, 349(11-12):621–624, 2011.

[3] A. Clemence, B. Tahar Imen, and M. Anis. An Extended Mean Field Game for Storage in Smart Grids. *ArXiv e-prints*, October 2017.

[4] R. Couillet, S.M. Perlaza, H. Tembine, and M. Debbah. Electrical vehicles in the smart grid: A mean field game analysis. *IEEE Journal on Selected Areas in Communications*, 30(6):1086–1096, 2012.

[5] A. De Paola, D. Angeli, and G. Strbac. Distributed control of microstorage devices with mean field games. *IEEE Transactions on Smart Grid*, 7(2):1119–1127, 2016.

[6] A. De Paola, V. Trovato, and D. Angeli. A mean field game approach for distributed control of thermostatic loads acting in simultaneous energy-frequency response markets. *IEEE Transactions on Smart Grid*, 2019.

[7] L. C. Evans. *Partial Differential Equations*. Graduate Studies in Mathematics. American Mathematical Society, 1998.

[8] D. Gomes, L. Lafleche, and L. Nurbekyan. A mean-field game economic growth model. *Proceedings of the American Control Conference*, 2016-July:4693–4698, 2016.

[9] D. Gomes and J. Saúde. A mean-field game approach to price formation. *To appear in Dynamic Games and Applications*, 2019.

[10] J. Graber and C. Mouzouni. Variational mean field games for market competition. *arXiv e-prints*, arXiv:1707.07853, Jul 2017.

[11] O. Guéant, J.-M. Lasry, and P.-L. Lions. Mean Field Games and Oil Production. 2010.

[12] A.C. Kizilkale and R.P. Malhame. A class of collective target tracking problems in energy systems: Cooperative versus non-cooperative mean field control solutions. *Proceedings of the IEEE Conference on Decision and Control*, 2015-February: 3493–3498, 2014.

[13] A.C. Kizilkale and R.P. Malhame. Collective target tracking mean field control for electric space heaters. *2014 22nd Mediterranean Conference on Control and Automation, MED 2014*, pages 829–834, 2014.

[14] A.C. Kizilkale and R.P. Malhame. Collective target tracking mean field control for markovian jump-driven models of electric water heating loads. *IFAC Proceedings Volumes (IFAC-PapersOnline)*, 19:1867–1872, 2014.

[15] J.-M. Lasry and P.-L. Lions. Mean field games. *Cahiers de la Chaire Finance et Développement Durable*, 2007.

[16] R. Malhamé and C.-Y. Chong. On the statistical properties of a cyclic diffusion process arising in the modeling of thermostat-controlled electric power system loads. *SIAM J. Appl. Math.*, 48(2):465–480, 1988.

[17] R. Malhamé, S. Kamoun, and D. Dochain. On-line identification of electric load models for load management. In *Advances in computing and control (Baton Rouge, LA, 1988)*, volume 130 of *Lect. Notes Control Inf. Sci.*, 290–304. Springer, Berlin, 1989.

[18] P. A. Markowich, N. Matevosyan, J.-F. Pietschmann, and M.-T. Wolfram. On a parabolic free boundary equation modeling price formation. *Math. Models Methods Appl. Sci.*, 19(10):1929–1957, 2009.

# Open and Closed Mirror Symmetry

*Lino Amorim*
Kansas State University
138 Cardwell Hall, 1228 N. 17th Street
Manhattan, KS 66506, USA
e-mail: lamorim@ksu.edu

**Resumo:** A simetria de espelho conjectura uma correspondência profunda entre a geometria simpléctica de um espaço e a geometria algébrica do seu "espelho". Existem várias versões desta correspondência, desde a igualdade de alguns invariantes numéricos, inicialmente conjecturada por físicos, a versões categóricas propostas por Kontsevich.

Este artigo revê algumas destas versões e ilustra-las num exemplo relativamente simples: uma esfera com três orbi-pontos (no lado simpléctico). Explicamos como construir o espaço "espelho", enunciamos as conjecturas de espelho e descrevemos uma abordagem à sua prova.

**Abstract** Mirror symmetry predicts a deep correspondence between the symplectic geometry of a space and the algebraic geometry of its "mirror". There are different versions of this correspondence, from the equality of some numerical invariants, first predicted by physicists, to categorical versions proposed by Kontsevich.

This paper reviews some of these versions and illustrates them on a relatively simple example: a sphere with three orbifold points (on the symplectic side). We explain how to construct the "mirror" space, state the mirror predictions and describe an approach to prove them.

**palavras-chave:** Simetria de espelho; categoria de Fukaya; orbi-variedade.

**keywords:** Mirror symmetry; Fukaya category; orbifold.

## 1 Introduction

### 1.1 A brief history

Mirror symmetry is a set of predictions from string theory relating the symplectic and complex geometry of certain pairs of Calabi-Yau manifolds. Superstring theory proposes that the space-time is (locally) of the form $\mathbb{R}^{1,3} \times X$, where $\mathbb{R}^{1,3}$ is the usual Minkowski space (that we see around us) and $X$ is a very (very) small Calabi-Yau three-fold. Meaning $X$ is a

Kähler manifold (therefore both complex and symplectic) of complex dimension three with a Ricci-flat metric. While looking for the $X$ that would help describe our universe, string theorists produced large lists of Calabi-Yau manifolds and found a surprising symmetry. There are many pairs of Calabi-Yau manifolds $X$ and $\check{X}$ which exchange Hodge numbers, that is $h^{1,1}(X) = h^{1,2}(\check{X})$ and $h^{1,2}(X) = h^{1,1}(\check{X})$.

The most famous example of this is the quintic threefold and its mirror. Let $X_a$ be the solution set in $\mathbb{CP}^4$ of the equation

$$X_a = \{z_0^5 + z_1^5 + z_2^5 + z_3^5 + z_4^5 - a z_0 z_1 z_2 z_3 z_4 = 0\} \subset \mathbb{CP}^4,$$

for some $a \in \mathbb{C}$. For most values of $a$ this is a smooth Calabi-Yau manifold. The group $(\mathbb{Z}/5)^3$ acts on $X_a$ and the quotient $X_a/G$ is singular but there is a resolution $\check{X}_a \to X_a/G$, which is smooth and the mirror partner of $X_a$.

Mirror symmetry predictions are much deeper than the equality of Hodge numbers. Candelas, de la Ossa, Green and Parkes [6] predicted the number of rational curves of degree $d$ in $X_a$ could be obtained from certain period integrals of the family $\check{X}_a$. This was remarkable since only the cases with $d \leq 3$ were known. This led to the development of Gromov-Witten invariants, a way to define precisely the counting of rational curves. Using Gromov-Witten invariants Givental [15] proved the predictions for the quintic. Both the (genus zero) Gromov-Witten invariants and the period integrals can be organized into *Frobenius* manifolds. Mirror symmetry then predicts an isomorphism between these two Frobenius manifolds. This is known as *closed-string mirror symmetry.*

In [18], Kontsevich proposed a new version of mirror symmetry at a categorical level. To a Calabi-Yau manifold $X$ one can associate two categories: the Fukaya category $Fuk(X)$ and the category of coherent sheaves $Coh(X)$. The Fukaya category is an $A_\infty$-category, which depends only on the symplectic structure of $X$, and whose objects are, roughly speaking, the Lagrangians submanifolds of $X$. The category of coherent sheaves $Coh(X)$ is an abelian category (which can be promoted to an $A_\infty$-category) which depends only on the complex structure of $X$. Kontsevich proposed that mirror symmetry exchanges these categories, that is, the derived categories of $Fuk(X)$ and $Coh(\check{X})$ are equivalent. This is usually called the homological mirror symmetry conjecture, or using physics terminology *open-string mirror symmetry.* This has been verified in several cases, see [23] for example.

Starting with the works of Givental and Batyrev it was suggested that mirror symmetry is not restricted to Calabi-Yau manifolds. It was conjectured that when $X$ is Fano [16] or when $X$ is of general type [17], there is also a mirror partner. In this case, the mirror is not simply a space, it's a non-compact manifold $\check{X}$ together with a holomorphic function $W : \check{X} \to \mathbb{C}$, which is called a Landau-Ginzburg model. For Landau-Ginzburg models one has to modify the mirror symmetry conjectures accordingly, for example replacing $Coh(\check{X})$ with the category of matrix factorizations $MF(\check{X}, W)$.

In 1996, Strominger-Yau-Zaslow [24] proposed a geometric explanation for mirror symmetry. Mirror (Calabi-Yau) pairs $X$ and $\check{X}$ should admit dual, special Lagrangian torus fibrations over the same base $B$. This is known as the SYZ conjecture. A proof of the SYZ conjecture seems to be out of reach and in fact these fibrations might only exist after deforming $X$. Nevertheless this conjecture has been very influential and inspired many important insights into mirror symmetry.

## 1.2 Family Floer theory

Mirror symmetry and the SYZ conjecture become more manageable if one takes a less symmetric approach. That is, if we consider the Calabi-Yau manifold $X$ just as a symplectic manifold and construct a variety (or rigid analytic space) $\check{X}$ over the (non-archimedian) Novikov field:

$$\Lambda := \left\{ \sum_{k=0}^{\infty} a_k T^{\lambda_k} | a_k \in \mathbb{C}, \lambda_k \in \mathbb{R}, \lambda_k \to \infty \right\}.$$

Then one can try to prove half of mirror symmetry, that is, to relate the symplectic geometry of $X$ (Gromov-Witten invariants or Fukaya category) to the algebraic/analytic geometry of $\check{X}$.

In this approach, proposed by Fukaya [11] (see also [19]), one starts with a (possibly singular) SYZ fibration or, more generally, some "interesting" family of Lagrangians in $X$ and constructs $\check{X}$ as the moduli space of objects in the Fukaya category supported on this family of Lagrangians. The fact that $Fuk(X)$ is a linear category over $\Lambda$ is then the reason why $\check{X}$ is not a complex manifold. This approach comes with an additional benefit: using family Floer cohomology (introduced by Fukaya), one has a canonically defined functor from $Fuk(X)$ to the category of coherent sheaves (or matrix factorizations) on $\check{X}$. This construction has been carried out for the case of smooth fibrations by Abouzaid [2].

In this note, we will illustrate these ideas in an example, first studied by Cho-Hong-Lau [9], which is as simple as possible: the family of Lagrangians consists of a single Lagrangian. The mirror will then be a Landau-Ginzburg model $(\check{X}, W)$ where $\check{X}$ is an affine space.

# 2 Orbifold spheres

## 2.1 Our example

Let $X := \mathbb{P}^1_{a,b,c}$ be an orbifold sphere with three orbifolds points with isotropy groups $\mathbb{Z}/a$, $\mathbb{Z}/b$, $\mathbb{Z}/c$, where $a, b, c \geq 2$. The orbifold Euler characteristic is given by $\chi\left(\mathbb{P}^1_{a,b,c}\right) = \frac{1}{a} + \frac{1}{b} + \frac{1}{c} - 1$. The orbifold $\mathbb{P}^1_{a,b,c}$ can be constructed as a global quotient of a Riemann surface $\Sigma$ by a finite group. If $\chi\left(\mathbb{P}^1_{a,b,c}\right) > 0$ then $\Sigma$ is a sphere, if $\chi\left(\mathbb{P}^1_{a,b,c}\right) = 0$ then $\Sigma$ is an elliptic curve and in the other cases $\Sigma$ is a surface of genus $\geq 2$. For example, $\mathbb{P}^1_{3,3,3} = E/(\mathbb{Z}/3)$, where $E$ is an elliptic curve with a $\mathbb{Z}/3$ symmetry.

We now introduce our Lagrangian: $\mathbb{L}$ which we call the Seidel Lagrangian, since it first appeared in [22]. This is an immersed circle $\mathbb{S}^1 \looparrowright \mathbb{P}^1_{a,b,c}$ with three transversal (double) self-intersections (see Figure 1). The three immersed points lie in the equator, determined by the three orbifold points. Moreover we assume that the image of $\mathbb{L}$ is invariant under reflection on the equator. The image of $\mathbb{L}$ and the equator divide the sphere into eight regions: two triangles and six bigons. We take $\mathbb{L}$ and scale the symplectic form so that each of these regions has area 1.

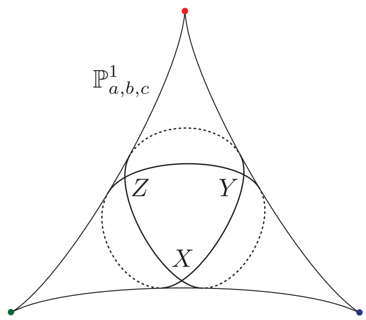



Figura 1: The orbifold sphere

## 2.2 The Fukaya algebra

We will start with a sketch of the construction of the Fukaya algebra of a Lagrangian submanifold. In fact, there is a family of Fukaya algebras parametrized by $H^*_{orb}(X)$ the orbifold cohomology of $X$. The orbifold cohomology is the singular cohomology of $\mathcal{I}X$ the inertia orbifold of $X$. In our example, we have

$$\mathcal{I}X = S^2 \bigcup_{i=1}^{a-1} pt \bigcup_{j=1}^{b-1} pt \bigcup_{k=1}^{c-1} pt,$$

that is a copy of $X$ (which as a topological space is a sphere) and one point for each non-trivial element in the isotropy groups of the three orbifold points in $X$. We define $H^*_{orb}(X) := H^*(\mathcal{I}X, \Lambda)$. The Novikov field $\Lambda$ has a real valuation $\nu : \Lambda \to \mathbb{R}$, given by the lowest power of $T$. Therefore $H^*_{orb}(X)$ inherits a valuation $\nu$. We fix $\tau \in H^*_{orb}(X)$ with $\nu(\tau) > 0$ and define the Fukaya algebra $\mathcal{F}_\tau(\mathbb{L})$.

Consider the fiber product $\mathbb{L} \times_X \mathbb{L}$, which in our example is

$$\mathbb{L} \times_X \mathbb{L} = S^1 \bigcup_{p=X,Y,Z} (p \cup p^-),$$

where $X, Y, Z$ are the self-intersections of $\mathbb{L}$. We define $\mathcal{F}_\tau(\mathbb{L}) := \Omega^*(\mathbb{L} \times_X \mathbb{L})\hat{\otimes}\Lambda$, where $\hat{\otimes}$ is the completed tensor product, with respect to the valuation induced by $\nu$. More concretely, $\mathcal{F}_\tau(\mathbb{L})$ consists of the de Rham complex of a circle plus two generators (one even, one odd) for each of the self-intersection points. We will now equip this space with a sequence of operations $\mathfrak{m}_k$ of arity $k \geq 0$.

Let $\Sigma$ be a orbifold which is topologically the closed unit disk in $\mathbb{C}$ and whose orbifold points lie in the interior. We take $k+1$ cyclically ordered marked points $z_0, \dots, z_k$ on the boundary of $\Sigma$ and $m$ marked points $w_1, \dots, w_m$ in the interior of $\Sigma$. We assume that each orbifold point is one of the $w_j$. Then we consider holomorphic maps $u : (\Sigma, \partial\Sigma) \to (X, \mathbb{L})$, with boundary on $\mathbb{L}$, in a fixed relative homology class $\beta \in H_2(X, \mathbb{L})$. We put a few more technical conditions on these maps, which in particular imply: 1) the restriction of $u$ to the boundary can only switch branches at self-intersections of $\mathbb{L}$ at one of the $z_i$'s; 2) orbifold points in $\Sigma$ are mapped to orbifold points in $X$, (see [8] for details). Then we consider the space of tuples $(\Sigma, z_0, \dots, z_k, w_1, \dots, w_m, u)$ modulo complex automorphisms of the domain. This space can be compactified by, roughly speaking, allowing the domain of the map $\Sigma$ to be a nodal disk, meaning a configuration of several disks and spheres attached at nodal points. For details see [13] for the manifold case and [7, 8] for the orbifold case. This is called the *stable map compactification* and we denote the resulting space by $\mathcal{M}_{k+1,m}(\beta)$. It follows from the work of Fukaya-Oh-Ohta-Ono [13] that the space $\mathcal{M}_{k+1,m}(\beta)$ is a compact Kuranishi space with boundary and corners. The definition of Kuranishi space is rather involved, we will use it as a black box to mean a space where we can pull-back and push-forward differential forms and the Stokes theorem works, in the same way as for manifolds.

It follows from the definition that these spaces have evaluation maps

$$\mathcal{I}X \xleftarrow{ev_{w_j}} \mathcal{M}_{k+1,m}(\beta) \xrightarrow{ev_{z_i}} \mathbb{L} \times_X \mathbb{L}.$$

For example, $ev_{z_i}(\Sigma, z_0, \ldots, z_k, w_1, \ldots, w_m, u) = u(z_i)$. We are now ready to define the $A_\infty$ maps $\mathfrak{m}_k^\tau : \mathcal{F}_\tau(\mathbb{L})^{\otimes k} \to \mathcal{F}_\tau(\mathbb{L})$ by the formula

$$\mathfrak{m}_k^\tau(h_1, \ldots, h_k) = \sum_{\beta, m \geq 0} \frac{T^{\omega(\beta)}}{m!} (ev_{z_0})_* (ev_{w_1}^* \tau \wedge \ldots \quad \wedge ev_{w_m}^* \tau \wedge ev_{z_1}^* h_1 \\ \wedge \ldots \wedge ev_{z_k}^* h_k).$$

The following theorem follows from the work of Fukaya–Oh–Ohta–Ono [12, 13] and further generalizations by Akaho–Joyce [3] and Cho–Poddar [8].

**Theorem 2.1** *$\mathcal{F}(\mathbb{L})$ with the operations $\{\mathfrak{m}_k^\tau\}_{k \geq 0}$ is a filtered $A_\infty$-algebra.*

What is a filtered $A_\infty$-algebra? Let's define this.

**Definition 2.2** *A filtered $A_\infty$-algebra is a $\mathbb{Z}/2$-graded $\Lambda$-vector space $\mathcal{A}$ of the form $\mathcal{A} = A_0 \hat{\otimes} \Lambda$, where $A_0$ is a complex vector space. There are maps $\mathfrak{m}_k : \mathcal{A}^{\otimes k} \to \mathcal{A}$ of degree $k \pmod 2$, for each $k \geq 0$, satisfying*

$$\sum_{\substack{0 \leq j \leq n \\ 0 \leq i \leq n-j}} (-1)^{|a_1| + \ldots + |a_i| + i} \mathfrak{m}_{n-j+1}(a_1, \ldots, \mathfrak{m}_j(a_{i+1}, \ldots, a_{i+j}), \ldots, a_n) = 0.$$

*Moreover $\nu(\mathfrak{m}_k(a_1, \ldots, a_k)) \geq \sum_i \nu(a_i)$ and $\nu(\mathfrak{m}_0) > 0$. We will also require that the $A_\infty$-algebra is unital: there is an even element $\mathbb{1}$ satisfying:*

$$\mathfrak{m}_2(\mathbb{1}, a) = (-1)^{|a|} \mathfrak{m}_2(a, \mathbb{1}) = a, \quad , \mathfrak{m}_k(\ldots, \mathbb{1}, \ldots) = 0, \ k \neq 2.$$

If we stare at the $A_\infty$ equation above for $n = 1, 2, 3$, we can easily see that when $\mathfrak{m}_0$ is a multiple of the unit $\mathbb{1}$, then $\mathfrak{m}_1$ is a differential and so we can consider the cohomology of the $A_\infty$-algebra. Moreover $\mathfrak{m}_2$ then defines an associative product on the cohomology. In general, filtered $A_\infty$-algebras are rather complicated objects, so when we have to work with one we try to deform it (when possible) to another where this condition holds. In order to do that we need to solve the *Maurer-Cartan* equation.

**Definition 2.3** *A Maurer-Cartan element in $\mathcal{A}$ is an odd element $b$ satisfying $\sum_{k \geq 0} \mathfrak{m}_k(b, \ldots, b) = \lambda \mathbb{1}$, for some $\lambda \in \Lambda$, called the potential of $b$.*

Note that the sum on the left hand side of the equation is in general an infinite sum, so we need to ensure convergence. The safest way to do this is to require that $\nu(b) > 0$, but as we will see in our example, this can sometimes be relaxed. Given a Maurer-Cartan element we can define a new $A_\infty$ structure on $\mathcal{A}$ by setting

$$\mathfrak{m}_k^b(a_1, \ldots, a_k) := \sum_{i_0, \ldots, i_k} \mathfrak{m}_{k+i_0+\ldots+i_k}(b, \ldots, b, a_1, b, \ldots, b, a_k, b, \ldots, b).$$

By construction $\mathfrak{m}_0^b = \lambda \mathbb{1}$. Maurer-Cartan elements for the Fukaya algebra $\mathcal{F}_\tau(L)$ of a Lagrangian $L$ are called bounding cochains. Objects in the Fukaya category $Fuk(X, \tau) := \oplus_\lambda Fuk_\lambda(X, \tau)$ are pairs $(L, b)$ where $L$ is a Lagrangian and $b$ is a Maurer-Cartan element in $\mathcal{F}_\tau(L)$ with potential $\lambda$. The endomorphism space of the object $(L, b)$ is then $H^*(\mathcal{F}_\tau(L), \mathfrak{m}_1^{\tau,b})$.

# 3 The mirror

## 3.1 Potential

Like we promised in the introduction, we will construct the mirror to $X = \mathbb{P}^1_{a,b,c}$ as the moduli space of objects in the Fukaya category of $X$ supported on the Seidel Lagrangian $\mathbb{L}$. More precisely we will construct a mirror for the pair $(\mathbb{P}^1_{a,b,c}, \tau)$, where $\tau \in H^*_{orb}(X)$. As explained in the previous section, the moduli space of these objects is exactly the space of Maurer-Cartan elements in $\mathcal{F}_\tau(\mathbb{L})$. In [4] we prove the following proposition.

**Proposition 3.1** *Let $X, Y, Z$ be the odd generators of $\mathcal{F}_\tau(\mathbb{L})$ corresponding to the three self-intersections. All elements of the form $b = T^{-3}(xX + yY + zZ)$, for elements $x, y, z \in \Lambda$ of non-negative valuation, are Maurer-Cartan elements with potential $W_\tau(x, y, z)$.*

At this point $W_\tau(x, y, z)$ is just a formal series on $x, y, z$, but in fact it is convergent in the following (non-archimedian) sense.

**Definition 3.2** *A convergent power series is an expression of the form $\sum_{i,j,k \in \mathbb{Z}_{\geq 0}} c_{i,j,k} x^i y^j z^k$, with $c_{i,j,k} \in \Lambda$ and $\lim_{i+j+k \to \infty} \nu(c_{i,j,k}) = +\infty$. The set of all convergent power series naturally forms a ring which we denote by $\Lambda\langle\langle x, y, z\rangle\rangle$.*

Let us explain the terminology here. We can define a non-archimedian norm on $\Lambda$ by setting $|v| := e^{-\nu(v)}$. Then one can see that $\Lambda\langle\langle x, y, z\rangle\rangle$ is exactly the ring of regular functions on the unit polydisk, see [5].

**Proposition 3.3 ([4])** *$W_\tau$ is a convergent power series. Moreover*

$$W_\tau = T^{-8}xyz + x^a + y^b + z^c + \textit{positive valuation in } T.$$

It turns out that when $\chi(\mathbb{P}^1_{a,b,c}) \geq 0$, $W_\tau$ is actually a polynomial. An explicit description of $W_\tau$ for arbitrary $\tau$ seems out of reach, but for our purposes knowing the leading order term in the above proposition is enough.

We are finally ready to define the mirror partner to $\mathbb{P}^1_{a,b,c}$.

**Definition 3.4** *The mirror to $(\mathbb{P}^1_{a,b,c}, \tau)$ is the Landau-Ginzburg model*

$$\check{X} = \mathcal{B} = \{(x, y, z), |x|, |y|, |z| \leq 1\} \subseteq \Lambda^3, \quad W_\tau : \mathcal{B} \to \Lambda.$$

One might ask why this is the correct mirror. Even assuming our philosophy that the mirror should be given as the moduli of objects in the Fukaya category supported in a certain family of Lagrangians in $X$, why is the Seidel Lagrangian the correct family? And even assuming that, how do we know we have "enough" bounding cochains? I don't believe there is a completely satisfactory answer to these questions. The short answer is that it works, meaning the closed-string mirror symmetry conjecture, that we will state in the next subsection, holds for this pair. Once we have established closed mirror symmetry, Abouzaid's generation criterion [1] tells us that, loosely speaking, our family of objects of the Fukaya category is "large" enough and therefore we have constructed the correct mirror and should expect open mirror symmetry to also hold for this pair.

## 3.2 Closed mirror symmetry

The closed mirror symmetry conjecture is an isomorphism of Frobenius manifolds. We will not define Frobenius manifold (see [20] for the complete definition), instead we will work at a more elementary level and consider it as a family of commutative algebras with a compatible inner product. In our situation, the families (on both sides of the mirror) are parameterized by $\tau \in H^*_{orb}(X)$.

On the symplectic side, the Frobenius manifold is the orbifold quantum cohomology, defined by Chen-Ruan [7]. The construction is similar to the construction of the Fukaya algebra. For each homology class $\alpha \in H_2(X)$, one constructs $\mathcal{M}^{sph}_{\ell+3}(\alpha)$ the moduli space of stable holomorphic *orbi-spheres* in $X$ with $\ell + 3$ marked points $w_1, \ldots, w_{\ell+3}$. Then we fix $\tau$ as before and define a product $\bullet_\tau$ on $H^*(\mathcal{I}X, \Lambda)$ as follows

$$A \bullet_\tau B := \sum_{\alpha,\ell\geq 0} \frac{T^{\omega(\alpha)}}{\ell!}(ev_{w_1})_*(ev_{w_2}^* A \wedge ev_{w_3}^* B \wedge ev_{w_4}^* \tau \wedge \ldots \wedge ev_{w_{\ell+3}}^* \tau).$$

**Theorem 3.5 ([7])** *The map $\bullet_\tau$ defines a commutative, associative product on $H_{orb}^*(X,\Lambda)$, compatible with the Poincaré pairing. We denote it by $QH_{orb}^*(X,\bullet_\tau)$.*

On the mirror, things are somewhat easier to define.

**Definition 3.6** *The Jacobian of $W_\tau$ is the ring obtained by taking the quotient of $\Lambda\langle\langle x,y,z\rangle\rangle$ by the ideal generated by the partial derivatives of $W_\tau$.*

$$Jac(W_\tau) = \frac{\Lambda\langle\langle x,y,z\rangle\rangle}{< \partial_x W_\tau, \partial_y W_\tau, \partial_z W_\tau >}.$$

In order to define an inner product in $Jac(W_\tau)$, one has to fix a volume form and then take the residue pairing. This is related to the choice of a primitive form as defined by Saito [21].

There is a natural map $\mathsf{KS}_\tau : QH^*(X,\bullet_\tau) \to Jac(W_\tau)$, called the Kodaira-Spencer map. We fix a basis $\{e_i\}_i$ of $H^*(\mathcal{I}X,\Lambda)$ and write $\tau = \sum_i \tau_i e_i$. We define the map by the formula $\mathsf{KS}_\tau(e_i) = \dfrac{\partial}{\partial \tau_i} W_\tau$.

This map was originally constructed by Fukaya–Oh–Ohta–Ono [14] for toric manifolds. They show that this is a well-defined, unital ring map. In fact, this is expected to be the case for a very wide class of symplectic manifolds/orbifolds. In [4], we extend their construction to our example and prove the following.

**Theorem 3.7** *The Kodaira-Spencer $\mathsf{KS}_\tau : QH^*(X,\bullet_\tau) \to Jac(W_\tau)$ is an unital, ring isomorphism.*

This theorem is not the complete closed mirror symmetry statement. The full-fledged statement requires an identification of the Euler vector fields, which we prove in [4]:

$$\mathsf{KS}_\tau\left(c_1(X) + \sum_i (1 - \frac{\deg e_i}{2})\tau_i e_i\right) = [W_\tau].$$

Moreover the Kodaira-Spencer should intertwine the Poincaré pairing with the residue pairing on $Jac(W_\tau)$ determined by some volume form $\omega_\tau$. A complete description of $\omega_\tau$ is still work in progress by Cho, Hong, Lau and myself.

### 3.3 Open mirror symmetry

The open (or homological) mirror symmetry conjecture in this example asserts that the derived categories of $Fuk(X, \tau)$ and $MF(W_\tau)$ the category of matrix factorizations of $W_\tau$ are equivalent. The matrix factorizations category is a dg-category, which captures some information about the singularities of $W_\tau$. We refer the reader to [10] for the definition.

One of the main advantages of this formalism, is that $\mathbb{L}$ determines, for each $\tau$, an $A_\infty$-functor $\mathcal{M}^{\mathbb{L}} : Fuk(X, \tau) \to MF(W_\tau)$. This is a version of the Yoneda embedding, see [9] for a full description. We expect the following to hold.

**Conjecture 3.8** *The functor $\mathcal{M}^{\mathbb{L}}$ induces an equivalence*

$$D^\pi Fuk_\lambda(X, \tau) \to D^\pi MF(W_\tau),$$

*where $D^\pi$ stands for the split-closed derived category.*

This conjecture was proved in some cases in [9] when $\tau = 0$ and is work in progress by Cho, Hong, Lau and myself. But we are not that far off from proving this. First note that the closed mirror symmetry statement that we saw in the previous subsection implies that $Jac(W_\tau)$ is finite dimensional, which implies that the critical points of $W_\tau$ are isolated. It then follows from Dyckerhoff [10] that $MF(W_\tau)$ has finitely many generators $P^\eta$, one for each critical point $\eta \in Crit(W_\tau)$. We prove in [4] that each $\eta$ also determines a bounding cochain $b_\eta$ for $\mathbb{L}$. Therefore it is not unreasonable to expect that $\mathcal{M}^{\mathbb{L}}$ sends the object $(\mathbb{L}, b_\eta)$ to $P^\eta$ and it is fully faithful (on cohomology) when restricted to these objects.

Assuming we can prove this, the only thing left to show is that the objects $(\mathbb{L}, b_\eta)$ split-generate the Fukaya category. This should follow from a suitable generalization of Abouzaid's generation criterion [1]. Let's explain how this criterion works. Let $\mathcal{A}$ be the subcategory of $Fuk(X, \tau)$ generated by the $(\mathbb{L}, b_\eta)$. There is a ring map

$$\mathcal{CO} : HH^*(\mathcal{A}) \to QH^*(X, \bullet_\tau),$$

whose domain is the Hochschild cohomology of $\mathcal{A}$. The generation criterion asserts that if the map $\mathcal{CO}$ is injective then $\mathcal{A}$ is derived equivalent to $Fuk(X, \tau)$. The reason this should hold in our example is the following. We expect the Hochschild cohomology $HH^*(\mathcal{A})$ to be isomorphic to the Jacobian $Jac(W_\tau)$, and under this isomorphism, the map $\mathcal{CO}$ should agree with the Kodaira-Spencer map. Therefore the condition needed for the generation criterion follows from the fact that $\mathsf{KS}$ is an isomorphism, in other words, it follows from closed mirror symmetry.

## References

[1] M. Abouzaid, "A geometric criterion for generating the Fukaya category", *Publ. Math. IHÉS*, Vol. 112 (2010), pp. 191–240.

[2] M. Abouzaid, "The family Floer functor is faithful", *J. Eur. Math. Soc. (JEMS)*, Vol. 19, no. 7 (2017), pp. 2139–2217.

[3] M. Akaho, D. Joyce, "Immersed Lagrangian Floer theory", *J. Differential Geom*, Vol. 86, no. 3 (2010), pp. 381–500.

[4] L. Amorim, C. Cho, H. Hong, S. Lau, "Big quantum cohomology of orbifold spheres", arXiv:2002.11180.

[5] S. Bosch, U. Güntzer, R. Remmert, *Non-Archimedean analysis. A systematic approach to rigid analytic geometry*. Grundlehren der Mathematischen Wissenschaften, 261. Springer-Verlag, Berlin, 1984.

[6] P. Candelas, X. de la Ossa, P. Green, L. Parkes, "A pair of Calabi–Yau manifolds as an exact soluble superconformal theory", *Nucl. Phys. B*, Vol. 359 (1991), pp. 21–74.

[7] W. Chen, Y. Ruan, "Orbifold Gromov-Witten theory", *Orbifolds in mathematics and physics*, Contemp. Math., Vol. 310, Amer. Math. Soc., Providence, RI, 2002, pp. 25–85.

[8] C. Cho; M. Poddar, "Holomorphic orbi-discs and Lagrangian Floer cohomology of symplectic toric orbifolds', *J. Differential Geom.*, Vol. 98, no. 1 (2014), pp. 21–116.

[9] C. Cho, H. Hong, S. Lau, "Localized mirror functor for Lagrangian immersions, and homological mirror symmetry for $\mathbb{P}^1_{a,b,c}$", *J. Differential Geom.*, Vol. 106, no. 1 (2017), pp. 45–126.

[10] T. Dyckerhoff, "Compact generators in categories of matrix factorizations", *Duke Math. J.*, Vol. 159, no. 2 (2011), pp. 223–274.

[11] K. Fukaya, "Multivalued Morse theory, asymptotic analysis and mirror symmetry", *Graphs and patterns in mathematics and theoretical physics*, Proc. Sympos. Pure Math., Vol. 73, Amer. Math. Soc., 2005, pp. 205–278.

[12] K. Fukaya, "Cyclic symmetry and adic convergence in Lagrangian Floer theory", *Kyoto J. Math.*, Vol. 50 (2010), pp. 521–590.

[13] K. Fukaya, Y. Oh, H. Ohta, K. Ono, *Lagrangian intersection Floer theory: anomaly and obstruction. Parts I and II*, Vol. 46, AMS/IP Studies in Advanced Mathematics, American Mathematical Society, Providence, RI, 2009.

[14] K. Fukaya, Y. Oh, H. Ohta, K. Ono, “Lagrangian Floer theory and Mirror symmetry on compact toric manifolds”, *Astérisque*, No. 376 (2016).

[15] A. Givental, “Equivariant Gromov-Witten invariants”, *Int. Math. Res. Not.*, no. 13 (1996), pp. 613–663.

[16] K. Hori, C. Vafa, “Mirror symmetry”, *arXiv:hep-th/0002222*.

[17] A. Kapustin, L. Katzarkov, D. Orlov, M. Yotov, “Homological mirror symmetry for manifolds of general type”, *Cent. Eur. J. Math.*, Vol. 7 (2009), pp. 571–605.

[18] M. Kontsevich, “Homological algebra of mirror symmetry”, *Proceedings of the International Congress of Mathematicians*, Vol. 1, 2 (Zürich, 1994), 120–139, Birkhäuser, Basel, 1995.

[19] M. Kontsevich, Y. Soibelman, “Affine structures and non-Archimedean analytic spaces”, *The unity of mathematics*, Progr. Math., Vol. 244, Birkhäuser, Boston, 2006, pp. 321–385.

[20] Y. Manin, *Frobenius manifolds, quantum cohomology, and moduli spaces.* American Mathematical Society Colloquium Publications 47, Providence, RI, 1999.

[21] K. Saito, “Period mapping associated to a primitive form”, *Publ. Res. Inst. Math. Sci.*, Vol. 19 , no. 3 (1983), pp. 1231–1264.

[22] P. Seidel, “Homological mirror symmetry for the genus two curve”, *J. Algebraic Geom.*, Vol. 20, no. 4 (2011), pp. 727–769.

[23] N. Sheridan, “Homological Mirror Symmetry for Calabi-Yau hypersurfaces in projective space”, *Inventiones Mathematicae*, Vol. 199, no. 1 (2015), pp. 1–186.

[24] A. Strominger, S. Yau, E. Zaslow, “Mirror Symmetry is T -duality”, *Nucl. Phys. B*, Vol. 479 (1996), pp. 243–259.

# Recent progress on the mathematical theory of plasmas

*Diogo Arsénio*
New York University Abu Dhabi
Abu Dhabi, United Arab Emirates
e-mail: diogo.arsenio@nyu.edu

**Resumo:** O sistema de Navier–Stokes–Maxwell incompressível é um modelo clássico que descreve a evolução de um plasma. Embora se saiba que existem pequenas soluções suaves para esse sistema (no espírito de Fujita–Kato), a existência de grandes soluções fracas (no espírito de Leray) no espaço de energia permanece desconhecida. Esse defeito pode ser atribuído à dificuldade de acoplar as equações de Navier–Stokes a um sistema hiperbólico. Nós descrevemos aqui resultados recentes, com o objetivo de criar soluções fracas para os sistemas de Navier–Stokes–Maxwell em grandes espaços funcionais. Em particular, explicamos como, para quaisquer dados iniciais com energia finita, uma condição de pequenez apenas no campo electromagnético é suficiente para garantir a existência de soluções globais.

**Abstract:** The incompressible Navier–Stokes–Maxwell system is a classical model describing the evolution of a plasma. Although small smooth solutions to this system (in the spirit of Fujita–Kato) are known to exist, the existence of large weak solutions (in the spirit of Leray) in the energy space remains unknown. This defect can be attributed to the difficulty of coupling the Navier–Stokes equations with a hyperbolic system. We describe here recent results aiming at building weak solutions to Navier–Stokes–Maxwell systems in large functional spaces. In particular, we explain how, for any initial data with finite energy, a smallness condition on the electromagnetic field alone is sufficient to grant the existence of global solutions.

**palavras-chave:** Equações de Navier–Stokes; equações de Maxwell; existência de soluções fracas; estimativas parabólicas; espaços de Besov.

**keywords:** Navier–Stokes equations; Maxwell's equations; existence of weak solutions; parabolic estimates; Besov spaces.

## 1 Introduction

Consider a gas made up of charged particles interacting microscopically through elastic collisions. At the macroscopic level, this gas behaves as a conducting fluid that will interact with any existing electromagnetic field. Moreover, the motion of the charged particles will also produce an electromagnetic field, in accordance with the laws of classical electrodynamics.

The magnetohydrodynamic evolution of the gas will therefore be conditioned by the complex interaction of an electrically conducting moving fluid with a self-induced electromagnetic force.

Such fluids are typically found in the core of nuclear fusion reactors in the form of plasmas, which are ionized gases. Another typical example of an electrically conducting fluid consists in liquid metals, such as the liquid iron found in the core of the earth, which is responsible for the geodynamo effect.

We give now an account of some recent mathematical developments, mainly from [1], concerning the study of plasmas (or conducting fluids). We make here the somewhat arbitrary choice of focusing exclusively on viscous incompressible regimes, because such physical characteristics lead to interesting mathematical properties. Of course, there are numerous other relevant regimes, but we will not discuss them.

We refer to [4] or [5] for a introduction to magnetohydrodynamics from a physical viewpoint.

## 2 The Navier–Stokes–Maxwell systems

The behavior of a viscous incompressible fluid is described by the Navier–Stokes equations

$$\partial_t u + u \cdot \nabla u - \mu \Delta u = -\nabla p + F, \quad \operatorname{div} u = 0, \tag{1}$$

where $\mu > 0$ is the viscosity, $t \in \mathbb{R}^+$ and $x \in \mathbb{R}^3$ are the time and space variables, $u(t,x)$ stands for the velocity field of the (incompressible) fluid, $F(t,x)$ is a given force field, and the scalar function $p(t,x)$ is the pressure and is also an unknown. Note that, for convenience, we ignore the effect of boundaries on the fluid by assuming the domain to be the whole space.

The validity of this model is well established at both physical and mathematical levels. We refer to [8] for a recent mathematical treatise on the incompressible Navier–Stokes equations.

In a conducting fluid, it is also important to take into account the influence of the Lorentz force produced by the charged particles. The relevant macroscopic field $F$ is therefore the Lorentz force

$$F = nE + j \times B, \tag{2}$$

where $E(t,x)$ and $B(t,x)$ are the electric and magnetic fields respectively, $n(t,x)$ is the electric charge density and $j(t,x)$ is the electric current.

The electromagnetic field is determined classically through Maxwell's equations

$$\begin{cases} \partial_t E - \nabla \times B = -j, & \operatorname{div} E = n, \\ \partial_t B + \nabla \times E = 0, & \operatorname{div} B = 0, \end{cases} \tag{3}$$

or its quasi-static approximation

$$\begin{cases} \nabla \times B = j, & \operatorname{div} E = n, \\ \partial_t B + \nabla \times E = 0, & \operatorname{div} B = 0. \end{cases} \tag{4}$$

Generally speaking, the coupling given by combining (1), (2) and (3) (or (4)) provides now an incompressible Navier–Stokes–Maxwell system. Note, however, that such a system is not closed yet, as it contains more unknowns than equations. In fact, there remains to specify how the density $n$ and the current $j$ are generated by the fluid. This is performed by incorporating the so-called Ohm's law into the system.

It turns out that there is more than one way of closing the Navier–Stokes–Maxwell system, as there are several different Ohm's laws that are appropriate. We discuss now some of the available options.

## 2.1 Coupling I

The quasi-static system (4) is an approximation of (3) that is relevant in many physical regimes. Indeed, in many practical situations, it is physically reasonable to neglect the so-called displacement current density $\partial_t E$ in Maxwell's equations (see [5]).

Furthermore, observe that the continuity equation

$$\partial_t n + \operatorname{div} j = 0 \tag{5}$$

is expected to hold universally, for $n$ and $j$ respectively represent the density and the flux of the same particles. Since $j$ is necessarily solenoidal (i.e. $\operatorname{div} j = 0$) in the quasi-static approximation due to Ampère's law $j = \nabla \times B$, one deduces that $n$ should be constant in time. The density $n$ is therefore fixed by the initial data and we might as well assume $n = 0$, for simplicity.

Now, recall that, in classical electrostatics, Ohm's law simply states that $E$ and $j$ are colinear. Here, accounting for the motion of the fluid and the effect of Galilean transformations in Faraday's equation $\partial_t B + \nabla \times E = 0$, Ohm's law becomes (see [5])

$$j = \sigma(E + u \times B), \tag{6}$$

where the electrical conductivity $\sigma > 0$ is assumed to be constant throughout the fluid.

All in all, combining (1), (2), (4) with (6), setting $n = 0$, and eliminating $j$ and $E$, leads to the magnetohydrodynamic system

$$\begin{cases} \partial_t u + u \cdot \nabla u - \mu \Delta u = -\nabla p + (\nabla \times B) \times B, & \operatorname{div} u = 0, \\ \partial_t B - \dfrac{1}{\sigma} \Delta B = \nabla \times (u \times B), & \operatorname{div} B = 0. \end{cases}$$

This system couples the Navier–Stokes system with a parabolic equation on the magnetic field $B$ and has been studied extensively. As far as the existence of global weak solutions is concerned, it does not present with any additional difficulty when compared to the classical incompressible Navier–Stokes system.

Indeed, one readily computes the formal energy inequality, for any $t \geq 0$ and any initial data $(u_0, B_0)$,

$$\begin{aligned} \frac{1}{2} \left( \|u\|_{L^2}^2 + \|B\|_{L^2}^2 \right)(t) + \int_0^t \left( \mu \|\nabla u\|_{L^2}^2 + \frac{1}{\sigma} \|\nabla B\|_{L^2}^2 \right)(s) ds \\ \leq \frac{1}{2} \left( \|u_0\|_{L^2}^2 + \|B_0\|_{L^2}^2 \right). \end{aligned} \tag{7}$$

This energy inequality yields strong dissipative properties on both $u$ and $B$. In particular, the ensuing a priori bounds are suitable for the application of Leray's method of construction of global weak solutions (see [8]). More precisely, it is possible to show that, for any suitable initial data $(u_0, B_0) \in L^2(\mathbb{R}^3)$, there exists a global weak solution

$$(u, B) \in L^\infty(\mathbb{R}^+, L^2(\mathbb{R}^3)) \cap L^2(\mathbb{R}^+, \dot{H}^1(\mathbb{R}^3)).$$

The uniqueness of such solutions remains unknown, though.

## 2.2 Coupling II

The reduced form of Maxwell's equations (4) may not be appropriate for every physical setting, and there may be situations where one is led to consider the evolution of the electromagnetic field $(E, B)$ governed by the full set of Maxwell's equations (3). In this case, one may combine (1), (2) and (6), with (3), which yields the Navier–Stokes–Maxwell system

$$\begin{cases} \partial_t u + u \cdot \nabla u - \mu \Delta u = -\nabla p + j \times B, & \operatorname{div} u = 0, \\ \partial_t E - \nabla \times B = -j, & j = \sigma (E + u \times B), \\ \partial_t B + \nabla \times E = 0, & \operatorname{div} B = 0, \end{cases} \tag{8}$$

where we have neglected the contribution of the Coulombian force $nE$ in the Lorentz force for physical reasons (see [5]). This system couples now the Navier–Stokes equations with a hyperbolic wave system, which significantly changes the nature of solutions.

More precisely, formally computing the corresponding energy inequality, one finds that, for any initial data $(u_0, E_0, B_0)$,

$$\begin{aligned}\frac{1}{2}\left(\|u\|_{L^2}^2 + \|E\|_{L^2}^2 + \|B\|_{L^2}^2\right)(t) + \int_0^t \left(\mu \|\nabla u\|_{L^2}^2 + \frac{1}{\sigma}\|j\|_{L^2}^2\right)(s)ds \\ \leq \frac{1}{2}\left(\|u_0\|_{L^2}^2 + \|E_0\|_{L^2}^2 + \|B_0\|_{L^2}^2\right).\end{aligned} \tag{9}$$

When compared to (7), this energy inequality only provides a rather weak control on the solutions, for there is no control on the regularity of the magnetic field.

This lack of compactness prevents us from applying Leray's method of construction of weak solutions, because it is impossible to show the weak stability of the non-linear term $j \times B$ solely based on the a priori bounds given by the energy inequality. As a matter of fact, it is not yet known whether, for any suitable initial data $(u_0, E_0, B_0) \in L^2(\mathbb{R}^3)$, there exists a global weak solution to the Navier–Stokes–Maxwell system (8).

It should be emphasized now that, even though the above system (8) elegantly combines the Navier–Stokes equations with the full Maxwell system, it contains a disturbing physical inconsistency. Indeed, as previously mentioned, we have neglected the term $nE$ in (2), which suggests that $n$ should be zero. However, in this model, the electric current $j$ is in general not solenoidal, which violates the continuity equation (5).

This inconsistency will be resolved in the coming couplings (10) and (11) below, which achieve to combine the Navier–Stokes equations with Maxwell's equations without breaking the continuity equation (5).

### 2.3 Coupling III

A systematic and rigorous study of hydrodynamic limits of Vlasov–Maxwell–Boltzmann systems, in a viscous incompressible regime, has been conducted in [2], where the following incompressible Navier–Stokes–Maxwell system was derived:

$$\begin{cases} \partial_t u + u \cdot \nabla u - \mu \Delta u = -\nabla p + j \times B, & \operatorname{div} u = 0, \\ \partial_t E - \nabla \times B = -j, & \operatorname{div} B = 0, \\ \partial_t B + \nabla \times E = 0, & \operatorname{div} E = 0, \\ j = \sigma \left( -\nabla \bar{p} + E + u \times B \right), & \operatorname{div} j = 0, \end{cases} \tag{10}$$

where the electromagnetic pressure $\bar{p}(t,x)$ is a new unknown. Observe that the introduction of the pressure $\bar{p}$ allows us to add a solenoidal condition on both $E$ and $j$ to the system. As a result, the fluid is neutral $n = 0$ and the continuity equation (5) holds.

As before, this system combines the incompressible Navier–Stokes equations with a hyperbolic system. One easily finds that solutions of (10) formally verify the same energy inequality (9), which fails to provide the necessary compactness to apply Leray's method of proof of existence of weak solutions. Again, it is unfortunately not yet known whether, for any suitable initial data $(u_0, E_0, B_0) \in L^2(\mathbb{R}^3)$, there exists a global weak solution to the Navier–Stokes–Maxwell system (10).

### 2.4 Coupling IV

Yet another incompressible Navier–Stokes–Maxwell system was derived in [2]. This new model turns out to be the most complete of them all, since it involves all electromagnetic variables (including a non-trivial charge density $n$). It takes the following form:

$$\begin{cases} \partial_t u + u \cdot \nabla u - \mu \Delta u = -\nabla p + nE + j \times B, & \operatorname{div} u = 0, \\ \partial_t E - \nabla \times B = -j, & \operatorname{div} B = 0, \\ \partial_t B + \nabla \times E = 0, & \operatorname{div} E = n, \\ j - nu = \sigma \left( -\nabla n + E + u \times B \right). \end{cases} \tag{11}$$

Here, again, it is to be noted that the continuity equation (5) holds true.

It is possible to show, at least formally, that solutions of the above system satisfy the energy inequality, for any initial data $(u_0, E_0, B_0, n_0)$,

$$\begin{aligned} \frac{1}{2} \left( \|u\|_{L^2}^2 + \|E\|_{L^2}^2 + \|B\|_{L^2}^2 + \|n\|_{L^2}^2 \right)(t) & \\ + \int_0^t \left( \mu \|\nabla u\|_{L^2}^2 + \frac{1}{\sigma} \|j - nu\|_{L^2}^2 \right)(s) ds & \\ \leq \frac{1}{2} \left( \|u_0\|_{L^2}^2 + \|E_0\|_{L^2}^2 + \|B_0\|_{L^2}^2 + \|n_0\|_{L^2}^2 \right). & \end{aligned}$$

As previously, the ensuing a priori bounds fail to provide enough control to apply Leray's method of proof of existence of weak solutions. It is therefore not yet known whether, for any suitable initial data $(u_0, E_0, B_0, n_0) \in L^2(\mathbb{R}^3)$, there exists a global weak solution to the Navier–Stokes–Maxwell system (11).

## 3 A global existence result

We believe that the system (8) captures the essential mathematical difficulties related to the coupling of the Navier–Stokes equations with Maxwell's system. We therefore present below the main result from [1] on the existence of weak solutions to the Navier–Stokes–Maxwell system (8).

It should be mentioned here, though, that the two-dimensional case has been previously succesfully handled in [10] (some subtle questions remain open; see also [1] and [6] for some two-dimensional results). We will therefore focus now exclusively on the three-dimensional setting of (8).

The existence of three-dimensional global mild solutions to (8), for small initial data, has also been previously addressed in [6] (and some previous works), where it was shown that, for any sufficiently small initial data $(u_0, B_0, E_0) \in \dot{H}^{\frac{1}{2}}(\mathbb{R}^3)$, there exists a global mild solution $(u, E, B) \in C(\mathbb{R}^+, \dot{H}^{\frac{1}{2}})$ to (8) (uniqueness of solutions is also available in this setting).

As for weak solutions, the following theorem from [1] provides the existence of global solutions to (8), for any initial data $(u_0, B_0, E_0) \in L^2(\mathbb{R}^3)$, provided the high frequencies of the electromagnetic field are controlled in some suitable norm.

**Theorem 1** ([1]). *There is a constant $C_* > 0$ such that, if the initial data $(u_0, E_0, B_0) \in L^2 \times (H^{\frac{1}{2}})^2$, with $\operatorname{div} u_0 = \operatorname{div} B_0 = 0$, satisfies*

$$\|(E_0, B_0)\|_{\dot{H}^{\frac{1}{2}}} C_* e^{C_* \|(u_0, E_0, B_0)\|_{L^2}^2} \leq 1 \,,$$

*then there is a global weak solution to (8) satisfying the energy inequality (9).*

The strategy of proof of this result follows the usual procedure of approximating (8) with a regularized system, in order to justify all formal a priori bounds, and then passing to the limit by showing the weak stability of the system.

As previously explained, the a priori bounds provided by the energy inequality (9) are not enough to deduce the weak stability of the non-linear term $j \times B$. However, in Theorem 1, the hyperbolic structure of Maxwell's equations is used to propagate the bound on the initial electromagnetic field in $\dot{H}^{\frac{1}{2}}$. In fact, it is shown therein that the electromagnetic field is uniformly bounded in $L^\infty(\mathbb{R}^+, \dot{H}^{\frac{1}{2}})$, which is then sufficient to establish the weak stability of $j \times B$.

All relevant a priori bounds on (8) are obtained through non-linear energy estimates performed in Besov spaces. Even though the general strategy remains rather standard, these estimates are complex and sometimes technical. They rely heavily on a precise use of paraproduct estimates, a careful analysis of the damped wave flow produced by Maxwell's system (3) and, most importantly, on crucial endpoint parabolic estimates to control the Stokes flow.

These endpoint parabolic estimates provide a new fundamental tool for the analysis of partial differential equations, particularly for models from fluid dynamics. We are therefore going to give a self-contained account of the main ideas behind such estimates in the next section.

We refer to [1] for the full justification of the above theorem.

## 4 Endpoint parabolic estimates

We show now how to derive the crucial parabolic estimates that are used in the proof of Theorem 1. Such estimates hold in any dimension $d \geq 1$ and show that solutions to the heat equation can gain up to two derivatives with respect to the source terms in Besov spaces, without resorting to the usual Chemin–Lerner spaces (see [1] for a definition of such spaces). In fact, we believe that this is an important principle that could be useful beyond its application to the proof of Theorem 1.

We introduce now a standard dyadic decomposition

$$\mathrm{Id} = \sum_{k\in\mathbb{Z}} \Delta_k,$$

where the Fourier multiplier operators $\Delta_k$ act on a function by localizing its frequencies $\xi$ to a domain $\{2^{k-1} \leq |\xi| \leq 2^{k+1}\}$.

Recall that the homogeneous Besov space $\dot{B}^s_{p,q}(\mathbb{R}^d)$, for any $s \in \mathbb{R}$ and $1 \leq p, q \leq \infty$, is then defined by the norm

$$\|f\|_{\dot{B}^s_{p,q}(\mathbb{R}^d)} = \left( \sum_{k=-\infty}^{\infty} 2^{ksq} \|\Delta_k f\|^q_{L^p(\mathbb{R}^d)} \right)^{\frac{1}{q}},$$

if $q < \infty$, and with the obvious modifications in case $q = \infty$. We refer to [1] for a precise definition of these spaces using the same notation.

We consider solutions of the forced heat equation

$$\partial_t w - \Delta w = f, \quad w_{|t=0} = 0. \tag{12}$$

Such solutions can be expressed by the Duhamel representation formula

$$w(t) = \int_0^t e^{(t-\tau)\Delta} f(\tau) d\tau. \tag{13}$$

Our first result provides a sharp estimate showing how the heat flow provides a gain of regularity of at most (but not equal to) two derivatives.

**Lemma 2.** *Let $\sigma \in \mathbb{R}$, $1 < r < m < \infty$ and $p \in [1, \infty]$. If $f$ belongs to $L^r([0,T], \dot{B}^{\sigma+\frac{2}{r}}_{p,\infty})$, then the solution of the heat equation (12) satisfies*

$$\|w\|_{L^m([0,T],\dot{B}^{\sigma+2+\frac{2}{m}}_{p,1})} \lesssim \|f\|_{L^r([0,T],\dot{B}^{\sigma+\frac{2}{r}}_{p,\infty})}.$$

*Proof.* First, observe that, employing the representation formula (13), there is an independent constant $C > 0$ such that

$$\|\Delta_k w(t)\|_{L^p} \lesssim \int_0^t e^{-C(t-\tau)2^{2k}} \|\Delta_k f(\tau)\|_{L^p} d\tau. \tag{14}$$

In particular, we obtain that

$$\begin{aligned}\|w(t)\|_{\dot{B}^{\sigma+2+\frac{2}{m}}_{p,1}} &\lesssim \int_0^t \sum_{k\in\mathbb{Z}} e^{-C(t-\tau)2^{2k}} 2^{k(\sigma+2+\frac{2}{m})} \|\Delta_k f(\tau)\|_{L^p} d\tau \\ &\lesssim \int_0^T h(t-\tau) \|f(\tau)\|_{\dot{B}^{\sigma+\frac{2}{r}}_{p,\infty}} d\tau,\end{aligned}$$

where we denoted

$$h(\lambda) = \sum_{k\in\mathbb{Z}} \mathbb{1}_{\{\lambda>0\}} e^{-C\lambda 2^{2k}} 2^{2k(1+\frac{1}{m}-\frac{1}{r})},$$

which is a well-defined convergent series whenever $1 + \frac{1}{m} - \frac{1}{r} > 0$.

Next, for any $\lambda > 0$, choosing $j \in \mathbb{Z}$ so that $2^{2j} \leq \lambda < 2^{2(j+1)}$, observe that

$$\begin{aligned}h(\lambda) &\leq \sum_{k\in\mathbb{Z}} e^{-C2^{2(j+k)}} 2^{2k(1+\frac{1}{m}-\frac{1}{r})} \\ &= 2^{-2j(1+\frac{1}{m}-\frac{1}{r})} \sum_{k\in\mathbb{Z}} e^{-C2^{2k}} 2^{2k(1+\frac{1}{m}-\frac{1}{r})} \lesssim \lambda^{-(1+\frac{1}{m}-\frac{1}{r})}.\end{aligned}$$

It therefore follows that, since $0 < 1 + \frac{1}{m} - \frac{1}{r} < 1$ and $1 < m, r < \infty$, by virtue of the Hardy–Littlewood–Sobolev inequality,

$$\|w(t)\|_{L^m \dot{B}^{\sigma+2+\frac{2}{m}}_{p,1}} \lesssim \left\| \int_0^T |t-\tau|^{-(1+\frac{1}{m}-\frac{1}{r})} \|f(\tau)\|_{\dot{B}^{\sigma+\frac{2}{r}}_{p,\infty}} d\tau \right\|_{L^m} \lesssim \|f\|_{L^r \dot{B}^{\sigma+\frac{2}{r}}_{p,\infty}}$$

which concludes the proof of the lemma. □

Note that the gain of regularity in the preceding result corresponds to $2 - 2(\frac{1}{r} - \frac{1}{m})$. In particular, the loss of $2(\frac{1}{r} - \frac{1}{m})$ is reminiscent of Bernstein inequalities in connection with the Littlewood–Paley theory (see [3, Section 2.1.1]) and Sobolev embeddings.

Further observe that, according to the preceding proof, the constant in the main estimate of Lemma 2 blows up as $r$ tends to $m$ with the same behavior as the sharp constant of the Hardy–Littlewood–Sobolev inequality (see [9] for a characterization of this sharp constant). However, we do not know whether this behavior is sharp for Lemma 2.

We are now particularly interested in the endpoint case $r = m$ of the preceding lemma, which would correspond formally to a gain of exactly two derivatives and is central to the proof of Theorem 1.

Unfortunately, the preceding proof fails miserably in this case, since it would require an endpoint application of the Hardy–Littlewood–Sobolev inequality, which is impossible. Instead, we are able to establish the following crucial endpoint lemma.

**Lemma 3** ([1])**.** *Let $\sigma \in \mathbb{R}$, $1 \leq q \leq r < \infty$ and $p \in [1, \infty]$. If $f$ belongs to $L^r([0,T], \dot{B}^\sigma_{p,q})$, then the solution of the heat equation (12) satisfies*

$$\|w\|_{L^r([0,T],\dot{B}^{\sigma+2}_{p,q})} \lesssim \|f\|_{L^r([0,T],\dot{B}^\sigma_{p,q})} .$$

The proof presented here is self-contained and is somewhat simpler than the one from [1] because it avoids abstract interpolation altogether.

*Proof.* By duality, it is enough to prove that, if $g$ is a function in $L^{a'}([0,T])$ with $a = \frac{r}{q} \geq 1$ and $\frac{1}{a} + \frac{1}{a'} = 1$, then

$$\int_0^T g(t) \|w(t)\|^q_{\dot{B}^{\sigma+2}_{p,q}} dt \lesssim \|f\|^q_{L^r([0,T],\dot{B}^\sigma_{p,q})} \|g\|_{L^{a'}([0,T])}.$$

To this end, we first write

$$\int_0^T g(t) \|w(t)\|^q_{\dot{B}^{\sigma+2}_{p,q}} dt = \sum_{k\in\mathbb{Z}} \int_0^T g(t) \|\Delta_k w(t)\|^q_{L^p} 2^{k(\sigma+2)q} dt.$$

Furthermore, we deduce from (14) that

$$\|\Delta_k w(t)\|_{L^p}^q \lesssim 2^{-k(2q-2)} \int_0^t e^{-C(t-\tau)2^{2k}} \|\Delta_k f(\tau)\|_{L^p}^q d\tau,$$

which implies that

$$\begin{aligned}\int_0^T & g(t)\|w(t)\|_{\dot B^{\sigma+2}_{p,q}}^q dt \\ &\lesssim \sum_{k\in\mathbb{Z}} \int_0^T \int_0^t |g(t)| e^{-C(t-\tau)2^{2k}} \|\Delta_k f(\tau)\|_{L^p}^q 2^{k(\sigma q+2)} d\tau dt.\end{aligned}$$

Next, we introduce a maximal operator defined by

$$Mg(\tau) = \sup_{\rho>0} \int_0^T \rho \mathbb{1}_{\{t-\tau\geq 0\}} e^{-(t-\tau)\rho} |g(t)| dt.$$

Classical results from harmonic analysis (see [7, Theorems 2.1.6 and 2.1.10]) establish that $M$ is bounded over $L^b([0,T])$, for any $1 < b \leq \infty$. One can now write that

$$\int_0^T g(t)\|w(t)\|_{\dot B^{\sigma+2}_{p,q}}^q dt \lesssim \sum_{k\in\mathbb{Z}} \int_0^T Mg(\tau)\|\Delta_k f(\tau)\|_{L^p}^q 2^{k\sigma q} d\tau,$$

whence, by definition of $\dot B^{\sigma}_{p,q}$,

$$\int_0^T g(t)\|w(t)\|_{\dot B^{\sigma+2}_{p,q}}^q dt \lesssim \int_0^T Mg(\tau)\|f(\tau)\|_{\dot B^{\sigma}_{p,q}}^q d\tau.$$

We finally conclude, by Hölder's inequality, that

$$\begin{aligned}\int_0^T Mg(\tau)\|f(\tau)\|_{\dot B^{\sigma}_{p,q}}^q d\tau &\lesssim \|Mg\|_{L^{a'}([0,T])} \|f\|_{L^r([0,T],\dot B^{\sigma}_{p,q})}^q \\ &\lesssim \|g\|_{L^{a'}([0,T])} \|f\|_{L^r([0,T],\dot B^{\sigma}_{p,q})}^q,\end{aligned}$$

which completes the proof of the lemma. □

## References


[1] Diogo Arsénio and Isabelle Gallagher. Solutions of Navier–Stokes–Maxwell systems in large energy spaces. *Trans. Amer. Math. Soc.*, 2019. In production.

[2] Diogo Arsénio and Laure Saint-Raymond. *From the Vlasov–Maxwell–Boltzmann system to incompressible viscous electro-magneto-hydrodynamics. Vol. 1.* EMS Monographs in Mathematics. European Mathematical Society (EMS), Zürich, 2019.

[3] Hajer Bahouri, Jean-Yves Chemin, and Raphaël Danchin. *Fourier analysis and nonlinear partial differential equations*, volume 343 of *Grundlehren der Mathematischen Wissenschaften [Fundamental Principles of Mathematical Sciences].* Springer, Heidelberg, 2011.

[4] Dieter Biskamp. *Nonlinear magnetohydrodynamics*, volume 1 of *Cambridge Monographs on Plasma Physics.* Cambridge University Press, Cambridge, 1993.

[5] Peter Alan Davidson. *An introduction to magnetohydrodynamics.* Cambridge Texts in Applied Mathematics. Cambridge University Press, Cambridge, 2001.

[6] Pierre Germain, Slim Ibrahim, and Nader Masmoudi. Well-posedness of the Navier-Stokes-Maxwell equations. *Proc. Roy. Soc. Edinburgh Sect. A*, 144(1):71–86, 2014.

[7] Loukas Grafakos. *Classical Fourier analysis*, volume 249 of *Graduate Texts in Mathematics.* Springer, New York, second edition, 2008.

[8] Pierre Gilles Lemarié-Rieusset. *The Navier–Stokes problem in the 21st century.* Boca Raton, FL: CRC Press, 2016.

[9] Elliott H. Lieb. Sharp constants in the Hardy-Littlewood-Sobolev and related inequalities. *Ann. of Math.*, 118(2):349–374, 1983.

[10] Nader Masmoudi. Global well posedness for the Maxwell-Navier-Stokes system in 2D. *J. Math. Pures Appl.*, 93(6):559–571, 2010.

# Resolutions of surfaces with big cotangent bundle and $A_2$ singularities

*Bruno De Oliveira*, *Michael Weiss*
University of Miami
1365 Memorial Dr
Miami, FL 33134
e-mail: bdeolive@math.miami.edu
weiss@math.miami.edu

**Resumo:** Neste trabalho obtemos um novo critério para garantir que a resolução duma superfície de tipo geral com singularidades canónicas tem o fibrado cotangente grande, e um novo limite inferior para os valores de $d$ para os quais existem superfícies lisas com o fibrado cotangente grande que são equivalentes por deformação a uma hipersuperfície lisa em $\mathbb{P}^3$ de grau $d$.

**Abstract** We give a new criterion for when a resolution of a surface of general type with canonical singularities has big cotangent bundle and a new lower bound for the values of $d$ for which there is a surface with big cotangent bundle that is deformation equivalent to a smooth hypersurface in $\mathbb{P}^3$ of degree $d$.

**palavras-chave:** fibrado cotangente grande; superfícies algébricas de tipo geral; singularidades canónicas

**keywords:** big cotangent bundle; surfaces of general type; canonical singularities.

## 1 Introduction and general theory

Symmetric differentials, i.e. sections of the symmetric powers of the cotangent bundle $S^m\Omega^1_X$, of a projective manifold $X$ play a role in obtaining hyperbolicity properties of $X$. Symmetric differentials give constraints on the existence of rational, elliptic and even entire curves in $X$ (nonconstant holomorphic maps from $\mathbb{C}$ to $X$), see for example [Dem15] and [Deb04].

The cotangent bundle of a projective manifold is said to be big if the order of growth of $h^0(X, S^m\Omega^1_X)$ with $m$ is maximal (i.e., $= 2\dim X - 1$). The work of Bogomolov [Bog77] and McQuillan [McQ98] gives that if a surface of general type has big $\Omega^1_X$, then $X$ satisfies the Green-Griffiths-Lang conjecture, i.e., there exists a proper subvariety $Z$ of $X$ such that any entire curve is contained in $Z$.

Smooth hypersurfaces $X \subset \mathbb{P}^3$ with degree $d \geq 5$ have $\Omega^1_X$ with strong positivity properties, such as $K_X$ being ample, but they have trivial cotangent algebra [Brj71],

$$S(X) := \bigoplus_{m=0}^{\infty} H^0(X, S^m\Omega^1_X) = H^0(X, S^0\Omega^1_X) = \mathbb{C}$$

see also [BDO08]. The absence of symmetric differentials on smooth hypersurfaces of $\mathbb{P}^3$ a priori prevents them from playing a role in obtaining hyperbolicity properties on smooth hypersurfaces of $\mathbb{P}^3$.

Previous work of the 1st author and Bogomolov [BDO06] showed that there are smooth surfaces $X$ with big $\Omega^1_X$ that are deformation equivalent to smooth hypersurfaces in $\mathbb{P}^3$. Hence symmetric differentials can still play a role in obtaining hyperbolicity properties for hypersurfaces of $\mathbb{P}^3$. In [BDO06] it was shown that there are nodal hypersurfaces $X \subset \mathbb{P}^3$ whose resolutions $\tilde{X}$ have big cotangent bundle. The simultaneous resolution result of Brieskorn [Bri70] implies that minimal resolutions $\tilde{X}$ of hypersurfaces $X \subset \mathbb{P}^3$ with only rational double points, i.e. canonical singularities, are deformation equivalent to smooth hypersurfaces of the same degree.

The results in this presentation are:

**Theorem 1.** *Let $X$ be a surface of general type with canonical singularities. Then the minimal resolution $\tilde{X}$ of $X$ has big cotangent bundle if*

$$\sum_{x \in Sing X} h^1(x) \geq -\frac{s_2(\tilde{X})}{3!}$$

See (2.1) for the definition of $h^1(x)$, it is an invariant of the singularity. Note that the left side encodes only information about the germs of the singularities of $X$, so it is local in nature. This result is stronger than the result in [RR14] stating that $\Omega^1_{\tilde{X}}$ is big if $s_2(\tilde{X}) + s_2(\mathcal{X}) > 0$, $s_2(\tilde{X})$ and $s_2(\mathcal{X})$ respectively the 2nd Segre number of $\tilde{X}$ and of the orbifold $\mathcal{X}$ associated to $X$, see section 2 for more details.

In section 2.2 we give a method to find $h^1(x)$ where $(X, x)$ is the germ of an $A_2$-singularity. In a later work [DOW20] we show how to extend this method to calculate $h^1(x)$ for other $A_n$ singularities. Then using theorem 1 and information on the possible number of canonical singularities of prescribed types allowed in a hypersurface $X \subset \mathbb{P}^3$ of degree $d$, we obtain

**Theorem 2.** *For $d = 9$ and $d \geq 11$, there are minimal resolutions of hypersurfaces $X \subset \mathbb{P}^3$ with canonical singularities and degree $d$ which have big cotangent bundle.*

The condition $s_2(\tilde{X}) + s_2(\mathcal{X}) > 0$ of [RR14] gives only $d \geq 13$ and there nodes are the best singularities. The above theorem uses $A_2$ singularities which due to theorem 1 are unexpectedly better than nodes, see 2.2 for more details.

### 1.1 Big Cotangent Bundle

The cotangent bundle $\Omega^1_X$ on a complex manifold of dimension $n$ is said to be big if

$$\lim_{m\to\infty} \frac{h^0(X, S^m\Omega^1_X)}{m^{2n-1}} \neq 0$$

(i.e., $h^0(X, S^m\Omega^1_X)$ has the maximal growth order possible with respect to $m$ for $\dim X = n$). The property of $\Omega^1_X$ being big is birational.

In the case of surfaces of general type there is a topologically sufficient condition for bigness of $\Omega^1_X$, $s_2(X) > 0$, where $s_2(X) = c_1^2(X) - c_2(X)$ is the 2nd Segre number of $X$. This follows from the asymptotic Riemann-Roch theorem for symmetric powers of $\Omega^1_X$:

$$h^0(X, S^m\Omega^1_X) - h^1(X, S^m\Omega^1_X) + h^2(X, S^m\Omega^1_X) = \frac{s_2(X)}{3!}m^3 + O(m^2) \quad (1.1)$$

and Bogomolov's vanishing for surfaces of general type, $h^2(X, S^m\Omega^1_X) = 0$ for $m > 2$ [Bog79].

Very few examples of minimal surfaces with $s_2(X) \leq 0$ are known to have $\Omega^1_X$ big, they appear in [BDO06] and [RR14]. In these examples, bigness of $\Omega^1_X$ follows from complex analytic and not topological properties of $X$. The complex analytic conditions are the presence of enough configurations of $(-2)$-curves associated with canonical singularities. In fact, these surfaces with big $\Omega^1_X$ are diffeomorphic to surfaces with trivial cotangent algebra, $S(X) \simeq \mathbb{C}$.

If $X$ is a smooth surface of general type, it follows from 1.1 and $h^2(X, S^m\Omega^1_X) = 0$ that $\Omega^1_X$ is big if and only if:

$$\lim_{m\to\infty} \frac{h^1(X, S^m\Omega^1_X)}{m^3} > -\frac{s_2(X)}{3!} \quad (1.2)$$

## 1.2 Quotient singularities and local asymptotic Riemann-Roch for orbifold $\hat{S}^m\Omega^1_X$

In this section we present the local asymptotic Riemann-Roch for the orbifold symmetric powers of the cotangent bundle of a normal surface with only quotient singularities. For references on this topic, see [Wah93], [Bla96], [Kaw92], [Miy08].

The germ of a normal surface singularity $(X, x)$ is a quotient singularity germ if it is biholomorphic to $(\mathbb{C}^2, 0)/G_x$, with $G_x \subset GL_2(\mathbb{C})$ finite and small, where $G_x$ is the local fundamental group. Canonical surface singularities are quotient singularities with $G_x \subset SL_2(\mathbb{C})$. Consider

$$\begin{array}{ccc} & & (\mathbb{C}^2, 0) \\ & \overset{\varphi}{\swarrow} & \downarrow \pi \\ (\tilde{X}, E) & \xrightarrow{\sigma} & (X, x) \end{array}$$

with $\pi : (\mathbb{C}^2, 0) \to (X, x)$, the quotient map by the local fundamental group, called the local smoothing of $(X, x)$ and $\sigma : (\tilde{X}, E) \to (X, x)$ a good resolution of $(X, x)$ where $(\tilde{X}, E)$ is the germ of a neighborhood of the exceptional locus $E$ with $E$ consisting of smooth curves intersecting transversally.

A reflexive coherent sheaf $\mathcal{F}$, i.e. $\mathcal{F}^{\vee\vee} = \mathcal{F}$, on $(X, x)$ is a locally free sheaf away from the singularity and satisfies $\mathcal{F} = i_*(\mathcal{F}|_{X\backslash\{x\}})$, $i : X\backslash\{x\} \hookrightarrow X$. Associated to a reflexive sheaf $\mathcal{F}$ on the quotient surface germ $(X, x)$ there are locally free sheaves $\tilde{\mathcal{F}}$ on $(\tilde{X}, E)$ (not uniquely determined) and $\hat{\mathcal{F}}$ on $(\mathbb{C}^2, 0)$ (uniquely determined) satisfying $\mathcal{F} \cong (\sigma_*\tilde{\mathcal{F}})^{\vee\vee} \cong (\pi_*^{G_x})\hat{\mathcal{F}}$, where $(\pi_*^{G_x})\hat{\mathcal{F}}$ is a maximal subsheaf of $\pi_*\hat{\mathcal{F}}$ on which $G_x$ acts trivially, ([Bla96] section 2).

The previous paragraph implies that reflexive coherent sheaves on normal surfaces with only quotient singularities $X$ are orbifold vector bundles on $X$ (also called $\mathbb{Q}$-vector bundles or locally $V$-free bundles over $X$). The orbifold $m$-symmetric power of the cotangent bundle on a normal surface $X$ with only quotient singularities is $\hat{S}^m\Omega^1_X := (S^m\Omega^1_X)^{\vee\vee}$ with $\Omega^1_X = i_*(\Omega^1_{X_{reg}})$. If $\tilde{X} \xrightarrow{\sigma} X$ is a good resolution $\hat{S}^m\Omega^1_X = (\sigma_* S^m\Omega^1_{\tilde{X}})^{\vee\vee}$.

In the proof of theorem 1 a lower bound for $h^1(\tilde{X}, S^m\Omega^1_{\tilde{X}})$ is given using only information on the singularities of $X$. Each $x_i$ contributes with $h^1(\tilde{U}_{x_i}, S^m\Omega^1_{\tilde{X}})$ where $\tilde{U}_{x_i}$ is the minimal resolution of an affine neighborhood $U_{x_i}$ of $x_i$ with $U_{x_i} \cap \mathrm{Sing}(X) = \{x_i\}$. The bigness of $\Omega^1_{\tilde{X}}$ depends on the asymptotics of $h^1(\tilde{X}, S^m\Omega^1_{\tilde{X}})$, see section (1.2), and hence on the combined asymptotics of the $h^1(\tilde{U}_{x_i}, S^m\Omega^1_{\tilde{X}})$.

Let $(\tilde{X}, E) \xrightarrow{\sigma} (X, x)$ be a good resolution of the germ of a quotient surface singularity and $\tilde{\mathcal{F}}$, $\mathcal{F}$ be sheaves such that $\tilde{\mathcal{F}}$ is locally free of rank $r$ on $\tilde{X}$ and $\mathcal{F} = (\sigma_*\tilde{\mathcal{F}})^{\vee\vee}$ a reflexive sheaf on $X$. In comparing the Euler characteristics $\chi(X, \mathcal{F})$ and $\chi(\tilde{X}, \tilde{\mathcal{F}})$ one has $\chi(X, \mathcal{F}) = \chi(\tilde{X}, \tilde{\mathcal{F}}) + \chi(x, \tilde{\mathcal{F}})$ with

$$\chi(x, \tilde{\mathcal{F}}) = \dim(H^0(\tilde{X} \setminus E, \tilde{\mathcal{F}})/H^0(\tilde{X}, \tilde{\mathcal{F}})) + h^1(\tilde{X}, \tilde{\mathcal{F}}) \tag{1.3}$$

called the modified Euler characteristic of $\tilde{\mathcal{F}}$ ([Wah93], [Bla96] 3.9). The asymptotics of 1.3 are described via a local asymptotic Riemann-Roch theorem ([Bla96] 4.1)

$$\lim_{m\to\infty} \frac{\chi(x, S^k\tilde{\mathcal{F}})}{m^{2+r-1}} = -\frac{1}{(2+r-1)!} s_2(x, \tilde{\mathcal{F}}) \tag{1.4}$$

with $s_2(x, \tilde{\mathcal{F}}) := c_1^2(x, \tilde{\mathcal{F}}) - c_2(x, \tilde{\mathcal{F}})$, the local 2nd Segre number of $\tilde{\mathcal{F}}$ and $c_i(x, \tilde{\mathcal{F}}) \in H^{2i}_{dRc}(\tilde{X}, \mathbb{C})$ the $i$-th local Chern class of $\tilde{\mathcal{F}}$. The local Chern classes appear when comparing the pullback of orbifold Chern classes of an orbifold vector bundle $\mathcal{F}$ on an orbifold $X$ and the Chern classes of the vector bundle $\tilde{\mathcal{F}}$ on $\tilde{X}$, a good resolution $\sigma : \tilde{X} \to X$ of $X$, satisfying $\mathcal{F} = (\sigma_*\tilde{\mathcal{F}})^{\vee\vee}$.

We are only concerned with good resolutions $\sigma : (\tilde{X}, E) \to (X, x)$ of canonical surface singularities and $\tilde{\mathcal{F}} = \Omega^1_{\tilde{X}}$, one has $c_1^2(x, \Omega^1_{\tilde{X}}) = 0$ and:

$$s_2(x, \Omega^1_{\tilde{X}}) = -c_2(x, \Omega^1_{\tilde{X}}) = -(e(E) - \frac{1}{|G_x|}) \tag{1.5}$$

with $e(E)$ the topological Euler characteristic of the exceptional locus and $|G_x|$ the order of the local fundamental group ([Bla96] 3.18). We will use the invariant of the singularity:

$$s_2(x, X) := s_2(x, \Omega^1_{\tilde{X}_{min}}) \tag{1.6}$$

where $\sigma : (\tilde{X}_{min}, E) \to (X, x)$ is the minimal good resolution.

# 2 Theorems

## 2.1 Resolutions with big cotangent bundle

We consider minimal resolutions $\sigma : \tilde{X} \to X$ of normal surfaces $X$ with only canonical singularities. The minimality condition has several advantages: i) the local 2nd Segre numbers $s_2(x, \tilde{\Omega}^1_{\tilde{X}})$ being considered are $s_2(x, X)$ which depend only on the singularity (since the resolution is fixed); ii) in section 2.2 the simultaneous resolution results used involve minimal resolutions of canonical singularities. Also, blowing up $b : \hat{X} \to X$ a smooth surface $X$ at a point does not affect inequality (1.2) determining bigness of the cotangent bundle, since

$$\lim_{m\to\infty} \frac{h^1(\hat{X}, S^m\Omega^1_{\hat{X}})}{m^3} + \frac{s_2(\hat{X})}{3!} = \lim_{m\to\infty} \frac{h^1(X, S^m\Omega^1_X)}{m^3} + \frac{s_2(X)}{3!}$$

Let $\sigma : \tilde{U}_x \to U_x$ be the minimal resolution of an affine normal surface $U_x$ with a single canonical singularity at the point $x \in U_x$. Set:

$$h^1(x) := \lim_{m\to\infty} \frac{h^1\left(\tilde{U}_x, S^m\Omega^1_{\tilde{X}}\right)}{m^3} \tag{2.1}$$

$$h^0(x) := \lim_{m\to\infty} \frac{\left[H^0\left(\tilde{U}_x \setminus E, S^m\Omega^1_{\tilde{U}_x}\right)/H^0\left(\tilde{U}_x, S^m\Omega^1_{\tilde{U}_x}\right)\right]}{m^3} \tag{2.2}$$

The local asymptotic Riemann-Roch equation (1.4) for the local modified Euler characteristic (1.3) for $\tilde{U}_x$ and $S^m\Omega^1_{\tilde{U}_x}$ gives:

$$h^1(x) = -\frac{1}{3!}s_2(x, X) - h^0(x). \tag{2.3}$$

with $s_2(x, \Omega^1_{\tilde{U}_x})$ an invariant of the canonical singularity $(U_x, x)$, since $\tilde{U}_x$ is its minimal resolution (and hence unique). In [DOW20] using local duality and local cohomology for the pair $(\tilde{X}, E)$, it is shown that $h^0(x) \leq h^1(X)$ holds, hence:

$$h^1(x) \geq -\frac{s_2(x, X)}{2 \cdot 3!} \tag{2.4}$$

**Theorem 1.** *Let $X$ be a normal projective surface of general type with only canonical singularities and $\sigma : \tilde{X} \to X$ a minimal resolution. Then $\Omega^1_{\tilde{X}}$ is big if and only if:*

$$\sum_{x \in SingX} h^1(x) \geq -\frac{s_2(\tilde{X})}{3!} \tag{2.5}$$

*Proof.* We saw in section 1.1 that $\Omega^1_{\tilde{X}}$ is big if and only if $\lim_{m\to\infty} \frac{h^1(\tilde{X}, S^m\Omega^1_{\tilde{X}})}{m^3} > -\frac{s_2(\tilde{X})}{3!}$.

From the Leray spectral sequence for $\sigma_*$ and Bogomolov's vanishing $H^2(\tilde{X}, S^m\Omega^1_{\tilde{X}}) = 0$ for $m > 2$, we obtain for $m > 2$:

$$\begin{aligned} 0 \longrightarrow H^1(X, \sigma_* S^m\Omega^1_{\tilde{X}}) \longrightarrow H^1(\tilde{X}, S^m\Omega^1_{\tilde{X}}) \longrightarrow H^0(X, R^1\sigma_* S^m\Omega^1_{\tilde{X}}) \\ \longrightarrow H^2(X, \sigma_* S^m\Omega^1_{\tilde{X}}) \longrightarrow 0 \end{aligned} \tag{2.6}$$

The 1st direct image sheaf $R^1\sigma_* S^m\Omega^1_{\tilde{X}}$ has support on the zero-dimensional singularity locus $\mathrm{Sing}(X) = \{x_1, \dots, x_k\}$ of $X$. Each $x_i$ has an affine neighborhood $U_{x_i}$ such that $U_{x_i} \cap \mathrm{Sing}(X) = \{x_i\}$. Using the Leray spectral sequence again for each $\tilde{U}_x = \sigma^{-1}(U_x)$, $\sigma : \tilde{U}_x \to U_{x_i}$ we obtain:

$$H^0\left(X, R^1\sigma_* S^m\Omega^1_{\tilde{X}}\right) = \bigoplus_{i=1}^{k} H^1\left(\tilde{U}_x, S^m\Omega^1_{\tilde{U}_x}\right)$$

Hence using the notation of section 2.1:

$$\sum_{x\in\mathrm{Sing}(X)} h^1(x) = \lim_{m\to\infty} \frac{h^0\left(X, R^1\sigma_* S^m\Omega^1_{\tilde{X}}\right)}{m^3} \tag{2.7}$$

**Claim**: $H^2(X, \sigma_* S^m\Omega^1_{\tilde{X}}) = 0$

*Proof.* Recalling that $\hat{S}^m\Omega^1_{\tilde{X}} := (\sigma_* S^m\Omega^1_{\tilde{X}})^{\vee\vee}$, consider the short exact sequence:

$$0 \to \sigma_* S^m\Omega^1_{\tilde{X}} \to \hat{S}^m\Omega^1_{\tilde{X}} \to Q_m \to 0.$$

Left injectivity holds since $\sigma_* S^m\Omega^1_{\tilde{X}}$ is torsion free. The support of $Q_m = \frac{(\sigma_* S^m\Omega^1_{\tilde{X}})^{\vee\vee}}{\sigma_* S^m\Omega^1_{\tilde{X}}}$ is again $\mathrm{Sing}(X)$, hence $H^2(X, \sigma_* S^m\Omega^1_{\tilde{X}}) \cong H^2(X, \hat{S}^m\Omega^1_{\tilde{X}})$.

The surface $X$ is an orbifold surface of general type with canonical singularities and $\hat{S}^m\Omega^1_{\tilde{X}}$ is the orbifold $m$-th symmetric power of the cotangent bundle of $X$. Bogomolov's vanishing $H^2(X, \hat{S}^m\Omega^1_{\tilde{X}}) = 0$ for $m > 2$ also holds in this setting, due to the existence of orbifold Kähler-Einstein metrics [Kob85], [TY86], see also [RR14]. □

The vanishing of $H^2\left(X, \sigma_* S^m\Omega^1_{\tilde{X}}\right) = 0$ for $m > 0$, (2.6) and (2.7) give:

$$\lim_{m\to\infty} \frac{h^1(\tilde{X}, S^m\Omega^1_{\tilde{X}})}{m^3} \geq \sum_{x\in\mathrm{Sing}(X)} h^1(x) \tag{2.8}$$

and the result follows from (1.2). □

**Remark:** Theorem 1 is stronger than the main theorem in [RR14] which states that $\Omega^1_{\tilde{X}}$ is big if $s_2(\tilde{X}) + s_2(X) > 0$. We have that ([Bla96] 3.14), $s_2(\tilde{X}) = s_2(X) + \sum_{x\in\mathrm{Sing}X} s_2(x, X)$, hence condition $s_2(\tilde{X}) + s_2(X) > 0$ can be reexpressed as:

$$- \sum_{x \in \mathrm{Sing} X} \frac{s_2(x, X)}{2} > -s_2(\tilde{X}) \tag{2.9}$$

It follows from (2.4) that the condition (2.5) in theorem 1 implies (2.9). In fact it gives much stronger results. In the next section we will show that if $(X, x)$ is the germ of an $A_2$ singularity, then $h^1(x) = \frac{67}{216}$ while $-\frac{s_2(x,X)}{2 \cdot 3!} = \frac{48}{216}$. This implies that our inequality (2.5) guarantees $\Omega^1_{\tilde{X}}$ is big for surfaces of general type $X$ with only $\frac{48}{67} \cdot \ell$ $A_2$-singularities, where $\ell$ is the number needed to satisfy inequality (2.9).

## 2.2 Deformations of smooth hypersurfaces with big $\Omega^1_X$

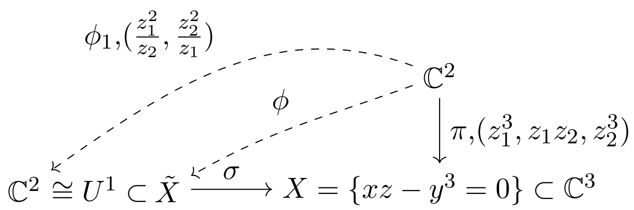
In this section we study for which $d$ there are (smooth) surfaces with big cotangent bundle that are deformation equivalent to smooth hypersurfaces in $\mathbb{P}^3$ of degree $d$. We do this by considering minimal resolutions $\tilde{X}$ of hypersurfaces $X \subset \mathbb{P}^3$ of degree $d$ with only $A_2$ singularities. A simultaneous resolution result of Brieskorn [Bri70] gives that $\tilde{X}$ is deformation equivalent to a smooth hypersurface of $\mathbb{P}^3$ of degree $d$. In [DOW20] other canonical singularities are also considered.

**Proposition 2.1.** *Let $\sigma : (\tilde{X}, E) \to (X, x)$ be the minimal resolution of the germ of an $A_2$ surface singularity. Then:*

$$h^0(x) := \lim_{m \to \infty} \frac{\dim[H^0(\tilde{X} \setminus E_i, S^m \Omega^1_{\tilde{X}}) / H^0(X, S^m \Omega^1_{\tilde{X}}))]}{m^3} = \frac{29}{216} \tag{2.10}$$

*Proof.* For the full proof see [DOW20].

We give here an extended description of what is involved in the proof. We use the affine model of an $A_2$-singularity $X = \{xz - y^3 = 0\} \subset \mathbb{C}^3$ with the minimal resolution $\tilde{X}$ obtained as the strict preimage of $X$ under $\sigma : \hat{\mathbb{C}}^3 \to \mathbb{C}^3$, the blow up of $\mathbb{C}^3$ at $(0, 0, 0)$.




where $\pi : \mathbb{C}^2 \to X$ gives the smoothing as in section 1.2. Let $U^1 = \tilde{X} \cap p^{-1}(U_1)$ with $p : \hat{\mathbb{C}}^3 \to \mathbb{P}^2$ the canonical projection and $U_1 = \{y \neq 0\} \subset \mathbb{P}^2$,

$[x:y:z]$ as homogeneous coordinates of $\mathbb{P}^2$. The exceptional locus of $\sigma$ is $E = E_1 + E_2$, $E_i$ $(-2)$-curves intersecting transversally. On $U^1$ put coordinates $(u_1, u_2)$ with $\phi_1^* u_1 = \frac{z_1^2}{z_2}$ and $\phi_1^* u_2 = \frac{z_2^2}{z_1}$ and $E \cap U^1 = \{u_1 u_2 = 0\}$.

The isomorphism $\phi^* : H^0(\tilde{X} \setminus E, S^m\Omega^1_{\tilde{X}}) \to H^0(\mathbb{C}^2, S^m\Omega^1_{\mathbb{C}^2})^{\mathbb{Z}_3}$ will be used to move the setting for finding $h^0(x)$ from $\tilde{X} \setminus E$ to $\mathbb{C}^2$. We need a good description of $G(m) := \phi^*(H^0(\tilde{X}, S^m\Omega^1_{\tilde{X}}))$. We use:

$$G(m) = \phi_1^*(H^0(\mathbb{C}^2, S^m\Omega^1_{\mathbb{C}^2})) \cap H^0(\mathbb{C}^2, S^m\Omega^1_{\mathbb{C}^2})$$

We call $z_1^{i_1} z_2^{i_2} dz_1^{m_1} dz_2^{m_2}$ a z-monomial of full type (f-type) $(i_1, i_2, m_1, m_2)_z$ and type $(i, m)_z$ with $i = i_1 + i_2$ the order and $m = m_1 + m_2$ the degree of the monomial. A monomial is holomorphic if $i_1, i_2 \geq 0$ and $\mathbb{Z}_3$-invariant if $i_1 + 2i_2 + m_1 + 2m_2 \equiv 0 \bmod 3$.

For each triple $(k, i, m)$ with $k \equiv -(m+i) \bmod 3$ there is a collection of z-monomials:

$$B(k,i,m)_z = \{(k-m+l, i+m-k-l, m-l, l)_z\}_{l=0,\ldots,m} \qquad (2.11)$$

These collections give a partition of the set of all $\mathbb{Z}_3$-invariant z-monomials of type $(i, m)$. Set $V(k,i,m)_z =$Span$(B(k,i,m)_z)$.

Let $B_h(k,i,m)_z$ be the subcollection of holomorphic z-monomials of $B(k,i,m)_z$. Set $V_h(k,i,m)_z$ := Span$(B_h(k,i,m)_z)$= $H^0(\mathbb{C}^2, S^m\Omega^1_{\mathbb{C}^2}) \cap V(k,i,m)$. Set $h_z(k,i,m) := \dim V_h(k,i,m)_z = \# B_h(k,i,m)_z$, from (2.11) it follows that $h_z(k,i,m) = \min(m+1, k+1, i+1, m-k+i+1)$. Note that $h_z(k,i,m) = 0$ unless $0 \leq k \leq m+i$.

Set $G(k,i,m) := G(m) \cap V(k,i,m) = G(m) \cap V_h(k,i,m)$. All the above gives (we will see below that $I(m) = 2m$):

$$\dim[H^0(\tilde{X} \setminus E, S^m\Omega^1_{\tilde{X}})/H^0(X, S^m\Omega^1_{\tilde{X}}))] = \dim[H^0(\mathbb{C}^2, S^m\Omega^1_{\mathbb{C}^2})^{\mathbb{Z}_3}/G(m)]$$
$$= \sum_{i=0}^{I(m)} \sum_{\substack{0 \leq k \leq m+i \\ k \equiv -(m+i) \bmod 3}} h_z(k,i,m) - \dim G(k,i,m) \qquad (2.12)$$

The reason to consider the collections $B(k,i,m)$ will now be examined. The rational map $\phi_1 : (\mathbb{C}^2, z_1, z_2) \dashrightarrow (\mathbb{C}^2, u_1, u_2)$ pulls back holomorphic u-monomials of type $(i, m)$ to rational $\mathbb{Z}_3$-invariant z-monomials of type $(i, m)$:

$$\phi_1^*(p, i-p, q, m-q)_u = \sum_{l=0}^{m} c_{ql}(3(p+q)-(i+2m)+l, -3(p+q)+2(i+m)-l, m-l, l)_z \qquad (2.13)$$

with the $c_{ql}$ given by $(2x-y)^q(-x+2y)^{m-q} = \sum_l c_{ql} x^{m-l} y^l$.

From (2.13) and (2.11) it follows that the pullback of a u-monomial of type $(i, m)$ lies in a single $V(k, i, m)$ and that the u-monomials whose pullback lie in $V(k, i, m)$ themselves form the collection $B(k, i, m)_u := \{(k'-m+l, i+m-k'-l, m-l, l)_u\}_{l=0,\dots,m}$ with $k' = \frac{i+m+k}{3}$. Let $B_h(k, i, m)_u$ be the subcollection of holomorphic u-monomials of $B(k, i, m)_u$ and set $V_h(k, i, m)_u$ =Span$(B_h(k, i, m)_u)$. Set $h_u(k, i, m) := \dim V_h(k, i, m)_u$, we have $h_u(k, i, m) = \min(m + 1, \frac{k+(i+m)}{3} + 1, i + 1, \frac{2(i+m)-k}{3} + 1)$.

We proceed to find $I(m)$ and $\dim G(k, i, m)$ and calculate (2.12). We have that $G(k, i, m) = \phi_1^*(V_h(k, i, m)_u) \cap V_h(k, i, m)_z$. By using information on the rank of relevant subblocks of matrix $[c_{ql}]$, with $c_{ql}$ as in (2.12) (see [DOW20] for details), we obtain that:

$$\dim G(k, i, m) = \max\left(h_z(k, i, m) + h_u(k, i, m) - (m + 1), 0\right)$$

From the formula for $h_u(k, i, m)$ above, it follows that $h_u(k, i, m) = m + 1$ and hence $G(k, i, m) = h_z(k, i, m)$ for all $0 \leq k \leq m + 1$ if $i \geq 2m$. This implies that all the terms in (2.12) for $i \geq 2m$ vanish, hence by setting $I(m) = 2m$ we can write the full sum and obtain:

$$h^0(x) = \lim_{m\to\infty} \frac{1}{m^3} \sum_{i=0}^{2m} \sum_{\substack{0\leq k\leq m+i \\ k\equiv -(m+i) \bmod 3}} \min(m+1-h_u(k, i, m), h_z(k, i, m)) = \frac{29}{216}$$

□

**Remark:** For $A_1$ singularities using the set up described in the 1st author's article [BDO08] the method to find $h^0(x)$ is substantially simpler and $h^0(x) = \frac{11}{108}$, see Jordan Thomas' thesis [Tho13]. For an approach in the lines of proposition 2.1 and valid for all $A_n$ singularities see [DOW20].

**Theorem 2.** *For $d = 9$ and $d \geq 11$ there are minimal resolutions of hypersurfaces in $\mathbb{P}^3$ with canonical singularities and degree $d$ which have big cotangent bundle.*

*Proof.* Let $X_{d,\ell} \subset \mathbb{P}^3$ denote a hypersurface of degree $d$ with $\ell$ $A_2$-singularities as its only singularities and $\tilde{X}_{d,\ell}$ its minimal resolution. The Brieskorn simultaneous resolution theorem, [Bri70] and Ehresmann's fibration theorem give that $\tilde{X}_{d,\ell}$ is diffeomorphic to a smooth hypersurface of degree $d$ in $\mathbb{P}^3$, hence $s_2(\tilde{X}_{d,\ell}) = -4d^2 + 10d$.

From sections 1.2 and 2.1 we have that $h^1(x) = -\frac{1}{3!}s_2(x, X) - h^0(x) = \frac{1}{3!}(e(E) - \frac{1}{|\mathbb{Z}_3|}) - h^0(x)$, where $(\tilde{X}, E)$ is a minimal resolution of the germ of the $A_2$-singularity $(X, x)$ $(e(E) = 3)$. Using proposition 2.1, it follows that:

$$h^1(x) = \frac{67}{216} \tag{2.14}$$

In Labs [Lab06] it is shown how to construct hypersurfaces in $\mathbb{P}^3$ with only $A_n$ singularities with $n$ fixed using Dessins d'Enfants. For $A_2$ singularities one has that there are hypersurfaces $X_{d,\ell}$ if:

$$\ell = \begin{cases} \frac{1}{2}d(d-1)\cdot\lfloor\frac{d}{3}\rfloor + \frac{1}{3}d(d-3)(\lfloor\frac{d-1}{2}\rfloor) - \lfloor\frac{d}{3}\rfloor) & d \equiv 0 \mod 3 \\ \frac{1}{2}d(d-1)\cdot\lfloor\frac{d}{3}\rfloor + \frac{1}{3}(d(d-3)+2)(\lfloor\frac{d-1}{2}\rfloor) - \lfloor\frac{d}{3}\rfloor) & \text{otherwise} \end{cases} \tag{2.15}$$

Theorem 1 and 2.14 give that $\Omega^1_{\tilde{X}_{d,\ell}}$ !is big if $\frac{67}{216}\ell > s_2(\tilde{X}_{d,\ell})$ or equivalently if:

$$\ell > \frac{72}{67}(2d^2 - 5d) \tag{2.16}$$

By 2.15 there are hypersurfaces $X_{d,\ell} \subset \mathbb{P}^3$ with $d$ and $\ell$ satisfying (2.16) if $d = 9$ or $d \geq 11$. □

**Remark:** 1) In Theorem 2 we can see the strength of theorem 1 when compared to the criterion for the cotangent bundle $\Omega^1_{\tilde{X}_{d,\ell}}$ to be big of [RR14], $s_2(\tilde{X}_{d,\ell})+s_2(X_{d,\ell}) > 0$. The criterion of [RR14] needs $\ell > \frac{3}{2}(2d^2-5d)$ instead of (2.16). The known upper bounds by Miyaoka or Varchenko, see [Lab06], for the number of $A_2$ singularities possible on a hypersurface in $\mathbb{P}^3$ of degree $d$ prevent $\ell > \frac{3}{2}(2d^2 - 5d)$ for $d \leq 11$. Moreover, one has to go to degree $d = 14$ for the known constructions to give enough $A_2$ singularities for the criterion of [RR14].

2) Following the method of theorem 2, if instead of using hypersurfaces in $\mathbb{P}^3$ with only $A_2$ singularities, one used hypersurfaces with only $A_1$ singularities (nodes), then one would need $\ell > \frac{9}{4}(2d^2 - 5d)$ nodes for the minimal resolution of an hypersurface with $\ell$ nodes to have big cotangent bundle. This would give surfaces with big cotangent bundle deformation equivalent to smooth hypersurfaces in $\mathbb{P}^3$ of degree $d \geq 10$. The known upper bounds for the number of nodes possible in hypersurfaces of a given degree, see [Lab06], give that for degree 9 you can not have more than 246 nodes, our criterion needs 264. So $A_2$ singularities give a better result.

## References

[BDO06] F. Bogomolov and B. De Oliveira, *Hyperbolicity of nodal hypersurfaces*, J. Reine Angew. Math. **596** (2006), 89–101.

[BDO08] ———, *Symmetric tensors and geometry of $\mathbb{P}^N$ subvarieties*, Geometric and Functional Analysis **18** (2008), 637–656.

[Bla96] R. Blache, *Chern classes and Hirzebruch-Riemann-Roch theorem for coherent sheaves on complex-projective orbifolds with isolated singularities*, Math. Z. **222** (1996), 7–57.

[Bog77] F. Bogomolov, *Families of curves on a surface of general type*, Dokl. Akad. Nauk SSSR **236** (1977), 1041–1044.

[Bog79] ———, *Holomorphic tensors and vector bundles on projective varieties*, Math. of the USSR-Izvestiya **13** (1979), 499–555.

[Bri70] E. Brieskorn, *Singular elements of semi-simple algebraic groups*, Actes du Congres International des Mathématiciens (Nice, 1970), vol. 2, 1970, pp. 279–284.

[Brj71] P. Brjukman, *Tensor differential forms on algebraic varieties*, Math. of the USSR-Izvestiya **5** (1971), 1021–1048.

[Deb04] O. Debarre, *Hyperbolicity of complex varieties*, Course Notes **1** (2004), 32.

[Dem15] J-P. Demailly, *Recent progress towards the kobayashi and green-griffiths-lang conjectures*, Manuscript Institut Fourier. (2015).

[DOW20] B. De Oliveira and M. Weiss, *Deformation of smooth hypersurfaces in $\mathbb{P}^3$ with big cotangent bundle*, To appear (2020).

[Kaw92] Y. Kawamata, *Abundance theorem for minimal threefolds*, Inventiones mathematicae **108** (1992), 229–246.

[Kob85] R. Kobayashi, *Einstein-kaehler metrics on open algebraic surfaces of general type*, Tohoku Math. J. (2) **37** (1985), 43–77.

[Lab06] O. Labs, *Dessins D'Enfants and Hypersurfaces with Many $A_j$-Singularities*, Journal of the London Mathematical Society **74** (2006), 607–622.

[McQ98] M. McQuillan, *Diophantine approximations and foliations*, Inst. Hautes Études Sci. Publ. Math. (1998), no. 87, 121–174.

[Miy08] Y. Miyaoka, *The orbibundle Miyaoka-Yau-Sakai inequality and an effective Bogomolov-McQuillan theorem*, Publ. Res. Inst. Math. Sci. **44** (2008), 403–417.

[RR14] X. Roulleau and E. Rousseau, *Canonical surfaces with big cotangent bundle*, Duke Math. J. **163** (2014), 1337–1351.

[Tho13] J. Thomas, *Contraction Techniques in the Hyperbolicity of Hypersurfaces of General Type*, Ph.D. thesis, New York University, 2013.

[TY86] G. Tian and S-T. Yau, *Existence of Kähler-Einstein metrics on complete Kähler manifolds and their applications to algebraic geometry*, Conference on Mathematical Aspects of String Theory **C8607214** (1986), 574–646.

[Wah93] J. Wahl, *Second Chern class and Riemann-Roch for vector bundles on resolutions of surface singularities*, Math. Ann. **295** (1993), 81–110.

# Symplectic geometry and the Alexander polynomial of a knot

*Luís Diogo*
Universidade Federal Fluminense
Instituto de Matemática e Estatística, Departamento de Matemática Aplicada
R. Prof. Marcos Waldemar de Freitas Reis, s/n, Bloco G - Campus do Gragoatá
São Domingos, Niterói, RJ, 24210-201, Brasil
e-mail: ldiogo@id.uff.br

**Resumo:** Apresentamos uma fórmula para o polinómio de Alexander clássico de um nó em termos de um invariante de nós introduzido recentemente, chamado polinómio de aumentação e definido a partir da homologia de contacto do nó. Damos uma ideia da prova, que parte de uma definição dinâmica do polinómio de Alexander e envolve a análise de vários espaços de moduli de curvas pseudoholomorfas.

**Abstract** We present a formula expressing the classical Alexander polynomial of a knot in terms of a very recent knot invariant, called the augmentation polynomial and defined using knot contact homology. We give an idea of the proof, which starts from a dynamical definition of the Alexander polynomial and involves analyzing several moduli spaces of pseudoholomorphic curves.

**palavras-chave:** geometria simplética, curvas pseudoholomorfas, invariantes de nós.

**keywords:** symplectic geometry, pseudoholomorphic curves, knot invariants.

## 1 Introduction

Knot theory and symplectic geometry have both seen a great development in recent years. In some instances, techniques from symplectic geometry have been successful in producing powerful new invariants of knots (like the knot Floer homology of Ozsváth–Szabó and Rasmussen [15, 16]), or in enhancing the understanding of previously known invariants (like a symplectic version of Khovanov homology [1]). In this note, we present a recent result obtained in collaboration with Tobias Ekholm, which yields a formula for the Alexander polynomial of a knot in terms of its augmentation polynomial. The former is a classical cornerstone of knot theory. The latter is a recently defined object, introduced in the context of knot contact homology. This is another very powerful new invariant of knots that was constructed using tools from symplectic geometry. Our result is saying that knot contact homology recovers the Alexander polynomial. Although this fact was already known from the work of Ng [14], the formula in terms of the augmentation polynomial appears to be new. It also has an unusual form for a relation between two polynomials. Our result will be stated as Theorem 5.1 below.

We will begin with a brief introduction to knots and the Alexander polynomial, including a dynamical definition of this invariant that will be useful for our purposes. Then, we change direction and give a quick introduction to symplectic geometry and pseudoholomorphic curves. After that, we explain how to use pseudoholomorphic curves to define knot contact homology, and how the latter yields the augmentation polynomial of a knot. Then, we state our result and give a terse presentation of the proof. Our goal will not be to convey the full logical structure of the argument (let alone its technical details), but only to give an idea of a practical and hopefully interesting application of pseudoholomorphic curves in symplectic geometry. We conclude with some directions for future work.

*Acknowledgements.* The author wishes to thank Jorge Milhazes de Freitas, Samuel Lopes and Diogo Oliveira e Silva, the organizers of the conference *Matemáticos Portugueses pelo Mundo* in 2019 in Porto, for the invitation to participate in that very intereresting event. Thanks are also due to Diogo Oliveira e Silva and the anonymous referees for many useful comments about this text.

# 2 Knots and the Alexander polynomial

## 2.1 Knots and links

A *knot* is a closed embedded curve in $\mathbb{R}^3$. This means that it is the image of a $C^\infty$-smooth injective map from the circle $S^1$ to $\mathbb{R}^3$, with non-zero derivative at every point. A *link* is a finite collection of knots that are all pairwise disjoint. We are interested in knots and links from the point of view of topology, in the sense that we don't want to distinguish those that differ by a smooth deformation causing no self-intersections, called an *isotopy*. Formally, this is a $C^\infty$-smooth map $f\colon [0,1]\times\mathbb{R}^3 \to \mathbb{R}^3$, such that $f_t \coloneqq f(t,.)$ is a diffeomorphism of $\mathbb{R}^3$ for every $t \in [0,1]$, and $f_0$ is the identity. Two links are *isotopic* if there is an isotopy $f$ such that the image under $f_1$ of one link is the other link.

In Figure 1 we have the two simplest examples of knots: the *unknot* and the *trefoil*. We can think of this picture as the result of projecting our knots

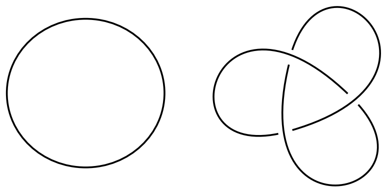


Figure 1: The unknot and the trefoil (with three crossings)

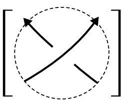
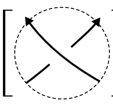
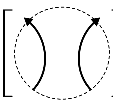
in $\mathbb{R}^3$ to a plane, so that the projection is injective at all but finitely many points, called *crossings*. A figure like this, encoding for each crossing which of the two strands is over the other, is called a *link diagram*.

It is intuitively clear that the unknot and the trefoil are not isotopic, but it is not entirely obvious how to prove this fact. The main problem in knot theory is to find an efficient way of deciding when two knots are isotopic.

## 2.2 The Alexander polynomial

One of the first tools that were created to distinguish knots and links is the *Alexander polynomial*. Given a link $L$, its Alexander polynomial $\mathrm{Alex}_L$ is a Laurent polynomial in one variable $\mu$. This means that the integer powers of $\mu$ are allowed to be negative. Define $\mathrm{Alex}_L$ as follows: pick an orientation for $L$, which is to say a direction for each of its component knots, and impose

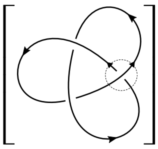
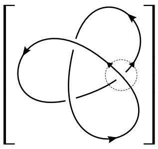
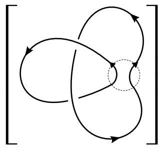
- $\mathrm{Alex}_{\mathrm{unknot}} = 1$.
- The *skein relation*:

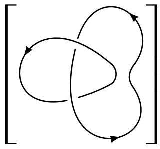
  Alex [  ] − Alex [  ] + $(\mu^{1/2} - \mu^{-1/2})$ Alex [  ] = 0

  This is a relation between Alexander polynomials of three links with link diagrams that are equal outside the depicted neighborhood of a crossing.
- *Isotopy invariance*: $\mathrm{Alex}_L = \mathrm{Alex}_{L'}$ if $L$ and $L'$ are isotopic links.

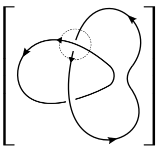
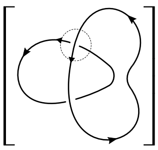
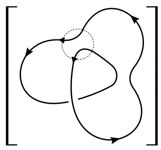
These three properties determine the Alexander polynomial for every link. Two non-obvious facts are that the Alexander polynomial is well-defined (in particular, the skein relation holds for every link diagram) and that it contains only integer powers of $\mu$, even though the skein relation involves fractional powers.

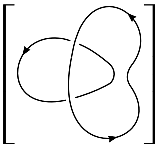
As an example, let us compute the Alexander polynomial of the trefoil. Applying the skein relation, we get

Alex  = Alex  $- (\mu^{1/2} - \mu^{-1/2})$ Alex 

$= 1 - (\mu^{1/2} - \mu^{-1/2})$ Alex 

where we used the fact that $\text{Alex}_{\text{unknot}} = 1$. The link we obtained on the right is called *Hopf link*. Let us apply the skein relation on a crossing of this link:

Alex  = Alex  $- (\mu^{1/2} - \mu^{-1/2})$ Alex 

= Alex  $- \mu^{1/2} + \mu^{-1/2}$

where we used again that $\text{Alex}_{\text{unknot}} = 1$. To finish the computation, we need to determine the Alexander polynomial of the link with two components on the right. Since the two components can be moved by an isotopy to lie in two disjoint open balls, the link is said to be trivial. We call it the *unlink with two components*. Its Alexander polynomial is zero, and we leave the proof of that to the reader as an exercise on the skein relation. We can now conclude the calculation of the Alexander polynomial of the trefoil:

$$\text{Alex}_{\text{trefoil}} = 1 - (\mu^{1/2} - \mu^{-1/2})(0 - \mu^{1/2} + \mu^{-1/2}) = \mu - 1 + \mu^{-1}.$$

Since this is different from the Alexander polynomial of the unknot, we can conclude that the unknot and the trefoil are not isotopic.

*Exercise* 1. Compute the Alexander polynomial of the *figure-eight knot* (the closure of the sailor's knot of the same name), depicted in Figure 2.

## 2.3 A dynamical definition of the Alexander polynomial

The definition of Alexander polynomial of a link that we gave in the previous section is very suitable for computations (at least for link diagrams without

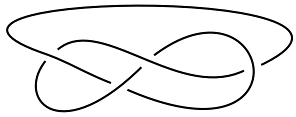


Figure 2: The figure-eight knot

too many crossings). Nevertheless, it is only one of many definitions of this invariant. We now present a different definition of the Alexander polynomial of a link, with a very different flavour. For simplicity, we will restrict our attention to the particular class of fibered knots, which we now define.

In this section, it will be convenient to think of the ambient space of a knot as the sphere $S^3$, instead of $\mathbb{R}^3$. This is reasonable, since we can identify $\mathbb{R}^3$ with the complement of a point in $S^3$. For this identification, two knots are isotopic in $\mathbb{R}^3$ if and only if they are isotopic in $S^3$.

We say that a knot $K$ is *fibered* if there is a $C^\infty$-smooth map $g\colon S^3\backslash K \to S^1$ with no critical points. This means that the knot complement $S^3\backslash K$ is a fiber bundle over $S^1$, with fiber a surface whose boundary is $K$ (such a surface is called a *Seifert surface*). The differential of the function $g$ is a 1-form $dg$. If we choose some Riemannian metric $\langle .,. \rangle$ on $S^3$, then the function $g$ also specifies a vector field in $S^3\backslash K$, called *gradient vector field* and denoted $\nabla g$, as follows: for every vector field $v$ on $S^3\backslash K$,

$$\langle \nabla g, v \rangle = dg(v).$$

Since the function $g$ has no critical points, the vector field $\nabla g$ has no zeros. A *gradient flow loop* is a path $\gamma\colon [0,R] \to S^3\backslash K$, for some $R > 0$, such that

- $\gamma(R) = \gamma(0)$ (which means that $\gamma$ closes up to a loop) and
- $\frac{d}{dt}\big(\gamma(t)\big) = (\nabla g)_{\gamma(t)}$ for every $t \in [0,R]$ (that is, the time-derivative of $\gamma$ coincides with $\nabla g$ at every point in $\gamma$).

Observe that if $\gamma\colon [0,R] \to S^3\backslash K$ is a gradient flow loop, then so is any multiple cover $\gamma_m\colon [0,mR] \to S^3\backslash K$, where $m$ is a positive integer. Here, $\gamma_m(t) = \gamma(t')$ for $t' \in [0,R]$ such that $t' \equiv t \mod R$. We say that a flow loop is *simple* if it is not multiply covered. Given a flow loop $\gamma$, we denote by $m(\gamma)$ its multiplicity with respect to its underlying simple loop. For every knot $K$, the homology group $H_1(S^3\backslash K;\mathbb{Z})$ is isomorphic to $\mathbb{Z}$. If we pick a generator $e$ for this homology group, then we can associate to a flow loop $\gamma$ its degree $d(\gamma)$, such that the class of $\gamma$ on homology is $d(\gamma)e$. Note that

$m(\gamma)$ divides $d(\gamma)$. To avoid flow loops too close to $K$, we will require the map $g$ to "grow near $K$".

**Theorem 2.1** (Milnor [13]). *The Alexander polynomial of a fibered knot $K$ is given by*

$$\mathrm{Alex}_K(\mu) = (1-\mu)\exp\left(\sum_{\gamma} \frac{\sigma(\gamma)}{m(\gamma)}\mu^{d(\gamma)}\right) \tag{1}$$

*where the sum is over all gradient flow loops (not only the simple ones). In the formula, $\sigma(\gamma) \in \{\pm 1\}$ is a sign (associated to the linearized return map of $\gamma$).*

This formula was generalized for all knots $K$ by Hutchings and Lee [11].

*Example* 2.2. Let us see how to recover the Alexander polynomial of the unknot from formula (1). The complement of the unknot in $S^3$ is diffeomorphic to $S^1 \times \mathbb{R}^2$ (if this is not clear, try to identify both spaces with the complement of the vertical axis in $\mathbb{R}^3$). In coordinates $(\theta, x, y)$ for $S^1 \times \mathbb{R}^2$, take $g(\theta, x, y) = \theta + x^2 + y^2$. For the standard product metric on $S^1 \times \mathbb{R}^2$, the only periodic orbits are the covers of the central circle $S^1 \times \{0\}$, and all the signs $\sigma(\gamma)$ in (1) are positive. The sum over flow loops becomes

$$\sum_{k>0} \frac{1}{k}\mu^k = -\ln(1-\mu)$$

hence

$$\mathrm{Alex}_{\mathrm{unknot}}(\mu) = (1-\mu)\exp\left(-\ln(1-\mu)\right) = 1$$

as we already knew.

# 3 Some symplectic geometry

## 3.1 Classical mechanics and symplectic geometry

Symplectic geometry is a recent area of mathematics, with its roots in classical mechanics, but with deep connections to other areas of mathematics and physics. In the Hamiltonian formulation of classical mechanics, a particle moving in $\mathbb{R}^3$ is described by its trajectory in the phase space $\mathbb{R}^6$, which keeps track of the position and momentum of the particle. If we denote position variables in $\mathbb{R}^3$ by $q_1, q_2, q_3$ and the corresponding momentum variables by $p_1, p_2, p_3$, then the trajectory of the particle in phase space satisfies Hamilton's equations

$$\begin{cases} \dot{q}_i = \partial_{p_i} H \\ \dot{p}_i = -\partial_{q_i} H \end{cases}$$

where the *Hamiltonian function* $H\colon \mathbb{R}^6 \to \mathbb{R}$ is the energy of the particle. This trajectory is a flow line of the *Hamiltonian vector field*, which we denote by $X_H$. If we define a differential 2-form on $\mathbb{R}^6$ by

$$\omega \coloneqq \sum_{i=1}^{3} dp_i \wedge dq_i, \tag{2}$$

then Hamilton's equations above tell us that the vector field $X_H$ is given by the condition

$$\omega(., X_H) = dH. \tag{3}$$

A reader unfamiliar with differential forms may find it mildly useful to think of $\omega$ as a way of prescribing signed areas to 2-dimensional oriented surfaces in $\mathbb{R}^6$ (where the signs depend on the orientations of the surfaces).

We can interpret equation (3) as saying that the 2-form $\omega$ allows us to do Hamiltonian mechanics for any function that we choose to call energy on $\mathbb{R}^6$. We can think of symplectic geometry as generalizing this point of view on mechanics to any differentiable manifold of even dimension $2n$, equipped with a differential 2-form $\omega$ whose properties mimic those of the form (2), namely:

- $\omega$ is closed: $d\omega = 0$, and
- $\omega$ is non-degenerate: the $n$-fold wedge product $\omega \wedge \ldots \wedge \omega$ is a volume form (which means that it vanishes nowhere).

For the application to knot theory that we present in this text, it will mostly suffice to think of the symplectic manifold $\mathbb{R}^6$. We refer to the article by Ana Cannas da Silva in this volume [4] for more on symplectic geometry.

### 3.2 Pseudoholomorphic curves

In 1985, Gromov introduced the notion of *pseudoholomorphic curve* [10], which was revolutionary in symplectic geometry. It gave a powerful tool to study symplectic manifolds, and eventually led to many deep relations to algebraic geometry and theoretical physics, in particular the so called *mirror symmetry* phenomenon (see Lino Amorim's article in this volume [3] for some background on mirror symmetry). Before we state one of the striking results in Gromov's paper, let us introduce some more terminology. First, we observe that an open subset of a symplectic manifold, equipped with the restriction of the symplectic form $\omega$, is also a symplectic manifold. Let $B^{2n}(r) \subset \mathbb{R}^{2n}$ denote the open ball of radius $r$ and centered at the origin. Consider also the open subset $B^2(R) \times \mathbb{R}^{2n-2} \subset \mathbb{R}^{2n}$, where it is crucial that $B^2(R)$ has coordinates $(p_1, q_1)$ (instead of $(p_1, p_2)$ or $(q_1, q_2)$, for instance) and $\mathbb{R}^{2n-2}$ has the remaining coordinates $p_2, \ldots, p_n, q_2, \ldots, q_n$. Given two symplectic manifolds $(M_1, \omega_1)$ and $(M_2, \omega_2)$, a smooth embedding $\varphi\colon M_1 \hookrightarrow M_2$ such that $\varphi^*\omega_2 = \omega_1$ is called a *symplectic embedding* (here, $\varphi^*$ is pullback by $\varphi$).

**Theorem 3.1** (Gromov's non-squeezing)**.** *If we have a symplectic embedding*

$$B^{2n}(r) \hookrightarrow B^2(R) \times \mathbb{R}^{2n-2},$$

*then $r \leqslant R$.*

Observe that a symplectic embedding is, by definition, volume-preserving. We can interpret Gromov's non-squeezing as saying that not all volume-preserving embeddings are symplectic.

Now that we have given a little indication of what pseudoholomorphic curves can achieve, let us define them. We need the auxiliary notion of an *almost complex structure* on an even-dimensional manifold $M^{2n}$, which is an endomorphism of the tangent bundle $J\colon TM \to TM$ (covering the identity map $M \to M$) such that $J^2 = -Id$. Given such a $J$ and a Riemann surface $(S, j)$, a pseudoholomorphic curve is a map $u\colon S \to M$ satisfying the Cauchy–Riemann equation

$$du \circ j = J \circ du.$$

If $M$ has a symplectic form $\omega$, then one can ask that $\omega(., J.)$ be a Riemannian metric on $M$, in which case $J$ is said to be *compatible* with $\omega$. Gromov's idea was to study $(M, \omega)$ by analyzing moduli spaces of pseudoholomorphic curves (modulo domain reparametrizations, and possibly with additional structures like fixing the homology class of the map, or equipping the domain with marked points). If $J$ is compatible with $\omega$, then we can control the $L^2$-norm of $u$ (the *energy*) by its $\omega$-area, which is crucial for obtaining compactness of moduli spaces. Gromov also observed that the space of $\omega$-compatible $J$ is contractible, which implies that the moduli spaces defined for two different $J$ are cobordant (that is, there is a manifold whose oriented boundary is the difference of the two moduli spaces). This allows for the definition of numerical invariants counting pseudoholomorphic curves (with appropriately chosen constraints) that depend on $\omega$ but not on the choice of $\omega$-compatible $J$. Those are called *Gromov–Witten invariants*, and they have many applications in symplectic and algebraic geometry.

# 4 Symplectic knot invariants

## 4.1 From knots to Lagrangians and Legendrians

Recent decades have seen many applications of pseudoholomorphic curves. We will focus on a particular application to knot theory, called *knot contact homology*. This is part of a broader packaging of pseudoholomorphic curve information that goes by the name of *symplectic field theory* [8], but we will focus on the specific case of interest to us. Let us begin with some geometric constructions.

Given a knot $K \subset \mathbb{R}^3$ , we can define its *conormal Lagrangian*

$$L_K \coloneqq \{(q,p) \in \mathbb{R}^6 \,|\, q \in K \text{ and } (\forall v \in T_q K)\, \langle p, v\rangle = 0\},$$

where $\langle .,.\rangle$ is the Euclidean inner product in $\mathbb{R}^3$. This is a submanifold of $\mathbb{R}^6$ that is diffeomorphic to $S^1 \times \mathbb{R}^2$, and whose intersection with $\mathbb{R}^3_q$ (the subspace of $\mathbb{R}^6$ where all $p_i = 0$) is the knot $K$. See Figure 3 for a geometric depiction that would greatly benefit from additional dimensions. Furthermore, $L_K$ is *Lagrangian*, in the sense that it has half the dimension of the ambient space $\mathbb{R}^6$, and the restriction of the symplectic form $\omega$ in (2) to $L_K$ vanishes. In addition, the Lagrangian $L_K$ is *exact*, which means the following. The symplectic form $\omega$ in $\mathbb{R}^6$ has a primitive $\lambda = \sum_{i=1}^3 p_i dq_i$ and the restriction $\lambda|_L$ admits a primitive $f \in C^\infty(L)$ (in this case, we can take $f$ to be any constant function). Other exact Lagrangians are $\mathbb{R}^3_q$ and $\mathbb{R}^3_p$.

We can identify $\mathbb{R}^6$ with the tangent bundle[1] of $\mathbb{R}^3_q$ and, with respect to the Euclidean inner product in $\mathbb{R}^3$, we can identify $\mathbb{R}^3 \times S^2 \subset \mathbb{R}^6$ with the unit tangent bundle of $\mathbb{R}^3_q$. Recall that the *geodesic flow* on the unit tangent bundle of a Riemannian manifold $Q$ takes a point $q \in Q$ and a unit vector $v \in T_q Q$ and follows the geodesic starting at $q$ in the direction prescribed by $v$. This is an example of what is called a *Reeb flow* in contact geometry (hence the name "knot contact homology"), but we will not go further in that direction in this note. The conormal Lagrangian $L_K$ intersects $\mathbb{R}^3 \times S^2$ in a 2-torus $\Lambda_K$, which we call *conormal Legendrian* (again borrowing terminology from contact geometry).

It will be useful to observe that $H_2(\mathbb{R}^3 \times S^2, \Lambda_K; \mathbb{Z})$ is isomorphic to $\mathbb{Z}^3$. We will explain this point, but a reader less familiar with homology groups might want to skip the details. Let us just mention that this is the reason why the augmentation polynomial below will have three variables.

[1] From the point of view of symplectic geometry, it would be preferable to think of the cotangent bundle, but we can ignore that point in this text.

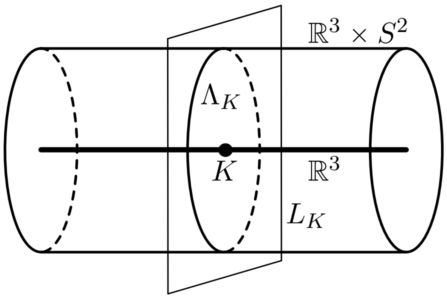



Figure 3: The Lagrangian $L_K$ in $\mathbb{R}^6$

In this paragraph, consider all homology groups with $\mathbb{Z}$ coefficients. The long exact sequence of the pair $(\mathbb{R}^3 \times S^2, \Lambda_K)$ includes the segment

$$H_2(\Lambda_K) \to H_2(\mathbb{R}^3 \times S^2) \to H_2(\mathbb{R}^3 \times S^2, \Lambda_K) \to H_1(\Lambda_K) \to 0.$$

The first map turns out to vanish. Since $H_2(\mathbb{R}^3 \times S^2) \cong H_2(S^2) \cong \mathbb{Z}$ and $H_1(\Lambda_K) \cong \mathbb{Z}^2$ ($\Lambda_K$ is a 2-torus), the sequence splits and we get the desired isomorphism with $\mathbb{Z}^3$. We get generators for this group from the choice of a generator $t$ for $H_2(S^2)$ and of generators $x, p$ for $H_1(\Lambda_K)$ (and a choice of splitting). It is customary to let $x$ be a longitude curve (projecting to $K$ under the restriction to $\Lambda_K$ of the projection $\mathbb{R}^3 \times S^2 \to \mathbb{R}^3$), and to let $p$ be a meridian curve (mapping to a constant under that same projection). Note that such a meridian curve $p$ lies in a cotangent fiber (that is, a 3-dimensional subspace of $\mathbb{R}^6$ with constant $q_i$ variables), hence the use of the letter associated with momentum.

## 4.2 Knot contact homology

We can now use pseudoholomorphic curves to associate a chain complex to the knot $K$. We will actually get a *differential graded algebra* (dga), which is a chain complex with a product satisfying the (graded) Leibniz rule. Our chain complex will be a tensor algebra generated by geodesic chords starting and ending in $\Lambda_K$. By this we mean paths $c\colon [a,b] \to \mathbb{R}^3 \times S^2$ that follow the geodesic flow and for which $c(a)$ and $c(b) \in \Lambda_K$. We don't want to get into details, but these chords are graded by a Maslov index (which is an integer).

Let us specify the ring over which we take the tensor algebra. This will be group ring (over $\mathbb{C}$) of $H_2(\mathbb{R}^3 \times S^2, \Lambda_K; \mathbb{Z})$, which, in light of the discussion at the end of the previous section, can be identified with the Laurent polynomial ring $R = \mathbb{C}[\lambda^{\pm 1}, \mu^{\pm 1}, Q^{\pm 1}]$, under the identifications $\lambda = e^x$, $\mu = e^p$ and $Q = e^t$.

The differential in the chain complex counts pseudoholomorphic curves in $\mathbb{R}\times(\mathbb{R}^3\times S^2)=\mathbb{R}^4\times S^2$ (which we can identify with the complement of $\mathbb{R}^3_q$ in $\mathbb{R}^6$), as follows. We define the differential for geodesic chords, and extend by linearity and the Leibniz rule. The differential of a geodesic chord $x$ is

$$\partial x = \sum_{y_1,\ldots,y_k} \left( \sum_{u\in\mathcal{M}(x;y_1,\ldots,y_k)} r(u) \right) y_1 \otimes \ldots \otimes y_k$$

where the first sum is over finite sequences of geodesic chords and the second sum is over elements of the moduli space of pseudoholomorphic curves $u$ in $\mathbb{R}^4 \times S^2$, whose domain is a disk with $k+1$ punctures on the boundary. The boundary components map to $\mathbb{R} \times \Lambda_K$. At the boundary punctures, $u$ is asymptotic to the fixed geodesic chords, with $x$ at $+\infty$ and the $y_i$ at $-\infty$ (both infinities in the first $\mathbb{R}$ summand of the target $\mathbb{R} \times (\mathbb{R}^3 \times S^2)$). See Figure 4 for an illustration of one such $u$. Finally, the coefficient $r(u) \in R$ keeps track of the relative homology class of $u$ in $H_2(\mathbb{R}^3 \times S^2, \Lambda_K; \mathbb{Z})$. We will not go into more details at this point, but the reader may have noticed that more choices are necessary, including of "capping half-disks" for the geodesic chords (to obtain a relative homology class).

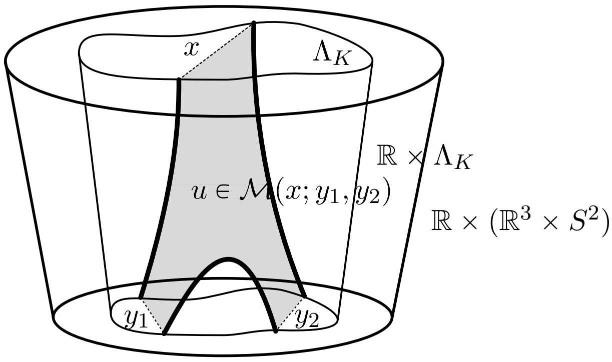



Figure 4: A contribution to $\partial x$

**Theorem 4.1** (Ekholm–Etnyre–Ng–Sullivan [6])**.** *The differential $\partial$ defined above squares to zero. The homology of this dga is an invariant of the knot.*

This homology is called knot contact homology. It is sometimes useful to keep track of the dga, denoted by $\mathcal{A}_K$, instead of passing to homology.

Although the technical details of the proof of Theorem 4.1 are quite involved, the idea is by now standard in symplectic geometry. To prove an algebraic identity like $\partial^2 = 0$, one interprets the contributions to $\partial^2$ as elements in the boundary of a suitably defined moduli space of pseudoholomorphic curves. By showing that this moduli space is a compact 1-dimensional oriented manifold, one concludes that the signed count of the elements in its boundary is zero.

Knot contact homology appears to be a strong knot invariant, but it is not yet clear just how strong. A recent result shows that a small but non-trivial enhancement of knot contact homology is a complete knot invariant (that is, two knots are isotopic if and only if their enhanced knot contact homologies are isomorphic) [7].

### 4.3 Augmentations

Although the definition of the dga $\mathcal{A}_K$ involves pseudoholomorphic curves, which can be very difficult to analyze, the dga turns out to admit a combinatorial model, which can be written down explicitly given a braid presentation for the knot $K$ [6]. Nevertheless, since the chain complex (a tensor algebra) is very large, it can be difficult to extract useful information from its homology. One way of obtining more treatable information about the dga is via its *augmentations*. An augmentation is a unital dga map

$$\varepsilon \colon \mathcal{A}_K \to \mathbb{C},$$

where the field $\mathbb{C}$ is thought of as a dga supported in degree zero and with trivial differential. In other words, $\varepsilon$ is a graded unital ring map (so, it is only non-trivial on the degree zero part of $\mathcal{A}_K$) satisfying $\varepsilon \circ \partial = 0$.

*Example* 4.2. One important source of augmentations is given by exact Lagrangians in $\mathbb{R}^6$ "which look like $\Lambda_K$ near infinity" (in some precise sense). A key example is the conormal $L_K$. Given such a Lagrangian, we can define an augmentation by assigning to each geodesic chord of degree 0 the count of pseudoholomorphic disks in $\mathbb{R}^6$ with boundary on the Lagrangian and one puncture on the boundary, where the disk is asymptotic to the geodesic chord "at infinity". The value of the augmentation on the coefficient ring $R$ is constrained by the topology of the Lagrangian and its ambient space. In the case of $L_K$, since the meridian $p$-curve is contractible in $L_K$ and the $t$-sphere is null-homologous in $\mathbb{R}^6$, it turns out that $\mu = 1 = Q$, but $\lambda$ is not constrained. So, for each value $\lambda \in \mathbb{C}\backslash\{0\}$ we get an augmentation of $\mathcal{A}_K$.

It turns out to be useful to also think of the space of augmentations geometrically. We define the *augmentation variety* of $K$, denoted by $V_K$, to consist of the union of maximal dimensional components of the Zariski closure of the set

$$\{(\varepsilon(\lambda), \varepsilon(\mu), \varepsilon(Q)) \in (\mathbb{C}\backslash\{0\})^3 \,|\, \varepsilon \text{ is an augmentation}\}.$$

The existence of the augmentations associated to $L_K$ in Example 4.2 implies that, for every knot $K$, the augmentation variety $V_K$ contains the line $\{(\lambda, 1, 1)\}$ (where $\lambda$ can be any element in $\mathbb{C}\backslash\{0\}$).

**Theorem 4.3** (Diogo–Ekholm [5])**.** *For every knot $K$, the augmentation variety $V_K$ is an affine algebraic subvariety of $(\mathbb{C}\backslash\{0\})^3$ of complex dimension at least 2.*

Conjecturally, $V_K$ is always 2-dimensional (so it is not all of $(\mathbb{C}\backslash\{0\})^3$). Define the *augmentation polynomial* of $K$ (denoted by $\operatorname{Aug}_K(\lambda, \mu, Q)$) as a polynomial with no repeated factors that generates the vanishing ideal of this variety: $V_K = V(\operatorname{Aug}_K)$.

*Example* 4.4. The augmentation polynomial of the unknot is

$$\operatorname{Aug}_U = 1 - \lambda - \mu + \lambda\mu Q$$

and that of the trefoil is

$$\operatorname{Aug}_T = \lambda^2(\mu-1)+\lambda(\mu^4-\mu^3Q+2\mu^2Q^2-2\mu^2Q-\mu Q^2+Q^2)+(\mu^3Q^4-\mu^4Q^3).$$

The augmentation polynomial has deep and surprising connections to string theory and to other knot invariants. It is conjecturally the same as the so-called *$Q$-deformed $A$-polynomial*, which is relevant for mirror symmetry and is related in a deep way with another important knot invariant called the *colored HOMFLYPT polynomial* [2, 9].

## 5 The Alexander polynomial from the augmentation polynomial

As we have seen, the Alexander polynomial and the augmentation polynomial are knot invariants defined in very different ways. Nevertheless, they are related in the following surprising manner.

**Theorem 5.1** (Diogo–Ekholm [5])**.** *Recall that $\lambda = e^x$, $\mu = e^p$ and $Q = e^t$. We have*

$$\operatorname{Alex}_K(\mu) = (1-\mu)\exp\left(\int -\frac{\partial_Q \operatorname{Aug}_K}{\partial_\lambda \operatorname{Aug}_K}\bigg|_{(\lambda,Q)=(1,1)} dp\right) \tag{4}$$

*if the denominator $\partial_\lambda \operatorname{Aug}_K|_{(\lambda,Q)=(1,1)}$ is not identically zero.*

In formula (4), the integral symbol represents an antiderivative. We will give a brief idea of why one might expect the formula to hold, at least for fibered knots. We will be very imprecise and will not justify most of our claims. Our goal is to illustrate how the study of moduli spaces of pseudoholomorphic curves can lead to meaningful algebraic identities (we already saw that this is also the idea of the proof that $\partial^2 = 0$ in the dga $\mathcal{A}_K$). Note that, according to Milnor's formula (1), we only need to argue that

$$\frac{d}{dp}\left(\sum_{\gamma \text{ in } S^3\backslash K} \frac{\sigma(\gamma)}{m(\gamma)}\mu^{d(\gamma)}\right) = -\frac{\partial_Q \operatorname{Aug}_K}{\partial_\lambda \operatorname{Aug}_K}\bigg|_{(\lambda,Q)=(1,1)}. \tag{5}$$

*Exercise* 2. Apply formula (4) to Example 4.4 to recover the Alexander polynomials of the unknot and the trefoil. The Alexander polynomial is often defined up to a power of $\mu$, and (4) should also be allowed that ambiguity.

### 5.1 From flow loops to pseudoholomorphic annuli

The left side of (5) involves orbits in $S^3\backslash K$, whereas the right side involves pseudoholomorphic curves in $\mathbb{R}^4 \times S^2$. To get a reformulation of the left side also in terms of pseudoholomorphic curves, we need another geometric ingredient. Recall that the conormal Lagrangian $L_K \subset \mathbb{R}^6$ intersects $\mathbb{R}^3_q$ in the knot $K$. There is a procedure called *Lagrangian surgery*, which produces another Lagrangian submanifold by smoothing out the union of $L_K$ with $\mathbb{R}^3_q$ (the version we need is described in [12]). Denote the new Lagrangian in $\mathbb{R}^6$ by $M_K$. This submanifold is diffeomorphic to $\mathbb{R}^3\backslash K$. Since $L_K$ and $\mathbb{R}^3_q$ are exact Lagrangians, one can ensure that $M_K$ is also exact. In particular, it has an associated family of augmentations $\varepsilon_{M_K}$, sending both generators $\lambda$ and $Q$ of the coefficient ring $R$ to 1, and the generator $\mu$ to any element of $\mathbb{C}\backslash\{0\}$. Hence, the line $\{(1,\mu,1)\}$ is also contained in the augmentation variety $V_K$ for every $K$. The key role of these augmentations is the reason behind taking $\lambda = Q = 1$ in formula (4).

In $\mathbb{R}^6$, we can consider *pseudoholomorphic annuli* between $\mathbb{R}^3_q$ and $M_K$. These are pseudoholomorphic maps $u\colon S^1 \times [0,A] \to \mathbb{R}^6$ (for some $A \geqslant 0$) such that the restriction of $u$ to $S^1 \times \{0\}$ maps to $\mathbb{R}^3_q$ and the restriction to $S^1 \times \{A\}$ maps to $M_K$. Denote the moduli space of such pseudoholomorphic annuli by $\mathcal{M}(\mathbb{R}^3_q; M_K)$.

The following result is stated in an overly simplified and somewhat imprecise manner.

**Proposition 5.2.** *For suitable choices of $g\colon S^3\backslash K \to S^1$, metric on $S^3$ and $J$ on $\mathbb{R}^6$, gradient flow orbits in $S^3\backslash K$ can be identified with pseudoholomor-*

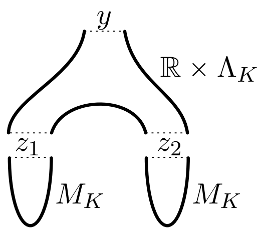



Figure 5: Definition of $F(\lambda, \mu, Q)$. The curve at the top could have arbitrarily many negative punctures capped by disks with boundary in $M_K$.

*phic annuli in $\mathcal{M}(\mathbb{R}^3_q; M_K)$. Therefore, the sum on the left side of (5) can be rewritten as*

$$A(\mu) := \sum_{u \in \mathcal{M}(\mathbb{R}^3_q; M_K)} \frac{\sigma(u)}{m(u)} \mu^{d(u)} \tag{6}$$

*for suitable signs $\sigma(u)$ and integers $m(u)$ and $d(u)$.*

Equation (5) is thus equivalent to

$$\frac{d}{dp}\left(A(\mu)\right) = -\left.\frac{\partial_Q \operatorname{Aug}_K}{\partial_\lambda \operatorname{Aug}_K}\right|_{(\lambda,Q)=(1,1)}, \tag{7}$$

where we recall again that $\mu = e^p$.

### 5.2 From pseudoholomorphic annuli to knot contact homology

Instead of showing equation (7) directly, we show that

$$\frac{d}{dp}\left(A(\mu)\right) = -\left.\frac{\partial_Q F}{\partial_\lambda F}\right|_{(\lambda,Q)=(1,1)}, \tag{8}$$

for a suitable holomophic function $F(\lambda, \mu, Q)$ such that

$$\left.\frac{\partial_Q F}{\partial_\lambda F}\right|_{(\lambda,Q)=(1,1)} = \left.\frac{\partial_Q \operatorname{Aug}_K}{\partial_\lambda \operatorname{Aug}_K}\right|_{(\lambda,Q)=(1,1)}. \tag{9}$$

The function $F$ is defined as follows. For an appropriately chosen generator $y$ of degree 1 of the dga $\mathcal{A}_K$ (actually, an $R$-linear combination of such generators), take its dga differential $\partial$, which is an expression in $\lambda$, $\mu$, $Q$ and other generators $z_1, \ldots, z_n$. Then, send the $z_i$ to their images under the augmentation $\varepsilon_{M_K}$. See Figure 5.

Now, consider the moduli space of pseudoholomorphic annuli in $\mathbb{R}^6$, with one boundary component in $\mathbb{R}^3_q$ and another in $M_K$ (as in $\mathcal{M}(\mathbb{R}^3_q; M_K)$

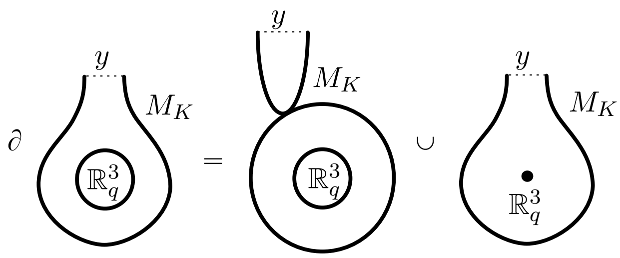



Figure 6: The boundary of a 1-dimensional moduli space

above), but with a puncture on the boundary component mapping to $M_K$. At this puncture, the curve is asymptotic to $y$. See the left side of Figure 6.

This moduli space is compact and 1-dimensional (since $y$ has degree 1) and (if $y$ is chosen carefully) its boundary has components of two types, which are depicted on the center and right in Figure 6. In the center configuration, the curve develops a node and breaks into a pseudoholomorphic plane asymptotic to $y$ and an annulus in $\mathcal{M}(\mathbb{R}^3_q; M_K)$. The boundaries of the plane and annulus intersect. In the rightmost configuration, the boundary loop in $\mathbb{R}^3_q$ shrinks to a point, so the punctured annulus becomes a plane.

A further study of the pseudoholomorphic planes in the center configuration reveals that the count of such broken curves (using $\mu$ to keep track of the homology of boundaries mapping to $M_K$) is given by

$$\frac{dA}{dp} \cdot \frac{\partial F}{\partial x}\bigg|_{(\lambda,Q)=(1,1)} = \frac{dA}{dp} \cdot \frac{\partial F}{\partial \lambda}\bigg|_{(\lambda,Q)=(1,1)},$$

recalling once more that $\lambda = e^x$. The derivatives in the formula keep track of the intersection of the boundaries of the disk and annulus. Similarly, the counts of curves in the configuration on the right turn out to be encoded by

$$\frac{\partial F}{\partial t}\bigg|_{(\lambda,Q)=(1,1)} = \frac{\partial F}{\partial Q}\bigg|_{(\lambda,Q)=(1,1)},$$

where $Q = e^t$. This time, the derivative keeps track of the fact that the disk intersects $\mathbb{R}^3_q$. Since these two configurations are the boundaries of a compact 1-dimensional manifold, the sum of their contributions (with appropriate signs) vanishes. This implies that

$$\frac{dA}{dp} \cdot \frac{\partial F}{\partial \lambda}\bigg|_{(\lambda,Q)=(1,1)} + \frac{\partial F}{\partial Q}\bigg|_{(\lambda,Q)=(1,1)} = 0,$$

which gives equation (8), as wanted.

For a brief justification of equation (9), let us just say that one can argue that $F$ vanishes on the augmentation variety $V_K$, so it should be of the form

$$F = g\ \mathrm{Aug}_K$$

for some analytic function $g(\lambda, \mu, Q)$. Equation (9) now follows from the product rule for derivatives and the fact that $\mathrm{Aug}_K$ vanishes along the line $\{(1, \mu, 1)\} \subset V_K$ (at least if we assume that $g|_{(\lambda,Q)=(1,1)}$ is not identically zero, which as it turns out we can).

### 5.3 Outlook

Theorem 5.1 should not be thought of as an efficient way of computing the Alexander polynomial of a knot, but rather as an unexpected relation between two very different knot invariants. It also suggests further investigation in a few directions. For example, one might not set $Q = 1$ in equation (4) and get a $Q$-deformed version of $\mathrm{Alex}_K$.

**Question 5.3.** *What is the significance of this deformation of the Alexander polynomial? Is it related to other deformations, coming for instance from knot Floer homology [15]?*

One might also wonder about the condition of non-vanishing of the denominator in the theorem. As it turns out, this condition cannot be neglected, as it does not hold, for instance, for the $8_{20}$ knot (as pointed out to us by Lenny Ng).

**Question 5.4.** *Is there an analogue of equation* (4) *when the denominator in the formula vanishes?*

It is likely that along some branch of the variety $V_K$, corresponding to the augmentation $M_K$, one could find such an analogue.

As a final note, the reader may have wondered about interpreting the integrand in formula (4) via implicit differentiation. Indeed, since $V_K$ is the vanishing locus of $\mathrm{Aug}_K$, that integrand is the partial derivative $\frac{\partial \lambda}{\partial Q}$ along the line $\{(1, \mu, 1)\} \subset V_K$. This leads to an alternative interpretation of the right side in the formula, related to curve counts in the resolved conifold (the total space of the bundle $\mathcal{O}_{\mathbb{C}P^1}(-1) \oplus \mathcal{O}_{\mathbb{C}P^1}(-1)$), in the spirit of [2]. That is another interesting story, but unfortunately it is beyond the scope of this discussion.

## References


[1] M. Abouzaid and I. Smith, “Khovanov homology from Floer cohomology”, *Journal of the American Mathematical Society*, Vol. 32 (2019), pp. 1–79.

[2] M. Aganagic, T. Ekholm, L. Ng and C. Vafa, “Topological strings, D-model, and knot contact homology”, *Advances in Theoretical and Mathematical Physics*, Vol. 18, No. 4 (2014), pp. 827–956.

[3] L. Amorim, “Open and closed mirror symmetry”, *Boletim da Sociedade Portuguesa de Matemática*, N. 77 (2019), pp. 15–26.

[4] A. Cannas da Silva, “An invitation to symplectic toric manifolds”, *Boletim da Sociedade Portuguesa de Matemática*, N. 77 (2019), pp. 119–132.

[5] L. Diogo, T. Ekholm, “Augmentations, annuli, and Alexander polynomials”, in preparation.

[6] T. Ekholm, J. Etnyre, L. Ng and M. Sullivan, “Knot contact homology”, *Geometry & Topology*, Vol. 17, No. 2 (2013), pp. 975–1112.

[7] T. Ekholm, L. Ng and V. Shende, “A complete knot invariant from contact homology”, *Inventiones mathematicae*, Vol. 211, No. 3 (2018), pp. 1149–1200.

[8] Y. Eliashberg, A. Givental and H. Hofer, “Introduction to symplectic field theory”, *Geometric and Functional Analysis Special Volume, Part II*, 2000, pp. 560–673.

[9] S. Garoufalidis, A. Lauda and T. Lê, “The colored HOMFLYPT function is $q$-holonomic”, *Duke Mathematical Journal*, Vol. 167, No. 3 (2016), pp. 397–447.

[10] M. Gromov, “Pseudo holomorphic curves in symplectic manifolds”, *Inventiones mathematicae*, Vol. 82, No. 2 (1985), pp. 307–347.

[11] M. Hutchings and Y.-J. Lee, “Circle-valued Morse theory and Reidemeister torsion”, *Geometry & Topology*, Vol. 3 (1999), pp. 369–396.

[12] C. Y. Mak and W. Wu, “Dehn twist exact sequences through Lagrangian cobordism”, *Compositio Mathematica*, Vol. 154, No. 12 (2018), pp. 2485–2533.

[13] J. Milnor, “A duality theorem for Reidemeister torsion”, *Annals of Mathematics*, Vol. 76, No. 2 (1962), pp. 137–147.

[14] L. Ng, “Framed knot contact homology”, *Duke Mathematical Journal*, Vol. 141, No. 2 (2008), pp. 365–406.

[15] P. Ozsváth and Z. Szabó, “Holomorphic disks and knot invariants”, *Advances in Mathematics*, Vol. 186, No. 1 (2004), pp. 58–116.

[16] J. Rasmussen, “Floer homology and knot complements”, PhD thesis, Harvard University, USA, 2003.

# Fermat's Last Theorem over Number Fields

*Nuno Freitas*
Universitat de Barcelona
Gran Via de Les Corts Catalanes 585
08007, Barcelona, Spain
e-mail: nunobarroso@ub.edu

**Resumo:** Discutimos a célebre demonstração do Último Teorema de Fermat e as dificuldades que surgem ao tentar aplicar a mesma estratégia de prova sobre corpos de números. Terminamos com uma amostra dos resultados conhecidos no caso de corpos quadráticos.

**Abstract** We overview the celebrated proof of Fermat's Last Theorem and the challenges that arise when trying to carry it over to number fields. We conclude with a sample of the known results for quadratic fields.

**palavras-chave:** Fermat, modularidade, curvas elípticas.

**keywords:** Fermat, modularity, elliptic curves.

## 1 Introduction

The search for a proof of Fermat's Last Theorem (FLT) is one of the richest and more romantic stories in the history of Mathematics. Remarkable progress in number theory as, for example, the origin of what is now algebraic number theory and some incredible breakthroughs in the Langlands program, have come to light due to this pursuit.

**Theorem 1 (FLT)** *The integer solutions to the Fermat equation*

$$x^n + y^n + z^n = 0 \tag{1}$$

*with $n \geq 3$ are trivial, i.e., they satisfy $xyz = 0$.*

The cases $n = 3$ and $n = 4$ of FLT were respectively solved by Euler and Fermat. From this it is easy to see that we only have to prove it for $n = p \geq 5$ a prime. If $(a, b, c)$ is a solution, then by scaling we can suppose that $\gcd(a, b, c) = 1$; we call such a solution *primitive*. The Fermat equation, viewed as defining a curve in $\mathbb{P}^2$, has genus $(p-1)(p-2)/2$, and a celebrated theorem of Faltings tells us that there are only finitely many primitive solutions to (1), for each fixed $n = p$. Despite the efforts of many great mathematicians through 350 years, it was only in 1995 that a complete proof was published. In this paper, we will discuss this modern approach to FLT due to Hellegouarch, Frey, Serre and Ribet which culminated in Wiles' proof [24] and created a new way of tackling Diophantine equations known as *the modular method*.

## 2 The modular method

The proof of FLT is based on three main pillars: Mazur's irreducibility theorem, Wiles' modularity theorem for semistable elliptic curves over $\mathbb{Q}$ and Ribet's level lowering theorem. Explaining these pillars will involve a detour into some of the most fascinating areas of modern number theory: elliptic curves, Galois representations, modular forms and modularity. For a comprehensive introduction to these topics we suggest [3, 21, 22]. For an overview and history of various methods to study the Generalized Fermat equation $x^r + y^q = z^p$ we refer to [1], which we follow closely in this section.

### 2.1 Elliptic curves

Let $K$ be a field. The simplest definition of an *elliptic curve* $E$ over $K$ is: a smooth curve in $\mathbb{P}^2$ given by an equation of the form

$$E \ : \ y^2 z + a_1 xyz + a_3 yz^2 = x^3 + a_2 x^2 z + a_4 xz^2 + a_6 z^3, \tag{2}$$

with $a_1$, $a_2$, $a_3$, $a_4$ and $a_6 \in K$. If the characteristic of $K$ is not 2 or 3, then we can transform to a much simpler model given by the affine equation

$$E \ : \ Y^2 = X^3 + aX + b, \tag{3}$$

where $a$ and $b \in K$, whose disriminant is

$$\Delta_E = -16(4a^3 + 27b^2).$$

We call (3) a *Weierstrass model of* $E$ with *discriminant* $\Delta_E$. The requirement that $E$ is smooth is equivalent to the assumption that $\Delta_E \neq 0$. There is another very important quantity attached to an elliptic curve called the *$j$-invariant* which can be computed from the model (3) by the formula

$$j_E = \frac{(-48a)^3}{\Delta_E}.$$

Note however that $j_E$ is an invariant of the isomorphism class of $E$ over $\overline{K}$, the algebraic closure of $K$, and so independent of the choosen model.

There is a distinguished $K$-point, the 'point at infinity', which we denote by $\infty$. Given a field $L \supseteq K$, the set of $L$-points on $E$ is given by

$$E(L) = \{(x, y) \in L^2 \,:\, y^2 = x^3 + ax + b\} \cup \{\infty\}.$$

It turns out that the set $E(L)$ has the structure of an abelian group with $\infty$ as the identity element. The group structure is easy to describe geometrically: three points $P_1$, $P_2$, $P_3 \in E(L)$ add up to the identity element if and only if there is a line $\ell$ defined over $L$ meeting $E$ in $P_1$, $P_2$, $P_3$ (with multiplicities counted appropriately). The classic Mordell–Weil Theorem states that for a number field $K$ the group $E(K)$ is finitely generated. For a model as in (3), the 2-torsion subgroup $E[2]$ consists of the points with $y = 0$ plus $\infty$. It turns out that the proofs by Euler and Fermat of FLT for $n = 3, 4$ are simply special cases of what are now standard Mordell–Weil group computations, as discussed in [1, Examples 1 and 2].

## 2.2 Modular forms

Let $\mathbb{H} = \{z \in \mathbb{C} \,:\, \mathrm{Im}(z) > 0\}$. Let $k$ and $N$ be positive integers and set

$$\Gamma_0(N) = \left\{ \begin{pmatrix} a & b \\ c & d \end{pmatrix} \in \mathrm{SL}_2(\mathbb{Z}) \,:\, c \equiv 0 \pmod{N} \right\},$$

which is a subgroup of $\mathrm{SL}_2(\mathbb{Z})$ of finite index. The group $\Gamma_0(N)$ acts on $\mathbb{H}$ via fractional linear transformations

$$\begin{pmatrix} a & b \\ c & d \end{pmatrix} \,:\, \mathbb{H} \to \mathbb{H}, \qquad z \mapsto \frac{az + b}{cz + d}.$$

The quotient $Y_0(N) = \Gamma_0(N)\backslash\mathbb{H}$ has the structure of a non-compact Riemann surface. This has a standard compactification denoted $X_0(N)$ and the difference $X_0(N) \setminus Y_0(N)$ is a finite set of points called the *cusps*.

A *modular form $f$ of weight $k$ and level $N$* is a function $f \,:\, \mathbb{H} \to \mathbb{C}$ that satisfies the following conditions:

(i) $f$ is holomorphic on $\mathbb{H}$;

(ii) $f$ satisfies the property

$$f\left(\frac{az + b}{cz + d}\right) = (cz + d)^k f(z), \tag{4}$$

for all $z \in \mathbb{H}$ and $\left(\begin{smallmatrix} a & b \\ c & d \end{smallmatrix}\right) \in \Gamma_0(N)$;

(iii) $f$ extends to a function that is holomorphic at the cusps.

It follows from these properties, and the fact that one of the cusps is the cusp at $i\infty$, that $f$ must have a Fourier expansion

$$f(z) = \sum_{n \geq 0} c_n q^n \quad \text{where} \quad q(z) = \exp(2\pi i z). \tag{5}$$

It turns out that the set of modular forms of weight $k$ and level $N$, denoted by $M_k(N)$, is a finite-dimensional vector space over $\mathbb{C}$. A *cusp form* of weight $k$ and level $N$ is an $f \in M_k(N)$ that vanishes at all the cusps. As $q(i\infty) = 0$ we see in particular that a cusp form must satisfy $c_0 = 0$. The cusp forms naturally form a subspace of $M_k(N)$ which we denote by $S_k(N)$.

There is a natural family of commuting operators $T_n : S_2(N) \to S_2(N)$ (with $n \geq 1$) called the *Hecke operators*. The *eigenforms* of level $N$ are the weight 2 cusp forms that are simultaneous eigenvectors for all the Hecke operators. Such an eigenform is called *normalized* if $c_1 = 1$ and thus its Fourier expansion has the form

$$f = q + \sum_{n \geq 1} c_n q^n.$$

## 2.3 Modularity

Let $E/\mathbb{Q}$ be given by a model (2) where the $a_i \in \mathbb{Z}$, and having (non-zero) discriminant $\Delta_E \in \mathbb{Z}$. Carrying out a suitable linear substitution, we generally work with a *minimal model*: that is one where the $a_i \in \mathbb{Z}$ and with discriminant having the smallest possible absolute value. Associated to $E$ is another, more subtle, invariant called the *conductor* $N_E$, which we shall not define precisely, but we merely point that it is a positive integer sharing the same prime divisors as the minimal discriminant; that it measures the 'bad behavior' of the elliptic curve $E$ modulo primes; and that it can be computed easily through *Tate's algorithm* [22, Chapter IV]. In particular, the primes $p$ not dividing $N_E$ are the primes of *good reduction* while those satisfying $p \| N_E$ are the primes of *multiplicative reduction*.

Now let $p \nmid \Delta_E$ be a prime. Reducing modulo $p$ a minimal equation (2) we obtain an elliptic curve $\tilde{E}$ over $\mathbb{F}_p$. The set $\tilde{E}(\mathbb{F}_p)$ is an abelian group as before, but now necessarily finite, and we denote its order by $\#\tilde{E}(\mathbb{F}_p)$. Let

$$a_p(E) = p + 1 - \#\tilde{E}(\mathbb{F}_p).$$

We are now ready to state a version of the modularity theorem due to Wiles, Breuil, Conrad, Diamond and Taylor [2, 23, 24]. This remarkable theorem was previously known as the Shimura–Taniyama conjecture.

**Theorem 2 (The Modularity Theorem)** *Let $E/\mathbb{Q}$ be an elliptic curve with conductor $N_E$. There exists a normalized eigenform $f = q + \sum c_n q^n$ of weight 2 and level $N_E$ with $c_n \in \mathbb{Z}$ for all $n$, and such that for every prime $p \nmid \Delta_E$ we have $c_p = a_p(E)$.*

For an elliptic curve $E$ and an eigenform $f$ as in this theorem we will also say that $f$ *corresponds to $E$ via modularity.*

## 2.4 Galois representations

Let $E$ be an elliptic curve over $\mathbb{C}$. The structure of the abelian group $E(\mathbb{C})$ is particularly easy to describe. There is a discrete lattice $\Lambda \subset \mathbb{C}$ of rank 2 (that is, as an abelian group $\Lambda \simeq \mathbb{Z}^2$) depending on $E$, and an isomorphism

$$E(\mathbb{C}) \simeq \mathbb{C}/\Lambda. \tag{6}$$

Let $p$ be a prime. By the $p$-torsion of $E(\mathbb{C})$ we mean the subgroup

$$E[p] = \{Q \in E(\mathbb{C}) \ : \ pQ = 0\}.$$

It follows from (6) that

$$E[p] \simeq (\mathbb{Z}/p\mathbb{Z})^2, \tag{7}$$

which can be viewed as 2-dimensional $\mathbb{F}_p$-vector space. Now let $E$ be an elliptic curve over $\mathbb{Q}$. Then we may view $E$ as an elliptic curve over $\mathbb{C}$, and with the above definitions obtain an isomorphism $E[p] \simeq (\mathbb{Z}/p\mathbb{Z})^2$. However, in this setting, the points of $E[p]$ have algebraic coordinates, and are acted on by $G_\mathbb{Q} := \mathrm{Gal}(\overline{\mathbb{Q}}/\mathbb{Q})$, *the absolute Galois group of the rational numbers.* Via the isomorphism (7), the group $G_\mathbb{Q}$ acts on $(\mathbb{Z}/p\mathbb{Z})^2$. Thus we obtain a 2-dimensional representation depending on $E/\mathbb{Q}$ and the prime $p$:

$$\overline{\rho}_{E,p} \ : \ G_\mathbb{Q} \to \mathrm{GL}_2(\mathbb{F}_p). \tag{8}$$

We say that the $\overline{\rho}_{E,p}$ is *reducible* if the matrices of the image $\overline{\rho}_{E,p}(G_\mathbb{Q})$ share some common eigenvector. Otherwise we say that $\overline{\rho}_{E,p}$ is irreducible. We have now given enough definitions to be able to state Mazur's theorem; this is often considered as the first step in the proof of FLT.

**Theorem 3 (Mazur [15])** *Let $E/\mathbb{Q}$ be an elliptic and $p$ a prime.*

*(i) If $p > 163$, then $\overline{\rho}_{E,p}$ is irreducible.*

*(ii) If $E$ has full 2-torsion (that is $E[2] \subseteq E(\mathbb{Q})$), square-free conductor and $p \geq 5$, then $\overline{\rho}_{E,p}$ is irreducible.*

## 2.5 Ribet's level lowering theorem

Let $E/\mathbb{Q}$ be an elliptic curve and associated mod $p$ Galois representation $\overline{\rho}_{E,p} : G_\mathbb{Q} \to \mathrm{GL}_2(\mathbb{F}_p)$ as above. Let $f$ be an eigenform. Deligne and Serre showed that such an $f$ gives rise, for each prime $p$, to a Galois representation $\overline{\rho}_{f,p} : G_\mathbb{Q} \to \mathrm{GL}_2(\mathbb{F}_{p^r})$, where $r \geq 1$ depends on $f$. If $E$ corresponds to $f$ via the Modularity Theorem, then $\overline{\rho}_{E,p} \sim \overline{\rho}_{f,p}$ (the two representations are isomorphic). Thus the representation $\overline{\rho}_{E,p}$ is *modular* in the sense that it arises from a modular eigenform. Recall also from the Modularity Theorem that, if $f$ corresponds to $E$ via modularity, then the conductor of $E$ is equal to the level of $f$. Sometimes it is possible to replace $f$ by another eigenform of smaller level which has the same mod $p$ representation. This process is called *level lowering*. We now state a special case of Ribet's level lowering theorem. For a prime $\ell$, we let $\upsilon_\ell(x)$ denote the $\ell$-adic valuation of $x \in \mathbb{Q}$.

**Theorem 4 (Ribet's level lowering theorem [18])** *Let $E/\mathbb{Q}$ be an elliptic curve with minimal discriminant $\Delta$ and conductor $N$. Let $p \geq 3$ be prime. Suppose that (i) the curve $E$ is modular and (ii) the mod $p$ representation $\overline{\rho}_{E,p}$ is irreducible. Let*

$$N_p = \frac{N}{M_p}, \quad \textit{where} \quad M_p = \prod_{\substack{\ell || N, \\ p \,|\, \upsilon_\ell(\Delta)}} \ell. \tag{9}$$

*Then $\overline{\rho}_{E,p} \sim \overline{\rho}_{g,p}$ for some eigenform $g$ of weight 2 and level $N_p$.*

We now know, by the Modularity Theorem that all elliptic curves over $\mathbb{Q}$ are modular, so condition (i) in Ribet's theorem is automatically satisfied. We include it here both for historical interest but also because analogous level lowering results are available over other fields and modularity of all elliptic curves is still an open question over general fields.

## 2.6 The proof of Fermat's Last Theorem

Suppose $p \geq 5$ is prime, and $a$, $b$ and $c$ are non-zero pairwise coprime integers satisfying (1) with $n = p$. We reorder $(a, b, c)$ so that

$$b \equiv 0 \pmod 2 \quad \text{and} \quad a^p \equiv -1 \pmod 4. \tag{10}$$

We consider the *Frey–Hellegouarch curve* which depends on $(a, b, c)$:

$$E \;:\; Y^2 = X(X - a^p)(X + b^p) \tag{11}$$

whose minimal discriminant and conductor are:

$$\Delta = \frac{a^{2p}b^{2p}c^{2p}}{2^8}, \qquad N = \prod_{\ell \mid \Delta} \ell.$$

Note that the conductor is square-free; conditions (10) ensure that $2 \mid\mid N$. The 2-torsion subgroup of $E$ is $E[2] = \{\infty,\ (0,0),\ (a^p,0),\ (-b^p,0)\} \subset E(\mathbb{Q})$. As $p \geq 5$, we know by part (ii) of Mazur's irreducibility theorem that $\overline{\rho}_{E,p}$ is irreducible. Moreover, $E$ is modular by the Modularity Theorem[1], and so the hypotheses of Ribet's theorem are satisfied. We compute $N_p = 2$ using the recipe in (9). It follows that $\overline{\rho}_{E,p} \sim \overline{\rho}_{g,p}$, where $g$ has weight 2 and level 2. But there are no eigenforms of weight 2 and level 2, a contradiction.

#### 2.6.1 Some Historical Remarks

In the early 1970s, Hellegouarch had the idea of associating to a non-trivial solution of the Fermat equation the elliptic curve (11); he noted that the number field generated by its $p$-torsion subgroup $E[p]$ has surprisingly little ramification. In the early 1980s, Frey observed that this elliptic curve enjoys certain remarkable properties that should rule out its modularity. Motivated by this, in 1985 Serre made precise his modularity conjecture and showed that it implies Fermat's Last Theorem. Serre's remarkable paper [20] also uses several variants of the Frey–Hellegouarch curve to link modularity to other Diophantine problems. Ribet announced his level-lowering theorem 1987, showing that modularity of the Frey–Hellegouarch curve implies FLT.

## 3 Fermat's Last Theorem over Number Fields

### 3.1 Historical background

Interest in the Fermat equation over various number fields goes back to the 19th and early 20th Century. For example, Dickson's *History of the Theory of Numbers* [4, pages 758 and 768] mentions extensions by Maillet (1897) and Furtwängler (1910) of classical ideas of Kummer to the Fermat equation $x^p + y^p = z^p$ ($p > 3$ prime) over the cyclotomic field $\mathbb{Q}(\zeta_p)$. However, the elementary, cyclotomic and Mordell–Weil approaches to the Fermat equation have had limited success. Indeed, even over $\mathbb{Q}$, no combination of these

[1]In fact, we only need modularity of *semistable* elliptic curves over $\mathbb{Q}$, i.e. those with square-free conductor, which was the original modularity result proved by Wiles.

approaches is known to yield a proof of FLT for infinitely many prime exponents $p$. It is therefore natural to attempt to carry Wiles' proof over to general number fields. The first work in this direction is due to Jarvis and Meekin [12] who showed FLT holds for the field $\mathbb{Q}(\sqrt{2})$. They further analyzed the situation over other real quadratic fields to conclude that

> *"...the numerology required to generalise the work of Ribet and Wiles directly continues to hold for $\mathbb{Q}(\sqrt{2})$... there are no other real quadratic fields for which this is true ..."*

### 3.2 The asymptotic Fermat's conjecture

Let $K$ be a number field and $\mathcal{O}_K$ its ring of integers. By the *Fermat equation with exponent $p$ over $K$* we mean

$$x^p + y^p + z^p = 0, \qquad x, y, z \in \mathcal{O}_K. \tag{12}$$

A solution $(a, b, c)$ of (12) is called *trivial* if $abc = 0$, otherwise *non-trivial*. Clearly, over any $K$ there are trivial solutions, such as $(1, -1, 0)$, but sometimes more, for example,

$$(18 + 17\sqrt{2})^3 + (18 - 17\sqrt{2})^3 = 42^3,$$

$$(1 + \sqrt{-7})^4 + (1 - \sqrt{-7})^4 = 2^4,$$

showing that the exact same statement as of FLT does not hold over $\mathbb{Q}(\sqrt{2})$ or $\mathbb{Q}(\sqrt{-7})$. Instead it makes sense to consider the question only for large enough exponents. More precisely, we will say that the **asymptotic Fermat's Last Theorem over $K$ holds** if there is some bound $B_K$ such that for prime $p > B_K$, all solutions to the Fermat equation (12) are trivial.

Now let $K = \mathbb{Q}(\sqrt{-3})$ and consider the element $\omega = \frac{\sqrt{-3}-1}{2}$. We have that $\omega^3 = 1$ and it is easy to see that, for all primes $p \geq 5$, the equality

$$\omega^p + (\omega^2)^p + 1^p = 0,$$

holds, hence the asymptotic FLT does not hold over $\mathbb{Q}(\sqrt{-3})$.

**Conjecture 1 (Asymptotic Fermat's Conjecture)** *Let $K$ be a number field. If $\omega \notin K$ then the asymptotic FLT over $K$ holds.*

## 4 The modular method over totally real fields

We restrict ourselves to *totally real fields*, i.e., number fields such that all embeddings into $\mathbb{C}$ have image in $\mathbb{R}$. This is a natural restriction, because modularity related objects and questions are very poorly understood for fields with at least one complex embedding, consequently all results about FLT for such fields are conditional on two deep conjectures of the Langlands program (see [19] for details). In contrast, for a totally real field $K$ there is a well established theory of *Hilbert modular forms* which are the natural replacement for the modular forms over $\mathbb{Q}$; it is not our objective to discuss details of this theory here. The only thing to keep in mind is that they satisfy the analogous properties over $K$ to those described in §2.2 and that modularity of elliptic curves over $K$ can be defined by a correspondence with Hilbert eigenforms, similar to the discussion in §2.3.

In particular, since $K$ is totally real, we have $\omega \notin K$ and we expect the Asymptotic Fermat Conjecture to hold for $K$. To properly discuss the challenges we face it helps to break the method into the following steps:

1. **Constructing a Frey curve.** Attach a Frey elliptic curve $E/K$ to a putative solution of (12).

2. **Modularity.** Prove modularity of $E/K$.

3. **Irreducibility.** Prove irreducibility of $\overline{\rho}_{E,p}$, the mod $p$ Galois representation attached to $E$.

4. **Level lowering.** Conclude that $\overline{\rho}_{E,p} \sim \bar{\rho}_{\mathfrak{f},\mathfrak{p}}$ where $\mathfrak{f}$ is a Hilbert eigenform over $K$ of (parallel) weight 2 and level among finitely many possibilities $N_i$. Here, $\bar{\rho}_{\mathfrak{f},\mathfrak{p}}$ denotes the mod $\mathfrak{p}$ Galois representation attached to $\mathfrak{f}$ for some $\mathfrak{p} \mid p$ in the field of coefficients $\mathbb{Q}_{\mathfrak{f}}$ of $\mathfrak{f}$.

5. **Contradiction.** Compute all the eigenforms $\mathfrak{f}$ predicted in Step 4 and show that $\overline{\rho}_{E,p} \nsim \bar{\rho}_{\mathfrak{f},\mathfrak{p}}$ for all of them.

Suppose that $a, b, c \in \mathcal{O}_K$ is a solution to (12) such that $abc \neq 0$. Since (12) is equation (1) over $K$ we consider in step 1 the classical Frey curve over $K$:

$$E_{a,b,c} \; : \; Y^2 = X(X - a^p)(X + b^p). \tag{13}$$

### 4.1 The case of $\mathbb{Q}(\sqrt{2})$

Recall that steps 2–4 in the proof of FLT over $\mathbb{Q}$ are covered respectively by the three remarkable theorems of Wiles, Mazur and Ribet. However, at the time of writing [12] only step 4 was known to hold over a general $K$ (due to the combined work of Jarvis, Rajae and Fijuwara [9, 10, 17]). To complete the modularity step Jarvis and Meekin showed that, under some non-restrictive assumptions on $a, b, c$ (analogous to (10)), the curve $E_{a,b,c}$ is semistable and its modularity followed by a result of Jarvis–Manoharmayum [11] stating that all semistable elliptic curves over $\mathbb{Q}(\sqrt{2})$ are modular. For the irreducibility part they applied a criterion of Kraus [14] for $p \geq 17$ and, finally, a contradiction follows because after completing step 4 there are again no eigenforms.

### 4.2 The contradiction step

By looking at the proofs of FLT over $\mathbb{Q}$ and $\mathbb{Q}(\sqrt{2})$, the reader may wonder why is there a step 5, as the contradiction happens automatically. It turns out these are the only cases where such convenient coincidence occurs. For example, if the class number of $K$ is $> 1$, we cannot assume coprimality of $a$, $b$, $c$, and the Frey curve will not be semistable, consequently the levels obtained after level lowering will have larger norms and, in general, there are eigenforms at these levels. In fact, step 5 is nowadays the most difficult part when applying the modular method to solve (12) or any other Diophantine equation assuming, of course, that an associated Frey curve exists (i.e. step 1 can be done); unfortunately, there are only a few Diophantine equations known to have attached Frey curves.

### 4.3 Modularity and Irreducibility

Although the modularity of the Frey curve was the hardest step in the proof of FLT, nowadays we know it holds in full generality due to a result of Freitas–Le Hung–Siksek [6].

**Theorem 5 (F.–Le Hung–Siksek)** *Let $K$ be a totally real field. Up to isomorphism over $\overline{K}$, there are at most finitely many non-modular elliptic curves $E$ over $K$.*

*Moreover, if $K$ is quadratic, then all elliptic curves over $K$ are modular.*

Furthermore, in the recent work of Derickx–Najman–Siksek [5], modularity of elliptic curves was extended to the case of totally real cubic fields.

**Theorem 6 (Derickx–Najman–Siksek)** *All elliptic curves over totally real cubic fields are modular.*

These modularity results have the following important consequence.

**Corollary 4.1** *Let $K$ be a totally real field. There is some constant $A_K$, depending only on $K$, such that for any non-trivial solution $(a, b, c)$ of the Fermat equation (12) with prime exponent $p > A_K$, the Frey curve $E_{a,b,c}$ given by (13) is modular.*

*Moreover, if $K$ is quadratic or cubic then $A_K = 0$.*

There is no irreducibility result for $\bar{\rho}_{E,p}$ over $K$ analogous to Mazur's theorem. Instead, we can derive the following result from the works of David and Momose, who build on Merel's Uniform Boundedness Theorem [16].

**Theorem 7** *Let $K$ be a totally real field. There is a constant $C_K$, depending only on $K$, such that the following holds. If $p > C_K$ is prime, and $E$ is an elliptic curve over $K$ with either good or multiplicative reduction at all $\mathfrak{q} \mid p$, then $\overline{\rho}_{E,p}$ is irreducible.*

## 4.4 A refined level lowering

The next step in the strategy is level lowering which is known to hold for general $K$ due to the combined work of Fujiwara, Jarvis and Rajaei. As explained above, after applying level lowering, we will not obtain a contradiction due to the presence of eigenforms. Instead, the idea is to use finer properties of the Frey curve $E_{a,b,c}$, to show that many of the eigenforms are not a real obstruction.

Before proceeding, it is helpful here to make a comparison with the equation $x^p + y^p + L^\alpha z^p = 0$ over $\mathbb{Q}$, with $L$ an odd prime and $\alpha$ a positive integer, considered by Serre and Mazur [20, p. 204]. A non-trivial solution to this latter equation gives rise, via modularity and level lowering, to a classical weight 2 newform $f$ of level $2L$; for $L \geq 13$ there are such eigenforms and we face the same difficulty. Mazur however shows that if $p$ is sufficiently large then $f$ corresponds to an elliptic curve $E'$ with full 2-torsion and conductor $2L$, and by classifying such elliptic curves concludes that $L$ is either a Fermat or a Mersenne prime. To be able to transfer and refine Mazur's argument to our setting, we need the following conjecture, which is the opposite direction to modularity and generalizes the Eichler–Shimura Theorem over $\mathbb{Q}$.

**Conjecture 2 ("Eichler–Shimura")** *Let $K$ be a totally real field. Let $\mathfrak{f}$ be a Hilbert newform of level $\mathcal{N}$ and parallel weight 2, and rational field of coefficients. Then there is an elliptic curve $E_\mathfrak{f}/K$ with conductor $\mathcal{N}$ having the same* L*-function as $\mathfrak{f}$.*

We will also need some more notation. For $K$ a totally real field, an element $x \in K$ and a prime ideal $\mathfrak{q}$ in $\mathcal{O}_K$ we write $\upsilon_{\mathfrak{q}}(x)$ to denote a $\mathfrak{q}$-adic valuation of $x$. Moreover, let

$$\begin{aligned} S &= \{\mathfrak{P} \;:\; \mathfrak{P} \text{ is a prime ideal of } \mathcal{O}_K \text{ dividing } 2\mathcal{O}_K\}, \\ T &= \{\mathfrak{P} \in S \;:\; \mathcal{O}_K/\mathfrak{P} = \mathbb{F}_2\}, \qquad U = \{\mathfrak{P} \in S \;:\; 3 \nmid \upsilon_{\mathfrak{P}}(2)\}. \end{aligned} \tag{14}$$

We choose a set $\mathcal{H}$ of prime ideals $\mathfrak{m} \notin S$ representing the elements in the class group of $K$. We also need an assumption, which we refer to as **(ES)**:

$$\textbf{(ES)} \qquad \begin{cases} \text{either } [K:\mathbb{Q}] \text{ is odd;} \\ \text{or } T \neq \emptyset; \\ \text{or Conjecture 2 holds for } K. \end{cases}$$

For a non-trivial solution $(a, b, c)$ to the Fermat equation (12), let

$$\mathcal{G}_{a,b,c} := a\mathcal{O}_K + b\mathcal{O}_K + c\mathcal{O}_K. \tag{15}$$

Now Mazur's argument adapted to our setting gives the following result.

**Theorem 8** *Let $K$ be a totally real field satisfying* **(ES)**. *There is a constant $B_K$, depending only on $K$, such that the following holds. Let $(a, b, c)$ be a non-trivial solution to (12) with prime exponent $p > B_K$, and rescale $(a, b, c)$ so that it remains integral and satisfies $\mathcal{G}_{a,b,c} = \mathfrak{m}$ for some $\mathfrak{m} \in \mathcal{H}$. Write $E$ for the Frey curve (13). Then there is an elliptic curve $E'$ over $K$ such that*

*(i) the conductor of $E'$ is divisible only by primes in $S \cup \{\mathfrak{m}\}$;*

*(ii) $\#E'(K)[2] = 4$;*

*(iii) $\overline{\rho}_{E,p} \sim \overline{\rho}_{E',p}$;*

*Write $j'$ for the $j$-invariant of $E'$. Then,*

*(a) for $\mathfrak{P} \in T$, we have $\upsilon_{\mathfrak{P}}(j') < 0$;*

*(b) for $\mathfrak{P} \in U$, we have either $\upsilon_{\mathfrak{P}}(j') < 0$ or $3 \nmid \upsilon_{\mathfrak{P}}(j')$;*

*(c) for $\mathfrak{q} \notin S$, we have $\upsilon_{\mathfrak{q}}(j') \geq 0$.*

*In particular, $E'$ has potentially good reduction away from $S$.*

# 5 Results over totally real fields

## 5.1 $S$-unit equations

Assuming **(ES)**, Theorem 8 implies that a non-trivial solution to the Fermat equation over $K$ with sufficiently large exponent $p$ yields an elliptic curve $E'/K$ with full 2-torsion and potentially good reduction away from the set $S$ of primes above 2. There are such elliptic curves over every $K$, for example the curve $Y^2 = X^3 - X$, and so we still do not get a simple contradiction. Note however that the latter elliptic curve does not satisfy conclusion (a) of Theorem 8 when the field $K$ is such that $T \neq \emptyset$, hence it is not an obstruction in that case. An element $x \in K$ is called an *S-unit* if $\upsilon_{\mathfrak{q}}(x) = 0$ for all $\mathfrak{q} \notin S$. Using the fact that elliptic curves with full 2-torsion and good reduction away from $S$ are classified by solutions to $S$-unit equations, we have the following result that describes when there are no $E'/K$ as in Theorem 8, and so no obstruction to the desired contradiction.

**Theorem 9 (F.–Siksek)** *Let $K$ be a totally real field satisfying (ES). Let $S$, $T$ and $U$ be as in (14). Write $\mathcal{O}_S^*$ for the group of $S$-units of $K$. Suppose that for every solution $(\lambda, \mu)$ to the $S$-unit equation*

$$\lambda + \mu = 1, \qquad \lambda,\, \mu \in \mathcal{O}_S^* \tag{16}$$

*there is*

*(A) either some $\mathfrak{P} \in T$ that satisfies* $\max\{|\upsilon_{\mathfrak{P}}(\lambda)|, |\upsilon_{\mathfrak{P}}(\mu)|\} \leq 4\,\upsilon_{\mathfrak{P}}(2)$*,*

*(B) or some $\mathfrak{P} \in U$ that satisfies both* $\max\{|\upsilon_{\mathfrak{P}}(\lambda)|, |\upsilon_{\mathfrak{P}}(\mu)|\} \leq 4\,\upsilon_{\mathfrak{P}}(2)$*, and* $\upsilon_{\mathfrak{P}}(\lambda\mu) \equiv \upsilon_{\mathfrak{P}}(2) \pmod 3$.

*Then the asymptotic Fermat's Last Theorem holds over $K$.*

For all fields $K$, equation (16) has solutions in $\mathbb{Q} \cap \mathcal{O}_S^*$, namely $(\lambda, \mu) = (2, -1), (-1, 2), (1/2, 1/2)$ which correspond to the elliptic curve $Y^2 = X^3 - X$, however these solutions satisfy (A) if $T \neq \emptyset$ and (B) if $U \neq \emptyset$.

## 5.2 The quadratic case

In view of Theorem 9, we have to solve the $S$-unit equation (16) and test the solutions in order to decide whether the asymptotic FLT holds over $K$. There are algorithms that, in principle, could do that for each particular $K$, but what is more interesting is to show that asymptotic FLT holds for infinite families of fields. For this we need to control the solutions to (16) over varying fields. This can be very hard but, in the case of real quadratic fields, we achieved considerable success, as illustrated by the following series of theorems taken from the joint works with Siksek [7, 8].

**Theorem 10** *Let $d \geq 2$ be square-free, such that $d \equiv 6, 10 \pmod{16}$ or $d \equiv 3 \pmod 8$. Then the asymptotic FLT holds over $\mathbb{Q}(\sqrt{d})$.*

*Moreover, for $d > 5$ satisfying $d \equiv 5 \pmod 8$, the same is true assuming that Conjecture 2 holds over $\mathbb{Q}(\sqrt{d})$.*

There are other explicit congruence conditions on $d$ for which asymptotic FLT is known to hold over $\mathbb{Q}(\sqrt{d})$ (see [7, Theorem 1]) and, moreover, there are also real quadratic fields $\mathbb{Q}(\sqrt{d})$ not given by a congruence condition on $d$ for which asymptotic FLT holds. We have the following density theorem.

**Theorem 11 (F.–Siksek)** *The asymptotic FLT holds for a set of real quadratic fields of density $5/6$. Assuming Conjecture 2 this density becomes 1.*

The following result shows that it is possible to optimize $B_K$ by making $K$ concrete. In this case the proof does not pass through Theorem 9, but instead one needs to optimize the exponent bound at every step of the strategy, which raises other challenges not discussed here.

**Theorem 12 (F.–Siksek)** *Let $3 \leq d \leq 23$ be square-free and $d \neq 5, 17$ or $d = 79$. Then, all solutions $(a, b, c)$ to the equation*

$$x^p + y^p + z^p = 0, \qquad a, b, c \in \mathbb{Q}(\sqrt{d}) \qquad p \geq 5 \text{ prime}$$

*satisfy $abc = 0$. Moreover, the same is true over $\mathbb{Q}(\sqrt{17})$ for half the exponents, more precisely, for all primes $p \geq 5$ such that $p \equiv 3, 5 \pmod 8$.*

### 5.2.1 FLT over $\mathbb{Q}(\sqrt{5})$

Note that the field $\mathbb{Q}(\sqrt{5})$ is not covered by any of the theorems above and indeed asymptotic FLT over $\mathbb{Q}(\sqrt{5})$ seems to be a very hard open problem. The main reason being that the modular method does not see the difference between the Fermat equation (12) and its variant with unit coefficients

$$\frac{1+\sqrt{5}}{2}x^p + \frac{1-\sqrt{5}}{2}y^p + z^p = 0$$

which has a solution $(1, 1, -1)$. Nevertheless, the following result gives evidence that FLT should be true over $\mathbb{Q}(\sqrt{5})$, as predicted by Conjecture 1.

**Theorem 13 (Kraus [13])** *Let $K = \mathbb{Q}(\sqrt{5})$ and $p < 10^7$ be a prime. Then the Fermat equation with exponent $p$ over $K$ has only the trivial solutions.*

## References

[1] M. A. Bennett, P. Mihăilescu and S. Siksek *The Generalized Fermat Equation*, pages 173–205 of Open Problems in Mathematics (J. F. Nash, Jr. and M. Th. Rassias eds), Springer, New York, 2016.

[2] C. Breuil, B. Conrad, F. Diamond, R. Taylor, *On the modularity of elliptic curves over $\mathbb{Q}$: wild 3-adic exercises*, Journal of the American Mathematical Society **14** (2001), 843–939.

[3] F. Diamond and J. Shurman, *A First Course on Modular Forms*, GTM **228**, Springer, 2005.

[4] L. E. Dickson, *History of the Theory of Numbers, Vol. II*, Chelsea, New York, 1971.

[5] M. Derickx, F. Najman and S. Siksek, *Elliptic curves over real cubic fields are modular*, preprint.

[6] N. Freitas, B. V. Le Hung and S. Siksek, *Elliptic curves over real quadratic fields are modular*, Inventiones Mathematicae **201** (2015), 159–206.

[7] N. Freitas and S. Siksek, *The asymptotic Fermat's last theorem for five-sixths of real quadratic fields*, Compos. Math. **151** (2015), no. 8, 1395–1415.

[8] N. Freitas and S. Siksek, *Fermat's Last Theorem over some small real quadratic fields*, Algebra & Number Theory **9** (2015), no. 4, 875–895.

[9] K. Fujiwara, *Level optimisation in the totally real case*, `arXiv:0602586v1`, 27 February 2006.

[10] F. Jarvis *Correspondences on Shimura curves and Mazur's principle at p*, Pacific J. Math., **213** (2), 2004, 267–280.

[11] F. Jarvis and J. Manoharmayum, *On the modularity of supersingular elliptic curves over certain totally real number fields*, Journal of Number Theory **128** (2008), no. 3, 589–618.

[12] F. Jarvis and P. Meekin, *The Fermat equation over $\mathbb{Q}(\sqrt{2})$*, J. Number Theory **109** (2004), 182–196.

[13] A. Kraus, *Sur le théorème de Fermat sur $\mathbb{Q}(\sqrt{5})$*, Annales Mathématiques du Québec 39.1 (2015), 49–59.

[14] A. Kraus, *Courbes elliptiques semi-stables et corps quadratiques*, J. of number theory **60** (1996), 245–253

[15] B. Mazur, *Rational isogenies of prime degree*, Inventiones Math. **44** (1978), 129–162.

[16] L. Merel, *Bornes pour la torsion des courbes elliptiques sur les corps de nombres*, Invent. Math. **124** (1996), 437–449.

[17] A. Rajaei, *On the levels of mod $\ell$ Hilbert modular forms*, J. reine angew. Math. **537** (2001), 33–65.

[18] K. A. Ribet, *On modular representations of* $\mathrm{Gal}(\overline{\mathbb{Q}}/\mathbb{Q})$ *arising from modular forms*, Invent. Math. **100** (1990), 431–476.

[19] M. H. Şengün and S. Siksek, *On the asymptotic Fermat's Last Theorem over numbers fields*, Commentarii Mathematici Helvetici **93** (2018), 359–375.

[20] J.-P. Serre, *Sur les représentations modulaires de degré* 2 *de* $\mathrm{Gal}(\overline{\mathbb{Q}}/\mathbb{Q})$, Duke Math. J. **54** (1987), 179–230.

[21] J. H. Silverman, *The Arithmetic of Elliptic Curves*, GTM **106**, Springer, 1986.

[22] J. H. Silverman, *Advanced Topics in the Arithmetic of Elliptic Curves*, GTM **151**, Springer, 1994.

[23] R. Taylor and A. Wiles, *Ring-theoretic properties of certain Hecke algebras*, Ann. of Math. **141** (1995), 553–572.

[24] A. Wiles, *Modular elliptic curves and Fermat's Last Theorem*, Ann. of Math. **141** (1995), 443–551.

# Concise notes on Special Holonomy with an emphasis on Calabi–Yau and $G_2$-Manifolds

*Gonçalo Oliveira*
Universidade Federal Fluminense IME–GMA
Niterói, Brazil
e-mail: galato97@gmail.com
sites.google.com/view/goncalo-oliveira-math-webpage/home

**Abstract:** These are notes for a very short introduction to some selected topics on special Riemannian holonomy with a focus on Calabi-Yau and $G_2$-manifolds. No material in these notes is original and more on it can be found in the papers/books of Bryant, Hitchin, Joyce and Salamon referenced during the text.

**keywords:** Special Holonomy, Calabi-Yau, $G_2$-manifold.

## 1 Introduction

Riemannian geometry is by now a well established and fundamental area of mathematics with most undergraduate degrees worldwide having an introductory course on it, such as one on curves and surfaces. Despite this there is still nothing like a classification of complete Riemannian manifolds and instead one attempts to understand them from secondary invariants such as their holonomy.

Given a n-dimensional Riemannian manifold $(X, g)$ its Levi-Civita connection yields a notion of parallel transport of tangent vectors along paths. This has the property that it preserves the length and angles between parallel transported vectors. When one fixes a point $p$ and a loop $\gamma_p$ based at that point, the parallel transport along $\gamma_p$ is an orthogonal linear transformation $\gamma_p : T_pX \to T_pX$ of the tangent space $T_pX$ to $X$ at $p$. The set of all such linear transformations $\mathrm{Hol}_p(X)$ is a subgroup of the group the orthogonal group $O(T_pX)$ called the holonomy group at $p$. If one fixes an orthogonal basis of $T_pX$, this may viewed as a subgroup of $O(n)$ which changes by conjugation upon changing the base point $p$. Thus, from now on we shall forget about the base point in the notation and simply refer to the holonomy group as $\mathrm{Hol}(X)$ which we think of as a conjugacy class in $O(n)$.

The classification of possible Riemannian holonomy groups was started by Cartan's algebraic classification of symmetric spaces [7, 8] in 1926. In the nosymmetric case one may, by a theorem of de Rham, restrict to the class of Riemannian manifolds for which the holonomy representation is irreducible, which are thus known as irreducible Riemannian manifolds. In 1953 Berger [2] compiled a set of restrictions which may be satisfied by any possible holonomy group of a simply connected, irreducible Riemannian manifold. The outcome is a list of these possible holonomy groups of these Riemannian manifolds. It is headed by $SO(n)$ which represents the generic holonomy group, and followed by some "rarer" subgroups of $SO(n)$ still acting on $\mathbb{R}^n$ in an irreducible manner. The full list is the following:

| Hol | n=dim(X) | Name |
|---|---|---|
| SO(n) | n | Orientable manifold |
| U(k) | 2k | Kähler manifold |
| SU(k) | 2k | Calabi–Yau manifold |
| Sp(k)·Sp(1) | 4k | Quaternion-Kähler manifold |
| Sp(k) | 4k | Hyperkähler manifold |
| G2 | 7 | $G_2$-manifold |
| Spin(7) | 8 | Spin(7) manifold |

These other possible holonomy groups are known as special holonomy groups and except for $G_2$ and Spin(7) they all appear in infinite families. For this reason $G_2$ and Spin(7) are also called as the exceptional holonomy groups.

Berger's technique to cut the list down to only these groups is quite indirect and consists in transforming what is apparently an integro-differential problem of computing all the holonomies round loops into a local differential problem. The idea is to instead, classify the Lie algebra of the possible Riemannian holonomy groups which by the Ambrose-Singer theorem can be obtained from the values of the Riemann curvature tensor. Its symmetries give restrictions on the possible Lie algebras and these are then integrated by a unique simply connected Lie group. Clearly, this approach solely puts restrictions on the possible holonomy groups and, at the time Berger's list appeared, it was not known whether all groups featuring it could actually be realized as Riemannian holonomy groups. Nowadays, due to the efforts of Aubin, Bryant, Calabi, Salamon and Yau together with several contributions from many others [4, 5, 20] we know that all these groups can actually be realized as the holonomy groups of complete Riemannian metrics. However, most intricacies of their geometry and internal classification remain to be understood at present yielding one of most active areas of research in Riemannian geometry.

In a somewhat perpendicular direction several of these geometries have also appeared in the physics literature. Since the 1990's, and also more recently, Calabi–Yau and $G_2$-manifolds have been attracting the interest of physicists working in string and M-theory respectively. The main reason for this is the possibility of using them in compactifications of these theories which are supposed to produce realistic 4-dimensional versions of the physical world including the standard model of particle physics together with a quantization of gravity.

These notes are a selected part of topics that are supposed to serve as a modern, very quick, introduction to both these classes of manifolds from a geometric structure point of view. In this setting, calibrations and stable forms appear naturally and we use these in our approach to both these classes of special holonomy Riemannian manifolds. This approach mixes the points of view of Salamon, Harvey–Lawson and Hitchin which I find very beautiful attractive. In trying to make the material as concise as possible I have left a lot of relevant material out. This can be found in the references given and finish this introduction by admitting that, perhaps, the best contribution of this note is its brevity and mixed viewpoint.

**Acknowledgments**

I would like to thank the very helpful comments of an anonymous referee which highly improved the readability of this text.
The author was supported by Fundação Serrapilheira 1812-27395, by CNPq grants 428959/2018-0 and 307475/2018-2, and FAPERJ through the program Jovem Cientista do Nosso Estado E-26/202.793/2019.

# 2 Calibrated Geometry and Holonomy

In these notes $X^n$ will denote a smooth real $n$-dimensional manifold and $Fr(M)$ its principal $GL(n,\mathbb{R})$-frame bundle. When $X$ is equipped with a Riemannian metric $g$ we will denote by $FO(n)$ its principal $O(n)$-bundle.

## 2.1 Geometric Structures

**Definition 1.** *Let $G \subset GL(n,\mathbb{R})$ be a Lie group, a $G$-structure on $X$, denoted by $P$, is a principal $G$-subbundle of $Fr(X)$.*

**Proposition 1** (weak Holonomy Principle)**.** *There is a one to one correspondence between sections of the bundle $Fr(X) \times_{GL(n,\mathbb{R})} GL(n,\mathbb{R})/G$ and $G$-structures on $X$.*

*Proof.* We shall only sketch the idea, for a full proof see page 11 in [18].
Let $x \in X$, then each point in the fibre $P_x$ gives an identification $T_xX \cong V := \mathbb{R}^n$. If $\eta_0 \in V^{\otimes r} \otimes (V^*)^{\otimes s}$ is $G$-invariant, we can define $\eta_x \in (T_xX)^{\otimes r} \otimes (T_x^*X)^{\otimes s}$ to equal $\eta_0$ using any of the identifications $T_xX \cong V$ given by the points of $P_x$. This gives a well defined tensor $\eta$ over the whole $X$.
Conversely, if $\eta$ is a section of the bundle $Fr(X) \times_{GL(n,\mathbb{R})} GL(n,\mathbb{R})/G$, then one can define the $G$-structure $P$ which stabilizes $\eta$. □

**Example 1.** *1. A Riemannian metric defines the $O(n)$-structure, denoted $FO(n)$.*

*2. An almost complex structure defines a $GL(n/2,\mathbb{C})$-structure.*

When $G \subset O(n)$ and $P$ is a $G$-structure one defines the $O(n)$-bundle $FO(n) = P \times_G O(n)$. A connection $\nabla$ on $P$ induces one on $TX$ whose torsion $T_\nabla \in \Omega^2(X,TX)$ is by definition

$$T_\nabla(V,W) = \nabla_V W - \nabla_W V - [V,W].$$

Given any two connections $\nabla, \nabla'$ as above, $\nabla' = \nabla + a$ with $a \in \Omega^1(X,\mathfrak{g}_P)$ where $\mathfrak{g}_P = P \times_G \mathfrak{g} \subset \mathfrak{so}(TX)$. Then it is easy to compute that $T_{\nabla'} = T_\nabla + \delta(a)$, where $\delta$ is a section of $\mathrm{Hom}(T^*X \otimes \mathfrak{g}_P, \Lambda^2 X \otimes TX)$.
Notice that since $\mathfrak{g} \subset \mathfrak{so}(\mathfrak{n}) \cong \Lambda^2\mathbb{R}^n$, the map $\delta$ is injective and in order to get rid of the dependence on the connection we can define the reduced map $[T_\nabla]$ with values in $\mathrm{coker}(\delta)$. This is usually called the intrinsic torsion (or the structure function of the $G$-structure $P$). The following result is an immediate consequence of this construction.

**Lemma 1.** *Let $(X,g)$ be a Riemannian manifold and $G \subset O(n)$ a $G$-structure $P \subset FO(n)$. Then, there is a connection $\nabla$ on $P$ inducing the Levi–Civita connection of $g$ on $M$ if and only if the reduced map $[T_\nabla]$ vanishes. Such a $G$-structure is said to be integrable.*

An immediate corollary of this construction is the next result, for which more details can be found in page 14 of [18].

**Corollary 1** (Holonomy Principle)**.** *Let $(X,g)$ be a Riemannian manifold and $x \in X$. Then, any $\eta_x \in \Omega^0(X,(T_xX)^{\otimes r} \otimes (T_x^*X)^{\otimes s})$ which is preserved by the holonomy at $x$ of the Levi–Civita connection $\nabla^{LC}$ is the value at $x$ of a $\nabla^{LC}$-parallel tensor field $\eta$.*
*Moreover, in this situation the $G = Hol$-structure $P$ determined by $\eta$ via the weak holonomy principle is equipped with a connection $\nabla$ inducing $\nabla^{LC}$. Equivalently, there is a $\nabla^{LC}$-parallel embedding of $P$ into $FO(n)$.*

**Remark 1.** *In general, a similar principle holds for any vector bundle with a connection.*

Let $G \subset SO(n)$ and $P$ a $G$-structure on $(X,g)$, then $G$ acts on the differential forms and splits these as irreducible representations as $\Lambda^k = \oplus_i \Lambda^k_i$. Moreover, the Hodge-$*$ is an isomorphism of $G$-representations $\Lambda^k_i \cong \Lambda^{n-k}_i$. These are the essential observations leading to the following Theorem of Chern, [9].

**Theorem 1.** *Let $P$ be a $G$-structure on $(X,g)$ as above and assume it has vanishing intrinsic torsion. Then, there is a metric $g$ such that if $\mathcal{H}^k$ denotes the harmonic $k$-forms, there is a splitting*

$$\mathcal{H}^k = \oplus_i \mathcal{H}^k_i,$$

*and isomorphisms $\mathcal{H}^k_i \cong \mathcal{H}^j_i$ if the corresponding $\Lambda^k_i \cong \Lambda^j_i$ are isomorphic representations.*

*Proof.* Since $P$ has vanishing intrinsic torsion, there is a metric $g$ whose Levi Civita connection $\nabla$ is induced by a connection on $P$. Thus, $\nabla$ preserves the embedding $P \hookrightarrow FSO(n) = P \times_G SO(n)$ and so for $\beta \in \Omega^k_i$, we have $\nabla\beta \in \Omega^0(X, T^*X \otimes \Lambda^k_i)$ and $\nabla^*\nabla\beta \in \Omega^k_i$. Having in mind that there is a Weitzenböck type formula

$$\Delta\beta = \nabla^*\nabla\beta + \mathcal{R}(\beta),$$

where $\mathcal{R}$ is an algebraic operator computed in terms of the curvature tensor $R \in \Omega^0(X, S^2\mathfrak{hol})$. Since $\mathfrak{hol} \subset \mathfrak{g}$ and the fact that $\mathcal{R} \in \Omega^0(X, \mathfrak{hol})$, it follows that $\mathcal{R}(\beta) \in \Omega^k_i$. Hence, the Laplacian $\Delta$ preserves the splitting into irreducible representations which then passes on to the harmonic forms. Moreover, one can show that $\nabla^*\nabla$ and $\mathcal{R}$ only depend on the representation in which they are acting and not on the specific degree of the differential form which concludes the proof of the statement. □

**Lemma 2.** *If $G \subset SO(n)$ is simply connected any a $G$-structure canonically lifts to a Spin-structure.*

*Proof.* Since $G$ is simply connected there is a unique lift of the inclusion of $G$ in $SO(n)$ to an inclusion $G \hookrightarrow Spin(n)$. Using this one can construct $\hat{F} = P \times_G Spin(n)$ and the projection $Spin(n) \to SO(n)$ gives a canonical map $\hat{F} \to FSO(n) = P \times_G SO(n)$. Hence $\hat{F}$ is a *Spin* structure on $(X,g)$. □

## 2.2 Stable Forms and Calibrations

In this section we shall review the notion of a stable form following Hitchin in [12] and [14]. Then, we shall see how some calibrations yield examples of such stable forms. Finally, we relate these to special Riemannian holonomy.

**Definition 2.** *Let $V^n$ be a real $n$ dimensional vector space, $\eta \in \Lambda^p V^*$ is a stable $p$-form if its $GL(V)$-orbit in $\Lambda^p V$ is open.*

**Example 2.**

1. *$n = 2m$, $m \in \mathbb{N}$ and $p = 2$. Then, $(V,\eta)$ is a symplectic vector space and the stabilizer of $\eta$ is $Sp(2m,\mathbb{R})$.*
2. *$n = 6$ and $p = 3$. There is an open orbit of $GL(6,\mathbb{R})$ on $\Lambda^3 V$ such than all $\eta$ lying on it have stabilizer $SL(3,\mathbb{C})$. Such an $\eta$ induces a complex structure on $V$ with respect to which $\eta$ is of type $(3,0)+(0,3)$.*
3. *$n = 7$ and $p = 3$. There are two open orbits of the $GL(7,\mathbb{R})$ on $\Lambda^3 V$, for $\eta$ in one of those the stabilizer is compact group $G_2$.*
4. *$n = 8$ and $p = 3$, there is an open orbit with stabilizer $PSU(3)$.*

In all the examples above the stabilizer preserves a volume form on the respective vector space. In fact, as observed by Hitchin in [12, 14], one has the following result.

**Proposition 2.** *There is a $GL(V)$-equivariant homogeneous function*

$$\phi : \Lambda^p V^* \to \Lambda^n V^*,$$

*of degree $\frac{n}{p}$. For each $\eta \in \Lambda^p V^*$, there is a unique $\hat{\eta}$, such that the derivative $d_\eta \phi : \Lambda^p V^* \to \Lambda^n V^*$ is given by*

$$d_\eta \phi(\dot{\eta}) = \hat{\eta} \wedge \dot{\eta},$$

*for $\dot{\eta} \in \Lambda^p V^*$ and moreover $\phi(\eta) = \frac{p}{n} \eta \wedge \hat{\eta}$.*

*Proof.* The existence of the $GL(V)$ equivariant function $\phi$ follows from the fact that all isotropy subgroups of such $\eta$ preserve a volume form on $V$. The $GL(V)$ invariance for scalar matrices $\lambda 1$, with $\lambda \in \mathbb{R}$, shows that $\phi(\lambda^p \eta) = \lambda^n \phi(\eta)$ and so $\phi$ is homogeneous of degree $n/p$.
The derivative $d\phi$ is linear and an element of $(\Lambda^p V^*)^* \otimes \Lambda^n V^* \cong \Lambda^{n-p} V^*$. Hence, there is a unique $\hat{\eta}$ with the properties stated, and the last statement that $\phi(\eta) = \frac{p}{n} \eta \wedge \hat{\eta}$ follows from Euler's formula

$$d\phi = \frac{n}{p} \phi$$

for homogeneous functions. □

**Example 3.** *1. $n=2m$ and $p=2$, $\eta$ is a symplectic form and $\hat{\eta}=\frac{\eta^{m-1}}{(m-1)!}$.*

*2. $n=6$ and $p=3$, $\eta+i\hat{\eta}$ is a form of type $(3,0)$, for the complex structure determined by $\eta$.*

*3. $n=7$ and $p=3$, the stabilizer of $\eta$ is $G_2$, which is a compact group. So the volume form $\phi(\eta)$ preserved by $G_2$ is the volume form of an invariant metric on $V$. Using this metric one obtains $\hat{\eta}=*\eta$.*

*4. $n=8$ and $p=3$, the stabilizer $PSU(3)$ is also compact and the same discussion goes on with $\hat{\eta}=-*\eta$.*

**Definition 3.** *If $g$ is a metric on an oriented vector space $V$ and $\{e_i\}_{i=1}^n$ an orthonormal basis, then a $p$-form $\theta\in\Lambda^pV^*$ is said to be a calibration if*

$$|\theta(e_{i_1},...,e_{i_k})|\leq 1,$$

*for all $i_1,...,i_k\in\{1,...,n\}$, i.e. if its comass is smaller or equal than 1.*

Equivalently, $\theta$ is a calibration on $(V,g)$, if and only if for all $p$-dimensional oriented subspaces $W\subset V$

$$\theta|_W\leq vol_W, \tag{1}$$

where $vol_W$ is the volume form of the metric $g|_W$ induced on $W$, by $g$.

**Definition 4.** *Let $(V,g)$ be a vector space with metric and $\theta\in\Lambda^pV^*$ a calibration on $V$. A subspace $W\subset V$ is said to be calibrated by $\theta$ is $\theta|_W=vol_W$, i.e. if equality is attained in the inequality 1.*

This discussion can be globalized in Harvey–Lawson's notion of a calibration [13].

**Definition 5.** *Let $X^n$ be a real $n$-dimensional smooth manifold and $\eta\in\Omega(X,\mathbb{R})$ a $p$-form is said to be stable if for all $p\in X$ $\eta_p\in\Lambda^pT_pX$ is a stable form.*
*If $(X,g)$ is an oriented Riemannian manifold and $\theta\in\Omega^p(X,\mathbb{R})$ is closed, then $\theta$ is called a calibration on $(X,g)$, if for all $x\in X$, $\theta_x$ is a calibration on $(T_xX,g_x)$. A submanifold $N\subset X$ is said to be calibrated by $\theta$ is for all $x\in N$, $T_xN\subset T_xX$ is calibrated by $\theta_x$.*

The construction from proposition 2 gives a volume form on $M$, whose volume defines the Hitchin functional

$$\Phi(\eta)=\int_X\phi(\eta)\in\mathbb{R}\cup\infty. \tag{2}$$

Notice that the existence of a stable $p$ form $\eta$ on $M^n$ reduces the structure group of the tangent bundle to the isotropy subgroup of the form $\eta$. A natural question is if there is any relation between these reductions and possible reductions of the holonomy group of a special metric on $M$, determined by $\eta$.

**Proposition 3.** *If $X$ is compact, and $[\eta] \in H^p(X,\mathbb{R})$ is a fixed cohomology class. Then Hitchin's functional gives a well defined function*

$$\Phi : [\eta] \to \mathbb{R},$$

*whose critical points are the $\eta \in [\eta]$ with $d\hat{\eta} = 0$.*

*Proof.* Let $\eta \in [\eta]$ be a critical point, since the variation is in the fixed cohomology class $[\eta]$ all tangent vectors are exact forms $d\alpha$. So for all $\alpha \in \Omega^{p-1}(X,\mathbb{R})$

$$0 = d\Phi_\eta(d\alpha) = \int_X \hat{\eta} \wedge d\alpha = \int_X d\hat{\eta} \wedge \alpha,$$

which shows that if $\hat{\eta}$ is a critical point then $d\hat{\eta} = 0$. Conversely, the same computation also shows that if $d\hat{\eta} = 0$, then $\hat{\eta}$ is a critical point. □

**Example 4.**
1. *$n = 2m$ and $p = 2$, $(X,\eta)$ is a symplectic manifold and $d\hat{\eta} = 0$ always. $(X,\omega)$ with $\omega = \eta$ can be equipped with a metric $g$ and compatible almost complex structure $I$. Then, for all $k \leq n$, $\frac{\omega^k}{k!}$ has comass $\leq 1$ and is closed and so a calibration. Submanifolds $N^{2k}$ calibrated by $\frac{\omega^k}{k!}$ are symplectic (or almost complex) submanifolds.*
   *If $\nabla I = 0$ the complex structure is integrable and the metric has holonomy contained in $U(n)$. Then $(X,I,\eta)$ is a Kähler manifold and the $\frac{\omega^k}{k!}$-calibrated submanifolds are complex submanifolds.*

2. *$n = 6$ and $p = 3$, then $\eta + i\hat{\eta}$ equips $X$ with an almost complex structure for which $\eta + i\hat{\eta}$ is of type $(3,0)$. If $\eta$ is a critical point of Hitchin's functional, then $\overline{\partial}(\eta + i\hat{\eta}) = 0$ and so the complex structure is integrable. Since $\eta + i\hat{\eta}$ is a nonvanishing holomorphic volume form, $X$ has trivial canonical bundle.*

   *If $(X,\omega,\Omega = \Omega_1 + i\Omega_2)$ is a Calabi–Yau 3-fold, then in particular it is Kähler and choosing $\eta = \omega$ the example above gives a reduction of the holonomy to $U(3)$. Moreover, choosing $\eta = \Omega_1$, gives this precise example and $\nabla\Omega = 0$, which reduces the holonomy to $SL(3,\mathbb{C})$ and so the holonomy of the metric is contained in $SU(3) = U(3) \cap SL(3,\mathbb{C})$.*

*In this case both $\Omega_1 = \eta$ and $\Omega_2 = \hat{\eta}$ are calibrations and submanifolds $N^3$ calibrated by them are called special Lagrangian submanifolds of phase $0, \frac{\pi}{2}$ respectively.*

3. *$n = 7$ and $p = 3$, the stable form $\eta$ is a critical point of Hitchin's functional if*

$$d\eta = d^*\eta = 0,$$

*for the metric on $M$ determined by $\eta$. Indeed, by a result of Fernández and Gray [11] this is equivalent to $\nabla\eta = 0$, which is to say the Holonomy of the metric is contained in $G_2$, by the Holonomy principle. In this case one usually uses the notation $\eta = \phi$, $\hat{\eta} = *\phi = \psi$ and $(X, \phi)$ is called a $G_2$-manifold. Both $\phi$ and $\psi$ are calibrations and submanifolds calibrated by them are respectively called associative and coassociative.*

In the examples above only for the case $n = 7$, the stable form $\eta$ determines a metric with reduced holonomy (in fact $G_2$ which an exceptional Lie group appearing in Berger's list). This is because the holonomy group of any oriented Riemannian manifold must be a subgroup of $SO(n)$ by the holonomy principle, and both $Sp(2m, \mathbb{R})$ and $SL(3, \mathbb{C})$ are non-compact groups.

**Proposition 4.** *Let $(X^n, g)$ be a Riemannian manifold equipped with a calibration $\theta \in \Omega^p(X, \mathbb{R})$. If $N^p \subset X$ is compact and calibrated by $\theta$, then $N$ is volume minimizing in its homology class $[N] \in H_p(X, \mathbb{R})$.*

*Proof.* Let $N' \in [N]$ be cohomologous to $N$, then there is $S^{p+1}$ with $\partial S = N \cup (-N')$ (with orientations) and Stokes theorem gives $\int_N \theta - \int_{N'} \theta = \int_S d\theta = 0$. Now the result follows from applying this and the definition of calibration to the following one line calculation

$$vol(N') = \int_{N'} vol_{N'} \geq \int_{N'} \theta = \int_N \theta = \int_N vol_N = vol(N). \tag{3}$$

□

Notice that the equation for a calibrated submanifold is a first order PDE, while being minimal is a second order one (the Euler Lagrange equations for critical points of the volume functional). This is an analogous situation to that of many gauge theories as for example the relation between ASD connections and the Yang Mills equations for connections on bundles over 4 manifolds. See [10, 19] for some higher dimensional gauge theories mimicking these.

We shall now change gears and focus on the more concrete cases of Calabi–Yau and $G_2$-manifolds. The interested reader can find a lot more about these for example in [6, 17, 18] and references therein.

## 3 Calabi–Yau Manifolds

On a Kähler manifold $(X^n, g, \omega)$ we shall implicitly always consider complex structure $I$ determined by $g$ and $\omega$. The next proposition relates the Ricci tensor and the holomorphic triviality of the canonical bundle $K_X = \Lambda^{n,0}_{\mathbb{C}} X$ to the holonomy of the underlying Kähler metric.

**Proposition 5.** *A Kähler manifold $(X^n, g, \omega)$ with $n = 2m$ is Ricci flat with trivial canonical bundle $K_X$ if and only if the holonomy of the Kähler metric on $X$ is contained in $SU(m)$.*

*In any case $X$ is a Ricci flat Kähler manifold with trivial canonical bundle $K_X$ and there is a unique (up to phase) holomorphic volume form $\Omega$ satisfying*

$$\frac{\omega^m}{m!} = (-1)^{\frac{m(m-1)}{2}} \left(\frac{i}{2}\right)^m \Omega \wedge \overline{\Omega}, \tag{4}$$

*which trivializes $K_X$.*

*Proof.* The Ricci form $\rho(\cdot,\cdot) = \mathrm{Ric}(\cdot, I\cdot)$ is the curvature of the connection on $K_X$ induced via the Levi Civita connection on the holomorphic tangent bundle. First suppose that $X$ is Ricci flat and $K_X$ trivial (this a necessary assumption if $X$ is not simply connected). Ricci flatness gives that $\rho = 0$, while the triviality of $K_X$ guarantees not only that $c_1(X) = 0 \in H^2(X, \mathbb{Z})$, but that the element in the Jacobian representing $K_X$ is trivial. Hence, the connection has no periods and there is an holomorphic trivializing section of $K_X$, i.e. there is a $(n,0)$-form $\Omega$ such that $\overline{\partial}\Omega = 0$. This implies it is parallel and so by the holonomy principle (corollary 1) the Kähler metric has holonomy contained in $SU(m)$.
The converse statement also follows from the holonomy principle since if the holonomy is contained in $SU(m)$, then there are nonzero parallel forms $\omega \in \Omega^2(X, \mathbb{R})$ and $\Omega \in \Omega^{3,0}(X, \mathbb{C})$ (unique up to phase) satisfying the relation 4 in the statement. Since $\nabla\Omega = 0$, then also $\overline{\partial}\Omega = 0$ and so it is holomorphic and trivializes $K_X$. Then, $c_1(X) = 0$ and the definition of curvature also gives $\rho(\Omega) = d_\nabla \nabla\Omega = 0$, and as $\Omega$ is nonvanishing $\rho = 0$, i.e. the metric is Ricci flat, which is the same thing as saying that the connection on $K_X$ induced by the Levi Civita one is flat. □

**Remark 2.** *If $X$ is Ricci-flat Kähler and simply connected, then $K_X$ is automatically trivial and the holonomy contained in $SU(m)$. This follows from the fact that $\rho = 0$ and so the Levi Civita connection equips $K_X$ with a flat connection. These are parametrized by $Hom(\pi_1(X), U(1))$, which vanishes as $X$ is simply connected. Then $K_X$ is trivial and proposition 5 shows the*

*holonomy is in $SU(m)$.*
*When $X$ is not simply connected there are counterexamples to this statement. For example an Enriques surface is a Ricci flat Kähler manifold with $c_1(X)$ a torsion class in $H^2(X,\mathbb{Z})$. In this case $K_X$ is not trivial and the flat connection can be seen as an element of $Hom(\pi_1(X),U(1)) = Hom(H_1(X),U(1)) = H^1(X,U(1))$, uniquely determined by the Hermitian metric on $K_X$ via Chern's construction.*

**Definition 6.** *A Calabi–Yau manifold $(X,\omega,\Omega)$ is a Ricci flat, Kähler manifold $(X,\omega)$ with trivial canonical bundle and a choice of holomorphic volume form $\Omega \in \Omega^{3,0}(X,\mathbb{C})$ satisfying equation 4.*

According to this definition Calabi–Yau manifolds will have holonomy contained in $SU(m)$. Some authors require the holonomy to be exactly $SU(m)$ and here these will be called irreducible Calabi–Yau manifolds. The question of existence of Calabi–Yau manifolds can be attacked directly by explicitly constructing the metric as is done in several noncompact examples or by PDE methods in both cases compact and noncompact. In line with the second of these, we have Yau's proof of the Calabi conjecture, [20], which states the following.

**Theorem 2.** *Let $X$ be a compact complex manifold with $c_1(X) = 0$ in $H^2(X,\mathbb{R})$, then in all Kähler classes in $X$, there is a unique Ricci-flat Kähler metric.*

The Calabi Conjecture, so called by having been proposed by Calabi years before Yau completed its proof in [20], asserts the existence of many compact Calabi–Yau manifolds. For example, if $X$ is a complex manifold with $c_1(X) = 0$, $\pi_1(X) = 0$ and which admits Kähler classes, then combining the Calabi conjecture 2 with proposition 5, there is a Calabi–Yau structure on each Kähler class of $X$.

**Remark 3.** *The Enriques surface from remark 2 is not a Calabi–Yau manifold according to definition 6, as it has nontrivial canonical bundle. However, the Calabi conjecture stated as Theorem 2, proves the existence of a Ricci-flat Kähler metric on the Enriques surface.*

The next results explore some properties of Calabi–Yau manifolds.

**Proposition 6.** *Let $(X,\omega,\Omega)$ be a compact Calabi–Yau manifold. Then there is a finite cover $\tilde{X}$ of $X$, which is biholomorphic to the product of $T^{2k} \times Y$, where $T^{2k}$ is a real $2k$ dimensional torus and $Y$ a complex manifold with $c_1(Y) = 0$.*

*If $(X,\omega,\Omega)$ is further assumed to be irreducible, then it has finite fundamental group.*

*Proof.* Calabi–Yau manifolds are Ricci flat and so the Cheeger Gromoll splitting theorem applies and for each $2k$ linearly independent parallel 1 forms, there is a $T^{2k}$ splitting off. As both the complex torus and $X$ have vanishing first Chern class so must be for $Y$.
Now suppose $X$ is irreducible, then it cannot have any parallel 1 form as this would make the Holonomy to be strictly contained in $SU(m)$. Moreover, for Ricci flat manifolds there is a Weitzenböck formula

$$\|\nabla\alpha\| = \langle\alpha,\Delta\alpha\rangle,$$

which shows that each harmonic 1 form gives rise to a parallel 1 form. Hence there can be no harmonic 1-forms and this forces the fundamental group of $X$ to be finite. □

**Remark 4.** *A version of this result also holds in the noncompact case, there one may have to let some of the torus directions to be noncompact (i.e. $\mathbb{R}$) and $Y$ may be noncompact as well.*

**Proposition 7.** *Let $(X,\omega,\Omega)$ be a compact Calabi–Yau manifold, then for $i\in\{1,...,m-1\}$*

$$dim(H^{i,0}(X,\mathbb{C})) \leq \frac{m!}{i!(m-i)!}.$$

*If $(X,\omega,\Omega)$ is further assumed to be irreducible, then*

$$\begin{aligned} H^{0,0}(X,\mathbb{C}) &= 1 \\ H^{m,m}(X,\mathbb{C}) &= 1 \\ H^{i,0}(X,\mathbb{C}) &= 0, \quad i\in 1,...,m-1. \end{aligned}$$

*Proof.* For $i=1,...,m-1$, let $\alpha\in\Omega^{i,0}$ be a representative of a class in $H^{i,0}(X,\mathbb{C})$ and recall that Calabi–Yau manifolds are Ricci flat. Then as in the proof of proposition 6, the Weitzenböck formula is $\|\nabla\alpha\| = \langle\alpha,\Delta\alpha\rangle$. Moreover, the Kähler identities imply that $\Delta\alpha = 2\overline{\partial}^*\overline{\partial}\alpha = 0$[1] and $\alpha$ is then parallel. In the general case, the maximum number of linearly independent of these is then the dimension of $\Lambda^{i,0}\mathbb{C}^n$, which is precisely $\frac{m!}{i!(m-i)!}$. In the irreducible case there can be no nonzero parallel $(i,0)$ forms as this would reduce the holonomy to be strictly contained in $SU(m)$. □

[1] Notice that $\overline{\partial}^*\alpha = 0$ as $\alpha$ is of type $(i,0)$.

**Remark 5.** *There is an alternative argument using the maximum principle which can be used in noncompact Ricci flat manifolds. This proves for example for noncompact irreducible Calabi–Yau manifolds there can be no decaying harmonic* $(i,0)$ *forms.*

**Proposition 8.** *Let* $(X,\omega,\Omega)$ *be a compact and irreducible Calabi–Yau manifold of real dimension* $n = 2m \geq 6$*. Then* $X^m$ *is a projective algebraic variety.*

*Proof.* Since $m \geq 3$ and $(X,\omega,\Omega)$ is irreducible, proposition 7 gives that $h^{2,0} = 0$ and so $H^2(X,\mathbb{C}) = H^{1,1}(X,\mathbb{C})$. Then, the image of $H^2(X,\mathbb{Z}) \to H^{1,1}(X,\mathbb{C})$ is nonempty and one can pick a positive class $\alpha$ there. Associated to this class there is a positive holomorphic line bundle $L$ with $c_1(L) = \alpha$ and the Kodaira Embedding theorem provides an embedding $X \hookrightarrow \mathbb{CP}^{h^0(X,L^k)-1}$, for sufficiently large $k \in \mathbb{N}$. □

## 4 $G_2$-Manifolds

Let $X^7$ be a 7 dimensional manifold and denote by $\Lambda^3_+$ the bundle of stable 3 forms over $X$ and by $\Omega^3_+$ its sections. Given $\phi \in \Omega^3_+$, then at any point $p \in X$ the stabilizer of $\phi_p$ in $GL(7,\mathbb{R})$ is conjugate to $G_2$ (as defined in the third item of example 2). Given such a section, it determines via the weak holonomy principle a $G_2$ structure, which itself determines via lemma 2 a Spin-structure on $(X^7,g)$. In fact, for $G_2$-structures the converse also holds.

**Proposition 9.** *A* 7*-dimensional oriented Riemannian manifold* $(X^7,g)$ *admits a* $G_2$*-structure if and only if it is Spin.*

*Proof.* Since $G_2$ is simply connected, given a $G_2$-structure lemma 2 guarantees the existence of a Spin structure $\hat{F}$. To prove the converse let $\hat{F}$ denote a Spin bundle and $\Delta$ the standard irreducible $Spin(7)$ representation, then $\mathrm{rk}_{\mathbb{R}} \Delta = 8$. Moreover, $Spin(7)$ acts transitively on $\mathbb{S}^7$ with stabilizer $G_2$, so it is enough to find a unit section of the bundle of spinors $\mathcal{S} = \hat{F} \times_{Spin(7)} \Delta$. Since this bundle has rank $8 > 7$ there is a nowhere vanishing section of $\mathcal{S}$, which we can normalize to have norm 1. Then the weak holonomy principle determines a $G_2$-structure. □

**Definition 7.** *Let* $(X^7,\phi)$ *be as above and* $\phi \in \Omega^3_+$*. Then* $\phi$ *and* $g$ *are compatible if for all vector fields* $V,W$*,* $\iota_V \phi \wedge \iota_W \phi \wedge \phi = 6g(V,W)_g dvol_g$*.*

**Definition 8.** *A $G_2$-manifold $(X,\phi)$ is a real 7 dimensional Riemannian manifold $(X^7,g)$, equipped with a compatible $G_2$-structure $\phi \in \Omega^3_+$ such that $\nabla\phi = 0$.*

From the Holonomy Principle a $G_2$-manifold has holonomy contained in $G_2$, when the holonomy is the full $G_2$ one says that $(X,\phi)$ is an irreducible $G_2$-manifold. We shall now go on to investigate some topological and geometric properties of (irreducible) $G_2$-manifolds starting with the following of Fernández and Gray [11].

**Theorem 3.** *Let $(X^7,g)$ be a Riemannian 7 dimensional manifold equipped with a stable 3 form $\phi$ compatible with $g$, the following are equivalent*

1. $\nabla\phi = 0$,
2. $d\phi = d^*\phi = 0$,
3. *The holonomy of $g$ is contained in $G_2$.*

*Proof.* The holonomy principle (corollary 1) implies that the holonomy of $g$ is in $G_2$ if and only if $\nabla\phi = 0$; and so it is enough to prove that the first two items are equivalent. In one direction this is obvious since $\nabla\phi \in \Omega^0(X, T^*X \otimes T^*X)$ and both $d\phi$ and $d^*\phi$ are obtained from $\nabla\phi$ by composition with algebraic operators, respectively the anti-symmetrization map $\wedge \in \mathrm{Hom}(T^*X \otimes T^*X, \Lambda^2 X)$ and the trace with respect to metric $g$, $\mathrm{tr}_g \in \mathrm{Hom}(T^*X \otimes T^*X, \Lambda^0)$. So if $\nabla\phi = 0$, then both $d\phi$ and $d^*\phi$ vanish.
In the opposite direction, suppose $d\phi = d^*\phi = 0$, and to proceed we need to investigate $\nabla\phi$ with more scrutiny. The intrinsic torsion of the $G_2$-structure determined by $\phi$ is $\nabla\phi$, seen as a section of $\mathrm{coker}(\delta)$, where $\delta$ is the map defined in the discussion preceeding Lemma 1. Recall that this bundle is modeled on $V^* \otimes \mathfrak{g}_2^\perp$, where $\mathfrak{g}_2^\perp \subset \mathfrak{so}(7)$ and $V \cong \mathbb{R}^7$ is the standard 7 dimensional representation of $\mathfrak{g}_2$. By an abuse of language we shall say $\nabla\phi$ is modeled on $V^* \otimes \mathfrak{g}_2^\perp$. Notice that $\mathfrak{so}(7) \cong \Lambda^2 V \cong \Lambda^2_7 \oplus \Lambda^2_{14}$ with $\Lambda^2_7 \cong V^*$ and $\Lambda^2_{14} \cong \mathfrak{g}_2$. We conclude that $\mathfrak{g}_2^\perp \cong V^*$ and so $\nabla\phi$ is a section of $V^* \otimes V^* \cong \Lambda^2 V \oplus S^2 V$. In fact this further decomposes into

$$V^* \otimes V^* \cong (V^* \oplus \mathfrak{g}_2) \oplus \left(S^2_0 V \oplus \mathbb{R}\right), \tag{5}$$

where $\mathbb{R}$ is the trivial representation, and it follows from highest weight theory that $S^2_0 V$ is irreducible of dimension 27. Hence, the decomposition above is irreducible.
Next, $d\phi$ is modeled on $\Lambda^4 V \cong \Lambda^3 V$, which decomposes into irreducible

components as $\mathbb{R} \oplus V \oplus S_0^2 V$. Since $0 = d\phi = \wedge \circ \nabla\phi$ and $\wedge$ is a morphism of representations and is surjective, it follows that $\nabla\phi$ has values in the the $\mathfrak{g}_2$ component of the decomposition in 5.
Next we analyse the vanishing of $d^*\phi$ which is modeled on $\Lambda^2 V \cong V \otimes \mathfrak{g}_2$. In the same way as before $0 = d^*\phi = \mathrm{tr}_g(\nabla\phi)$ and since $\mathrm{tr}_g$ is also a surjective morphism of representations the component of $\nabla\phi$ in $\mathfrak{g}_2$ also vanishes. Combined with $d\phi = 0$ this shows that $\nabla\phi = 0$ and completes the proof of the statement. □

Comparing the second point above with proposition 3, more specifically the third item in example 4 shows that $G_2$-manifolds are (in the compact case) critical points of Hitchin's functional. In fact, they have maximal volume with respect to local variations of the 3 form $\phi$. Next, we shall give a modern proof of the following Theorem of Bonan [3].

**Theorem 4.** *Let $(X^7, g)$ be a $G_2$-manifold, then $g$ is Ricci flat.*

*Proof.* Denote by $P \subset FSO(n)$ the $G_2$ structure and by $R \in \Omega^0(X, S^2\mathfrak{g}_P)$ the Riemann curvature tensor of $g$. Using highest weight theory we can decompose the space of algebraic curvature tensors into irreducible representations. We start by decomposing

$$S^2\mathfrak{g}_2 \cong W_{0,0} \oplus W_{2,0} \oplus W_{0,2}, \tag{6}$$

where $W_{0,0} \cong \mathbb{R}$ is the trivial irreducible representation and the $(n,k) \in \mathbb{Z}^2$ are labeling the weights, so that $W_{1,0} \cong V$ and $W_{0,1} \cong \mathfrak{g}_2$. Moreover, the first Bianchi identity states that $R \in \ker(b)$, where

$$b : S^2(V^*) \to \Lambda^3 V^* \otimes V^*$$

is the Bianchi map which antisymmetrizes the first three entries. However $\ker(b) = \ker(b : S^2(\mathfrak{g}_2) \to \Lambda^4 V)$. Decompose the right hand side into irreducibles $\Lambda^4 V \cong W_{0,0} \oplus W_{1,0} \oplus W_{2,0}$ and compare with the relation 6. In fact, the Bianchi map is a morphism of $G_2$-representations and is injective on $W_{0,0}$ and $W_{2,0} \cong S_0^2 V^*$, so we conclude that the kernel of the Bianchi map is the 77 simensional piece $W_{0,2}$. Hence the Riemannian curvature tensor has values on $W_{0,2}$ (this result is attributed to Alexeevski [1]).
We now use this information in order to analyze the Ricci tensor $Ric$, which has values on $S^2(V^*)$. It is obtained from $R$ via $Ric = r(R)$, where

$$r : W_{0,2} \to S^2(V^*)$$

is the Ricci contraction, mapping a curvature tensor to a symmetric, bilinear form. This is also a morphism of $G_2$ representations and since $S^2(V^*)$ decomposes into irreducible components as $W_{0,0} \oplus W_{2,0}$, $r$ must vanish identically and so does *Ric*. □

$G_2$-manifolds are Ricci flat (theorem 4) and a similar application of the Cheeger-Gromoll splitting theorem and the Böchner technique, to the one used for Calabi–Yau manifolds in Prposition 6 gives the following two propositions

**Proposition 10.** *Let $(X,g)$ be a compact $G_2$-manifold. Then, there is a finite cover $\tilde{X}$ of $X$, which is isometric to $T^{7-k} \times Y^k$, where $T^{7-k}$ is a torus and $Y^k$ is $k$ dimensional manifold equipped with a Ricci flat metric. Moreover, if $(X,\phi)$ is further supposed to be irreducible, then it has finite fundamental group.*

**Proposition 11.** *Let $(X,g)$ be a simply connected $G_2$-manifold, then $(X,g)$ is irreducible, i.e. $Hol = G_2$ if and only if there are no parallel 1-forms.*

*Proof.* Since $(X,g)$ is a $G_2$-manifold the holonomy is contained in $G_2$ and $g$ is Ricci flat. Hence, if there is a parallel one form one can use the flow of the associated Killing field, which is parallel by the Bochner formula, to find a line and use the Cheeger-Gromoll splitting theorem to write $X = \mathbb{R}_t \times Y^6$ with the cylindrical metric $g = dt^2 + g_6$. In this case $Hol(g) = Hol(g_6) \subset G_2 \cap (1 \times SO(6)) \cong 1 \times SU(3)$, which is properly contained in $G_2$.
In the opposite direction we prove that if the holonomy $Hol$ is properly contained in $G_2$ then there is a parallel 1-form. First we analyze the case where $(X,g)$ is locally symmetric. If this is the case, then since from Bonan's theorem 4 is Ricci flat and locally symmetric it must actually be flat. If $(X,g)$ is not locally symmetric and $Hol$ is a proper subgroup of $G_2$ we can invoke Berger's theorem [2] to conclude that $Hol$ is either $1 \times SU(3)$, $SO(3) \times SU(2)$, $1_3 \times SU(2)$ or $1_7$. In each of these cases there is a local splitting $U = U_1 \times U_2$ and $g|_U = g_1 + g_2$, where $U_1$ is at most 3 dimensional and Ricci flat and so flat. So the case $SO(3) \times SU(2)$ actually has to reduce to $1_3 \times SU(2)$ and in all the cases there is a locally flat factor, then since $X$ is simply connected there is a global parallel one form. □

**Remark 6.** *Notice that in the first direction the condition that $X$ is simply connected is not used. Hence it is true that for $(X,g)$ an irreducible $G_2$-manifold there are no parallel 1-forms.*

**Proposition 12.** *Let $(X, \phi)$ be a $G_2$-manifold, then the exterior bundle splits orthogonally as*

$$\begin{aligned} \Lambda^1 &= \Lambda^1_7 \\ \Lambda^2 &= \Lambda^2_7 \oplus \Lambda^2_{14} \\ \Lambda^3 &= \Lambda^3_1 \oplus \Lambda^3_7 \oplus \Lambda^3_{27}, \end{aligned}$$

*where the subscript indicates the rank of the component and these components are such that for $\psi = *\phi$*

$$\begin{aligned} \Lambda^2_7 &= \{\iota_V \phi, V \in \Gamma(TX)\} = \{\beta \mid *(\beta \wedge \phi) = 2\omega\} \\ \Lambda^2_{14} &= \{\beta \mid \beta \wedge \psi = 0\} = \{\beta \mid *(\beta \wedge \phi) = -\beta\} \\ \Lambda^3_1 &= \langle \phi \rangle \\ \Lambda^2_7 &= \{\iota_V \psi, V \in \Gamma(TX)\} \\ \Lambda^3_{27} &= \{\beta \mid \beta \wedge \psi = 0 \quad and \quad \beta \wedge \phi = 0\}. \end{aligned}$$

*Moreover if $\beta$ is a 2-form and $\pi_7, \pi_{14}$ denote the respective projections on the irreducible components, then the following algebraic identities hold*

$$*(*(\beta \wedge \psi) \wedge \psi) = 3\pi_7(\beta) \quad (7)$$

$$*(\beta \wedge \phi) = 2\pi_7(\beta) - \pi_{14}(\beta). \quad (8)$$

It follows from Chern's theorem 1 that on a $G_2$-manifold the Laplacian $\Delta_\phi$ preserves the decomposition of the spaces of differential forms into irreducible $G_2$ representations. Hence, the decomposition above still holds at the level of Harmonic forms.

**Corollary 2.** *Let $(X, \phi)$ be a $G_2$-manifold, then the spaces of harmonic forms $\mathcal{H}^*$ decompose into irreducible representations as*

$$\begin{aligned} \mathcal{H}^1 &= \mathcal{H}^1_7 \\ \mathcal{H}^2 &= \mathcal{H}^2_7 \oplus \mathcal{H}^2_{14} \\ \mathcal{H}^3 &= \mathcal{H}^3_1 \oplus \mathcal{H}^3_7 \oplus \mathcal{H}^3_{27}, \end{aligned}$$

*and there are isomorphisms $\mathcal{H}^1 \cong \mathcal{H}^2_7 \cong \mathcal{H}^3_7$. In particular, if $X$ is compact this induces a splitting of the de Rham cohomology.*

We can combine corollary 2 to Chern's theorem with proposition 11 to investigate further the topology of irreducible $G_2$-manifolds.

**Proposition 13.** *Let $(X^7, g)$ be an irreducible $G_2$-manifold, then the spaces of harmonic forms $\mathcal{H}^*$ decompose into irreducible representations as*

$$\begin{aligned} \mathcal{H}^1 &= 0 \\ \mathcal{H}^2 &= \mathcal{H}^2_{14} \\ \mathcal{H}^3 &= \mathcal{H}^3_1 \oplus \mathcal{H}^3_{27}. \end{aligned}$$

*In particular, if $X$ is compact then $b^1 = 0$, $b^2 = b^2_{14}$ and $b^3 = 1 + b^3_{27}$.*

*Proof.* The irreducibility condition, i.e. that $Hol = G_2$ implies via remark 6 that there are no parallel 1 forms. Since $Ric = 0$ by Bonan's theorem 4, there is a Weitzenböck formula $\nabla^*\nabla\alpha = \Delta\alpha$ for all 1-forms $\alpha$. Combining this with corollary 2 gives the decomposition of the harmonic forms in the statement. In the particular case when $X$ is compact, the result follows from Hodge theory. □

**Remark 7.** *In particular, this further proves that a compact, irreducible $G_2$-manifold has finite fundamental group.*

Now we will focus on compact $G_2$-manifolds which were first constructed by Dominic Joyce [15, 16], see also [17] for a summary of this first construction. On these we shall construct a quadratic form on the second cohomology which can be used to identify a constraint on the first Pontryagin class $p_1(X) \in H^4(X, \mathbb{R})$ of a compact, irreducible $G_2$-manifold.

**Definition 9.** *Let $(X, g)$ be a compact $G_2$-manifold and define the bilinear form $Q$ on $H^2(X, \mathbb{R})$ given by*

$$Q(\alpha, \beta) = \langle \alpha \cup \beta \cup [\phi], [X] \rangle.$$

**Lemma 3.** *Let $(X, g)$ be a compact, irreducible $G_2$-manifold. Then, the quadratic form on $H^2(X, \mathbb{R})$ given by $\alpha \mapsto Q(\alpha, \alpha)$ is negative definite.*

*Proof.* Let $a \in \alpha \neq 0$ be the harmonic representative, then by proposition 13 it follows that $a = \pi_{14}(a)$, i.e. $\pi_7(a) = 0$. Moreover, using equation 8 one has

$$a \wedge a \wedge \phi = -\pi_{14}(a) \wedge *\pi_{14}(a) = -|a|^2 \,\mathrm{dvol},$$

hence $Q(\alpha, \alpha) = -\int_X |a|^2 \,\mathrm{dvol} < 0$. □

**Proposition 14.** *Let $(X, g)$ be a compact, irreducible $G_2$-manifold, then $\langle p_1(X) \cup [\phi], [X] \rangle \leq 0$.*

*Proof.* Let $R$ denote the curvature of the Levi-Civita connection of $g$. In a local trivialization $R \in \Omega^2_{14} \otimes \mathfrak{g}_2$ and $\mathfrak{g}_2 \subset \mathfrak{so}(4)$, i.e. it is represented by an antisymmetric matrix $R_{ij}$ of forms in $\Omega^2_{14}$. Then, $p_1(X) \cup [\phi]$ is represented by

$$\mathrm{tr}(R \wedge R) \wedge \phi = \sum_{i,j} R_{ij} \wedge R_{ji} \wedge \phi = -\sum_{i,j} R_{ij} \wedge R_{ij} \wedge \phi = -|R|^2 \,\mathrm{dvol}.$$

Hence as in the previous lemma (or rather as in its proof) $\langle p_1(X) \cup [\phi], [X] \rangle = -\int_X |R|^2 \,\mathrm{dvol} \leq 0$. □

## References

[1] D. V. Alekseevsky, *Riemannian spaces with unusual holonomy groups.* Funct. Anal. Appl 2.2 (1968): 1–10.

[2] M. Berger, *Sur les groupes d'holonomie homogènes de variétés a connexion affine et des variétés riemanniennes*, Bull. Soc. Math. France, 83 (1955): 279–330.

[3] E. Bonan, *Sur des verietes Riemanniennes a groupe dholonomie G2 ou Spin(7).* Comptes Rendus Hebdomadaires des seances de l'Academie des Sciences Serie A 262.2 (1966): 127–129.

[4] Robert L. Bryant, *Metrics with exceptional holonomy*, Annals of Mathematics, 126 (1987): 525–576.

[5] Robert L. Bryant, Simon M. Salamon, *On the construction of some complete metrics with exceptional holonomy.* Duke Math. J. 58, (1989): 829–850

[6] Robert L. Bryant, *Some Remarks on G2-Structures*, Gökova Geometry 12th Topology Conference (GGT), 2003.

[7] Élie Cartan, *Sur une classe remarquable d'espaces de Riemann. I* Bulletin de la Société Mathématique de France 54 (1926): 214–264.

[8] Élie Cartan, *Sur une classe remarquable d'espaces de Riemann. II.* Bulletin de la Société Mathématique de France 55 (1927): 114–134.

[9] Shiing-Shen Chern, *On a generalization of Kähler geometry*, Algebraic geometry and topology. A symposium in honour of S. Lefschetz, Princeton University Press, Princeton, N. J., 1957, 103–121.

[10] Simon Donaldson, Ed. Segal, *Gauge theory in higher dimensions, II.* Surveys in differential geometry. Volume XVI. Geometry of special holonomy and related topics (2013).

[11] Marisa Fernández, Alfred Gray, *Riemannian manifolds with structure group G2.* Annali di matematica pura ed applicata 132.1 (1982): 19–45.

[12] Nigel Hitchin, *The geometry of three-forms in six dimensions.* Journal of Differential Geometry 55.3 (2000): 547–576.

[13] Harvey, Reese, and H. Blaine Lawson. *Calibrated geometries.* Acta Mathematica 148.1 (1982): 47-157.

[14] Nigel Hitchin, *Stable forms and special metrics.* Contemporary mathematics 288 (2001): 70–89.

[15] D. Joyce, *Compact Riemannian 7-manifolds with holonomy G2. I*, Journal of Differential Geometry 43 (1996), 291–328.

[16] D. Joyce, *Compact Riemannian 7-manifolds with holonomy G2. II*, Journal of Differential Geometry 43 (1996), 329–375.

[17] D. Joyce, *Compact manifolds with special holonomy.* Oxford University, 2000.

[18] Simon Salamon, *Riemannian geometry and holonomy groups*, Pitman Research Notes in Math., 201. Longman, Harlow, 1989.

[19] Gang Tian, *Gauge theory and calibrated geometry, I.* Annals of Mathematics 151.1 (2000): 193–268.

[20] Shing-Tung Yau, *Calabi's conjecture and some new results in algebraic geometry.* Proceedings of the National Academy of Sciences 74.5 (1977): 1798–1799.

# Dynamics of Planar Piecewise Isometries: Recent Advances

*Pedro Peres*, *Ana Rodrigues*
University of Exeter
Department of Mathematics
Exeter EX4 4QF, UK
e-mail: pp334@exeter.ac.uk
a.rodrigues@exeter.ac.uk

**Resumo:** Neste trabalho revemos alguns resultados relacionados com o estudo de isometrias por pedaços. Introduziremos *embeddings* de uma transformação de troca de intervalos numa isometria por pedaços, discutiremos a renormalização de uma isometria por pedaços particular e provaremos a existência de curvas invariantes para estas transformações.

**Abstract:** In this survey we review recent results on the study of the dynamics of piecewise isometries. We will introduce embeddings of an interval exchange transformation into a piecewise isometry, discuss the renormalization of a particular piecewise isometry and finally show that invariant curves exist for such transformations.

**palavras-chave:** Renormalização; curvas invariantes.

**keywords:** Renormalization; invariant curves.

## 1 Introduction

An *interval exchange transformation (IET)* is a bijective piecewise order preserving isometry $f$ of an interval $I \subset \mathbb{R}$, where $I$ is partitioned into subintervals $\{I_\alpha\}_{\alpha\in\mathcal{A}}$, indexed over a finite alphabet $\mathcal{A}$ of $d \geq 2$ symbols, so that the restriction of $f$ to each subinterval is a translation. IETs were studied for instance in [20, 28]. Masur [22] and Veech [28] proved independently that a typical IET is uniquely ergodic while Avila and Forni [12] established that a typical IET is either weakly mixing or an irrational rotation.

*Piecewise isometries (PWIs)* are higher dimensional generalizations of one dimensional IETs. They have been defined on higher dimensional spaces and Riemannian manifolds [5, 17]. In this paper we consider orientation preserving planar piecewise isometries with respect to the standard euclidean metric. Let $X$ be a subset of $\mathbb{C}$ and $\mathcal{P} = \{X_\alpha\}_{\alpha\in\mathcal{A}}$ be a finite partition of $X$ into convex sets (or *atoms*), that is $\bigcup_{\alpha\in\mathcal{A}} X_\alpha = X$ and $X_\alpha \cap X_\beta = \emptyset$

for $\alpha \neq \beta$. Given a *rotation vector* $\theta \in \mathbb{T}^{\mathcal{A}}$ (with $\mathbb{T}^{\mathcal{A}}$ denoting the torus $\mathbb{R}^{\mathcal{A}}/2\pi\mathbb{Z}^{\mathcal{A}}$) and a *translation vector* $\eta \in \mathbb{C}^{\mathcal{A}}$, we say $(X,T)$ is a *piecewise isometry* if $T$ is such that

$$T(z) := T_\alpha(z) = e^{i\theta_\alpha} z + \eta_\alpha, \text{ if } z \in X_\alpha,$$

so that $T$ is a piecewise isometric rotation or translation (see [16]).

For a given PWI we may partition $X$ into a *regular* and an *exceptional set* [7]. If we consider the zero measure set given by the union $\mathcal{E}$ of all preimages of the set of discontinuities $D$, then its closure $\overline{\mathcal{E}}$ (which may be of positive measure) is called the *exceptional set* for the map. The complement of the exceptional set is called the *regular set* for the map and consists of disjoint polygons or disks that, if $X$ is compact, are periodically coded by their itinerary through the atoms of the PWI. There is numerical evidence that the exceptional set may have positive Lebesgue measure for typical PWIs [5]. In [18], the author shows that this is the case for certain rectangle-exchange transformations.

Even when the exceptional set has positive Lebesgue measure, there is numerical evidence that Lebesgue measure on the exceptional set may not be ergodic - there can be invariant curves that prevent trajectories from spreading across the whole of the exceptional set [7]. In [3, 7], the existence of a large number of these invariant curves, apparently nowhere smooth, are investigated.

In [1] Adler, Kitchens and Tresser found renormalization operators for three rational rotation parameters for a non ergodic piecewise affine map of the Torus. Lowenstein and Vivaldi [21] gave a computer assisted proof of the renormalization of a family of piecewise isometries of a rhombus with one translation parameter and a fixed rational rotation parameter. Hooper [19] investigated a two dimensional parameter space of polygon exchange maps, a family of PWIs with no rotation, invariant under a renormalization operation. In [2] the authors showed how to construct minimal rectangle exchange maps, associated to Pisot numbers, using a cut-and-project method and prove that these maps are renormalizable. The maps described in these papers are PWIs with no rotational component, exhibiting very particular behaviour among more general PWIs, making it difficult to generalize their techniques.

In this survey we present recent results on the study of the dynamics of planar piecewise isometries. We introduce a new notion of renormalization to study a class of PWIs called Translation Cone Exchange Transformations. We also introduce the notion of embedding IETs into PWIs and use IET renormalization techniques to establish the existence of invariant curves for PWIs which are not the union of line segments or circle arcs.

## 2 Interval exchange transformations

In this section we recall some notions of the theory of interval exchange transformations following [13], [27] and [29].

As in [13, 29], let $\mathcal{A}$ be an alphabet on $d \geq 2$ symbols, and let $I \subset \mathbb{R}$ be an interval having 0 as left endpoint. In what follows we use the notation $\mathbb{R}^{\mathcal{A}} \simeq \mathbb{R}^d$ and $\mathbb{R}_+^{\mathcal{A}} \simeq \mathbb{R}_+^d$. We choose a partition $\{I_\alpha\}_{\alpha \in \mathcal{A}}$ of $I$ into subintervals which we assume to be closed on the left and open on the right. An *interval exchange transformation* (IET) is a bijection of $I$ defined by

(1) A vector $\lambda = (\lambda_\alpha)_{\alpha \in \mathcal{A}} \in \mathbb{R}_+^{\mathcal{A}}$ with coordinates corresponding to the lengths of the subintervals, that is, for all $\alpha \in \mathcal{A}$, $\lambda_\alpha = |I_\alpha|$. We write $I = I(\lambda) = [0, |\lambda|)$, where $|\lambda| = \sum_{\alpha \in \mathcal{A}} \lambda_\alpha$.

(2) A pair $\pi = \begin{pmatrix} \pi_0 \\ \pi_1 \end{pmatrix}$ of bijections $\pi_\varepsilon : \mathcal{A} \to \{1, ..., d\}$, $\varepsilon = 0, 1$, describing the ordering of the subintervals $I_\alpha$ before and after the application of the map. This is represented as

$$\pi = \begin{pmatrix} \alpha_1^0 & \alpha_2^0 & ... & \alpha_d^0 \\ \alpha_1^1 & \alpha_2^1 & ... & \alpha_d^1 \end{pmatrix}.$$

We call $\pi$ a *permutation* and identify it, at times, with its *monodromy invariant* $\tilde{\pi} = \pi_1 \circ \pi_0^{-1} : \{1, ...d\} \to \{1, ...d\}$. In algebra literature it is common to reserve the term permutation for the monodromy invariant $\tilde{\pi}$, however, unlike the present notation, this would not be invariant under the induction and renormalization algorithms used in the study of IETs. We denote by $\mathfrak{S}(\mathcal{A})$ the set of irreducible permutations, that is $\pi \in \mathfrak{S}(\mathcal{A})$ if and only if $\tilde{\pi}(\{1, ..., k\}) \neq \{1, ..., k\}$ for $1 \leq k < d$.

Define a linear map $\Omega_\pi : \mathbb{R}^{\mathcal{A}} \to \mathbb{R}^{\mathcal{A}}$ by

$$(\Omega_\pi(\lambda))_{\alpha \in \mathcal{A}} = \sum_{\pi_1(\beta) < \pi_1(\alpha)} \lambda_\beta - \sum_{\pi_0(\beta) < \pi_0(\alpha)} \lambda_\beta. \tag{1}$$

Given a permutation $\pi \in \mathfrak{S}(\mathcal{A})$ and $\lambda \in \mathbb{R}_+^{\mathcal{A}}$ the interval exchange transformation associated is the map $f_{\lambda,\pi}$ that rearranges $I_\alpha$ according to $\pi$, that is $f_{\lambda,\pi}(x) = x + \upsilon_\alpha$, for any $x \in I_\alpha$, where $\upsilon_\alpha = (\Omega_\pi(\lambda))_\alpha$. We write $f = f_{\lambda,\pi}$ and also denote an IET by the pair $(I, f_{\lambda,\pi})$.

Given $(\lambda, \pi) \in \mathbb{R}_+^{\mathcal{A}} \times \mathfrak{S}(\mathcal{A})$ and for $\varepsilon = 0, 1$, denote by $\beta_\varepsilon$ the last symbol in the expression of $\pi_\varepsilon$. Assume the intervals $I_{\beta_0}$ and $I_{\beta_1}$ have different lengths. Let $I^{(1)}$ be the interval obtained by removing the smallest of these intervals from $I$. The first return map of $f_{\lambda,\pi}$ to $I^{(1)}$ is again an IET, $f_{\lambda^{(1)},\pi^{(1)}}$. This defines a map $\mathcal{R}(\lambda,\pi) = (\lambda^{(1)},\pi^{(1)})$ called *Rauzy induction*. We assume the *infinite distinct orbit condition (IDOC)*, introduced by Keane in [20], which assures that the iterates $\mathcal{R}^n$ are defined for all $n \geq 0$. We denote $\mathcal{R}^n(\lambda,\pi) = (\lambda^{(n)},\pi^{(n)})$ and by $\{I_\alpha^{(n)}\}_{\alpha\in\mathcal{A}}$ the partition of the domain $I^{(n)}$ of $f_{\lambda^{(n)},\pi^{(n)}}$.

The *Rauzy class* (see [29]) of a permutation $\pi \in \mathfrak{S}(\mathcal{A})$, is the set $\mathfrak{R}(\pi)$ of all $\pi^{(1)} \in \mathfrak{S}(\mathcal{A})$ such that there exist $\lambda, \lambda^{(1)} \in \mathbb{R}_+^{\mathcal{A}}$ and $n \in \mathbb{N}$ such that $\mathcal{R}^n(\lambda,\pi) = (\lambda^{(1)},\pi^{(1)})$. A Rauzy class $\mathfrak{R}$ can be visualized in terms of a directed labelled graph, the *Rauzy graph* (see [27]). Its vertices are in bijection with $\mathfrak{R}$ and it is formed by edges that connect permutations which are obtained one from another by $\mathcal{R}$ and are labeled respectively by 0 or 1 according to the type of the induction. A *path* $\varrho = (\varrho_1,...,\varrho_n)$ is a sequence of compatible edges of the Rauzy graph, that is, such that the starting vertex of $\varrho_{i+1}$ is the ending vertex of $\varrho_i$, $i = 1,...,n-1$. We say a path is *closed* if the starting vertex of $\varrho_1$ is the ending vertex of $\varrho_n$. The set of all paths in this graph is denoted by $\Pi(\mathfrak{R})$.

The *Rauzy cocycle* $B_R(\lambda,\pi)$ is a matrix function such that each entry $(B_R^{(n)}(\lambda,\pi))_{\alpha,\beta}$ of $B_R^{(n)}(\lambda,\pi)$ counts the number of visits of $I_\alpha^{(n)}$ to $I_\beta$ during the Rauzy induction time.

The *projection of the Rauzy cocycle* on the Torus $\mathbb{T}^{\mathcal{A}} \simeq \mathbb{R}^{\mathcal{A}}/2\pi\mathbb{Z}^{\mathcal{A}}$ is given by

$$B_{\mathbb{T}^{\mathcal{A}}}(\lambda,\pi)\cdot\theta = B_R(\lambda,\pi)\cdot\theta \mod 2\pi,$$

for any $(\lambda,\pi) \in \mathbb{R}_+^{\mathcal{A}} \times \mathfrak{R}$, $n \geq 0$ and $\theta \in \mathbb{T}^{\mathcal{A}}$.

A *translation surface* (see for instance [12], [28]), is a surface with a finite number of conical singularities endowed with an atlas such that coordinate changes are given by translations in $\mathbb{R}^2$. Given an IET it is possible to associate, via a suspension construction, a translation surface, with genus $g(\mathfrak{R})$ only depending on the combinatorial properties of the underlying IET (see [28]).

## 3 Translated cone exchange transformations

In this section we present a result on the renormalization of a particular family of PWIs which we designate by *Translated cone exchange transformations* following [24].

Consider a family of dynamical systems $\mathcal{F} = \{f_\mu : X \to X\}$ parametrized by $\mu \in \mathcal{P}$, where $\mathcal{P}$ is called the parameter space of $\mathcal{F}$. A *renorma-*

*lization scheme* for $\mathcal{F}$ is a decreasing chain of subsets of $X$, $X = Y_0(\mu) \supset Y_1(\mu) \supset Y_2(\mu) \supset ...$, together with a renormalization operator $\mathcal{R} : \mathcal{P} \to \mathcal{P}$ such that the first return map of a point in $Y_{n+1}(\mu)$ under iteration by $f_{\mathcal{R}^n(\mu)} : Y_n(\mu) \to Y_n(\mu)$ is given by $f_{\mathcal{R}^{n+1}(\mu)} : Y_{n+1}(\mu) \to Y_{n+1}(\mu)$. Renormalization is a powerful tool in the study of nonlinear maps (see [10]), such as diffeomorphisms of the circle [26], one-frequency Schrödinger cocycles [11] and analytic unimodal maps [15].

Set $\omega = (\omega_1, ..., \omega_d) \in \mathbb{W}$, where $\mathbb{W}$ is the open polytope defined by

$$\mathbb{W} = \left\{ \omega \in \mathbb{R}^d_+ : 0 < \sum_{j=1}^{d} \omega_j < \pi \right\}, \tag{2}$$

and let $\vartheta = \dfrac{\pi}{2} - \dfrac{|\omega|}{2}$, where $|\omega|$ is the $\ell_1$ norm of $\omega$.

In order to introduce the family of TCEs, consider a partition of the upper half plane $\mathbb{H}$ into $d+2$ cones $\mathcal{P} = \{P_0, P_1, \ldots, P_d, P_{d+1}\}$, where $P_j = \{z \in \mathbb{C} : \arg(z) \in W_j\}$, and $W_j$ for $j = 0, \ldots, d+1$ are defined as

$$W_j = \begin{cases} [0, \vartheta), & \text{for } j = 0, \\ [\vartheta, \vartheta + \omega_1], & \text{for } j = 1, \\ (\vartheta + \sum_{k=1}^{j-1} \omega_k, \vartheta + \sum_{k=1}^{j} \omega_k], & \text{for } j \in \{2, ..., d\}, \\ (\pi - \vartheta, \pi], & \text{for } j = d+1. \end{cases}$$

We set $\nu = \tan(\vartheta)$. Note that $\nu$ depends on $|\omega|$, and when necessary to stress this dependence we write $\nu = \nu(|\omega|)$.

Let $G : \mathbb{H} \to \mathbb{H}$ be the following family of translation maps

$$G(z) = \begin{cases} z - 1, & z \in P_0, \\ z - \eta', & z \in P_j, \ j \in \{1, ..., d\}, \\ z + \eta, & z \in P_{d+1}, \end{cases}$$

depending on the parameters $\vartheta, \eta$ and $\eta'$ with $\vartheta > 0$, $\eta \in \mathbb{R}^+ \backslash \mathbb{Q}$ and $0 < \eta' < \eta$.

Consider a permutation $\pi \in \mathfrak{S}(\{1, ..., d\})$ with a monodromy invariant $\tilde{\pi}$, and let $\theta_j(\omega, \tilde{\pi})$ be the angle associated to the monodromy invariant $\tilde{\pi}$ for the cone $P_j$ for $j = 1, \ldots, d$. We have

$$\theta_j(\omega, \tilde{\pi}) = \sum_{\tilde{\pi}(k) < \tilde{\pi}(j)} \omega_k - \sum_{k<j} \omega_k. \tag{3}$$

Let $E : \mathbb{H} \to \mathbb{H}$ be the following family of maps

$$E(z) = \begin{cases} z, & z \in P_0 \cup P_{d+1}, \\ z e^{i\theta_j(\omega, \tilde{\pi})}, & z \in P_j, \ j \in \{1, ..., d\}, \end{cases}$$

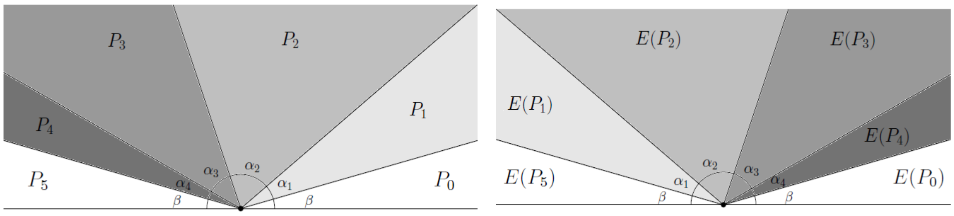



Figure 1: On the left a partition $\mathcal{P}$ with $d = 5$. On the right the action of map $E$ on this partition with $\tilde{\pi}(1) = 4$, $\tilde{\pi}(2) = 3$, $\tilde{\pi}(3) = 2$ and $\tilde{\pi}(4) = 1$.

depending on $\theta_j(\omega, \tilde{\pi})$. This map also depends on $\omega$ and $\vartheta$ as the partition elements $P_j$ depend on these parameters. Note that we have

$$\vartheta + \arg(E(z)) = f_{\omega,\pi}(\arg(z) - \vartheta),$$

for $z \in P_j$, $j = 1, ..., d$. Hence $E$ exchanges these cones according to the monodromy invariant $\tilde{\pi}$.

From the translation and exchange families of maps we get our family of TCEs, $F : \mathbb{H} \to \mathbb{H}$, given by

$$F(z) = G \circ E(z).$$

We define the *central cone* $P_c$ of $F$ as

$$P_c = P_1 \cup ... \cup P_d,$$

the *first hitting time of* $z \in \mathbb{H}$ *to* $P_c$, as the map $k : \mathbb{H} \to \mathbb{N}$ given by

$$k(z) = \inf\{n \geq 1 : F^n(z) \in P_c\}, \tag{4}$$

and the *first return map of* $z \in P_c$ *to* $P_c$, as the map $F_c : P_c \to P_c$ such that

$$F_c(z) = F^{k(z)}(z). \tag{5}$$

The typical notion of renormalization may not capture all possible self similar behaviour in PWIs. TCEs apparently exhibit invariant regions on which the dynamics is self similar after rescaling. Thus, we say a TCE is renormalizable if $F_c$, the first return map to $P_c$ described above, is conjugated to itself by a scaling map.

**Theorem 3.1** ([24]) For all $\omega \in \mathbb{W}$, $\eta = 1/(k + \Phi)$ and $\eta' = 1 - k\eta$ with $k \in \mathbb{N}$, there is an open set $U$ containing the origin such that $F$ is renormalizable for all $z \in U$, that is

$$F_c(\Phi^2 z) = \Phi^2 F_c(z). \tag{6}$$

The proof of this theorem uses a one dimensional approach to the study of these TCEs. We define sequences coding information related to the first return map of a given line contained in the cone $P_c$. We are then able to relate the renormalizability of a map of this family with the periodicity of these sequences and indeed, for the parameters in the statement of the theorem, these are proved to be periodic. As a consequence of this we show that for these parameters $F_c$ is a PWI with respect to a partition $\mathcal{P}_{F_c}$ of countably many atoms.

We say that a collection of atoms $\mathcal{B} \subseteq \mathcal{P}$ is a *barrier* for a PWI $(T, \mathcal{P})$ if $X \setminus \bigcup_{B\in\mathcal{B}} B$ is the union of two disjoint connected components $A_1$, $A_2$ such that

$$A_1 \cap T(A_2) = T(A_1) \cap A_2 = \emptyset,$$

and for any $P \in \mathcal{P}$ such that $P \subseteq A_j$ and $T(P) \cap (\bigcup_{B\in\mathcal{B}} \overline{B}) \cap \overline{A_j} = \emptyset$ then $T(P) \cap (\bigcup_{B\in\mathcal{B}} B) = \emptyset$, for $j = 1, 2$.

Denote the ray in $\mathbb{H}$ passing through the origin and with slope $a \in \mathbb{R}$ by

$$L_a = \{z \in \mathbb{H} : \mathrm{Im}(z) = a\mathrm{Re}(z)\}, \tag{7}$$

and by $\partial\mathcal{P}$ the union of the boundaries of the elements of the partition $\mathcal{P}$ and by $L_\nu$ and $L_{-\nu}$, respectively, the rays $\overline{P_0} \cap \overline{P_1}$ and $\overline{P_d} \cap \overline{P_{d+1}}$.

For $\omega \in \mathbb{W}$, $\eta = 1/(k + \Phi)$ and $\eta' = 1 - k\eta$, $k \in \mathbb{N}$, we denote by $\mathfrak{A}(\eta, \eta')$ the subset of $\mathbb{W}$ such that for all $\omega \in \mathfrak{A}(\eta, \eta')$ there are $d' \geq 2$, $\lambda \in \mathbb{R}_+^{d'}$, $\pi \in \mathfrak{S}(\{1, ..., d'\})$ and a continuous embedding $\gamma$ of $f_{\lambda,\pi} : I \to I$ into $F_c : P_c \to P_c$ such that

i) the collection $\mathcal{B} = \{P \in \mathcal{P}_{F_c} : P \cap \gamma(I) \neq \emptyset\}$, is a barrier for $F_c$,

ii) $\gamma(0) \in L_{-\nu}$ and $\lim_{a\to|\lambda|} \gamma(a) \in L_\nu$,

iii) $\gamma(I) \subset \Phi^2 U$, where $U$ is the open set from Theorem 3.1.

In the next theorem we show, as a consequence of Theorem 3.1, that the existence of one continuous embedding of an IET into a first return map $F_c$ of a TCE, satisfying the property that the image of the embedding is contained in a barrier, implies the existence of infinitely many embeddings of the same IET into $F_c$, as well as infinitely many bounded and forward invariant regions.

**Theorem 3.2** ([24]) Let $\eta = 1/(k + \Phi)$, $\eta' = 1 - k\eta$ with $k \in \mathbb{N}$ and assume that $\mathfrak{A}(\eta, \eta')$ is non-empty. For all $\omega \in \mathfrak{A}(\eta, \eta')$,

i) There exist sets $V_1, V_2, ...,$ which are forward invariant for $F_c$ and $y^* > 0$ such that for all $z \in P_c$, satisfying $0 < \mathrm{Im}(z) < y^*$, there is an $n \in \mathbb{N}$ for which $z \in V_n$.

ii) For all $n \in \mathbb{N}$ there exist constants $0 < \underline{b}_n < \bar{b}_n$ such that for all $z \in V_n$ and $k \in \mathbb{N}$,

$$\underline{b}_n < |F^k(z)| < \bar{b}_n. \tag{8}$$

iii) There exist infinitely many continuous embeddings of IETs into $F_c$.

The proof of Theorem 3.2 relies on the Jordan curve Theorem, and on the properties of the barrier containing the image of the embedding in order to prove the existence of one invariant set $V_1$. Then the renormalizability of $F$ implies the existence of infinitely many such sets.

# 4 Embedding interval exchange transformations into piecewise isometries

Recently [9], we developed a new mechanism that allow us to study the dynamics of PWIs using tools from IETs - embeddings - and we used combinatorial properties of IETs to prove that in order for a PWI to realize a continuous embedding of an IET with the same permutation its parameters must satisfy a particular condition. In this section we give an overview of these mathematical tools.

It is commonly accepted that the phase space of typical Hamiltonian systems is divided into regions of regular and chaotic motion [14]. Area preserving maps which can be obtained as Poincaré sections of Hamiltonian systems, exhibit this property as well, with KAM curves splitting the domain into regions of chaotic and periodic dynamics (see for instance [23]). A general and rigorous treatment of this has been however missing. Area preserving PWIs that have been studied as linear models for the standard map (see [4]), can exhibit a similar phenomenon. Unlike IETs which are typically ergodic, there is numerical evidence, as noted in [7], that Lebesgue measure on the exceptional set is typically not ergodic in some families of PWIs - there can be non-smooth invariant curves that prevent trajectories from spreading across the whole of the exceptional set. These curves were first observed in [3] for an isolated parameter and later found in [7] to be apparently abundant for a large family of PWIs.

We now relate the existence of invariant curves to the general problem of embedding IET dynamics within PWIs. We start by introducing some definitions.

An injective map $\gamma : I \to X$ is a *piecewise continuous embedding* of $(I, f)$ into $(X, T)$ if $\gamma|_{I_\alpha}$ is a homeomorphism for each $\alpha \in \mathcal{A}$ such that $\gamma(I_\alpha) \subset X_\alpha$

and

$$\gamma \circ f(x) = T \circ \gamma(x), \tag{9}$$

for all $x \in I$. In this case note that $\gamma(I) \subset X$ is an invariant set for $(X, T)$.

If $\gamma$ is a piecewise continuous embedding that is continuous on $I$, we say it is a *continuous embedding* (or *embedding* when this does not cause any ambiguity). Otherwise we say it is a *discontinuous embedding*.

We say $\gamma$ is a *differentiable embedding* if it is a piecewise continuous embedding and $\gamma|_{I_\alpha}$ is continuously differentiable. We characterize certain differentiable embeddings as, in some sense, trivial: given $I' \subseteq I$ we say a map $\gamma : I' \to \mathbb{C}$ is an *arc map* if there exists $\xi \in \mathbb{C}$, $r, a > 0$ and $\varphi \in [0, 2\pi)$ such that for all $x \in I'$,

$$\gamma(x) = re^{i(ax+\varphi)} + \xi.$$

We say an embedding $\gamma : I \to \mathbb{C}$ of an IET into a PWI is an *arc embedding* if there exists a finite partition of $I$ into subintervals such that the restriction of $\gamma$ to each subinterval is an arc map. We say an embedding $\gamma$ of an IET into a PWI is a *linear embedding* if $\gamma$ is a piecewise linear map. Moreover an embedding is *non-trivial* if it is not an arc embedding or a linear embedding. Figure 4 shows an illustration of a non-trivial embedding.

From the definitions it is clear that the image $\gamma(I)$ of an embedding is an invariant curve for the underlying PWI and that if the embedding is non-trivial this curve is not the union of line segments or circle arcs. For any IET it is straightforward to construct a PWI in which it is trivially embedded. The same is not true for non-trivial embeddings, for which results have been much scarcer.

We say a $d$-PWI is a PWI with a partition of $d$ atoms. Similarly, a $d$-IET is an IET with a partition of $d$ subintervals. In [9] we showed that there are no non-trivial continuous embeddings of minimal 2-IETs into orientation preserving planar PWIs.

**Theorem 4.1** ([9]) A minimal 2-IET has no non-trivial continuous embedding into a 2-PWI.

The next theorem states that a 3-PWI has at most one non-trivially continuously embedded minimal 3-IET with the same underlying permutation.

**Theorem 4.2** ([9]) A 3-PWI has at most one non-trivially continuously embedded minimal 3-IET with the same underlying permutation.

The proofs of Theorems 4.1 and 4.2 rely on the use of combinatorial properties of IETs to prove that in order for a PWI to realize a continuous

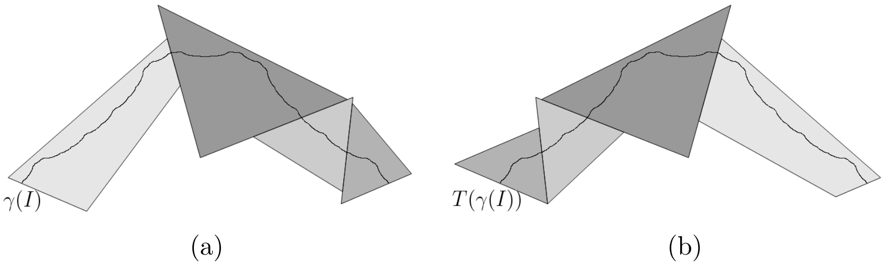



Figure 2: An illustration of the action of a PWI $T$ with rotation vector $\theta \approx (4.85, 0.92, 1.31, 1.28)$ on its partition and on an invariant curve $\gamma(I)$. The map $\gamma$, estimated using technical tools from [25], is a non-trivial embedding of a self-inducing IET associated to the monodromy invariant $\tilde{\pi}(j) = 4 - (j-1)$, $j = 1, ..., 4$ and a translation vector of algebraic irrationals $\lambda \approx (0.43, 0.34, 0.12, 0.11)$.

embedding of an IET with the same permutation, its parameters must satisfy a necessary condition which may be found in [9].

## 5 Existence of invariant curves

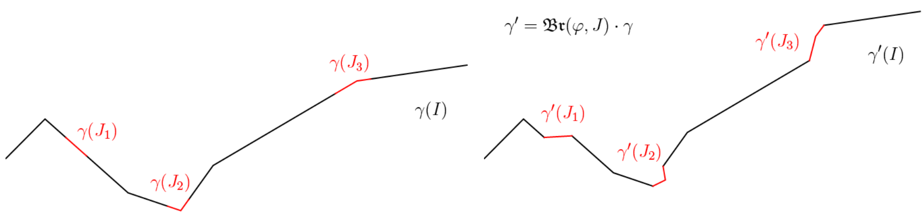
In this section we show that almost every IET with an associated translation surface of genus $g \geq 2$ can be non-trivially and isometrically embedded in a family of piecewise isometries giving an overview of the technical tools used to prove the main results following our work in [25] .

In order to prove the main result presented in this section, we need to define the *Breaking operator* $\mathfrak{Br}$: given an ordered sequence $J = \{J_k\}_k$ of subintervals of $I$, an angle $\varphi \in [-\pi, \pi)$ and a piecewise linear map $\gamma : I \to \mathbb{C}$ the image of $\mathfrak{Br}(\varphi, J) \cdot \gamma$ is a piecewise linear curve, obtained from $\gamma(I)$ by rotating the segments $\gamma(J_k)$ by $\varphi$,

$$\mathfrak{Br}(\varphi, J) \cdot \gamma(x) = \begin{cases} \gamma(x) \cdot e^{i\varphi} + \overline{\epsilon}_k(\varphi, J), & x \in J_k, \\ \gamma(x) + \underline{\epsilon}_k(\varphi, J), & x \in L_k, \end{cases}$$

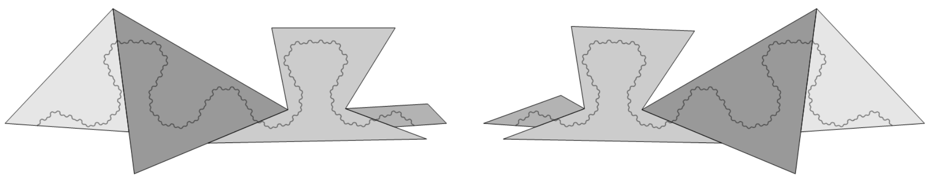
where $\overline{\epsilon}_k(\varphi, J)$ and $\underline{\epsilon}_k(\varphi, J)$ are determined by continuity and $L = \{L_k\}_k$ is the ordered sequence of subintervals determined by $I \backslash J$.

Given $(\lambda, \pi) \in \mathbb{R}_+^{\mathcal{A}} \times \mathfrak{S}(\mathcal{A})$, consider the sequence $J^{(n)} = \{J_k^{(n)}\}_{k<r(n-1)}$ obtained by ordering the collection of sets $\{f_{\lambda,\pi}^k(I^{(n-1)} \backslash I^{(n)})\}_{k<r(n-1)}$, where $r(n-1)$ is the smallest $r \geq 1$ such that $f_{\lambda,\pi}^k(I^{(n-1)} \backslash I^{(n)}) \subset I^{(n)}$. Recall that




(a) (b)

Figure 3: Action of the operator $\mathfrak{Br}$: (a) shows the image $\gamma(I)$ of a piecewise linear curve; (b) shows the image $\mathfrak{Br}(\varphi, J) \cdot \gamma(I)$ , with $\varphi = \frac{\pi}{4}$ and $J = \{J_1, J_2, J_3\}$.



(a) (b)

Figure 4: Action of a 4-PWI $\theta$-adapted to a self-similar 4-IET. The curve depicted is $\gamma_\theta(I)$ and it is the image of a non-trivial embedding of the IET into this PWI.

$B_{\mathbb{T}^\mathcal{A}}$ is the projection of the Rauzy cocycle on the Torus defined in the introduction. Given $\theta \in \mathbb{T}^\mathcal{A}$ let

$$\theta^{(0)} = \theta, \quad \theta^{(n)} = B^{(n)}_{\mathbb{T}^\mathcal{A}}(\lambda, \pi) \cdot \theta,$$

With $\beta_{1,m} = (\pi_1^{(m)})^{-1}(d)$, we define the **breaking sequence** of curves $\{\gamma_\theta^{(n)}(x)\}$, by

$$\gamma_\theta^{(0)}(x) = x, \quad \gamma_\theta^{(n)}(x) = \mathfrak{Br}\left(\theta^{(n-1)}_{\beta_{1,n-1}}, J^{(n)}\right) \cdot \gamma_\theta^{(n-1)}(x), \quad x \in I.$$

Denote by $\Theta'_{\lambda,\pi}$ the set of all $\theta \in \mathbb{T}^\mathcal{A}$ such that:

- for all $n \geq 0$, $\gamma_\theta^{(n)} : I \to \mathbb{C}$ is an injective map;
- there exists a topological embedding $\gamma_\theta : I \to \mathbb{C}$ such that

$$\gamma_\theta(x) = \lim_{n \to +\infty} \gamma_\theta^{(n)}(x), \quad x \in I.$$

Given $\theta \in \Theta'_{\lambda,\pi}$, we say that a PWI $T : X \to X$ together with a partition $\{X_\alpha\}_{\alpha \in \mathcal{A}}$ is **$\theta$-adapted to** $(\lambda, \pi)$ if for all $\alpha \in \mathcal{A}$,

1. $X_\alpha \supseteq \gamma_\theta(I_\alpha)$ ;

2. For any $z \in \mathbb{C}$, we have $T(z) = T_\alpha(z)$, for all $z \in X_\alpha$, where

$$T_\alpha(z) = e^{i\theta_\alpha} \left( z - \gamma_\theta \left( \sum_{\pi_0(\beta)<\pi_0(\alpha)} \lambda_\beta \right) \right) + \gamma_\theta \left( f_{\lambda,\pi} (\sum_{\pi_0(\beta)<\pi_0(\alpha)} \lambda_\beta) \right).$$

Given $(\lambda, \pi) \in \mathbb{R}_+^{\mathcal{A}} \times \mathfrak{R}$ it is possible to associate, via a suspension construction, a translation surface, with genus $g(\mathfrak{R}) \geq 1$ depending only on the Rauzy class $\mathfrak{R}$.

The next theorem states the existence of invariant curves for PWIs which are not unions of circle arcs or line segments.

**Theorem 5.1** ([25]) For almost every IET $(I, f_{\lambda,\pi})$ with a Rauzy class $\mathfrak{R}$ satisfying $g(\mathfrak{R}) \geq 2$, there exists a set $\mathcal{W} \subseteq \mathbb{T}^{\mathcal{A}}$ of dimension $g(\mathfrak{R})$ such that for all $\theta \in \mathcal{W}$ there exists a map $\gamma_\theta : I \to \mathbb{C}$, which is a non-trivial embedding of $(I, f_{\lambda,\pi})$ into any PWI that is $\theta$-adapted to $(\lambda, \pi)$.

## References

[1] Adler, R., Kitchens, B., Tresser, C. (2001). *Dynamics of non-ergodic piecewise affine maps of the torus.* Ergodic Theory and Dynamical Systems **21**, 959–999.

[2] Alevy, I., Kenyon, R., Yi, R., (2018). *A family of minimal and renormalizable rectangle exchange maps.* arXiv:1803.06369

[3] Ashwin, P. (1996). *Non-smooth invariant circles in digital overflow oscillations.* Proceedings of the 4th Int. Workshop on Nonlinear Dynamics of Electronic Systems, Sevilla, 417–422.

[4] Ashwin, P., (1997). *Elliptic behaviour in the sawtooth standard map*, Phys. Lett. A, 232, pp. 409–416.

[5] Ashwin, P., Fu, (2002). *On the Geometry of Orientation-Preserving Planar Piecewise Isometries.* Journal of Nonlinear Science, 12, no. 3, 207–240.

[6] Ashwin, P. , Goetz, A. (2006). *Polygonal invariant curves for a planar piecewise isometry.* Trans. Amer. Math. Soc. **358** no. 1, 373–390.

[7] Ashwin, P. , Goetz, A. (2005). *Invariant curves and explosion of periodic islands in systems of piecewise rotations*. SIAM J. Appl. Dyn. Syst., 4(2), 437–458.

[8] Ashwin, P., Goetz, A. (2010). *Cone exchange transformations and boundedness of orbits*. Cambridge University Press, 30(5), 1311–1330.

[9] Ashwin P., Goetz A., Peres P., Rodrigues A., (2018). *Embeddings of Interval Exchange Transformations into Planar Piecewise Isometries*, Ergodic Theory and Dynamical Systems, 1-27. doi:10.1017/etds.2018.112

[10] Avila, A. (2010) *Dynamics of Renormalization Operators*. Proceedings of the International Congress of Mathematicians, Hyderabad, India.

[11] Avila, A. (2015) *Global theory of one-frequency Schrödinger operators*. Acta Math. Volume 215, Number 1, 1–54.

[12] Avila, A., Forni, G. (2007) *Weak mixing for interval exchange transformations and translation flows*, Ann. of Math. (2), 165(2), 637–664.

[13] Avila, A., Leguile, M. (2016) *Weak mixing properties of interval exchange transformations and translation flows*, Bulletin de la Societe mathematique de France, **146**.

[14] Chirikov B.V. (1983) *Chaotic dynamics in Hamiltonian systems with divided phase space*. In: Garrido L. (eds) Dynamical System and Chaos. Lecture Notes in Physics, vol 179. Springer, Berlin, Heidelberg.

[15] Coullet, P., Tresser, C. (1978). *Itérations d?endomorphismes et groupe de renormalisation*. C. R. Acad. Sci. Paris 287A, 577–580.

[16] Goetz, A. (2000). *Dynamics of piecewise isometries*. Illinois journal of mathematics, **44**, 465–478.

[17] Gutkin, E., Haydn, N., (1997) *Topological entropy of polygon exchange transformations and polygonal billiards*. Ergod. Th. & Dynam. Sys. 17, 849–867.

[18] Haller, H. (1981). *Rectangle Exchange Transformations*. Monatsh. Math. **91**(3), 215–232.

[19] Hooper, W. Patrick (2013) *Renormalization of polygon exchange maps arising from corner percolation*. Inventiones mathematicae, 191(2), 255-320.

[20] Keane, M. S., (1975) *Interval exchange transformations*, Math. Z. 141, 25–31.

[21] Lowenstein, J., Vivaldi, F., (2016). *Renormalization of a one-parameter family of piecewise isometries*, Dynamical Systems, 31:4, 393–465.

[22] Masur, H., (1982) *Interval exchange transformations and measured foliations*, Ann. of Math. 115, 169–200.

[23] MacKay, R., Percival, I., (1985). *Converse KAM - theory and practice*, Comm. Math. Phys. 98(4) 469–512.

[24] Peres, P., Rodrigues, A. (Preprint 2019) *On the dynamics of translated cone exchange transformations* https://arxiv.org/pdf/1809.05496.pdf

[25] Peres, P., Rodrigues, A. (Preprint 2019) *Existence of non-trivial embeddings of interval exchange transformations into piecewise isometries* https://arxiv.org/pdf/1901.10406.pdf

[26] Sinai, Ya. G., Khanin, K. M. (1989). *Smoothness of conjugacies of diffeomorphisms of the circle with rotations.* Uspekhi Mat. Nauk 44, 69–99.

[27] Sinai, Y., Ulcigrai, C. (2005). *Weak mixing in interval exchange transformations of periodic type.* Letters in Mathematical Physics, 74, 2, 111–133.

[28] Veech, W. A., (1982) *Gauss measures for transformations on the space of interval exchange maps*, Ann. of Math. 115, 201–242.

[29] Viana, M., (2006) *Ergodic Theory of interval exchange maps*, Revista Matematica Complutense 19, 7–100.

# An Invitation to Symplectic Toric Manifolds

*Ana Cannas da Silva*[1]
ETH Dep. Mathematics
Raemistrasse 101
8092 Zurich, Switzerland
e-mail: `ana.cannas@math.ethz.ch`

**Resumo:** Este artigo é uma introdução a variedades simplécticas tóricas para não especialistas, começando com uma breve síntese de variedades simplécticas e acções hamiltonianas. As variedades simplécticas tóricas formam já um tema extenso, ao qual a modesta lista de referências abaixo não faz justiça – o objectivo deste texto não é ser exaustivo ou justo, mas simplesmente deixar entrever o que são estes espaços e a razão pela qual o leitor poderá querer adicioná-los ao seu repertório de objectos geométricos.

**Abstract** This is an elementary introduction to symplectic toric manifolds for nonspecialists, starting with a brief review of symplectic manifolds and hamiltonian torus actions. Symplectic toric manifolds are by now a vast subject, for which the undersized list of references below does no justice – the aim of this text is not to be exhaustive or fair, but simply to give a glimpse into what these spaces are like and why it can be a good idea to add them to your repertoire of geometric objects.

**palavras-chave:** Variedade simpléctica; acção hamiltoniana; polítopo.

**keywords:** Symplectic manifold; hamiltonian action; polytope.

## 1 What is Symplectic Geometry?

Geometry concerns the study and measure of space. *Symplectic* refers to an additional structure that can be put on some even-dimensional spaces. Symplectic geometry is intrinsically related to complex geometry and, just like complex geometry, is sometimes counterintuitive. Whereas local complex geometry is basically modelled on $\mathbb{C}$, $\mathbb{C}^2$, $\mathbb{C}^3$, etc, local symplectic geometry is basically modelled on $\mathbb{R}^2$, $\mathbb{R}^4$, $\mathbb{R}^6$, etc.

[1]This text is based on the lecture titled *Symplectic Toric Manifolds* at the conference *Matemáticos Portugueses pelo Mundo* (Portuguese Mathematicians around the World) in Porto on 24/June/2019. Many thanks to Diogo Oliveira e Silva, Jorge Milhazes de Freitas and Samuel Lopes for organizing this excellent mathematical *S. João* celebration!

A **symplectic form** $\omega$ at a point $p$ of a manifold $M$ is a special type of differential 2-form, i.e., a device that takes two tangent vectors $\vec{u}, \vec{v} \in T_pM$ as input and returns a real number as output, that may be interpreted as

$$\omega(\vec{u}, \vec{v}) = \text{ kind of } \textit{signed area} \text{ of parallelogram spanned by } \vec{u} \text{ and } \vec{v}.$$

By *signed area* we mean, in particular, a number that may be positive, negative, or zero, contrasting with usual (euclidean, riemannian, ...) geometries.

In the case of the basic model of $\mathbb{R}^2$ with its so-called **standard symplectic form**,

$$\omega_0 := dx \wedge dy \ ,$$

this *signed area* is

$$\omega_0(\vec{u}, \vec{v}) = \det \begin{pmatrix} u_1 & v_1 \\ u_2 & v_2 \end{pmatrix} = u_1 v_2 - u_2 v_1 \ ,$$

thus actually equal to plus or minus the euclidean area of the parallelogram spanned by $\vec{u}$ and $\vec{v}$. The sign depends on the orientation of the basis $\vec{u}, \vec{v}$ and $\omega_0(\vec{v}, \vec{u}) = -\omega_0(\vec{u}, \vec{v})$. Moreover, there is only *zero* as output in just one dimension, since $\omega_0(\vec{v}, \vec{v}) = 0$ for all $\vec{v}$.

In the next case of $\mathbb{R}^4$, the standard symplectic form,

$$\omega_0 := dx_1 \wedge dy_1 + dx_2 \wedge dy_2 \ ,$$

just adds up the contributions from the projections onto the two coordinate planes $x_1, y_1$ and $x_2, y_2$. If we have vectors $\vec{u} = \vec{u}_1 + \vec{u}_2$ and $\vec{v} = \vec{v}_1 + \vec{v}_2$ (where $\vec{u}_1, \vec{v}_1$ and $\vec{u}_2, \vec{v}_2$ denote the projections onto the coordinate planes $x_1, y_1$ and $x_2, y_2$), then

$$\omega_0(\vec{u}, \vec{v}) = \underbrace{(dx_1 \wedge dy_1)\,(\vec{u}_1, \vec{v}_1)}_{\text{signed area of } A_1} + \underbrace{(dx_2 \wedge dy_2)\,(\vec{u}_2, \vec{v}_2)}_{\text{signed area of } A_2}$$

can be thought of as a sum of signed areas for the projections $A_1$ and $A_2$ onto each of the coordinate planes $x_1, y_1$ and $x_2, y_2$. Other projections are not taken into account.

The higher cases $\mathbb{R}^{2n}$ are analogous. In particular, in $\mathbb{R}^6$ we have the standard symplectic form

$$\omega_0 := dx_1 \wedge dy_1 + dx_2 \wedge dy_2 + dx_3 \wedge dy_3 \ .$$

Physicists often regard $(x_1, x_2, x_3)$ as position coordinates and $(y_1, y_2, y_3)$ as momentum (kind of velocity) coordinates of a particle in 3-dimensional space. The symplectic form $\omega_0$ encodes the mutual entanglement of position and momentum in a somewhat implausible way that actually fits reality. In Section 2, we will describe the motion of a classical mechanical system via Hamilton's equations for position and momentum in terms of a flow on a symplectic manifold.

Historical remark:

> *Symplectic geometry* is a branch of mathematics, that could be viewed as emerging in the XIX century from classical mechanics. The mathematicians William Rowan Hamilton (1805-1865) and Sofia Kovalevskaya (1850-1891) were at the onset of this field and worked on problems related to the motion of rigid bodies. Symplectic geometry experienced a vigorous expansion in the last 50 years and deals nowadays with many other geometric problems, stimulated by interactions with diverse areas of mathematics and physics. The adjective *symplectic* in mathematics is a calque[2] coined by Hermann Weyl, by substituting the Latin root in *complex* by the corresponding Greek root, in order to label the symplectic group.

In general, a **symplectic manifold** is a pair $(M,\omega)$ where $M$ is a manifold (necessarily even-dimensional, say $\dim M = 2n$) and $\omega$ is a closed nondegenerate 2-form on $M$. Whereas closedness is a natural differential condition from analysis, nondegeneracy is an algebraic condition saying that at each point any nonzero tangent vector admits a nontrivial pairing with some other tangent vector – this is what forces the evenness of the dimension.

One of the fundamental theorems in symplectic geometry goes back to Darboux [6] in the late XIX century in the context of differential systems. What is now known as **Darboux's theorem** states that any symplectic manifold looks *locally* near any of its points like a neighborhood of the origin in $\mathbb{R}^{2n}$ equipped with

$$\omega_0 := dx_1 \wedge dy_1 + \ldots + dx_n \wedge dy_n \ .$$

We hence refer to $(\mathbb{R}^{2n},\omega_0)$ as the *local model*. Although this shows that there are no local invariants in symplectic geometry besides the dimension, the local symplectic geometry, i.e. the symplectic geometry of $(\mathbb{R}^{2n},\omega_0)$, is already quite interesting and there remain deep open questions about

[2] A *calque* or *loan translation* is a word or phrase that is introduced through translation of the constituents into another language.

it. Normal form theorems like Darboux's play a central role in symplectic geometry.

On a symplectic manifold $(M,\omega)$, the top power of the symplectic form, $\omega^n$, is necessarily a volume form, called the **symplectic volume**. This follows from the nondegeneracy of $\omega$, and may be also seen through Darboux's theorem with $\omega_0^n = n!\, dx_1 \wedge dy_1 \wedge \ldots \wedge dx_n \wedge dy_n$. Therefore, a symplectic manifold is *symplectically* oriented, and nonorientable manifolds cannot be symplectic.

On a symplectic manifold $(M,\omega)$, we are able to integrate the symplectic form $\omega$ over a surface $A \subset M$:

$$\int_A \omega = \textit{ symplectic area } \text{of } A.$$

In the case of $(\mathbb{R}^4,\omega_0)$, this yields again a sum of contributions from the two projections onto each of the coordinate planes $x_1, y_1$ and $x_2, y_2$:

$$\int_A \omega = \underbrace{\int_{A_1} dx_1 \wedge dy_1}_{\text{signed area of } A_1} + \underbrace{\int_{A_2} dx_2 \wedge dy_2}_{\text{signed area of } A_2} .$$

Such a measurement is *anisotropic* in the sense that (multiple-dimensional) directions are not all the same. For instance, a nontrivial surface in the $x_1, x_2$-plane has one-dimensional projection onto the $x_1, y_1$ and $x_2, y_2$ planes, hence has *zero symplectic area.* Such a surface in a four-dimensional manifold is called *lagrangian.* On the other hand, a nontrivial surface in the $x_1, y_1$ plane already has a *nonzero symplectic area.* Such a surface is called *symplectic.*

In general, we distinguish different important types of submanifolds in a $2n$-dimensional symplectic manifold $(M,\omega)$. A **symplectic submanifold** is a submanifold where the restriction of the symplectic form is nondegenerate, hence still a symplectic form. Such submanifolds are again even-dimensional. When $n = 1$, these submanifolds turn out to be related to *complex curves.* An **isotropic submanifold** is a submanifold where the restriction of the symplectic form vanishes identically. Any one-dimensional submanifold is isotropic and isotropic submanifolds are at most half-dimensional; this follows from linear algebra. A **lagrangian submanifold** is an $n$-dimensional isotropic submanifold. Lagrangian submanifolds are thus the largest isotropic submanifolds and turn out to be related to dynamics.

Examples and nonexamples:

(0) As mentioned, the examples $(\mathbb{R}^{2n}, \omega_0)$ above are the local prototypes of symplectic manifolds.

(1) Any oriented surface may be equipped with a symplectic structure by choosing any area form to take the role of symplectic form. In particular, a unit sphere in $\mathbb{R}^3$ equipped with the standard (euclidean) area form is automatically a symplectic manifold. This area form may be written away from the poles as

$$\omega_{\text{std}} := d\theta \wedge dh \ ,$$

where $h$ is a height function and $\theta$ the angle around that height axis, giving total area $4\pi$; cf. Section 4.

(2) Some of the simplest 4-dimensional symplectic manifolds are products of oriented surfaces, such as $S^2 \times S^2$ equipped with a sum of area forms (eventually different on each factor), and complex projective space $\mathbb{CP}^2$, that is, the space of complex lines in $\mathbb{C}^3$. The standard symplectic form in $\mathbb{CP}^2$ (or, for that matter, in $\mathbb{CP}^n$) is called *Fubini-Study* form and we will give some insight into it in Section 3. In general, products of symplectic manifolds are symplectic.

(3) The only spheres that may be symplectic are the 2-dimensional ones. Let us see why. In a sphere $S^k$ of any other dimension, closed 2-forms are always exact (this topological fact is usually encoded as $H^2(S^k) = 0$ for $k \neq 2$). Now, by Stokes' theorem, a symplectic form cannot be exact on a compact manifold without boundary, because if it were $\omega = d\alpha$, then its top power $\omega^n = d(\alpha \wedge \omega^{n-1})$ would also be exact, which is impossible for a volume form on such a manifold:

$$\int_M \omega^n = \int_M d(\alpha \wedge \omega^{n-1}) = \int_{\partial M} \alpha \wedge \omega^{n-1} = 0 \text{ contradicting } \int_M \omega^n > 0 \ .$$

By now there are a number of texts on symplectic geometry, a subset of which is [11, 12, 4]. For a beautiful overview geared towards symplectic topology, see McDuff's lecture [10].

## 2 What are Hamiltonian Torus Symmetries?

The definition of symplectic form contains exactly what is needed for the following general assertion: *On a symplectic manifold* $(M, \omega)$*, any smooth* *function $H : M \to \mathbb{R}$ generates (in a nontrivial way) a flow that preserves both the symplectic structure $\omega$ and the function $H$.*

Such a flow is called the **hamiltonian flow** generated by $H$ and then $H$ is called a corresponding **hamiltonian function**. The asserted property refers to the *existence and uniqueness* (by nondegeneracy of $\omega$) of a vector field $X_H$ defined by

$$\omega(X_H, \cdot) = dH(\cdot) \ . \qquad \star$$

This vector field $X_H$ satisfies the following equations where we use Cartan's magic formula, $\mathcal{L}_X = d\imath_X + \imath_X d$, for the Lie derivative with respect to a vector field $X$:

$$\mathcal{L}_{X_H}\omega = d \underbrace{\imath_{X_H}\omega}_{dH} + \imath_{X_H} \underbrace{d\omega}_{0} = 0 \qquad \text{and} \qquad \mathcal{L}_{X_H} H = \imath_{X_H} \underbrace{dH}_{\imath_{X_H}\omega} = 0 \ .$$

This vector field $X_H$ integrates (by the theorem of Picard-Lindelöf) into a local time evolution, a.k.a. *flow*, and the equations $\mathcal{L}_{X_H}\omega = 0$ and $\mathcal{L}_{X_H} H = 0$ amount infinitesimally to this flow preserving $\omega$ and $H$. The vector field $X_H$ is called the **hamiltonian vector field** of $H$.

<u>Examples and nonexamples:</u>

(0) For euclidean space $(\mathbb{R}^6, \omega_0)$ and any function $H : \mathbb{R}^6 \to \mathbb{R}$, equation $\star$ for the flow generated by $H$ translates into **Hamilton's equations**:

$$\begin{cases} \dfrac{dx_k}{dt} = \dfrac{\partial H}{\partial y_k} \\ \dfrac{dy_k}{dt} = -\dfrac{\partial H}{\partial x_k} \ . \end{cases}$$

(1) For the unit sphere $(S^2, \omega_{\text{std}} = d\theta \wedge dh)$ and hamiltonian function $H$ equal to the height function $h$, equation $\star$ yields as hamiltonian vector field

$$X_H = \frac{\partial}{\partial \theta} \ ,$$

so the corresponding flow rotates around the height axis. This clearly preserves area $\omega_{\text{std}}$ and height $H$. Notice how this contrasts with the gradient flow of $H$, which is basically perpendicular and preserves neither $\omega_{\text{std}}$ nor $H$.

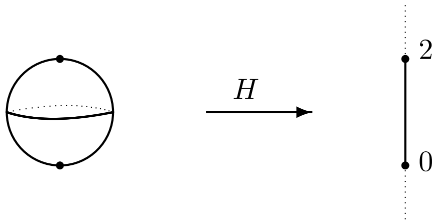



Figure 1: Hamiltonian function for the rotation of $S^2$ around the height axis.

(2) For the 2-torus $(\mathbb{T}^2, \omega := d\theta_1 \wedge d\theta_2)$, we have that the rotation given by the vector field $\frac{\partial}{\partial\theta_1}$ preserves area, yet is *not* hamiltonian, since the contraction

$$\omega(\tfrac{\partial}{\partial\theta_1}, \cdot) = d\theta_2(\cdot)$$

is closed yet *not* exact, i.e., there is no corresponding global hamiltonian function.

The flow in Example (1) is also an example of $S^1$*-action*. Indeed, the time-$t$ evolution $\varphi_t$ is given, with respect to these coordinates, by $\varphi_t : (\theta, h) \longmapsto (\theta + t, h)$, so it is $2\pi$-periodic (i.e., $\varphi_{t+2\pi} \equiv \varphi_t$) and satisfies the group law (i.e., $\varphi_{t_1} \circ \varphi_{t_2} \equiv \varphi_{t_1+t_2}$). Because it is also hamiltonian, we call it a **hamiltonian $S^1$-action**.

Analogously, for a $d$-dimensional torus $\mathbb{T}^d = S^1 \times \ldots \times S^1$ we define a **hamiltonian $\mathbb{T}^d$-action** to be an action of $\mathbb{T}^d$ for which each of the $S^1$-factors acts in a hamiltonian fashion, say with hamiltonian function $H_k$, and each of these $H_k$ is invariant by the rest of the action. By collecting these hamiltonian functions, we build an invariant function

$$H := (H_1, \ldots, H_d) : M \to \mathbb{R}^d \ .$$

This upgraded version of hamiltonian function is known as a (special case of) **moment map**. The concept of *moment map* for hamiltonian actions of arbitrary Lie groups has recently become central in geometry and topology.

Atiyah [2] and, independently, Guillemin and Sternberg [9] proved in the 80's, that the image of such a function $H : M \to \mathbb{R}^d$ on a compact, connected symplectic manifold $(M, \omega)$ corresponding to a hamiltonian $\mathbb{T}^d$-action is always a convex polytope. Moreover, they showed that that image is simply the convex hull of the images of the fixed points of the action. This deep and key theorem is known as the **convexity theorem**.

To get rid of lazy factors in that action, we concentrate on **faithful (i.e. effective) actions** for which only the identity group element gives rise to the identity diffeomorphism. We think of effective hamiltonian $\mathbb{T}^d$-actions as **hamiltonian torus symmetries**. Now, if a $d$-dimensional torus acts in a faithful and hamiltonian fashion on a $2n$-dimensional symplectic manifold, then it must be $d \leq n$. This follows from the fact that the orbits are isotropic, that isotropic submanifolds are at most half-dimensional, and that Lie theory tells us that a faithful action of a $d$-dimensional Lie group always admits orbits equivariantly diffeomorphic to the group itself, the so-called *principal* orbits. Therefore, a maximal hamiltonian torus symmetry is of the form $\mathbb{T}^n$ acting on $M^{2n}$.

## 3 What are Symplectic Toric Manifolds?

A **symplectic toric manifold** is a compact connected symplectic manifold $(M, \omega)$ with a maximal hamiltonian torus symmetry, meaning, with a faithful hamiltonian action of a half-dimensional torus. If $\dim M = 2n$, then we have the $n$-dimensional torus $\mathbb{T}^n$ acting faithfully and with a moment map

$$H : M \to \mathbb{R}^n \ .$$

Examples and nonexamples:

(0) Examples with $(\mathbb{R}^{2n}, \omega_0)$ are ruled out by lack of compactness. However, most of the theory could be, and often is, extended to such examples.

(1) The unit sphere $(S^2, \omega_{\text{std}} = d\theta \wedge dh)$ together with the $S^1$-action generated by the height function $H = h$ is a symplectic toric manifold. We point out some of the features of this example, to which we will come back in more general set-ups:

   (a) The image interval $[0, 2]$ is the orbit space, i.e., there is exactly one $S^1$-orbit per height value. The endpoints of this interval correspond to the two fixed points (singular orbits), South pole and North pole.

   (b) The best coordinates to understand this system are the *angle* coordinate $\theta$ where the rotation occurs and the function $H = h$ encoding the hamiltonian *action*, valid away from the poles. Such coordinates are called **action-angle coordinates**. With respect to such coordinates, the symplectic form is simply a product form $d\theta \wedge dH$, just like a form in the local model space $(\mathbb{R}^2, dx \wedge dy)$.

(c) The area of an invariant strip on $S^2$ corresponding to a subinterval of $[0,2]$ of height $\Delta h$ is equal to $2\pi \cdot \Delta h$. This result goes back more than two millenia; see Section 4.

(1') We revisit the previous example from a complex viewpoint. Regarding $S^2$ as a Riemann sphere, we denote by $[z_0 : z_1]$ the point given by the complex line in $\mathbb{C}^2$ through $(z_0, z_1)$ and $(0,0)$. The South pole is $[1:0]$ and the North pole is $[0:1]$. Now we recast that example as $(\mathbb{CP}^1, \omega_{\text{FS}})$, where the *Fubini-Study* symplectic form $\omega_{\text{FS}}$ is equal to $\frac{1}{4}\omega_{\text{std}}$, the element $e^{it}$ of the circle acts by multiplication on the coordinate $z_1$,

$$e^{it} \cdot [z_0 : z_1] = [z_0 : e^{it} z_1] \ ,$$

which, on a chart, is again a simple shift of the angle coordinate, and the corresponding hamiltonian function is

$$H_1 := \frac{|z_1|^2}{2\left(|z_0|^2 + |z_1|^2\right)} \ .$$

(2) Consider now complex projective space $\mathbb{CP}^n$ (as a $2n$-dimensional real manifold) with a diagonal action of $\mathbb{T}^n$ by

$$\left(e^{i\theta_1}, \dots, e^{i\theta_n}\right) \cdot [z_0 : z_1 : \dots : z_n] = [z_0 : e^{i\theta_1} z_1 : \dots : e^{i\theta_n} z_n] \ .$$

The **Fubini-Study symplectic form** is a globally well-defined form, which, away from the hyperplanes $z_k = 0$, is given by the Darboux-type formula

$$\omega_{\text{FS}} = d\theta_1 \wedge dH_1 + \dots + d\theta_n \wedge dH_n \ ,$$

where the component $H_k$ of the moment map $H : \mathbb{CP}^n \to \mathbb{R}^n$ is

$$H_k := \frac{|z_k|^2}{2\left(|z_0|^2 + \dots + |z_n|^2\right)} \ .$$

For instance, when $n = 3$ we get the following picture:

We list again the earlier features, some of which now take more thought to check:

(a) The image simplex is the orbit space, i.e., there is exactly one $\mathbb{T}^n$-orbit per point on the $n$-simplex. The vertices of this simplex

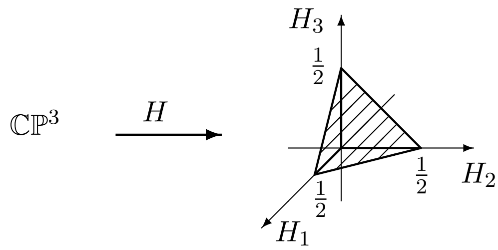



Figure 2: Moment map for the standard action on $\mathbb{CP}^3$.

correspond to the $n+1$ fixed points, $[1:0:\ldots:0],\ldots[0:\ldots:0:1]$. The interior points correspond to orbits through points of the form $[z_0:z_1:\ldots:z_n]$ with all coordinates $z_k$ nonzero.

(b) Best to understand this system are the *action-angle coordinates*, $H_1,\ldots,H_n$ and $\theta_1,\ldots,\theta_n$. With respect to these coordinates, and in points mapping by $H$ to the interior of the simplex, the symplectic form is just like a form in the local model space $(\mathbb{R}^{2n}, dx_1 \wedge dy_1 + \ldots + dx_n \wedge dy_n)$.

(c) The (symplectic) volume of a $\mathbb{T}^n$-invariant subset $H^{-1}(S)$ is simply equal to $(2\pi)^n \cdot |S|$, where $|S|$ is the (euclidean-)volume of the subset $S$ of the simplex.

By the convexity theorem, we already know that the moment map image of a $2n$-dimensional symplectic toric manifold is a polytope in $\mathbb{R}^n$. One can show that such a polytope enjoys special properties: it is *simple*, i.e., there are $n$ edges meeting at each vertex, it is *rational*, i.e., the edges meeting at each vertex $\tau$ are of the form $\tau + tu_j$, $t \geq 0$, with each $u_j \in \mathbb{Z}^n$, and it is *smooth*, i.e., for each vertex, the corresponding $u_1,\ldots,u_n$ can be chosen to form a $\mathbb{Z}$-basis of $\mathbb{Z}^n$; see, for instance, [5].

As first proved by Delzant [7], it turns out that this polytope encodes enough information to reconstruct its originating symplectic toric manifold, and that all such *simple, rational, smooth* polytopes occur as moment map images of symplectic toric manifolds. **Delzant's theorem** is a celebrated result classifying symplectic toric manifolds in terms of polytopes:

$$\left\{ \begin{array}{c} 2n\text{-dim'l symplectic} \\ \text{toric manifolds} \end{array} \right\} \longleftrightarrow \left\{ \begin{array}{c} \text{simple rational smooth} \\ \text{polytopes in } \mathbb{R}^n \end{array} \right\}$$

where this one-to-one correspondence takes a symplectic toric manifold, $(M,\omega,H)$ where the $\mathbb{T}^n$-action admits $H: M \to \mathbb{R}^n$ as moment map, to the polytope which is the *image of this moment map*:

$$(M, \omega, H) \quad \longleftrightarrow \quad H(M) \; .$$

For such a correspondence, there are underlying notions of equivalence of the objects involved. In the simplest version, polytopes in $\mathbb{R}^n$ are identified up to translation, and symplectic toric manifolds are identified up to equivariant diffeomorphism preserving the symplectic forms: $(M_1, \omega_1, H_1)$ and $(M_2, \omega_2, H_2)$ with actions of $\mathbb{T}^n$ are *equivalent* if and only if there is a diffeomorphism $\varphi : M_1 \to M_2$ such that $\varphi^* \omega_2 = \omega_1$ and $\varphi(g \cdot p) = g \cdot \varphi(p)$ for all $g \in \mathbb{T}^n$ and $p \in M_1$.

Note that the problem of classifying compact symplectic manifolds in dimension 4 or higher is completely open. The presence of a hamiltonian torus symmetry significantly helps.

Since there is just one 1-dimensional polytope of length $\ell$ up to translation, we see that the only 2-dimensional symplectic toric manifolds are scaled spheres $(S^2, \frac{\ell}{2}\, \omega_{\text{std}})$ with rotation action as above. The panorama for 2-dimensional polytopes is much more rich. Still, up to translation, the 2-dimensional simple, rational, smooth polytopes with only three vertices are the triangles with vertices $(0,0)$, $(\ell, 0)$ and $(0, \ell)$ or their transforms by $\mathrm{GL}(2; \mathbb{Z})$. This is saying that the corresponding symplectic toric manifolds are $(\mathbb{CP}^2, 2\ell\, \omega_{\text{FS}})$ with standard $\mathbb{T}^2$-action or their transforms by an isomorphism of $\mathbb{T}^2$.

The upshot is that any such symplectic toric manifold is given *combinatorially* in terms of a polytope in an euclidean space of half the dimension that of the manifold. Hence, all questions pertaining to such manifolds should admit an answer in terms of polytopes – a mathematician's dream! In particular, the earlier properties admit generalizations to all symplectic toric manifolds $(M, \omega, H)$ as follows:

(a) The polytope image is the orbit space, so $H$ is also the point-orbit projection, and the vertices of the polytope correspond to the fixed points. There are precise descriptions of the isotropy subgroups in terms of the face-stratification.

(b) There are *action-angle coordinates*, $H_1, \ldots, H_n$ and $\theta_1, \ldots, \theta_n$, valid at points mapping to the interior of the polytope, which are the best coordinates to understand this system. With respect to them, the symplectic form is $\omega = d\theta_1 \wedge dH_1 + \ldots + d\theta_n \wedge dH_n$.

(c) The (symplectic) volume of a $\mathbb{T}^n$-invariant subset is equal to $(2\pi)^n$

times the (euclidean) volume of the corresponding subset in the polytope.

A lot of the geometry of symplectic toric manifolds has already been understood, yet many interesting questions remain. Currently, these manifolds are used as test grounds for theories or conjectures in topology, geometry and mathematical physics, such as *mirror symmetry*.

Many open questions for these manifolds relate to their lagrangian submanifolds. We can see that connected lagrangian submanifolds invariant by $\mathbb{T}^n$ are principal $\mathbb{T}^n$-orbits, i.e., those corresponding to the interior points of the image polytope. We might now ask about other lagrangian submanifolds that *fit nicely* with respect to the torus action, in the sense that they are invariant by some subgroup of $\mathbb{T}^n$ and they intersect $\mathbb{T}^n$-orbits in a *clean* way. The image under the moment map of such a lagrangian submanifold of $(M,\omega,H)$ lies in the intersection of the polytope $H(M)$ with an affine subspace. Examples are all principal $\mathbb{T}^n$-orbits, the standard *real part* submanifolds like $\mathbb{RP}^n$ in $\mathbb{CP}^n$, lagrangian submanifolds like the one presented in [3], and many lagrangian submanifolds sitting in level sets of components of the moment map.

## 4 Epilogue – all the way from Archimedes

We close by going back more than two millenia to Archimedes' supposedly favourite work on measuring spheres and cylinders. In around 200 BC, Archimedes was the first to realize that *the surface area of a sphere between two parallel planes intersecting it depends only on the distance between those planes* and not on the height where they intersect the sphere. Moreover, Archimedes asserted that the surface area on the sphere is the same as that of a cylinder with the radius of that sphere and height given by the distance between the planes, as the following figure illustrates. This is exactly the feature that allows us to write the standard area form as $\omega_{\text{std}} = d\theta \wedge dh$.

Nowadays, if you know first-year calculus, you may check Archimedes result by computing an appropriate surface integral using, for instance, cylindrical coordinates $(\theta, z)$ to write points on the sphere as $(x,y,z) = (\sqrt{R^2-z^2}\cos\theta, \sqrt{R^2-z^2}\sin\theta, z)$:

$$\text{Area} = \int_0^{2\pi}\int_h^{h+\Delta h} R\,dz\,d\theta = 2\pi R\cdot\Delta h \ ,$$

or else use some approximation method and then take the limit [1]. However, Archimedes did not know calculus. It seems that he used an approximation

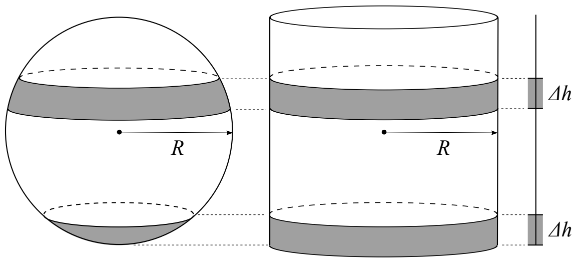



Figure 3: Spherical and cylindrical strips all with the same area: $2\pi R \cdot \Delta h$; image kindly reproduced from [1].

argument, for which a relevant reference is the *palimpsest*[3] discovered in the XX century after some quite adventurous history.

In the 80's, Duistermaat and Heckman [8] showed powerful results for symplectic manifolds with hamiltonian torus actions, which may be viewed as a vast generalization of Archimedes' theorem for the 2-sphere. Just like Archimedes might have had no idea that, more than two millennia later, his spirit would be at the origin of new mathematics, one wonders what other leaps await mankind starting from symplectic toric manifolds.

## References

[1] Abreu, M., Cannas da Silva, A., 22 séculos a medir área (*22 centuries measuring area*), paper contributed for the *Klein project* in Portuguese (2011) available at `klein.sbm.org.br/?s=cannas`

[2] Atiyah, M., Convexity and commuting Hamiltonians, *Bull. London Math. Soc.* **14** (1982) 1–15.

[3] Cannas da Silva, A., A Chiang-type lagrangian in $\mathbb{CP}^2$, *Lett. Math. Phys.* **108** (2018) 699–710.

[4] Cannas da Silva, A., *Lectures on Symplectic Geometry*, Lecture Notes in Mathematics 1764, Springer-Verlag (2008).

[5] Cannas da Silva, A., Seminar on Symplectic Toric Manifolds, lecture notes, ETH Spring 2019, available at `people.math.ethz.ch/~acannas/Papers/stm2019.pdf`.

[3] A *palimpsest* is a manuscript page, either from a scroll or a book, from which the text has been scraped or washed off so that the page can be reused for another document.

[6] Darboux, G., Sur le problème de Pfaff, *Bull. Sci. Math. et Astr.* **6** (1882) 14–36 and 49–62.

[7] Delzant, T., Hamiltoniens périodiques et images convexes de l'application moment, *Bull. Soc. Math. France* **116** (1988) 315–339.

[8] Duistermaat, J. J., G. J. Heckman, On the variation in the cohomology of the symplectic form of the reduced phase space, *Invent. Math.* **69** (1982) 259–268.

[9] Guillemin, V., Sternberg, S., Convexity properties of the moment mapping, *Invent. Math.* **67** (1982) 491–513.

[10] McDuff, D., Symplectic Geometry Today, talk at ETH Zurich, 14/March/2019, available at `math.ethz.ch/news-and-events/events/gomath-women-in-math/symposium`.

[11] McDuff, D., Salamon, D., *Introduction to Symplectic Topology*, Oxford Mathematical Monographs, Oxford University Press, New York, 1995.

[12] Weinstein, A., *Lectures on Symplectic Manifolds*, Regional Conference Series in Mathematics **29**, Amer. Math. Soc., Providence, 1977.

# A sharp inequality in Fourier restriction theory

*Diogo Oliveira e Silva*
School of Mathematics, University of Birmingham
Edgbaston, Birmingham B15 2TT, UK
e-mail: d.oliveiraesilva@bham.ac.uk

*René Quilodrán*
Santiago, Chile
e-mail: rquilodr@dim.uchile.cl

**Resumo:** Recentemente, os autores provaram em [15, 16] que *as funções constantes são os únicos maximizantes reais da desigualdade $L^2 \to L^6$ de extensão de Fourier na 2-esfera.* Isto é um caso particular de [16, Teorema 1.1], cuja prova contém vários dos métodos e ideias-chave. Neste artigo, descrevemos a prova deste caso particular, e apresentamos algumas generalizações e problemas em aberto.

**Abstract:** We focus on the proof of the following recent result [15, 16] in Sharp Fourier Restriction Theory: *Constant functions are the unique real-valued maximizers for the $L^2 \to L^6$ adjoint Fourier restriction inequality on the 2-sphere.* This is a special case of [16, Theorem 1.1] which already relies on several of the key methods and ideas. We discuss generalizations, extensions, and present a few open problems.

**palavras-chave:** Teoria de restrição de Fourier; desigualdade de Tomas–Stein; constantes ótimas; maximizantes.

**keywords:** Sharp Fourier Restriction Theory; Tomas–Stein inequality; optimal constants; maximizers.

## 1 Introduction

We start with a collection of three apparently unrelated problems from geometry, probability theory, and algebra.

**Question 1** *Given $d \geq 2, 0 < k < d$, what is the maximal volume of the intersection of the unit cube $[-\frac{1}{2}, \frac{1}{2}]^d$ with a $k$-dimensional subspace of $\mathbb{R}^d$?*

**Question 2** *Given $d, n \geq 2$, what is the probability distribution of an $n$-step uniform random walk in $\mathbb{R}^d$?*

**Question 3** *Given $d \geq 2$, what is the minimal codimension of a proper subalgebra of the special orthogonal Lie algebra $\mathfrak{so}(d)$?*

One of the goals of the present note is to describe how each of these questions played a very natural role in the recent solution of an extremal problem from harmonic analysis, to which we now turn our attention.

## 1.1 Fourier restriction theory

The Fourier transform is one of the most ubiquitous tools in mathematics. By decomposing a general function $f : \mathbb{R}^d \to \mathbb{C}$ into a superposition of simpler, "symmetric" functions,

$$\widehat{f}(\xi) = \int_{\mathbb{R}^d} f(x) e^{-ix\cdot\xi}\, \mathrm{d}x,$$

it opens the door to powerful analytic arguments that have shaped the history of mathematics for the last two centuries. Despite its paramount importance, fundamental questions about the Fourier transform remain open.

A consequence of the classical Hausdorff–Young inequality is that the Fourier transform $\widehat{f}$ of an $L^p$ function $f : \mathbb{R}^d \to \mathbb{C}$ is defined almost everywhere on $\mathbb{R}^d$, provided $1 \leq p \leq 2$. It is a striking observation of E. M. Stein from the late 1960s that, for a special range of $p$'s, the function $\widehat{f}$ can be meaningfully defined on submanifolds of $\mathbb{R}^d$ possessing some degree of curvature. The simple yet fundamental observation that *curvature causes the Fourier transform to decay* links geometry to analysis, and lies at the base of Fourier restriction theory. Take, for instance, the example of the unit sphere, $\mathbb{S}^{d-1} := \{\omega \in \mathbb{R}^d : |\omega| = 1\}$, a compact manifold with positive Gaussian curvature which inherits its surface measure $\mathrm{d}\sigma_{d-1}$ from the ambient space $\mathbb{R}^d$ in the natural way. The celebrated *Fourier restriction conjecture* predicts the validity of the estimate[1]

$$\int_{\mathbb{S}^{d-1}} |\widehat{g}(\omega)|^q \,\mathrm{d}\sigma_{d-1}(\omega) \leq C \|g\|^q_{L^p(\mathbb{R}^d)}, \text{ if } 1 \leq p < \frac{2d}{d+1}, q \leq \frac{d-1}{d+1} p', \quad (1)$$

and is remarkable in its numerous connections and applications. It exhibits deep links to Bochner–Riesz summation methods and to decoupling phenomena for the Fourier transform, and is known to imply the Kakeya conjecture. Despite the great deal of attention received by this circle of problems during the past four decades, the restriction conjecture remains open

[1]Here, $p'$ denotes the conjugate exponent to $p$, given by $\frac{1}{p} + \frac{1}{p'} = 1$.

in dimensions $d \geq 3$. For further details, we refer the interested reader to the classical survey [20], and the very recent, exciting account from [18].

If $d \geq 2$ and $q \geq 2\frac{d+1}{d-1}$, then the cornerstone Tomas–Stein inequality [17, 21] states that there exists $C = C(d,q) < \infty$, such that

$$\|\widehat{f\sigma}_{d-1}\|_{L^q(\mathbb{R}^d)} \leq C\|f\|_{L^2(\mathbb{S}^{d-1})}, \tag{2}$$

for every function $f : \mathbb{S}^{d-1} \to \mathbb{C}$ which is square-integrable with respect to $\mathrm{d}\sigma_{d-1}$. Here, $\widehat{f\sigma}_{d-1}(x) := \mathcal{E}(f)(x) := \int_{\mathbb{S}^{d-1}} f(\omega)e^{i\omega\cdot x}\,\mathrm{d}\sigma_{d-1}(\omega)$, $x \in \mathbb{R}^d$, denotes the Fourier extension operator, which is the adjoint of the restriction operator, $\mathcal{E}^*(g) := \widehat{g}\,|_{\mathbb{S}^{d-1}}$, considered in (1). Inequality (2) finds deep applications in harmonic analysis and PDE. In particular, it underlies most of the early progress towards the Fourier restriction conjecture; see [20]. The Tomas–Stein argument directly implies some of the foundational Strichartz estimates for various dispersive partial differential equations, e.g. the Schrödinger, wave, and Klein–Gordon equations; see [19]. Moreover, inequality (2) has been generalized to a variety of contexts, and found surprising applications ranging from fractal geometry [14] to number theory [10], among many others.

### 1.2 Sharp Fourier Restriction Theory

A class of problems which is the subject of some exciting ongoing research goes under the name of *Sharp Fourier Restriction Theory*. For a gentle introduction to this fascinating topic, we refer the reader to the recent survey [7], and proceed to describe a few concrete examples.

Associated to (2), we have the functional

$$f \mapsto \Phi_{d,q}(f) := \frac{\|\widehat{f\sigma}_{d-1}\|^q_{L^q(\mathbb{R}^d)}}{\|f\|^q_{L^2(\mathbb{S}^{d-1})}}.$$

A very natural problem is to determine the value of the best (smallest) constant in inequality (2),

$$\mathbf{T}^q_{d,q} := \sup_{\mathbf{0}\neq f\in L^2} \Phi_{d,q}(f),$$

i.e. the operator norm of the extension operator. A related, but typically harder, problem is to characterize all the *maximizers* of $\Phi_{d,q}$, that is to say, the nonzero functions which realize the best constant $\mathbf{T}_{d,q}$. The mere *existence* of maximizers is a highly non-trivial question, which for $\Phi_{d,q}$ happens to be open at the endpoint $q = 2\frac{d+1}{d-1}$ in all dimensions $d \geq 4$; see [8] for a conditional result in this direction.

### 1.2.1 A sharp $L^2$–$L^4$ result

A remarkable recent result of D. Foschi [6] establishes that constant functions are the unique real-valued maximizers for the endpoint Tomas–Stein inequality in three-dimensional space,

$$\|\widehat{f\sigma}_2\|_{L^4(\mathbb{R}^3)} \le \mathbf{T}_{3,4}\|f\|_{L^2(\mathbb{S}^2)}. \tag{3}$$

In particular, $\mathbf{T}_{3,4} = \|\widehat{\sigma}_2\|_{L^4(\mathbb{R}^3)}\|\mathbf{1}\|^{-1}_{L^2(\mathbb{S}^2)} = 2\pi$. The proof is short, simple, and relies on an elegant geometric identity,

$$|\omega+\nu|^2 + |\nu+\zeta|^2 + |\zeta+\omega|^2 = 4,$$

which holds for any triple of unit vectors $(\omega,\nu,\zeta) \in (\mathbb{S}^2)^3$ satisfying $|\omega+\nu+\zeta| = 1$. Additional ingredients that play a key role in [6] are some symmetry considerations, a natural spectral analysis, and two fortuitous coincidences.

The first coincidence is that in the three-dimensional case some calculations simplify considerably in comparison with other dimensions. Technically, this is seen at the level of the convolution measure $\sigma_{d-1} * \sigma_{d-1}$, which finds its simplest form when $d = 3$; see (23) below. The difficulties inherent to the higher dimensional cases were partially overcome in [4], thereby extending Foschi's $L^2 \to L^4$ sharp result to dimensions $4 \le d \le 7$. If $d = 8$, then a new phenomenon emerges, and the identification of one single maximizer of $\Phi_{d,4}$ is a challenging open problem in all dimensions $d \ge 8$.

The second coincidence is that $4 = 2 \times 2$. In particular, since the Fourier transform intertwines multiplication and convolution,

$$|\widehat{f\sigma}_2|^4 = (\widehat{f\sigma}_2\overline{\widehat{f\sigma}_2})^2 = (\widehat{f\sigma_2 * \bar{f}\sigma_2})^2.$$

An application of Plancherel's identity, $\|\widehat{F}\|_{L^2(\mathbb{R}^3)} = (2\pi)^{3/2}\|F\|_{L^2(\mathbb{R}^3)}$, then reveals that (3) can be equivalently recast as a convolution inequality,

$$\|f\sigma_2 * \bar{f}\sigma_2\|_{L^2(\mathbb{R}^3)} \le (2\pi)^{-3/2}\mathbf{T}^2_{3,4}\|f\|^2_{L^2(\mathbb{S}^2)}.$$

The 2-fold convolution measure $f\sigma_2 * \bar{f}\sigma_2$ turns out to be a relatively simple object of study, even though the function $f \in L^2$ may be quite rough. The situation changes dramatically if instead we consider the $k$-fold convolution $(f\sigma_2)^{*k}$, for $k \ge 3$. In fact, prior to our very recent work [15, 16], no sharp instance of inequality (2) was known if $q \in (4,\infty)$, in any dimension $d \ge 2$.

### 1.2.2 A sharp $L^2$–$L^6$ result

In [16], we proved that constant functions are the unique real-valued maximizers of the functional $\Phi_{d,2n}$, whenever $d \in \{3,4,5,6,7\}$ and $n \geq 3$ is an integer. The following particular case of [16, Theorem 1.1] will be the focus of our attention.

**Theorem 1** *Constants are the unique real-valued maximizers of* $\Phi_{3,6}$.

This of course translates into a sharp inequality

$$\|\widehat{f\sigma_2}\|_{L^6(\mathbb{R}^3)} \leq \mathbf{T}_{3,6}\|f\|_{L^2(\mathbb{S}^2)}, \tag{4}$$

with $\mathbf{T}_{3,6} = \|\widehat{\sigma}_2\|_{L^6(\mathbb{R}^3)}\|\mathbf{1}\|^{-1}_{L^2(\mathbb{S}^2)} = (2\pi)^{5/6}$. We choose to delve into the proof of Theorem 1 because it already contains several of the main themes which were introduced in [15, 16]. On the other hand, the convenient choice of parameters $(d,q) = (3,6)$ causes several technicalities to disappear, and makes us hopeful that the key ideas may be conveyed in the course of this short note.

### 1.3 Notation

The constant function is denoted $\mathbf{1} : \mathbb{S}^{d-1} \to \{1\}$, $\mathbf{1}(\omega) \equiv 1$, and the zero function is denoted $\mathbf{0} : \mathbb{S}^{d-1} \to \{0\}$, $\mathbf{0}(\omega) \equiv 0$. If there is no danger of confusion, we sometimes write $L^2 = L^2(\mathbb{S}^{d-1})$. Since we will mostly be working in dimension $d = 3$, we simplify the forthcoming notation by setting $\Phi_q := \Phi_{3,q}$, $\mathbf{T}_q := \mathbf{T}_{3,q}$, and $\mathrm{d}\sigma := \mathrm{d}\sigma_2$. Finally, if $x, y$ are real numbers, we write $x \lesssim y$ if there exists a finite absolute constant $C$ such that $|x| \leq C|y|$.

### 1.4 Outline

We organize the exposition in five steps, each of them bringing in tools from the calculus of variations (§2), symmetrization techniques (§3), operator theory (§4), Lie theory (§5), and probability theory (§6). These ingredients are then combined in §7, yielding a short proof of Theorem 1. In §8, we discuss some extensions, generalizations, and open problems.

## 2 Step 1: Calculus of variations

Let $f$ be a maximizer[2] for $\Phi_6$, and normalize it so that $\|f\|_{L^2} = 1$. Recall the operators $\mathcal{E}, \mathcal{E}^*$ which were defined immediately after (2). The following

[2]The existence of maximizers for $\Phi_6$ follows from [5, Theorem 1.1].

chain of inequalities holds:

$$\begin{aligned}\|\mathcal{E}\|^6_{L^2\to L^6} &= \|\mathcal{E}(f)\|^6_{L^6(\mathbb{R}^3)} = \langle|\mathcal{E}(f)|^4\mathcal{E}(f),\mathcal{E}(f)\rangle = \langle\mathcal{E}^*(|\mathcal{E}(f)|^4\mathcal{E}(f)), f\rangle_{L^2(\mathbb{S}^2)}\\ &\leq \|\mathcal{E}^*(|\mathcal{E}(f)|^4\mathcal{E}(f))\|_{L^2(\mathbb{S}^2)} \leq \|\mathcal{E}^*\|_{L^{6/5}\to L^2}\||\mathcal{E}(f)|^4\mathcal{E}(f)\|_{L^{6/5}(\mathbb{R}^3)}\\ &= \|\mathcal{E}^*\|_{L^{6/5}\to L^2}\|\mathcal{E}(f)\|^5_{L^6(\mathbb{R}^3)} = \|\mathcal{E}\|^6_{L^2\to L^6},\end{aligned}\tag{5}$$

where $\langle\cdot,\cdot\rangle$ denotes the $L^{6'}-L^6$ pairing in $\mathbb{R}^3$, and $\langle\cdot,\cdot\rangle_{L^2(\mathbb{S}^2)}$ denotes the $L^2$ pairing on $\mathbb{S}^2$. The only steps which are not entirely trivial amount to an application of the Cauchy–Schwarz inequality, and the fact that the operators norms $\|\mathcal{E}\|_{L^2\to L^6} = \|\mathcal{E}^*\|_{L^{6/5}\to L^2}$ coincide.[3] Since the first and the last terms in the chain of inequalities (5) are the same, all inequalities have to be equalities. In particular, equality holds in the application of the Cauchy–Schwarz inequality, which forces the two functions in question to be constant multiples of each other. In other words, $\mathcal{E}^*(|\mathcal{E}(f)|^4\mathcal{E}(f)) = \lambda f$, for some $\lambda\in\mathbb{C}$. Recalling the definition of the extension and restriction operators, this boils down to

$$(|\widehat{f\sigma}|^4\widehat{f\sigma})^\vee|_{\mathbb{S}^2} = \lambda f.\tag{6}$$

By Plancherel's identity, the latter equality can be written in convolution form,

$$(f\sigma * f_\star\sigma * f\sigma * f_\star\sigma * f\sigma)|_{\mathbb{S}^2} = (2\pi)^{-3}\lambda f.\tag{7}$$

Here, $f_\star = \bar{f}(-\cdot)$ denotes the *conjugate reflection* of $f$, and accounts for the complex conjugates that appear on the left-hand side of (6). Identity (7) is the *Euler–Lagrange equation* associated to the functional $\Phi_6$, and any nonzero, square integrable solution of (7) is called a *critical point* of $\Phi_6$.

The Euler–Lagrange equation (7) can be used to show that any maximizer of (4), and more generally any critical point of $\Phi_6$, is an infinitely differentiable function. This is a manifestation of the general phenomenon that convolution operators are smoothing, but the actual proof entails a number of technical difficulties. We omit the details, and encourage the interested reader to take a look at [15].

## 3 Step 2: Symmetrization

This step is more elementary than the previous one, but plays an equally important part in the analysis. Inequality (4) can be equivalently rewritten

[3]If $T: L^p\to L^q$ is a bounded linear operator, then its adjoint $T^*$ defines a bounded linear operator from $L^{q'}$ to $L^{p'}$, with the same operator norm. Also, $6'=\frac{6}{5}$.

in convolution form as

$$\|f\sigma * f\sigma * f\sigma\|_{L^2(\mathbb{R}^3)} \leq (2\pi)^{-3/2}\mathbf{T}_6^3\|f\|_{L^2(\mathbb{S}^2)}^3.$$

Since $|f\sigma * f\sigma * f\sigma| \leq |f|\sigma * |f|\sigma * |f|\sigma$ holds pointwise, it follows that[4]

$$\|f\sigma * f\sigma * f\sigma\|_{L^2(\mathbb{R}^3)} \leq \||f|\sigma * |f|\sigma * |f|\sigma\|_{L^2(\mathbb{R}^3)}. \tag{8}$$

Further define the *antipodally symmetric rearrangement* $f_\sharp$ of $f$ via

$$f_\sharp := \sqrt{\frac{|f|^2 + |f_\star|^2}{2}},$$

where $f_\star$ denotes the conjugate reflection of $f$ as above. Note that the $L^2$-norms of $f_\sharp$ and $f$ (or $f_\star$) coincide. A straightforward application of the elementary inequality between the arithmetic and geometric means reveals that

$$\|f\sigma * f\sigma * f\sigma\|_{L^2(\mathbb{R}^3)} \leq \|f_\sharp\sigma * f_\sharp\sigma * f_\sharp\sigma\|_{L^2(\mathbb{R}^3)},$$

with equality if and only if $f = f_\star = f_\sharp$. These considerations imply that, in the search for maximizers of $\Phi_6$, we may limit our attention to non-negative, antipodally symmetric functions. In other words,

$$\mathbf{T}_6^6 = \max_{0 \leq f = f_\star \in L^2 \setminus \{\mathbf{0}\}} \Phi_6(f). \tag{9}$$

This is a key simplification which enables several of the subsequent steps to work.

# 4 Step 3: Operator theory

In this section, we explore some of the compactness inherent to the problem. Given a nonzero function $f \in L^2(\mathbb{S}^2)$, normalized so that $\|f\|_{L^2} = 1$, consider the integral operator

$$T_f : L^2(\mathbb{S}^2) \to L^2(\mathbb{S}^2),\ T_f(g) = g * K_f,$$

with convolution kernel given by

$$K_f(\xi) = (|\widehat{f\sigma}|^4)^\vee(\xi) = (2\pi)^3(f\sigma * f_\star\sigma * f\sigma * f_\star\sigma)(\xi). \tag{10}$$

[4] For a characterization of the cases of equality in (8), see [4, Lemma 8].

The relevance of this operator is easy to highlight. In fact, the Euler–Lagrange equation (6) is nothing but the eigenvalue problem for $T_f$, namely $T_f(f) = \lambda f$. Observe that $\lambda$ is entirely dictated by $f$: From $\lambda f = T_f(f)$ one has that $\lambda \int |f|^2 = \int T_f(f)\bar{f} = \int |\widehat{f\sigma}|^6$, whence $\lambda = \Phi_6(f)$ (since $\|f\|_{L^2} = 1$).

We proceed to study $T_f$ from the operator theoretic point of view. First of all, the function $K_f$ from (10) satisfies $K_f(0) = \|\widehat{f\sigma}\|_{L^4}^4$. Moreover, $K_f$ defines a bounded, continuous function on $\mathbb{R}^3$, satisfying $K_f(\xi) = \overline{K_f(-\xi)}$, for all $\xi$. As a consequence, the operator $T_f$ is self-adjoint, $T_f = T_f^*$, and positive definite: $\langle T_f(g), g\rangle_{L^2} > 0$, for every nonzero $g \in L^2$. The operator $T_f$ is also Hilbert–Schmidt (and therefore compact), since the companion kernel $K_f^\flat$ defined by $K_f^\flat(\omega,\nu) := K_f(\omega-\nu)$ belongs to $L^2(\mathbb{S}^2\times\mathbb{S}^2)$. But more is true: the operator $T_f$ is actually *trace class*. To see this, let $\{\lambda_j\}_{j=0}^\infty \subset (0,\infty)$ denote the eigenvalues of $T_f$ in non-increasing order, counted with multiplicity, with corresponding $L^2$-normalized eigenfunctions $\{\varphi_j\}_{j=0}^\infty$. By the classical theorem of Mercer (see e.g. [22, §VI.4]),

$$K_f^\flat(\omega,\nu) = \sum_{j=0}^\infty \lambda_j \varphi_j(\omega)\overline{\varphi_j(\nu)},$$

where the series converges absolutely and uniformly. The trace of $T_f$ can be then estimated as follows:

$$\begin{aligned}\operatorname{tr}(T_f) = \sum_{j=0}^\infty \langle T_f(\varphi_j), \varphi_j\rangle_{L^2(\mathbb{S}^2)} = \sum_{j=0}^\infty \lambda_j = \int_{\mathbb{S}^2} K_f^\flat(\omega,\omega)\, \mathrm{d}\sigma(\omega) \\ = 4\pi K_f(0) = 4\pi \|\widehat{f\sigma}\|_{L^4(\mathbb{R}^3)}^4 \lesssim \|f\|_{L^2(\mathbb{S}^2)}^4 < \infty,\end{aligned} \tag{11}$$

where in the last line we invoked the endpoint Tomas–Stein inequality (3).

## 5 Step 4: Lie theory

It is natural to expect the symmetries of the sphere to enter the picture at some point. The symmetry group of $\mathbb{S}^2$, including reflections, is the orthogonal group, O(3). The subgroup of rotations, i.e. orthogonal $3 \times 3$ matrices with unit determinant, is the so-called *special orthogonal group*, SO(3). As a Lie group, SO(3) is compact, connected, and of dimension 3. Its Lie algebra, $\mathfrak{so}(3)$, consists of skew-symmetric $3 \times 3$ matrices with real entries. The exponential map, $\exp : \mathfrak{so}(3) \to \mathrm{SO}(3), A \mapsto \exp(A)$, is surjective onto SO(3). For more information on the Lie group SO(3) and its Lie algebra, see [11].

Given a matrix $A \in \mathfrak{so}(3)$, define the vector field $\partial_A$ acting on sufficiently smooth functions $f : \mathbb{S}^2 \to \mathbb{C}$ via $\partial_A f := \lim_{t\to 0} t^{-1}(f(\exp(tA)\cdot) - f)$. The functional $\Phi_6$ enjoys the following symmetries:

$$\Phi_6(f \circ \exp(tA)) = \Phi_6(f) = \Phi_6(e_\xi f),$$

for all $t \in \mathbb{R}$, $A \in \mathfrak{so}(3)$, and $\xi \in \mathbb{R}^3$, where $e_\xi$ stands for the character $e_\xi(\omega) = e^{i\xi\cdot\omega}$. These symmetries naturally give rise to new eigenfunctions for the operator $T_f$ considered in §4, as the following result indicates. We write $\omega = (\omega_1, \omega_2, \omega_3) \in \mathbb{S}^2$, and by $\omega_j f$ we mean the function defined via $(\omega_j f)(\omega) = \omega_j f(\omega)$.

**Lemma 1** *Let $f : \mathbb{S}^2 \to \mathbb{R}$ be non-constant, continuously differentiable, antipodally symmetric, and such that $\|f\|_{L^2} = 1$. Assume $T_f(f) = \lambda f$. Then:*

$$T_f(\omega_j f) = \tfrac{\lambda}{5}\,\omega_j f, \quad \text{for every } j \in \{1,2,3\}, \tag{12}$$

$$T_f(\partial_A f) = \tfrac{\lambda}{5}\,\partial_A f, \quad \text{for every } A \in \mathfrak{so}(3). \tag{13}$$

*Moreover, there exist $A, B \in \mathfrak{so}(3)$, such that the set $\{\omega_1 f, \omega_2 f, \omega_3 f, \partial_A f, \partial_B f\}$ is linearly independent over $\mathbb{C}$.*

*Sketch of proof.* We omit the derivation of the identities (12), (13), and instead refer the reader to the proof of [16, Prop. 5.2]. The functions $\omega_1 f, \omega_2 f, \omega_3 f$ are linearly independent[5] – this is elementary.

Since $f \in C^1(\mathbb{S}^2)$ is non-constant, there exist $A, B \in \mathfrak{so}(3)$, such that $\partial_A f, \partial_B f$ are linearly independent. To see why this is necessarily the case, consider the linear map $D : \mathfrak{so}(3) \to C^0(\mathbb{S}^2)$, $D(A) = \partial_A f$. Let $r := \dim\ker D$. By the Rank-Nullity Theorem, the image of $D$ has dimension $\dim \mathfrak{so}(3) - \dim\ker D = 3 - r$, and so it suffices to show that $r \leq 1$. Aiming at a contradiction, suppose that $r \geq 2$. In this case, there exist linearly independent matrices $X, Y \in \mathfrak{so}(3)$, such that $\partial_X f = \partial_Y f \equiv 0$. The matrices $X, Y$ correspond to infinitesimal rotations around certain unit vectors $\omega, \nu \in \mathbb{S}^2$, respectively. Since $X, Y$ are linearly independent, then so are $\omega, \nu$. But $\partial_X f \equiv 0$ implies that $f$ is constant along all $\omega$-*latitudes*, i.e. circles determined by intersecting $\mathbb{S}^2$ with the translates of a 2-plane orthogonal to $\omega$. In a similar way, $f$ is constant along all $\nu$-*latitudes*. Since $\omega, \nu$ are linearly independent, it follows that any two points on $\mathbb{S}^2$ can be joined by a path consisting of the alternating concatenation of a certain (finite)

[5] Since $f$ is real-valued, linear independence of the set $\{\omega_1 f, \omega_2 f, \omega_3 f, \partial_A f, \partial_B f\}$ over $\mathbb{C}$ is equivalent to that over $\mathbb{R}$.

number of $\omega$-latitudes and $\nu$-latitudes. Since $f$ is constant along each such latitude, it is constant along the whole path. It follows that $f$ is identically constant, which is absurd.

An alternative approach, which is perhaps less intuitive but has the advantage of generalizing to higher $d \geq 3$, uses the fact[6] that the dimension of a proper, nontrivial subalgebra of $\mathfrak{so}(3)$ is equal to 1 (think of the embedding $\mathfrak{so}(2) \subseteq \mathfrak{so}(3)$). As a consequence, if $r \geq 2$, then the Lie algebra generated by $\ker D$ equals the whole of $\mathfrak{so}(3)$. In turn, this together with the fact that the action of SO(3) on $\mathbb{S}^2$ is transitive, can be used to show that $f$ is constant, which again yields the desired contradiction.

Finally, the linear independence of the set $\{\omega_1 f, \omega_2 f, \omega_3 f, \partial_A f, \partial_B f\}$ follows from the facts that $\omega_1 f, \omega_2 f, \omega_3 f$ are real-valued, antipodally anti-symmetric functions, whereas $\partial_A f, \partial_B f$ are real-valued, antipodally symmetric functions. $\square$

The conclusion is that, given a sufficiently smooth, non-constant eigenfunction $f = f_\star$ of $T_f$ with eigenvalue $\lambda$, we can always find *five* further eigenfunctions of $T_f$, each with eigenvalue $\frac{\lambda}{5}$, and with the crucial property that the set $\{\omega_1 f, \omega_2 f, \omega_3 f, \partial_A f, \partial_B f\}$ is linearly independent over $\mathbb{C}$.

## 6 Step 5: Probability theory

Consider three independent, identically distributed random variables $X_1, X_2, X_3$, taking values on $\mathbb{S}^2$ with uniform distribution. In this case, the random variable $Y_3 = X_1 + X_2 + X_3$ corresponds to the so-called *uniform 3-step random walk* in $\mathbb{R}^3$, and is distributed according to the 3-fold convolution of the normalized surface measure on $\mathbb{S}^2$. In other words, if $\bar{\sigma} := \sigma(\mathbb{S}^2)^{-1}\sigma$ and $\Omega \subseteq \mathbb{R}^3$ is a Borel subset, then

$$\mathbb{P}(Y_3 \in \Omega) = \int_\Omega (\bar{\sigma} * \bar{\sigma} * \bar{\sigma})(\xi)\, \mathrm{d}\xi.$$

Let $p_3$ denote the probability density associated to the random variable $|Y_3|$. For any measurable subset $E \subseteq (0, \infty)$, we then have that

$$\mathbb{P}(|Y_3| \in E) = \int_E p_3(r)\, \mathrm{d}r.$$

A straightforward computation in spherical coordinates further reveals that $(\sigma * \sigma * \sigma)(r) = \sigma(\mathbb{S}^2)^2 p_3(r) r^{-2}$. Similar considerations apply to the simpler

[6] Incidentally, this provides an answer to Question 3 when $d = 3$.

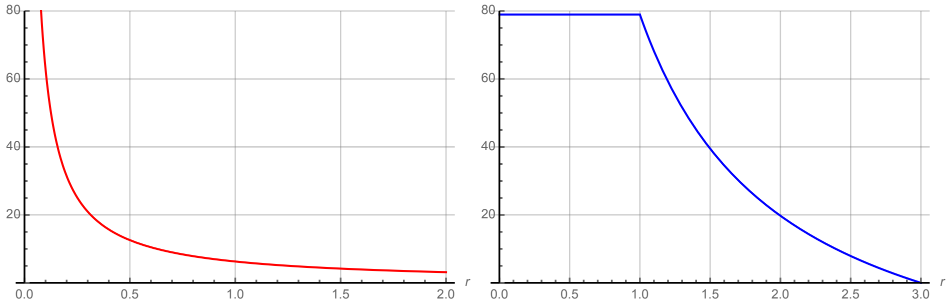



Figura 1: Left: Plot of the function $r \mapsto (\sigma * \sigma)(r)$ for $0 \leq r \leq 2$. Pairs of antipodal points on $\mathbb{S}^2$ contribute towards the singularity at $r = 0$. Right: Plot of the function $r \mapsto (\sigma * \sigma * \sigma)(r)$ for $0 \leq r \leq 3$.

uniform 2-step random walk in $\mathbb{R}^3$, $Y_2 = X_1 + X_2$, in which case we let $p_2$ denote the density of $|Y_2|$.

Random walks have been the subject of active investigation for more than a century, and as such it comes as no surprise that explicit formulae for $p_2, p_3$ are well-known, thereby providing an answer to Question 2 when $d = 3$ and $n \in \{2, 3\}$; see [3, 9]. They translate into the following result for convolutions; see also Figure 1.

**Lemma 2** *The following identities hold:*

$$(\sigma * \sigma)(\xi) = \frac{2\pi}{|\xi|}, \quad \text{if } |\xi| \leq 2, \tag{14}$$

$$(\sigma * \sigma * \sigma)(\xi) = \begin{cases} 8\pi^2, & \text{if } |\xi| \leq 1, \\ 4\pi^2\Big(-1 + \dfrac{3}{|\xi|}\Big), & \text{if } 1 \leq |\xi| \leq 3. \end{cases} \tag{15}$$

As an immediate consequence of Lemma 2, we may compute the quantities

$$\Phi_4(\mathbf{1}) = (2\pi)^3 \|\mathbf{1}\|^{-4}_{L^2(\mathbb{S}^2)} \|\sigma * \sigma\|^2_{L^2(\mathbb{R}^3)} = 16\pi^4, \tag{16}$$

$$\Phi_6(\mathbf{1}) = (2\pi)^3 \|\mathbf{1}\|^{-6}_{L^2(\mathbb{S}^2)} \|\sigma * \sigma * \sigma\|^2_{L^2(\mathbb{R}^3)} = 32\pi^5, \tag{17}$$

which will be of use in the next section.

## 7 Proof of Theorem 1

Armed with the tools developed in §2–§6, the proof of Theorem 1 is now quite short.

*Proof of Theorem 1.* It will suffice to prove that any real-valued, continuously differentiable, non-constant critical point $f$ of $\Phi_6$ satisfies $\Phi_6(f) < \Phi_6(\mathbf{1})$. In view of (9), we may further assume that $f = f_\star$, and naturally that $\|f\|_{L^2} = 1$. Multiplying both sides of the Euler–Lagrange equation, $T_f(f) = \lambda f$, by $f$, and then integrating, one checks as in §4 that $\lambda = \Phi_6(f)$. It then follows that

$$\Phi_6(f) = \lambda = \tfrac{1}{2}(\lambda + 5 \times \tfrac{\lambda}{5}) < \frac{1}{2}\sum_{j=0}^{\infty} \lambda_j = \frac{1}{2}\int_{\mathbb{S}^2} K_f^\flat(\omega,\omega)\,\mathrm{d}\sigma(\omega) = 2\pi K_f(0), \tag{18}$$

where the strict inequality is a consequence of Lemma 1, together with the fact that all eigenvalues of $T_f$ are positive. The remaining identities in (18) have already appeared in (11). On the other hand, we have that

$$K_f(0) = \|\widehat{f\sigma}\|^4_{L^4(\mathbb{R}^3)} = \Phi_4(f) \le \Phi_4(\mathbf{1}), \tag{19}$$

where the last inequality follows from Foschi's result [6], discussed in §1.2.1. From (16), (17), we further have that

$$2\pi\Phi_4(\mathbf{1}) = \Phi_6(\mathbf{1}), \tag{20}$$

and so from (18) and $2\pi\times$ (19), it then follows that $\Phi_6(f) < \Phi_6(\mathbf{1})$. This completes the proof of the theorem. □

## 8 Extensions, generalizations, and open problems

In the last section, we discuss the extension of Theorem 1 to other exponents $q \ge 6$, its generalization to higher dimensions $d \ge 3$, and the corresponding questions for complex-valued maximizers. We conclude with a list of open problems.

### 8.1 Other exponents.

We have already hinted at the very special role played by even integers. It is reassuring to observe that all the steps from §2–§6 work, *mutatis mutandis*, whenever $q \ge 6$ is an even integer. In fact, the whole proof strategy can be made to work, for any $q \in \{6, 8, 10, \ldots\}$. However, one encounters some difficulties along the way. Perhaps most significantly, the natural substitute of (20) boils down to the inequality

$$\Phi_q(\mathbf{1}) \le \frac{1}{\sigma(\mathbb{S}^2)}\frac{q+6}{q+1}\Phi_{q+2}(\mathbf{1}), \tag{21}$$

which needs to be checked for each of the relevant values of $q$. If $q \geq 4$ is an even integer, then

$$\sigma(\mathbb{S}^2)^{q/2}\Phi_q(\mathbf{1}) = (2\pi)^3\|\sigma^{*(q/2)}\|^2_{L^2(\mathbb{R}^3)} = \|\widehat{\sigma}\|^q_{L^q(\mathbb{R}^3)}.$$

The Fourier transform of the surface measure $\sigma$ on $\mathbb{S}^2$ is given by[7] $\widehat{\sigma}(x) = 4\pi\,\mathrm{sinc}(|x|)$, and so (21) holds if and only if

$$\int_0^\infty |\mathrm{sinc}(r)|^q\, r^2\, \mathrm{d}r \leq \frac{q+6}{q+1}\int_0^\infty |\mathrm{sinc}(r)|^{q+2}\, r^2\, \mathrm{d}r. \tag{22}$$

Three natural paths to tackle inequality (21) present themselves. In fact, one can proceed via:

(a) explicit formulae for uniform random walks;

(b) rigorous numerical integration;

(c) asymptotic analysis of the weighted integrals in (22).

Path (a) is quite elegant, path (b) is very robust, and path (c) gathers elements from both. By construction, paths (a), (b) are able to provide a solution to a finite number of exponents only. On the other hand, path (c) relies on asymptotics, and as such it naturally misses a few initial cases. Therefore each of the paths is useful on its own, and the three of them intertwine nicely together.

The integrals in (22) are related to the *cube slicing problem*, addressed in Question 1. To see why this is the case, consider the unit cube $Q_d := [-\frac{1}{2}, \frac{1}{2}]^d \subset \mathbb{R}^d$, and let $H \subset \mathbb{R}^d$ be a linear subspace of codimension 1. Then the volume of the $(d-1)$-dimensional section $H \cap Q_d$ is at least 1, and at most $\sqrt{2}$. The lower bound is best possible, and attained if and only if $H$ is parallel to a face of $Q_d$. The upper bound is also best possible, and attained if and only if $H$ contains a $(d-2)$-dimensional face of $Q_d$. These results were obtained by K. Ball [1], as a consequence of the key inequality

$$\frac{1}{\pi}\int_{-\infty}^\infty |\mathrm{sinc}(r)|^p\, \mathrm{d}r \leq \sqrt{\frac{2}{\pi}},$$

which holds for every $p \geq 2$, with equality if and only $p = 2$. Even though many partial results are known, the complete answer to Question 1 for generic values of $d, k$ remains a topic of current research interest; see [13] and the references therein.

[7] The sinc function is defined as $\mathrm{sinc}(r) := \frac{\sin r}{r}$.

## 8.2 Higher dimensions

The sharp form of inequality (2) for $q = 4$ is unknown if $d \geq 8$. Without this starting point, our bootstrapping approach to proving Theorem 1 seems condemned from the very start. On the other hand, the whole proof strategy can be made to work in dimensions $d \in \{4, 5, 6, 7\}$, but at least three new difficulties arise.

Firstly, the results in §2–§5 can all be adapted to the higher dimensional case, even though the discussion in §5 (in particular, the proof of Lemma 1) requires some care. In fact, a complete answer to Question 3 is known, and reveals a curious difference that occurs in the *four-dimensional* case: The minimal codimension of a proper subalgebra of $\mathfrak{so}(d)$ equals $d - 1$ if $d \geq 3, d \neq 4$, but equals 2 if $d = 4$; see [12]. This stems from the fact that the group $\mathrm{SO}(4)/\{\pm I\}$ is *not* simple, whereas all other groups $\mathrm{SO}(d)$, $d \neq 4$, are simple (after modding out by $\{\pm I\}$ is $d$ if even). In turn, this relates back to the existence of quaternions, and partly accounts for some exotic aspects of the geometry of 4-manifolds.

Secondly, the computations from §6 rely on a solution to Question 2, which for general values of $n$ was obtained recently, but only *under the additional assumption that $d$ is odd*; see [2, 9]. This can be partly explained by the formula which generalizes (14) to all dimensions $d \geq 2$:

$$(\sigma_{d-1} * \sigma_{d-1})(\xi) = \frac{\sigma_{d-2}(\mathbb{S}^{d-2})}{2^{d-3}} \frac{1}{|\xi|} (4 - |\xi|^2)_+^{\frac{d-3}{2}}, \tag{23}$$

together with the realization that the right-hand side of (23) defines a polynomial expression in the variables $|\xi|, |\xi|^{-1}$ if and only if $d$ is odd. For a generalization of (15) to dimensions $d \in \{3, 5, 7, 9\}$, see Figure 2. To the best of our knowledge, a complete answer to Question 2 in even dimensions remains a fascinating, largely open problem, which via the theory of hypergeometric functions and modular forms exhibits some deep connections to number theory; see [3] and the references therein.

Thirdly, the higher-dimensional generalization of (21) boils down to the inequality[8]

$$\Phi_{d,q}(\mathbf{1}) \leq \frac{1}{\sigma_{d-1}(\mathbb{S}^{d-1})} \frac{q + 2d - \delta_{d,4}}{q+1} \Phi_{d,q+2}(\mathbf{1}), \tag{24}$$

which needs to be checked for each of the relevant values of $d, q$. An explicit formula for the Fourier transform $\widehat{\sigma}_{d-1}$ is known in all dimensions $d \geq 2$, but

[8] The Kronecker delta satisfies $\delta_{d,4} = 1$ if $d = 4$, and $\delta_{d,4} = 0$ if $d \neq 4$. The introduction of $\delta_{d,4}$ is justified by the distinct behaviour of $\mathfrak{so}(4)$ discussed above.

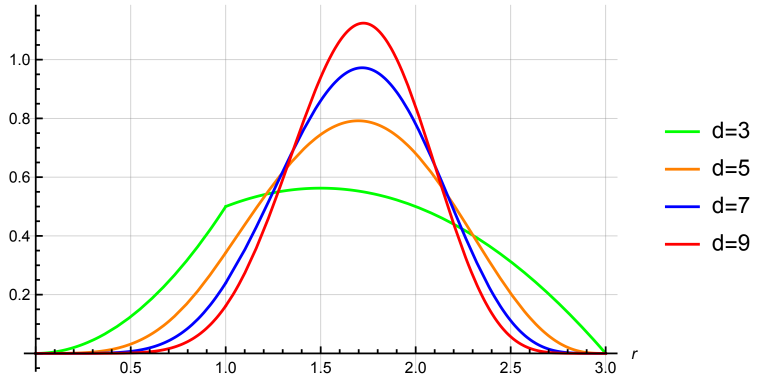



Figura 2: Plot of the function $r \mapsto r^{d-1}(\sigma_{d-1} * \sigma_{d-1} * \sigma_{d-1})(r)$ for $0 \leq r \leq 3$, when $d \in \{3,5,7,9\}$. Multiplication by $r^{d-1}$ distorts the picture but clarifies the behavior, because the surface area of a sphere of radius $r$ in $\mathbb{R}^d$ is proportional to $r^{d-1}$; in particular, one-dimensional integrals of the plotted function are proportional to integrals of the function in $\mathbb{R}^d$.

it involves the Bessel function $J_{(d-2)/2}$, which is *not* an elementary function whenever $d$ is even; see [17, Ch. VIII, §3]. In fact, setting $\nu = (d-2)/2$, we have that

$$\widehat{\sigma}_{d-1}(x) = (2\pi)^{\frac{d}{2}} |x|^{-\nu} J_\nu(|x|),$$

and consequently (24) holds if and only if

$$\int_0^\infty |J_\nu(r)|^q \, r^{d-1-q\nu} \, \mathrm{d}r \leq \frac{((\frac{d}{2})!)^2}{2^{2-d}} \frac{q+2d-\delta_{d,4}}{q+1} \int_0^\infty |J_\nu(r)|^{q+2} \, r^{d-1-(q+2)\nu} \, \mathrm{d}r. \tag{25}$$

A careful combination of the paths (a), (b), (c) outlined in §8.1 above can be used to verify inequality (25), and therefore (24), in the appropriate range of exponents and dimensions. Details can be consulted in [16, §7].

### 8.3 $\mathbb{C}$-valued maximizers

It is natural to ask about general complex-valued maximizers of $\Phi_{d,q}$, for $d \geq 2$ and even $q \geq 2\frac{d+1}{d-1}$. In [16, Theorem 1.2], we show that in this case any $\mathbb{C}$-valued maximizer of $\Phi_{d,q}$ is of the form $ce^{i\xi\cdot\omega}F(\omega)$, for some $\xi \in \mathbb{R}^d$, some $c \in \mathbb{C} \setminus \{0\}$, and some nonnegative, antipodally symmetric maximizer $F$ of $\Phi_{d,q}$. Given the discussion in §8.1 and §8.2, all $\mathbb{C}$-valued maximizers of $\Phi_{d,q}$ are then given by $ce^{i\xi\cdot\omega}$, for some $\xi \in \mathbb{R}^d$ and $c \in \mathbb{C} \setminus \{0\}$, provided $d \in \{3,4,5,6,7\}$ and $q \geq 4$ is an even integer.

### 8.4 Open problems

We collect some of the outstanding problems which have been mentioned throughout the present note, and add a few others to the list.

1. Do constant functions maximize $\Phi_{2,6}$? If this is indeed the case, then [16, Theorem 1.1] implies that constant functions maximize $\Phi_{2,q}$ as well, for every even integer $q \geq 6$.

2. Are non-zero solutions of the Euler–Lagrange equation which generalizes (6) to arbitrary dimensions $d \geq 2$ and exponents $q \geq 2\frac{d+1}{d-1}$,

$$(|\widehat{f\sigma}_{d-1}|^{q-2}\widehat{f\sigma}_{d-1})^{\vee}|_{\mathbb{S}^{d-1}} = \lambda f,$$

   necessarily $C^\infty$-smooth even when $q$ is *not* an even integer?

3. Do maximizers of $\Phi_{d,q}$ exist at the endpoint $q = 2\frac{d+1}{d-1}$ if $d \geq 4$? See [8, Theorem 1.1] for a conditional result along these lines.

4. Assuming the answer to the question in (3) to be affirmative, do constant functions maximize $\Phi_{d,q}$ if $q = 2\frac{d+1}{d-1}$, in all dimensions $d \geq 4$? Conversely, are all real-valued maximizers of $\Phi_{d,2\frac{d+1}{d-1}}$ constant?

## Acknowledgements

D.O.S. is supported by the EPSRC New Investigator Award "Sharp Fourier Restriction Theory", grant no. EP/T001364/1, and expresses his gratitude to Edgar Costa, José Mourão, and Gonçalo Oliveira for organizing the first edition of the conference *Global Portuguese Mathematicians* at Instituto Superior Técnico, Lisbon, July 2017, whose hospitality is greatly appreciated. The authors would like to thank Luís Diogo for the careful and critical reading of a preliminary version of this article, and the anonymous referee for helpful comments.

## References

[1] K. Ball, "Cube slicing in $\mathbb{R}^n$", *Proc. Amer. Math. Soc.*, Vol. 97, No. 3 (1986), pp. 465–473.

[2] J. Borwein and C. Sinnamon, "A closed form for the density functions of random walks in odd dimensions", *Bull. Aust. Math. Soc.*, Vol. 93, No. 2 (2016), pp. 330–339.

[3] J. Borwein, A. Straub, J. Wan and W. Zudilin, "Densities of short uniform random walks. With an appendix by Don Zagier", *Canad. J. Math.*, Vol. 64, No. 5 (2012), pp. 961–990.

[4] E. Carneiro and D. Oliveira e Silva, "Some sharp restriction inequalities on the sphere", *Int. Math. Res. Not. IMRN*, No. 17 (2015), pp. 8233–8267.

[5] L. Fanelli, L. Vega and N. Visciglia, "On the existence of maximizers for a family of restriction theorems", *Bull. Lond. Math. Soc.*, Vol. 43, No. 4 (2011), pp. 811–817.

[6] D. Foschi, "Global maximizers for the sphere adjoint Fourier restriction inequality", *J. Funct. Anal.*, Vol. 268 (2015), pp. 690–702.

[7] D. Foschi and D. Oliveira e Silva, "Some recent progress on sharp Fourier restriction theory", *Anal. Math.*, Vol. 43, No. 2 (2017), pp. 241–265.

[8] R. Frank, E. H. Lieb and J. Sabin, "Maximizers for the Stein–Tomas inequality", *Geom. Funct. Anal.*, Vol. 26, No. 4 (2016), pp. 1095–1134.

[9] R. García-Pelayo, "Exact solutions for isotropic random flights in odd dimensions", *J. Math. Phys.*, Vol. 53, 103504 (2012).

[10] B. Green, "Roth's theorem in the primes", *Ann. of Math. (2)*, Vol. 161, No. 3 (2005), pp. 1609–1636.

[11] B. Hall, *Lie groups, Lie algebras, and representations. An elementary introduction*, Graduate Texts in Mathematics, 222. Springer-Verlag, New York, 2015.

[12] K. H. Hofmann, "Lie algebras with subalgebras of co-dimension one", *Illinois J. Math.*, Vol. 9 (1965), pp. 636–643.

[13] H. König and A. Koldobsky, "On the maximal perimeter of sections of the cube", *Adv. Math.*, Vol. 346 (2019), pp. 773–804.

[14] G. Mockenhaupt, "Salem sets and restriction properties of Fourier transforms", *Geom. Funct. Anal.*, Vol. 10, No. 6 (2000), pp. 1579–1587.

[15] D. Oliveira e Silva and R. Quilodrán, "Smoothness of solutions of a convolution equation of restricted-type on the sphere", Preprint at arXiv:1909.10220 [math.CA], 2019.

[16] D. Oliveira e Silva and R. Quilodrán, "Global maximizers for adjoint Fourier restriction inequalities on low dimensional spheres", Preprint at arXiv:1909.10230 [math.CA], 2019.

[17] E. M. Stein, *Harmonic Analysis: Real-Variable Methods, Orthogonality, and Oscillatory Integrals*, Princeton University Press, Princeton, 1993.

[18] B. Stovall, "Waves, spheres, and tubes: a selection of Fourier restriction problems, methods, and applications", *Notices Amer. Math. Soc.*, Vol. 66, No. 7 (2019), pp. 1013–1022.

[19] R. S. Strichartz, "Restrictions of Fourier transforms to quadratic surfaces and decay of solutions of wave equations", *Duke Math. J.*, Vol. 44, No. 3 (1977), pp. 705–714.

[20] T. Tao, "Some recent progress on the restriction conjecture", Fourier analysis and convexity, 217–243, *Appl. Numer. Harmon. Anal.*, Birkhäuser Boston, Boston, MA, 2004.

[21] P. Tomas, "A restriction theorem for the Fourier transform", *Bull. Amer. Math. Soc.*, Vol. 81, No. 2 (1975), pp. 477–478.

[22] D. Werner, *Funktionalanalysis*, Springer-Verlag, Berlin, 2018.

# Quantity vs. size in representation theory

*Jorge Vitória*
Università degli Studi di Cagliari
Palazzo delle Scienze, Via Ospedale, 72
09123 Cagliari, Italy
e-mail: jorge.vitoria@unica.it

**Resumo:** Neste texto revemos dois resultados na teoria das representações de álgebras de dimensão finita nos quais a quantidade de um certo tipo de estruturas está intimamente ligado ao tamanho dessas mesmas estruturas. Mais concretamente, discutimos os seguintes factos: (1) uma álgebra de dimensão finita admite apenas um número finito de módulos indecomponíveis a menos de isomorfismo se e só se todos os módulos indecomponíveis são de dimensão finita; (2) a categoria de módulos de uma álgebra de dimensão finita admite apenas um número finito de classes de torsão se e só se todas as classes de torsão são geradas por módulos de dimensão finita.

**Abstract:** In this note, we survey two instances in the representation theory of finite-dimensional algebras where the quantity of a type of structures is intimately related to the size of those same structures. More explicitly, we review the fact that (1) a finite-dimensional algebra admits only finitely many indecomposable modules up to isomorphism if and only if every indecomposable module is finite-dimensional; (2) the category of modules over a finite-dimensional algebra admits only finitely many torsion classes if and only if every torsion class is generated by a finite-dimensional module.

**palavras-chave:** categorias de módulos; módulos indecomponíveis; classes de torsão.

**keywords:** module categories; indecomposable modules; torsion classes.

## 1 Introduction

In representation theory we strive to understand how the properties of a given ring are reflected on that ring's actions on abelian groups. These actions are formalised by the notion of a *module over a ring*, and they are often difficult, if not impossible, to classify completely. Hence, questions in the subject area typically include the problem of classifying modules with certain common properties (simplicity, indecomposability, projectivity, injectivity, ...) up to isomorphism. One could call this a *microscopic* approach to representation theory, in which the main actors are the actual modules over a given ring. These are difficult problems and, more often than not, their solutions involve a fair amount of combinatorics. Another strand of representation theory takes instead a bird's eye view of the subject, i.e. a *macroscopic* point of view, considering the category of all modules and its subcategories as central objects of study. Typical questions within this line of thought include classification problems for subcategories of modules subject to certain properties. Here the tools are more of a homological and categorical nature.

These different points of view are used to study a broad range of rings, from group rings (in representation theory of finite groups) to universal enveloping algebras (in Lie theory) from commutative rings (often occurring in algebraic geometry) to path algebras of quivers (central to the study of finite-dimensional algebras). Recall that an (associative) algebra over a field $\mathbb{K}$ is nothing but a ring with a compatible $\mathbb{K}$-vector space structure. Here, we will be focused on some aspects of the representation theory of $\mathbb{K}$-algebras which are finite-dimensional as $\mathbb{K}$-vector spaces. Our aim is to survey two results that attempt to answer the following type of question:

**Question:** To which extent do finite-dimensional modules over a finite-dimensional algebra $\Lambda$ control *the structure* of the category of all $\Lambda$-modules?

To make this question precise, we need to establish what we mean by *structure*. In this paper this expression will have two meanings. The first surveyed result is a classical theorem in representation theory, and it discusses when is it true that any given module can be built from finite-dimensional modules using the simplest operation available: direct sums. On the second result, however, *the structures* we want to have under control are distinguished classes of modules, called torsion classes. We then aim to answer the question of whether any torsion class in the category of all modules is determined by the finite-dimensional modules contained in it.

In both settings, however, a remarkable pattern arises: the *fewer* the objects under consideration in the finite-dimensional world (may they be modules or torsion classes), the tighter is the grip that finite-dimensional modules have on the whole category. In other words, *quantity* controls *size*. Moreover, and perhaps equally surprising, the converse also holds.

This note is structured as follows. In Section 2 we discuss some basic notions from representation theory of finite-dimensional algebras. In Section 3 we discuss, without proof, the first surveyed result: a well-known, classical theorem in representation theory, due to Auslander ([5, 6]), Fuller-Reiten ([11]) and Ringel-Tachikawa ([15]) concerning algebras of finite representation type. We illustrate the properties under study through some examples. In Section 4 we look at torsion classes and discuss some examples. Finally, in Section 5 we survey a recent result from [3] that builds on the work of Demonet-Iyama-Jasso ([10]). It focuses on algebras whose module category admits only finitely many torsion classes. We sketch a proof of the theorem, leaving out some technical facts that we state without proof. In that respect, the choice made in this paper is to present the results from a torsion-theoretic point of view, leaving outside of the exposition the (intrinsic) relation of the arguments to $\tau$-tilting theory ([1]) or silting modules ([2]).

**Acknowledgements:** The author is very grateful for many insightful discussions with Lidia Angeleri Hügel, Frederik Marks and Rosanna Laking on these topics, and for their comments on a preliminary version of the article.

## 2 Representations of finite-dimensional algebras

Throughout, let $\mathbb{K}$ be an **algebraically closed field** and $\Lambda$ a **finite-dimensional $\mathbb{K}$-algebra**. In this section we discuss finite-dimensional algebras and their representations. For a thorough introduction to the subject, we refer the reader to [4] or [13].

**Example 2.1.** *Let $Q$ be a finite directed graph (usually called a **quiver**). Consider the $\mathbb{K}$-vector space spanned by all oriented paths in $Q$, and endow it with a multiplication defined by concatenation of paths when possible, and zero when concatenation is not possible. This yields a $\mathbb{K}$-algebra, denoted by $\mathbb{K}Q$, called the **path algebra** of $Q$. If $Q$ has no oriented cycles then $\mathbb{K}Q$ is finite-dimensional. For example, if $Q$ is the quiver*

$$1 \xrightarrow{\alpha} 2 \xrightarrow{\beta} 3$$

*then $\Lambda := \mathbb{K}Q$ is a 6-dimensional $\mathbb{K}$-vector space spanned by the stationary paths $e_1$, $e_2$, $e_3$, and the paths $\alpha$, $\beta$ and $\beta\alpha$. Multiplication comes by concatenation as explained above: $\beta \cdot \alpha = \beta\alpha$, while $\alpha \cdot \beta = 0$. This algebra is in fact isomorphic to the algebra of $3 \times 3$ lower triangular matrices over $\mathbb{K}$.*

We will be looking at two categories associated to $\Lambda$.

- $\mathsf{Mod}(\Lambda)$: the category of left $\Lambda$-modules;
- $\mathsf{mod}(\Lambda)$: the category of finite-dimensional left $\Lambda$-modules.

If two finite-dimensional algebras have equivalent categories of modules, they are indistinguishable from a representation-theoretic standpoint. Algebras related in this way are said to be **Morita equivalent**. We will recall that any finite-dimensional algebra $\Lambda$ is Morita equivalent to a quotient of a path algebra. As a consequence, the category of modules over $\Lambda$ is equivalent to the category of bound representations of a quiver.

A **representation of a quiver** Q over $\mathbb{K}$ is the assignment of a vector space to each vertex of $Q$ and a (compatibly chosen) linear map to each arrow of $Q$. A representation of $Q$ is **bound** by an ideal $I$ of $\mathbb{K}Q$ if the linear maps chosen for the arrows of the quiver compose and add up to the zero map when following a linear combination of paths contained in $I$. We denote by $\mathsf{Rep}(Q, I)$ the category of representations of $Q$ bound by $I$ (where morphisms between representations are given by linear maps at every vertex making the obvious diagrams commute). We refer to [4, Corollary I.6.10 and Theorem II.3.7] for the following theorem.

**Theorem 2.2.** *Let $\Lambda$ be a finite-dimensional $\mathbb{K}$-algebra. Then there is a quiver $Q$ and an ideal $I$ of $\mathbb{K}Q$ such that $\mathsf{Mod}(\Lambda) \cong \mathsf{Mod}(\mathbb{K}Q/I) \cong \mathsf{Rep}(Q, I)$.*

The equivalence above restricts to an equivalence between finite-dimensional $\Lambda$-modules and finite-dimensional representations of $Q$ bound by $I$. In essence, the theorem indicates that the study of modules over any finite-dimensional $\mathbb{K}$-algebra is an upgrade of classical linear algebra over $\mathbb{K}$.

**Remark 2.3.** *The classical linear algebra theorem on Jordan normal forms can be regarded as a classification of the isomorphism classes of finite-dimensional representations over a quiver $Q$ with one vertex and one loop. Indeed, any square matrix $n \times n$ corresponds to an endomorphism of $\mathbb{K}^n$, and any conjugation by invertible matrices corresponds to an isomorphism between the associated representations of $Q$. Moreover, Jordan blocks of a given matrix correspond to the indecomposable summands of the associated representation (see Section 3 for the notion of indecomposability). Since $\mathbb{K}Q$ is isomorphic to the polynomial algebra $\mathbb{K}[X]$, much of classical linear algebra can be regarded as the study of finite-dimensional $\mathbb{K}[X]$-representations.*

## 3 Finite representation type

It is not very surprising that over a finite-dimensiona $\mathbb{K}$-algebra $\Lambda$, the structure of the category of finite-dimensional modules, $\mathsf{mod}(\Lambda)$, is much better understood than the structure of the category of all modules $\mathsf{Mod}(\Lambda)$. In fact, as stated in the following theorem (see for example [4, Theorem I.4.10]), objects in the category $\mathsf{mod}(\Lambda)$ can be completely described by a set of fundamental blocks: the finite-dimensional indecomposable modules. A $\Lambda$-module $M$ is said to be **indecomposable** if, whenever $M \cong M_1 \oplus M_2$, then either $M_1$ or $M_2$ must be zero. In other words, a module is indecomposable if it admits no nontrivial direct summands.

**Theorem 3.1** (Krull-Remak-Schmidt)**.** *Every finite-dimensional $\Lambda$-module $M$ is a direct sum of indecomposable $\Lambda$-modules, which are uniquely determined by $M$ up to isomorphism.*

The same statement, however, does not hold in general for infinite-dimensional $\Lambda$-modules. In fact, infinite-dimensional modules may exhibit a striking pathological property: not having **any** indecomposable direct summands! Such modules are called **superdecomposable** (see Example 3.3).

Let us consider the following four *dream* properties for the representation theory of a finite-dimensional algebra $\Lambda$.

(RF1) Every $\Lambda$-module is a direct sum of indecomposable $\Lambda$-modules.

(RF2) Every indecomposable $\Lambda$-module is finite-dimensional.

(RF3) There are only finitely many indecomposable finite-dimensional $\Lambda$-modules up to isomorphism.

(RF4) There are only finitely many indecomposable $\Lambda$-modules up to isomorphism.

The property (RF1) is called **pure semisimplicity** and the property (RF3) is called **representation-finiteness**. While (RF1) gives structural information on the category $\mathsf{Mod}(\Lambda)$, one can consider (RF2) a property regarding *size* and both (RF3) and (RF4) properties regarding *quantity*. We will establish a connection between all of these very soon.

**Example 3.2.** *The algebra $\Lambda$ from Example 2.1 is representation-finite, i.e. $\Lambda$ satisfies (RF3). Up to isomorphism, there are precisely six indecomposable finite-dimensional $\Lambda$-modules and there are no infinite-dimensional indecomposables (thus $\Lambda$ satisfies also (RF2) and (RF4)). Using Theorem 2.2, we can use quiver representations to describe these indecomposable modules:*

$$P_3 := (\,0 \longrightarrow 0 \longrightarrow \mathbb{K}\,) \qquad S_2 := (\,0 \longrightarrow \mathbb{K} \longrightarrow 0\,)$$

$$S_1 := (\,\mathbb{K} \longrightarrow 0 \longrightarrow 0\,) \qquad P_2 := (\,0 \longrightarrow \mathbb{K} \xrightarrow{1} \mathbb{K}\,)$$

$$P_1 := (\,\mathbb{K} \xrightarrow{1} \mathbb{K} \xrightarrow{1} \mathbb{K}\,) \qquad I_2 := (\,\mathbb{K} \xrightarrow{1} \mathbb{K} \longrightarrow 0\,).$$

*It can also be shown that every $\Lambda$-module is isomorphic to a direct sum of copies of these six $\Lambda$-modules, thus proving that $\Lambda$ also satisfies (RF1).*

**Example 3.3.** *It is very easy to produce examples of finite-dimensional algebras that do not satisfy (RF3) (and that, therefore, do not satisfy (RF4) either). For example, if* $\Lambda$ *is the path algebra over* $\mathbb{K}$ *of the quiver*

$$Q\colon\ 1 \ \begin{smallmatrix} \xrightarrow{\ \alpha\ } \\ \xrightarrow{\ \beta\ } \\ \xrightarrow[\ \gamma\ ]{} \end{smallmatrix} \ 2$$

*we can produce infinitely many pairwise non-isomorphic indecomposable finite-dimensional* $\Lambda$*-modules. Given* $\lambda$ *in* $\mathbb{K}$*, the representation of* $Q$

$$M_\lambda\colon\ \mathbb{K} \ \begin{smallmatrix} \xrightarrow{\ \lambda\ } \\ \xrightarrow{\ 0\ } \\ \xrightarrow[\ 1\ ]{} \end{smallmatrix} \ \mathbb{K}$$

*is indecomposable. Since* $\mathbb{K}$ *is infinite (it is algebraically closed), we immediately get infinitely many indecomposable finite-dimensional* $\Lambda$*-modules. Moreover, it is easy to check that* $M_\lambda$ *is isomorphic to* $M_\mu$ *if and only if* $\lambda = \mu$*. We can also produce an indecomposable infinite-dimensional* $\Lambda$*-module, showing that* $\Lambda$ *does not satisfy (RF2). An explicit example is*

$$G\colon\ \mathbb{K}(X) \ \begin{smallmatrix} \xrightarrow{\ X\ } \\ \xrightarrow{\ 0\ } \\ \xrightarrow[\ 1\ ]{} \end{smallmatrix} \ \mathbb{K}(X)$$

*where* $\mathbb{K}(X)$ *is the field of rational functions in one variable. This module belongs to an important family of indecomposable infinite-dimensional modules: the* ***generic modules****. These are indecomposable infinite-dimensional modules that, as modules over their own endomorphism ring, have finite length. They play an important role in controlling the overall representation theory of a finite-dimensional algebra. We refer to [14] and references therein for further information on generic modules.*

*Finally, we produce a superdecomposable* $\Lambda$*-module, thus showing (in a rather extreme way!) that (RF1) is not satisfied by* $\Lambda$*. This example can be found in [14]. We need two nontrivial ingredients: the existence of injective envelopes (see for example [4, Corollary I.5.14]) and the existence of a particularly nice functor between two categories of modules (see below).*

- *First consider the free algebra* $\mathbb{K}\langle X, Y\rangle$ *in two variables over* $\mathbb{K}$*. Note that this infinite-dimensional* $\mathbb{K}$*-algebra is isomorphic to the path algebra over* $\mathbb{K}$ *of the quiver with one vertex and two loops* $X$ *and* $Y$ *on that vertex. Let* $I$ *be the injective envelope of* $\mathbb{K}\langle X, Y\rangle$ *in* $\mathsf{Mod}(\mathbb{K}\langle X, Y\rangle)$*.*

*We show that $I$ is superdecomposable. If $N \neq 0$ is a summand of $I$, then it intersects $\mathbb{K}\langle X, Y\rangle$ nontrivially since $I$ is an essential extension of $\mathbb{K}\langle X, Y\rangle$. Let $a \neq 0$ be an element in $N \cap \mathbb{K}\langle X, Y\rangle$ and consider the injective envelope $J$ of $\mathbb{K}\langle X, Y\rangle Xa$. It follows that $J$ is a nonzero summand of $N$ (by the injectivity of $J$). It remains to see that $J \neq N$, and this follows from the fact that the element $Ya$ of $N \cap \mathbb{K}\langle X, Y\rangle$ cannot lie in $J$ since $\mathbb{K}\langle X, Y\rangle Xa \cap \mathbb{K}\langle X, Y\rangle Ya = \{0\}$ and since $J$ is an essential extension of $\mathbb{K}\langle X, Y\rangle Xa$. Thus $N$ is not indecomposable.*

- *There is a functor $F\colon \mathsf{Mod}(\mathbb{K}\langle X, Y\rangle) \longrightarrow \mathsf{Mod}(\Lambda)$ sending a (left) $\mathbb{K}\langle X, Y\rangle$-module $M$ to the representation of $Q$ given by*

$$F(M)\colon\ M \xrightarrow[\;1\;]{\overset{X}{\longrightarrow}\;\;Y} M$$

*where $X$ and $Y$ are the linear maps resulting from the left action of $X$ and $Y$ on $M$. This functor is full and exact ([14]) and, as a consequence, $F(I)$ is a superdecomposable $\Lambda$-module.*

These examples suggest that the properties (RF1)-(RF4) come in a single package and cannot be satisfied separately. The following fundamental theorem in the representation theory of finite-dimensional algebras states that, indeed, these properties are equivalent. In other words, a finite-dimensional algebra has *very few* (= finitely many) indecomposables up to isomorphism if and only if all indecomposables are *small* (= finite-dimensional).

**Theorem 3.4.** *[5, 6, 11, 15] For a finite-dimensional $\mathbb{K}$-algebra $\Lambda$, the conditions (RF1), (RF2), (RF3) and (RF4) are equivalent.*

## 4 Torsion pairs

Sometimes it is useful to have a birds' eye view of a category of modules and, rather than analysing the category module by module, organise collections of modules which share certain properties into certain classes.

**Example 4.1.** *Every abelian group has a subgroup given by the elements that have finite order. This is called the torsion subgroup. The quotient of an abelian group by its torsion subgroup yields an abelian group where no element has finite order. We may therefore say that the classes of torsion abelian groups and torsionfree abelian groups give us some valuable information on the category of abelian groups.*

The following definition is an abstraction of the example above to an arbitrary abelian category.

**Definition 4.2.** *A pair $(\mathcal{T}, \mathcal{F})$ of full subcategories of an abelian category $\mathcal{A}$ is a **torsion pair** if*

1. $\mathsf{Hom}_{\mathcal{A}}(T, F) = 0$ *for any $T$ in $\mathcal{T}$ and any $F$ in $\mathcal{F}$.*
2. *For any $X$ in $\mathcal{A}$, there are objects $t(X)$ and $f(X)$ in $\mathcal{T}$ and $\mathcal{F}$ respectively, and a short exact sequence of the form*

$$0 \longrightarrow t(X) \longrightarrow X \longrightarrow f(X) \longrightarrow 0$$

*Given a torsion pair $(\mathcal{T}, \mathcal{F})$ in $\mathcal{A}$, we say that $\mathcal{T}$ is a **torsion class** and $\mathcal{F}$ a **torsionfree class**.*

It follows from the definition that the torsion class determines the torsionfree class. For a subcategory $\mathcal{X}$ of $\mathsf{Mod}(\Lambda)$, denote by $\mathcal{X}^{\perp}$ the full subcategory of $\mathsf{Mod}(\Lambda)$ whose objects are the modules $Y$ for which $\mathsf{Hom}_{\Lambda}(X, Y) = 0$ for all $X$ in $\mathcal{X}$. Dually, one may also define ${}^{\perp}\mathcal{X}$. Given a torsion pair $(\mathcal{T}, \mathcal{F})$, we have $\mathcal{F} = \mathcal{T}^{\perp}$ and $\mathcal{T} = {}^{\perp}\mathcal{F}$. It can also be shown that in the category $\mathsf{Mod}(\Lambda)$, a full subcategory $\mathcal{X}$ is a torsion class if and only if it is closed under coproducts (i.e. for any family of objects in $\mathcal{X}$, its coproduct lies also in $\mathcal{X}$), quotients (i.e. any quotient of a module in $\mathcal{X}$ also lies in $\mathcal{X}$) and extensions (i.e. in any short exact sequence with outer terms in $\mathcal{X}$, the middle term must also belong to $\mathcal{X}$). Dually, torsionfree classes are those full subcategories closed under products, submodules and extensions.

**Example 4.3.** *Consider $\mathsf{Mod}(\mathbb{Z})$, the category of abelian groups. Let $\mathcal{T}$ be the class of abelian groups for which every element has finite order and $\mathcal{F}$ the class of abelian groups that have no elements of finite order. It is easy to check the axioms listed above showing that $(\mathcal{T}, \mathcal{F})$ is a torsion pair. It can, furthermore be shown that $\mathcal{F}$ is the subcategory of abelian groups which are subgroups of a product of a (possibly infinite) number of copies of $\mathbb{Q}$ and that $\mathcal{T} = {}^{\perp}\mathbb{Q}$. We then say that this torsion pair is **cogenerated** by $\mathbb{Q}$.*

**Example 4.4.** *Let $\Lambda$ be the algebra from Example 2.1. Since every module in $\mathsf{Mod}(\Lambda)$ is isomorphic to a direct sum of the indecomposable $\Lambda$-modules listed in Example 3.2, and since torsion classes are closed under coproducts and summands, every torsion class is determined by the set of indecomposable modules lying in it. Hence, every torsion class in $\mathsf{Mod}(\Lambda)$ is the closure under coproducts and their direct summands of some set of indecomposable $\Lambda$-modules $\mathcal{X}$ that is closed under quotients and extensions. We denote* *this additive closure by* $\mathsf{Add}(\mathcal{X})$*, and the complete list of torsion classes in* $\mathsf{Mod}(\Lambda)$ *(excluding* $\{0\}$ *and* $\mathsf{Mod}(\Lambda)$*) is*

$$\mathcal{T}_1 := \mathsf{Add}(\{S_2\}) \quad \mathcal{T}_2 := \mathsf{Add}(\{P_3\}) \quad \mathcal{T}_3 := \mathsf{Add}(\{S_1\})$$

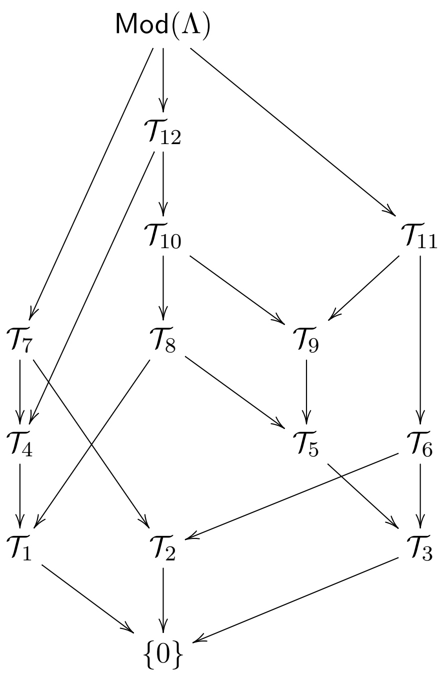
$$\mathcal{T}_4 := \mathsf{Add}(\{P_2, S_2\}) \quad \mathcal{T}_5 := \mathsf{Add}(\{I_2, S_1\}) \quad \mathcal{T}_6 := \mathsf{Add}(\{P_3, S_1\})$$

$$\mathcal{T}_7 := \mathsf{Add}(\{P_3, P_2, S_2\}) \quad \mathcal{T}_8 := \mathsf{Add}(\{S_2, I_2, S_1\}) \quad \mathcal{T}_9 := \mathsf{Add}(\{P_1, I_2, S_1\})$$

$$\mathcal{T}_{10} := \mathsf{Add}(\{S_2, P_1, I_2, S_1\}) \quad \mathcal{T}_{11} := \mathsf{Add}(\{P_3, P_1, I_2, S_1\})$$

$$\mathcal{T}_{12} := \mathsf{Add}(\{P_2, P_1, S_2, I_2, S_1\})$$

*We can order these classes by inclusion, obtaining the following Hasse quiver, where an arrow* $A \to B$ *denotes a strict inclusion* $A \supsetneq B$ *with no element* $C$ *such that* $A \supsetneq C \supsetneq B$*. The arrows in the Hasse diagram are linked to a process called* ***mutation*** *(see, for example, [1]). While we will not discuss this process, it plays an important role in the proof of Theorem 5.4.*




*Note also that this Hasse quiver depicts a well-known object: the three-dimensional associahedron. For further details on the combinatorics of torsion pairs of path algebras of quivers we refer to [12].*

**Example 4.5.** *For a $\Lambda$-module $M$, the subcategory $M^{\perp}$ of $\mathsf{Mod}(\Lambda)$ is closed under products, submodules and extensions and, thus, $M^{\perp}$ is a torsionfree class. The corresponding torsion class is then necessarily given by ${}^{\perp}(M^{\perp})$, and it is clear that $M$ lies in ${}^{\perp}(M^{\perp})$. In general it is not easy to describe which modules lie in ${}^{\perp}(M^{\perp})$. However, if $M$ is, for example, a projective $\Lambda$-module, then one can show that ${}^{\perp}(M^{\perp})$ coincides with the subcategory $\mathsf{Gen}(M)$ formed by all $\Lambda$-modules which are quotients of some coproduct of copies of $M$. In Section 5 we describe the torsion pairs that are of the form $(\mathsf{Gen}(M), M^{\perp})$ for a finite-dimensional $\Lambda$-module $M$.*

## 5 Torsion-finiteness

In this section we will discuss categories of modules that have only finitely many torsion classes. Let us first look at torsion classes of $\mathsf{Mod}(\Lambda)$, for a finite-dimensional $\mathbb{K}$-algebra $\Lambda$, which are of the form $\mathsf{Gen}(M)$ for some $\Lambda$-module $M$. They satisfy the following useful property.

**Lemma 5.1.** *[10, Lemma 3.10] Suppose that $\mathcal{T} = \mathsf{Gen}(M)$ is a torsion class. Suppose that there is an ascending sequence of torsion classes*

$$\mathcal{T}_1 \subseteq \mathcal{T}_2 \subseteq \mathcal{T}_3 \subseteq \cdots \subseteq \mathcal{T}_n \subseteq \mathcal{T}_{n+1} \subseteq \cdots$$

*such that $\bigcup_{i\geq 1} \mathcal{T}_i = \mathcal{T}$. Then the sequence stabilises.*

*Proof.* Given $n \geq 1$ such that $M$ lies in the torsion class $\mathcal{T}_n$, all coproducts of copies of $M$ and their quotients must lie in $\mathcal{T}_n$, proving that $\mathcal{T}_n = \mathcal{T}$. □

In the category $\mathsf{mod}(\Lambda)$ of finite-dimensional modules, quotients of finite coproducts of a finite-dimensional module $M$ sometimes also form a torsion class. Such a subcategory is denoted by $\mathsf{gen}(M)$. Clearly such subcategories also satisfy the property of Lemma 5.1. Torsion pairs in $\mathsf{mod}(\Lambda)$ and torsion pairs in $\mathsf{Mod}(\Lambda)$ are related by the following theorem. Given a subcategory $\mathcal{X}$ of $\mathsf{mod}(\Lambda)$, denote by $\varinjlim \mathcal{X}$ the subcategory of $\mathsf{Mod}(\Lambda)$ whose objects are direct limits of direct systems with terms in $\mathcal{X}$. Note that, since a direct limit of a direct system is a quotient of the coproduct of the terms in that system, torsion classes in $\mathsf{Mod}(\Lambda)$ are closed under direct limits.

**Theorem 5.2.** *[7, Lemma 4.4] If $(\mathcal{U}, \mathcal{V})$ is a torsion pair in $\mathsf{mod}(\Lambda)$, then $(\varinjlim \mathcal{U}, \varinjlim \mathcal{V})$ is a torsion pair in $\mathsf{Mod}(\Lambda)$, and*

$$(\varinjlim \mathcal{U}, \varinjlim \mathcal{V}) = (\mathsf{Gen}(\mathcal{U}), \mathcal{U}^{\perp}).$$

*This assignment is injective since*

$$(\mathcal{U}, \mathcal{V}) = ((\varinjlim \mathcal{U}) \cap \mathsf{mod}(\Lambda), (\varinjlim \mathcal{V}) \cap \mathsf{mod}(\Lambda)),$$

*and a torsion pair* $(\mathcal{T}, \mathcal{F})$ *in* $\mathsf{Mod}(\Lambda)$ *arises in this way if and only if* $\mathcal{F}$ *is closed under direct limits in* $\mathsf{Mod}(\Lambda)$*.*

The theorem above establishes a close relation between torsion classes of the form $\mathsf{gen}(M)$ and $\mathsf{Gen}(M)$ in $\mathsf{mod}(\Lambda)$ and in $\mathsf{Mod}(\Lambda)$ respectively, for a finite-dimensional $\Lambda$-module $M$. In fact, it follows that $\mathsf{gen}(M)$ is a torsion class in $\mathsf{mod}(\Lambda)$ if and only if $\mathsf{Gen}(M)$ is a torsion class in $\mathsf{Mod}(\Lambda)$, in which case $\mathsf{Gen}(M) = \varinjlim \mathsf{gen}(M)$. The following result characterises the torsion pairs that are of the form $(\mathsf{Gen}(M), M^{\perp})$. Recall that a **pure submodule** $Y$ of a $\Lambda$-module $X$ is a submodule such that for any right $\Lambda$-module $Z$, we have that $Z \otimes_\Lambda Y$ is still a submodule of $Z \otimes_\Lambda X$. For example, it is easy to check that if $X/Y$ is a flat module, then $Y$ is a pure submodule of $X$.

**Proposition 5.3.** *Let* $\Lambda$ *be a finite dimensional* $\mathbb{K}$*-algebra and let* $(\mathcal{T}, \mathcal{F})$ *be a torsion pair in* $\mathsf{Mod}(\Lambda)$*. The following statements are equivalent.*

1. $\mathcal{T} \cap \mathsf{mod}(\Lambda) = \mathsf{gen}(M)$ *for a finite dimensional* $\Lambda$*-module* $M$*;*
2. $(\mathcal{T}, \mathcal{F}) = (\mathsf{Gen}(M), M^{\perp})$ *for a finite dimensional* $\Lambda$*-module* $M$*;*
3. $\mathcal{T}$ *is closed under products and* $\mathcal{F}$ *is closed under direct limits.*

*Proof.* Let $(\mathcal{T}, \mathcal{F})$ be a torsion pair in $\mathsf{Mod}(\Lambda)$.

(1) $\Rightarrow$ (2): Suppose that $\mathcal{T} \cap \mathsf{mod}(\Lambda) = \mathsf{gen}(M)$. Since $\mathcal{T} \cap \mathsf{mod}(\Lambda)$ is a torsion class in $\mathsf{mod}(\Lambda)$ it follows that, as seen above, $\mathsf{Gen}(M) = \varinjlim(\mathcal{T} \cap \mathsf{mod}(\Lambda))$ is a torsion class, and that $\mathsf{Gen}(M) \subseteq \mathcal{T}$. To prove that $\mathcal{T} \subseteq \mathsf{Gen}(M)$ we need the following facts.

(P1) [8, 2.2, Example 3] Every $\Lambda$-module is a pure submodule of the product of its finite-dimensional quotients.

(P2) [7, Theorem 4.2] For a torsion class $\mathcal{U}$ in $\mathsf{mod}(\Lambda)$, $\varinjlim \mathcal{U}$ is always closed under pure submodules and, moreover, it is closed under products if and only if $\mathcal{U} = \mathsf{gen}(N)$ for a finite-dimensional $\Lambda$-module $N$.

If $X$ is a module in $\mathcal{T}$ then, by (P1), $X$ is a pure submodule of its finite-dimensional quotients, all of which lie in $\mathcal{T} \cap \mathsf{mod}(\Lambda) = \mathsf{gen}(M)$. Since, by (P2), $\mathsf{Gen}(M) = \varinjlim \mathsf{gen}(M)$ is closed under products and pure submodules, $X$ lies in $\mathsf{Gen}(M)$.

(2) $\Rightarrow$ (3): Since $M$ is finite-dimensional, the functor $\mathsf{Hom}_\Lambda(M,-)$ commutes with direct limits and, thus, $\mathcal{F} = M^\perp$ is closed under direct limits. Moreover, since $\mathsf{Gen}(M) \cap \mathsf{mod}(\Lambda) = \mathsf{gen}(M)$, it follows from the statement (P2) that $\mathsf{Gen}(M)$ is closed under products.

(3) $\Rightarrow$ (1): If $\mathcal{F}$ is closed under direct limits then, by Theorem 5.2,

$$(\mathcal{T}, \mathcal{F}) = (\varinjlim(\mathcal{T} \cap \mathsf{mod}(\Lambda)), \varinjlim(\mathcal{F} \cap \mathsf{mod}(\Lambda))).$$

By the fact (P2) cited above, it then follows that there is a finite-dimensional $\Lambda$-module $M$ such that $\mathcal{T} \cap \mathsf{mod}(\Lambda) = \mathsf{gen}(M)$. □

If one wishes to classify all torsion pairs in $\mathsf{Mod}(\Lambda)$, the following four *dream* properties could be of help, just like in Section 3.

(TF1) Every torsion pair in $\mathsf{Mod}(\Lambda)$ is of the form $(\varinjlim \mathcal{U}, \varinjlim \mathcal{V})$ for a torsion pair $(\mathcal{U}, \mathcal{V})$ in $\mathsf{mod}(\Lambda)$;

(TF2) Every torsion class in $\mathsf{Mod}(\Lambda)$ is of the form $\mathsf{Gen}(M)$ for a finite-dimensional $\Lambda$-module $M$;

(TF3) There are only finitely many torsion classes in $\mathsf{mod}(\Lambda)$;

(TF4) There are only finitely many torsion classes in $\mathsf{Mod}(\Lambda)$.

Note that, once again, (TF1) is a property concerning the structure of torsion pairs in the category of $\Lambda$-modules, while (TF2) can be seen as a measure of *size* and (TF3) and (TF4) as measures of *quantity*. Just like in the previous section, it turns out that these properties (TF1)–(TF4) are equivalent to each other. In other words, a finite-dimensional algebra admits *very few* (= finitely many) torsion classes in its category of (finite-dimensional) modules if and only if all torsion classes in its module category are generated by *small* (= finite-dimensional) modules. The following theorem is essentially proved in [3], using the language of support $\tau$-tilting modules ([1]) and silting modules ([2]). The proof we present here purposefully avoids mentioning these modules, re-working the existing arguments from a torsion-theoretic point of view.

**Theorem 5.4.** *[3, Theorem 4.8] For a finite-dimensional $\mathbb{K}$-algebra $\Lambda$, the conditions (TF1), (TF2), (TF3) and (TF4) are equivalent.*

*Proof.* In order to prove this theorem we need the following additional fact.

(P3) [10, Proposition 3.8] There are only finitely many torsion classes in $\mathsf{mod}(\Lambda)$ if and only if every torsion class in $\mathsf{mod}(\Lambda)$ is of the form $\mathsf{gen}(M)$ for a finite-dimensional $\Lambda$-module $M$.

This fact partially relies on the combinatorial technique mentioned in the previous section: mutation. Indeed, if there are either infinitely many torsion classes in $\mathsf{mod}(\Lambda)$ or if there is a torsion class which is not of the form $\mathsf{gen}(M)$, one can build a non-stabilising ascending chain of torsion classes

$$\mathcal{T}_1 \subsetneq \mathcal{T}_2 \subsetneq \mathcal{T}_3 \subsetneq \cdots \subsetneq \mathcal{T}_n \subsetneq \mathcal{T}_{n+1} \subsetneq \cdots$$

which, with the help of Lemma 5.1, guarantees the other condition. We refer the reader to [1] and [10] for further details.

(TF1) $\Rightarrow$ (TF2): Let $(\mathcal{T}, \mathcal{F})$ be a torsion pair in $\mathsf{Mod}(\Lambda)$. By assumption, $\mathcal{T} = \varinjlim \mathcal{U}$ and $\mathcal{F} = \varinjlim \mathcal{V}$, where $\mathcal{U} = \mathcal{T} \cap \mathsf{mod}(\Lambda)$ and $\mathcal{V} = \mathcal{F} \cap \mathsf{mod}(\Lambda)$. Let us first show that $\mathcal{T} = {}^{\perp}\mathcal{V}$. Indeed, it is easy to check that ${}^{\perp}\mathcal{V}$ is a torsion class containing $\mathcal{U}$ and that the corresponding torsionfree class $({}^{\perp}\mathcal{V})^{\perp}$ contains $\mathcal{V}$. Since every torsion class is closed under direct limits we immediately conclude that $\mathcal{T} = \varinjlim \mathcal{U} \subseteq {}^{\perp}\mathcal{V}$. Since, by assumption, $\mathcal{F}$ is also closed under direct limits, we have that $\mathcal{F} = \varinjlim \mathcal{V} \subseteq ({}^{\perp}\mathcal{V})^{\perp}$ and, therefore, we can also conclude that $\mathcal{T} = {}^{\perp}\mathcal{F} \supseteq {}^{\perp}(({}^{\perp}\mathcal{V})^{\perp}) = {}^{\perp}\mathcal{V}$. This proves that $\mathcal{T} = {}^{\perp}\mathcal{V}$, as wanted.

Now, since $\mathcal{V}$ is a subcategory of finite-dimensional $\Lambda$-modules, ${}^{\perp}\mathcal{V}$ is closed under products (see, for example, [9, Example 2.3]). Therefore, (TF2) follows from Proposition 5.3.

(TF2) $\Rightarrow$ (TF3): If $\mathcal{U}$ is a torsion class in $\mathsf{mod}(\Lambda)$, then by Theorem 5.2, $\varinjlim \mathcal{U}$ is a torsion class in $\mathsf{Mod}(\Lambda)$. From (TF2) we conclude that $\varinjlim \mathcal{U} = \mathsf{Gen}(M)$ for a finite-dimensional $\Lambda$-module $M$ and, thus, $\mathcal{U} = \mathsf{Gen}(M) \cap \mathsf{mod}(\Lambda) = \mathsf{gen}(M)$. Finally, (TF3) follows (P3).

(TF3) $\Rightarrow$ (TF1): This follows from (P3) and Proposition 5.3.

(TF4) $\Rightarrow$ (TF3): This is a direct consequence of Theorem 5.2.

(TF2) $\Rightarrow$ (TF4): (TF2) implies that the assignment in Theorem 5.2 is a bijection (since $M^{\perp}$ is closed under direct limits for any finite-dimensional $\Lambda$-module $M$), i.e. there are as many torsion classes in $\mathsf{mod}(\Lambda)$ as in $\mathsf{Mod}(\Lambda)$. Since (TF2) is proved to be equivalent to (TF3), we conclude that there are finitely many torsion classes in $\mathsf{Mod}(\Lambda)$. $\square$

Note that the properties (P1), (P2) and (P3) are fundamental to our proof, and they depend heavily on the fact that we are working with finite-dimensional algebras. We should, therefore, be very careful with any attempts to naively generalise the result above to larger classes of rings. We finish this survey with an example of a commutative noetherian (but not artinian) ring where the philosophy of this last section fails.

**Example 5.5.** *Let $R$ be the (commutative, noetherian) ring of fractions of $\mathbb{Z}$ obtained by inverting all odd integers, i.e.*

$$R = \{\frac{a}{b} \in \mathbb{Q} \colon \mathsf{gcd}(b, 2) = 1\}.$$

*This ring is a principal ideal domain, and its nontrivial ideals are just the powers of the maximal ideal $\mathfrak{p}$ generated by 2. Finitely generated modules over a principal ideal domain are very well-understood. In our case, for every finitely generated $R$-module there is an isomorphism*

$$M \cong \bigoplus_{k\geq 0} (R/\mathfrak{p}^k)^{n_k(M)},$$

*where $n_k(M) \neq 0$ for only finitely many $k$ (and $\mathfrak{p}^0 = \{0\}$).*

- *We first show that $\mathsf{mod}(R)$ has only two nontrivial torsion classes. Let $\mathcal{T} \neq \{0\}$ be a torsion class in $\mathsf{mod}(R)$, the subcategory of finitely generated $R$-modules. If there is a module $M$ in $\mathcal{T}$ such that $M$ is faithful (i.e., such that $n_0(M) \neq 0$), then since $\mathcal{T}$ is closed under direct summands, $R$ lies in $\mathcal{T}$ and $\mathcal{T} = \mathsf{mod}(R)$. If, on the other hand, every module in $\mathcal{T}$ has a nonzero annihilator, it is easy to show that $R/\mathfrak{p}$ lies in $\mathcal{T}$ (since it is a quotient of any nonzero module). Finally, observe that since $R/\mathfrak{p}^n$ is an iterated extension of $R/\mathfrak{p}$, we get that*

$$\mathcal{T} = \{M \in \mathsf{mod}(R) \colon n_0(M) = 0\}.$$

- *We now produce a torsion class in $\mathsf{Mod}(R)$ that is not generated by a finitely generated $R$-module. Since $R$ is a principal ideal domain, an $R$-module is injective if and only if it is divisible. It can then be shown that the injective $R$-modules are those in $\mathsf{Gen}(\mathbb{Q} \oplus \mathbb{Q}/R)$, and that they form a torsion class. Note additionally that $\mathbb{Q}$ is not a finitely generated $R$-module. Finally, it can be shown that any other generator of this same torsion class must contain $\mathbb{Q}$ as a summand.*

*In conclusion, $\mathsf{mod}(R)$ satisfies the analogous condition to (TF3) and, yet, $\mathsf{Mod}(R)$ admits a torsion class that cannot be generated by a finitely generated $R$-module, therefore not satisfying the analogous condition to (TF2).*

## References


[1] T. Adachi, O. Iyama, I. Reiten, *τ-tilting theory*, Compos. Math. **150** (2014), 415–452.

[2] L. Angeleri Hügel, F. Marks and J. Vitória, *Silting modules*, Int. Math. Res. Not. IMRN **2016**, no. 4 (2016), 1251–1284.

[3] L. Angeleri Hügel, F. Marks and J. Vitória, *A characterisation of τ-tilting finite algebras*, in Model Theory of Modules, Algebras and Categories, Contemp. Math. **730**, Amer. Math. Soc. (2019), 75–89.

[4] I. Assem, D. Simson and A. Skowronski, *Elements in the representation theory of associative algebras, Volume 1*, London Mathematical Society Student Texts (2006).

[5] M. Auslander, *Representation theory of Artin algebras II*, Comm. in Algebra **1** (1974), 293–310.

[6] M. Auslander, *Large modules over artin algebras*, in Algebra, Topology and Category Theory: a collection of papers in honor of Samuel Eilenberg; Academic Press (1976), 1–17.

[7] W. Crawley-Boevey, *Locally finitely presented additive categories*, Comm. in Algebra **22**, no. 5 (1994), 1641–1674.

[8] W. W. Crawley-Boevey, *Infinite-dimensional modules in the representation theory of finite-dimensional algebras*, Algebras and modules, I (Trondheim, 1996), CMS Conf. Proc. **23** Amer. Math. Soc. (1998), 29–54.

[9] S. Dean, *Duality and contravariant functors in the representation theory of Artin algebras*, J. Algebra Appl. **18**, no. 6 (2019), 27 pp.

[10] L. Demonet, O. Iyama and G. Jasso, *τ-tilting finite algebras, bricks and g-vectors*, Int. Math. Res. Not. IMRN **2019** (2019), 852–892.

[11] K. Fuller and I. Reiten, *Note on rings of finite representation type and decompositions of modules*, Proc. Amer. Math. Soc. **50** (1975), 39–44.

[12] C. Ingalls and H. Thomas, *Noncrossing partitions and representations of quivers*, Compos. Math. **145**, no. 6 (2009), 1533–1562.

[13] R.S. PIERCE, *Associative algebras*, Graduate Texts in Mathematics, Springer (1982).

[14] C.M. RINGEL *Infinite length modules. Some examples as introduction*, Infinite Length Modules. Some Examples as Introduction. In: Krause H., Ringel C.M. (eds) Infinite Length Modules. Trends in Mathematics. Birkhäuser (2000), 1–74.

[15] C.M. RINGEL and H. TACHIKAWA, *QF-3 rings*, J. Reine Ang. Math. **272** (1975), 49–72.

**Informações aos autores**

O Boletim da SPM publica-se, em geral, duas vezes por ano.

O Boletim constitui um espaço diversificado de informação, promove a circulação de ideias e de opiniões, bem como troca de experiências entre os que ensinam, investigam ou aplicam a Matemática.

O Boletim não publica artigos de investigação especializada. Estes trabalhos poderão ser submetidos à Portugaliæ Mathematica, revista de prestígio internacional, também propriedade da SPM e editada pela European Mathematical Society.

Os artigos dedicados a assuntos de natureza pré-universitária deverão, de preferência, ser submetidos à Gazeta de Matemática.

As actividades da SPM, nomeadamente das suas Delegações Regionais e Secções, são noticiadas com regularidade no Boletim.

As opiniões expressas pelos autores dos artigos publicados no Boletim não representam necessariamente posições da SPM.

Os Editores das Secções são os únicos responsáveis pela aceitação de artigos nas Secções que dirigem. A Secção Opinião — Cartas ao Director é da exclusiva responsabilidade da Directora do Boletim. Todos os outros trabalhos serão enviados pelos editores a *Referees* especializados para aconselharem sobre a respectiva publicação (com eventuais alterações).

Os manuscritos devem ser submetidos em `http://revistas.rcaap.pt/BoletimSPM` ou enviados por correio electrónico para um dos editores. Agradece-se o envio da versão PDF do texto. Os autores devem indicar as respectivas instituições, bem como os seus endereços de correio electrónico. Os trabalhos submetidos devem incluir um sumário em português e em inglês, e uma lista de palavras-chave nestas duas línguas. Recomenda-se vivamente que os trabalhos sejam preparados em LaTeX. A bibliografia deve seguir o padrão habitual no LaTeX.

Endereço para correspondência:

Boletim da Sociedade Portuguesa de Matemática

Av. da República, 45–3º E, 1050–187 Lisboa

Boletim da Sociedade Portuguesa de Matemática

77

spm
SOCIEDADE PORTUGUESA DE MATEMÁTICA

FCT
Fundação para a Ciência e a Tecnologia
MINISTÉRIO DA EDUCAÇÃO E CIÊNCIA